Washington University in St. Louis

Olin Business School

Deep Learning for Prediction of Business Outcomes

Final Project:

Bitcoin Daily Closing Price Prediction

Vincent Kong

DAT 565E

Prof Salih Tutun, PhD

May 10, 2023

Contents

Abstract

With the increasing popularity of cryptocurrencies, Bitcoin has emerged as one of the most widely recognized digital assets, experiencing a significant surge in price in recent years. Given the interest of investors in predicting Bitcoin's daily price movements, we propose the development of a robust deep learning model for accurate price forecasting. Our approach involves leveraging Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) architectures, which have shown promise in capturing temporal dependencies.

To determine the best model, we conducted a comprehensive exploration of various hyperparameters and model architectures to identify the optimal combination. Additionally, we incorporated several techniques, including ModelCheckPoint, and Dropout, to mitigate the issue of overfitting, thereby enhancing the model's performance.

Overall, the results demonstrate both RNN and LSTM models achieving high accuracy in predicting Bitcoin's daily closing price, while our analysis reveals that the LSTM model surpasses the RNN model in accurately forecasting Bitcoin's daily close price, indicating its superior performance in this context.

1. Introduction

As a decentralized and virtual currency that operates independently of any government or central bank, Bitcoin exists as a digital file that can be securely shared and transferred between users. In recent years, Bitcoin has experienced a surge in popularity, attracting a growing number of investors. However, due to its volatile nature, with frequent price fluctuations, investors face the challenge of predicting its price in advance to mitigate risks and enhance capital gains. As the investment in Bitcoin continues to rise, the need for reliable price prediction tools becomes crucial. By accurately forecasting Bitcoin's price, investors can make informed decisions and adjust their investment strategies accordingly. This enables them to minimize the associated risks and maximize potential profits.

However, given the intricate nature of Bitcoi's price movement, simplistic machine learning methods may not adequately capture and comprehend the underlying patterns in the data. As a result, this report focuses on developing a sophisticated deep learning model utilizing RNN and LSTM algorithms. Our objective is to leverage these advanced techniques for precise predictions based on Bitcoin's historical data spanning from 2012 to 2021. By incorporating RNN and LSTM algorithms, which excel in capturing temporal dependencies, we aim to enhance the model's ability to discern complex patterns and trends in Bitcoin's price data. These deep learning architectures possess the capacity to process sequential information and learn from historical sequences, making them well-suited for forecasting tasks. To evaluate the performance of the developed model, we will employ Mean Squared Error (MSE) and Mean Absolute Error (MAE) as evaluation metrics.

Overall, a reliable Bitcoin price prediction model holds immense value for various stakeholders such as financial institutions, Cryptocurrency exchanges and risk management firms, and the development of such a model contributes to informed decision-making, risk mitigation, and improved financial performance in the dynamic realm of cryptocurrencies, benefitting multiple parties in the business ecosystem.

2. Literature Review

2.1 Methodologies for time series forecasting

In this article "*Bitcoin and Cryptocurrency: Challenges, Opportunities and Future Works*", Ahmed Tealab briefly talks about the development of methodologies used for time series forecasting, and introduces some common artificial neural network methods, such as ANN, AR, ARIMA and so on. Likewise, he analyzes their strengths and weaknesses, and he mentions remaining problems to be solved for new models in the future. From other researchers' papers, he addresses that many studies implement neural network models for forecasting, but most of them lack theoretical support and a methodical procedure. It reminds us to choose models reasonably and pay attention to the details of their implications. In addition, the algorithms of RNN and LSTM in our report are distinct from those included in the paper, and they may perform better in capturing the pattern of time-series data**.**

2.2 Bitcoin Price Prediction using Regression and Arima

For the Github project created by Dushyant Panchal, Ishan Kapur, and Ishan KapurSajag Prakash, it aims to build a time-series predive model and its objective is to make predictions of the Bitcoin daily prices too. They select specific machine learning models including Regression models and Arima models, and explain the reasons for using them. They discuss each model's advantages and disadvantages, and they conclude that ARIMA is the best model among these machine learning models they used in their project. It inspires us that it's important to see the overview and check the stationary of our dataset before building a model. Additionally, our approach is distinct from their approach, and we are motivated to apply Deep Learning techniques like RNN and LSTM to obtain accurate predictions.

2.3 Bitcoin Price Prediction with RNN

The Kaggle project was created by Mustafa Özetmiz, he proposes to build a predictive model using RNN to predict Bitcoin daily price. Even though he uses a relatively outdated dataset, he clearly demonstrates how to manipulate the data and apply SimpleRNN. He attempts to use the weighted price of each day, and build a five-layers RNN model to perform the task. Likewise, he adds the technique of Dropout, and the loss function he uses is mean squared error. Although each step is clearly explained, the model he built is relatively simple, and its construction still lacks thorough consideration from our perspective. What makes our approach unique is that we use RNN and LSTM algorithms for comparison, which provides a more comprehensive understanding of the strengths and weaknesses of each model. In addition, we use a more up-to-date dataset, which allows for more accurate predictions, and we apply more techniques, such as ModelCheckpoint and Regularization, to make it perform better.

3. Problem Description

Various existing works on price prediction often face challenges such as low accuracy and limited capability for long-term predictions. In our project, we propose a novel approach that leverages deep learning, specifically a Recurrent Neural Network (RNN) model, to forecast Bitcoin prices using time series data. Our primary objective is to improve the accuracy of long-term price predictions. By training and testing the RNN model on the available dataset, we aim to generate forecasts that extend beyond the immediate future, providing more reliable and precise predictions for Bitcoin prices.

To mitigate potential issues of overfitting during the training and testing process, we will implement several techniques in our model. First, we will utilize ModelCheckpoint, which allows us to save the weights of the best-performing model during training. This ensures that we retain the model with the optimal performance and avoid overfitting. Second, we will incorporate Dropout, a regularization technique, which randomly drops out a fraction of the model's units during training. This helps prevent over-reliance on specific units and encourages the network to learn more robust and generalized representations. By implementing these measures, we aim to enhance the generalization capability of our RNN model, thereby mitigating overfitting issues and improving the accuracy and reliability of long-term Bitcoin price predictions.

4. Database Background and Data Preprocessing

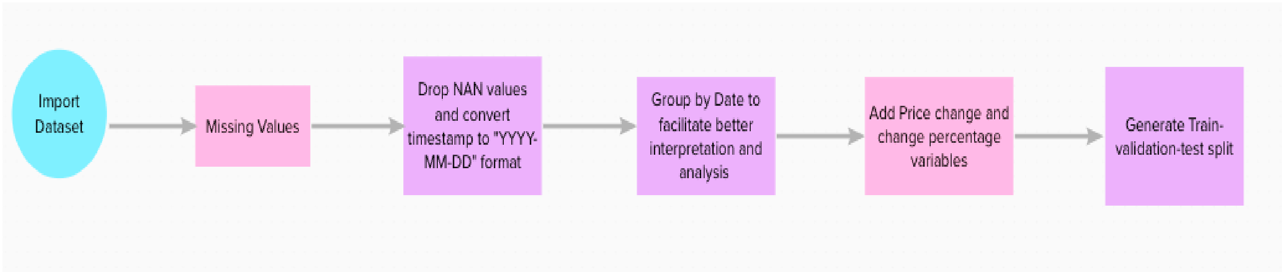


Figure 1 Data Processing Diagram

For our project, we utilized a dataset that encompassed Bitcoin's historical prices from January 1st, 2012 to March 31st, 2021. The original dataset consisted of 4,857,377 rows and 8 columns, featuring Timestamp, Open, High, Low, Close, Volume (BTC), Volume (Currency), and Weighted Price.

To prepare our dataset for modeling, we employed various techniques, and the process is shown in Figure 1. Initially, we used the df.dropna() function to handle missing values, resulting in a refined dataset with 3,376 rows and 9 columns. Subsequent data preparation steps were undertaken to enhance the dataset's suitability for analysis. Firstly, we transformed the 'Timestamp' column into a standardized date format, allowing us to convert the time units from seconds to days. This transformation facilitated better interpretation and analysis of the dataset. After the data is ready for analysis, we divide the dataset into training, validation, and testing sets based on the timeline. We classified Jan 1st, 2012 to Jan 1st, 2017 into a training dataset, assigned Jan 1st, 2017 to Jan 1st, 2019 into a validation dataset, and assigned Jan 1st, 2019 to Dec 31th, 2019 into a testing dataset. This division ensured that the model's training phase utilized historical data, while the testing phase evaluated the model's performance on later recent data. Considering the COVID-19 pandemic has significantly impacted global financial markets, as well as the cryptocurrency market, the data after 2020-01-01 was excluded in our analysis to increase model accuracy and mitigate the pandemic's influence.

Furthermore, we scaled the dataset and applied normalization to the Bitcoin price data. Given the wide range of Bitcoin prices, scaling the dataset brought all features to a similar scale, preventing any single feature from dominating the learning process. Additionally, scaling the data helped expedite the model training process.

Finally, to accommodate LSTM and RNN models, we reshaped the input sequences (x_train) and target values (y_train) into a 3D array format. LSTM and RNN models require input data in this format, with three dimensions: batch size (the number of samples processed in each batch), time steps (the number of historical data points considered for making predictions), and features (the number of variables considered in the input data). For our predictions, we selected 60 days as the time steps and 1 feature (close price).

By following these data preparation steps, we ensured that the dataset was appropriately cleaned, transformed, scaled, and reshaped, making it suitable for training and evaluating LSTM and RNN models on Bitcoin's historical price data.

5. Supervised Learning

Supervised learning techniques were applied to predict Bitcoin's price. By combining the knowledge learned from class, we incorporated Recurrent Neural Network model (RNN) and Long Short-Term Memory model (LSTM). Here are several main reasons for choosing the RNN and LSTM model.

1) Bitcoin price data is sequential.

   RNN and LSTM models are particularly well-suited for capturing and analyzing such sequential dependencies in the data. These models capture the temporal patterns and dynamics present in Bitcoin price data. While they do not directly incorporate factors like supply and demand or market competition, they can indirectly capture the influence of these factors through the analysis of historical price patterns and trends.

2) Bitcoin price displays nonlinear relationships.
   RNN and LSTM can capture the nonlinear relationships by learning from historical price patterns.

In the following models, the activation function "ReLu" will be applied due to its simplicity and ability to handle non-linear relationships, and the evaluation metrics "MAPE" will also be used to measure the average percentage difference between the predicted values and the actual values.

5.1 Recurrent Neural Network (RNN)

1) Method Explanation

We added three SimpleRNN layers in our model. By stacking multiple layers, the model can learn and extract increasingly complex patterns and dependencies from the sequential data. The third SimpleRNN layer with 32 units serves the purpose of reducing the dimensionality of the output. As the model progresses through the layers, the representation becomes more abstract and compressed. This reduction in dimensionality helps to capture the most relevant information and discard redundant or noise-like patterns. It can enhance the model's generalization ability and computational efficiency.

To prevent overfitting, we added a dropout layer with 0.2 dropout rate. Dropout is a regularization technique that randomly drops out a certain percentage of units (in our project, 20%) during training. By randomly dropping out 20% of units, the model is encouraged to prevent memorizing and relying too heavily on individual training examples.

The dense output layer is added to the RNN model to produce the final prediction. We used a dense output layer to predict the Bitcoin price.

2) Tuning Parameters for Finding Good Results

First, we revised the method for splitting the dataset into training, validation, and testing sets. Initially, we only had a training set and did not include a separate validation set. However, we observed that the mean absolute percentage error (MAPE) remained unchanged and did not decrease. To address this, we include a validation dataset into the analysis. The data from 2012-01-01 to 2017-01-01 was included into the training dataset, data from 2017-01-01 to 2019-01-01 was assigned into the validation dataset, and the data from 2019-01-01 to 2019-12-31 was put into the testing dataset.

Next, we focused on refining the architecture of the RNN model. We increased the number of RNN units in each layer to enhance the model's capacity to capture intricate patterns. Specifically, we changed the configuration from (64, 64, 2) to (500, 300, 200). This modification allowed our model to better handle the complex and volatile nature of Bitcoin price data.

Furthermore, we recognized the importance of providing the model with a longer historical context for prediction. Therefore, we expanded the number of time steps from 20 to 60. By including more historical data, the model captures more nuanced temporal patterns and improves its predictive accuracy.

Lastly, we calculate the loss for both the training and validation predictions and print the loss values. We found there is an overfitting problem, so we reduce the numbers of units in the layers.

5.2 Long Short-Term Memory (LSTM)

1) Method Explanation

To further improve the prediction accuracy, we try a different model of LSTM. The LSTM model implemented in the code incorporates multiple LSTM layers. LSTM layers are a specialized type of RNN layer designed to capture long-term dependencies and address the vanishing gradient problem often encountered in traditional RNNs.

Similar to SimpleRNN layers, each LSTM layer is capable of capturing sequential patterns and dependencies in the input data. By stacking multiple LSTM layers, the model can learn and extract increasingly complex patterns and dependencies from the sequential data.

2) Tuning Parameters for Finding Good Results

Firstly, we increased the number of epochs from 30 to 100, which would provide more training iterations for the model to learn from the data, and allow the model to further refine its weights and bias, potentially capturing more complex patterns in the data. Moreover, we adjusted the batch size from 32 to 50, a larger batch size can provide a more accurate estimate of the gradient, as it considers more samples before updating the model's weights. This can result in more stable and accurate weight updates, potentially improving the model's convergence and performance. Furthermore, we tried to switch the loss function from "mean_squared_error" to "mae" (mean absolute error) function. However, this change did not result in a significant decrease in MAPE.

At the end, these adjustments resulted in a reduction in the mean absolute percentage error (MAPE) in the validation set from 14 to 4, indicating improved accuracy in predicting Bitcoin's prices.

5.3 Drawbacks of Models

Firstly, incorporating additional variables into the dataset. Bitcoin prices are influenced by multiple factors, such as trading volume, market demand and supply, and competition. The expanded variables can improve the predictive capabilities and a more comprehensive analysis of Bitcoin price movements.

Secondly, expanding the dataset by including more historical price data. In current prediction, we transformed and aggregated the dataset, resulting in a reduced number of rows. By including a larger time span of historical data, the model can potentially capture more patterns in Bitcoin price movements, and improve the prediction accuracy.

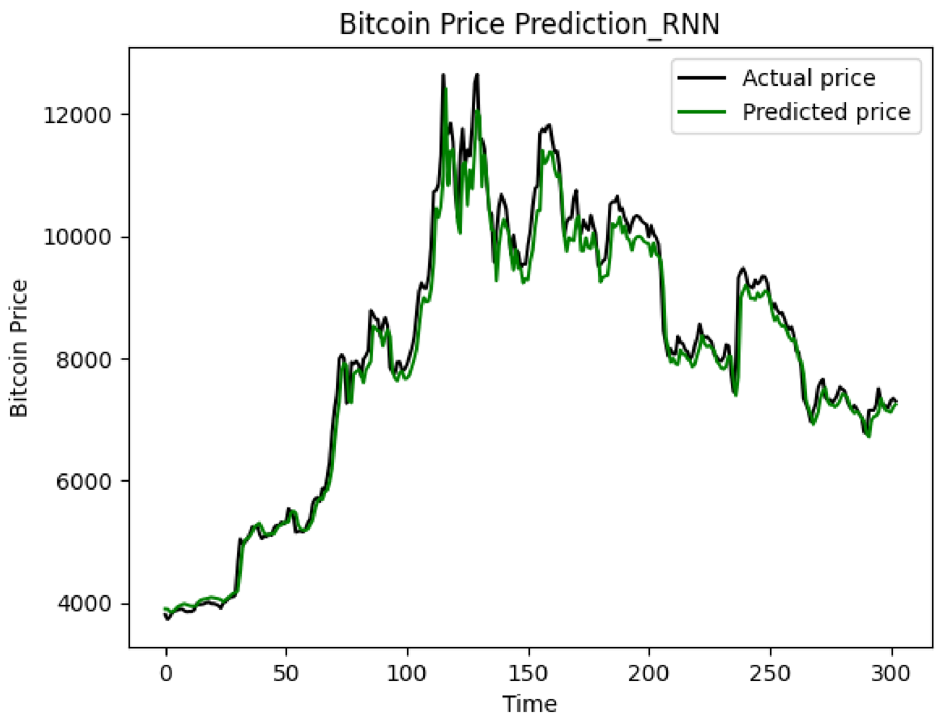
Furthermore, the COVID-19 pandemic has significantly impacted global financial markets, including the cryptocurrency market. To increase model accuracy and mitigate the pandemic's influence, data after 2020-01-01 was excluded. However, this reduction in the dataset limits the models' ability to capture the price movements on Bitcoin prices. Future research could use more powerful computing resources, such as GPUs, to analyze higher frequency price data and gain deeper insights into the Bitcoin prices' movement.

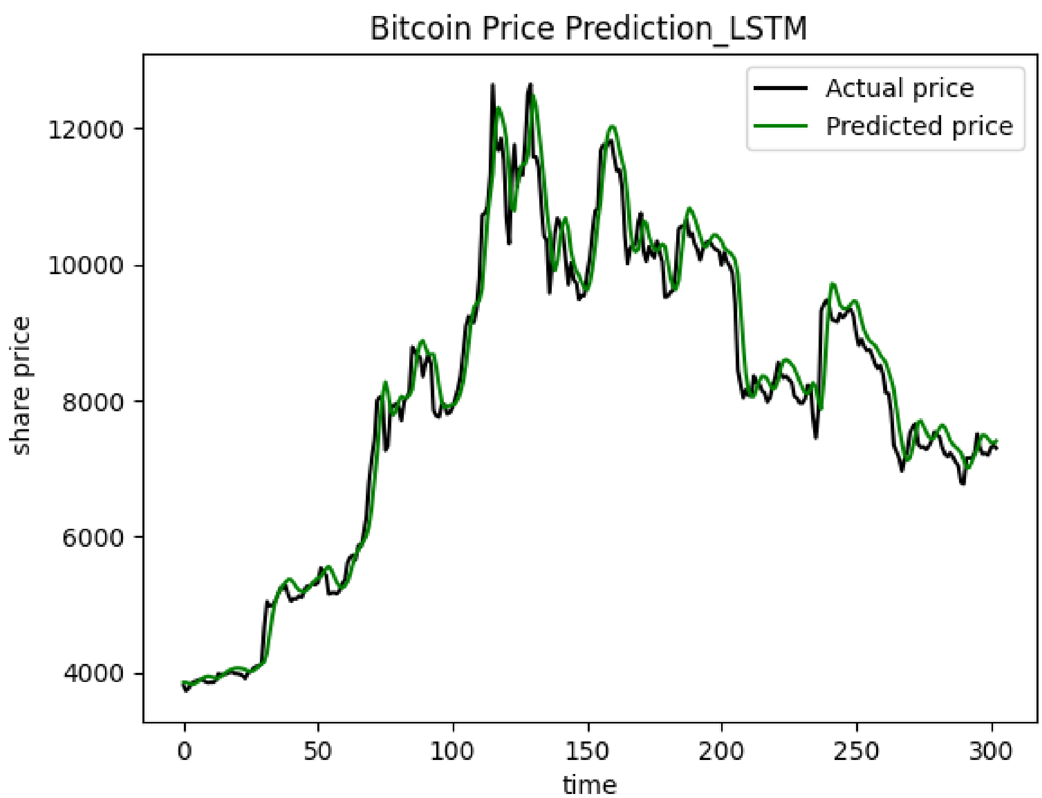
6. Models Performance

In evaluating the performance of the RNN and LSTM models on the validation dataset, the results show that the LSTM model outperformed the RNN model. The LSTM model achieved a lower mean absolute percentage error (MAE) of 0.215, compared to the RNN model's MAE of 0.317.

The superior performance of the LSTM model can be attributed to its ability to capture longer-term dependencies in the data. LSTM models have specialized memory cells and gating mechanisms that enable them to retain and utilize information over longer sequences. This characteristic makes them well-suited for tasks where capturing and understanding temporal patterns is crucial, such as time series forecasting.

The lower MAPE value of the LSTM model indicates that it had a smaller average percentage difference between its predictions and the actual values in the validation dataset. This suggests that the LSTM model was more accurate in predicting the target variable compared to the RNN model.





Model	Loss of the Validation	Mape	MAE

	dataset		
RNN	0.1075	79.0768	0.317
LSTM	9.6430e-04	4.3487	0.215

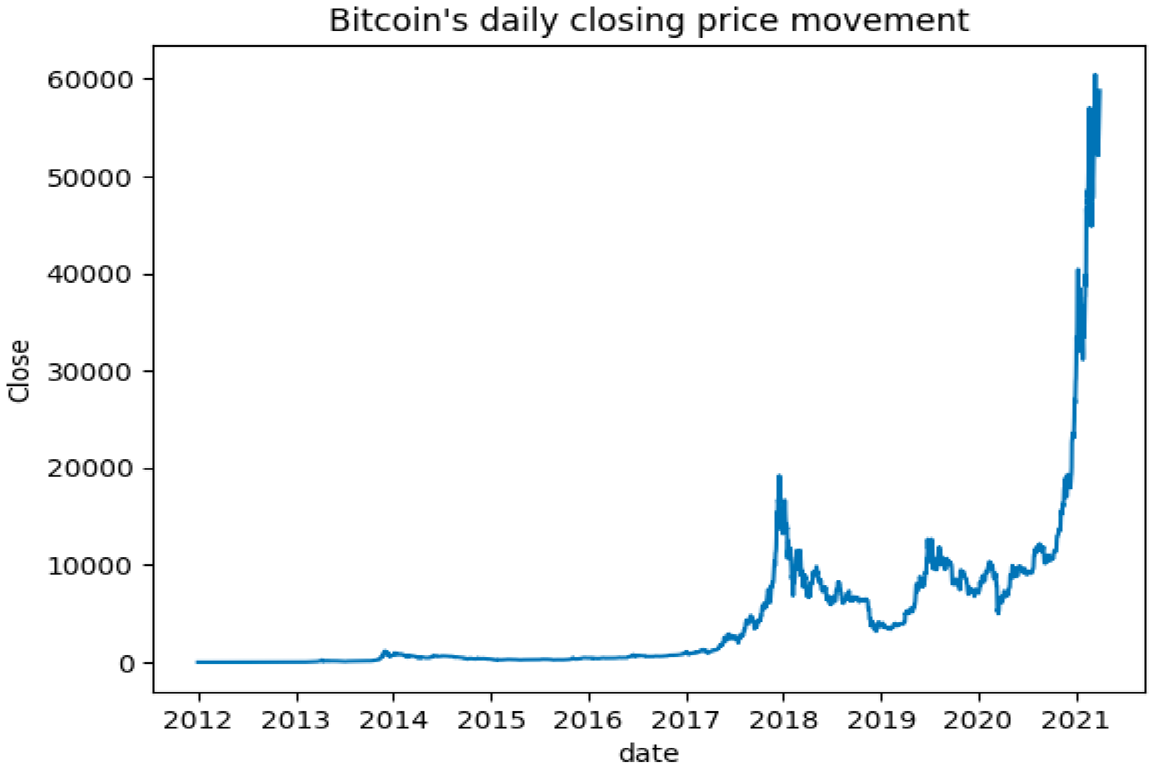
7. Conclusion, Discussion and Future Work

Based on our analysis, we found that both RNN model and LSTM model perform well, but in terms of accuracy, the LSTM model outperformed the RNN model. The MAPE rate in the simpleRNN model is 79.0768, while the MAPE rate in the LSTM model is 4.3487, indicating LSTM's ability to make more accurate predictions.

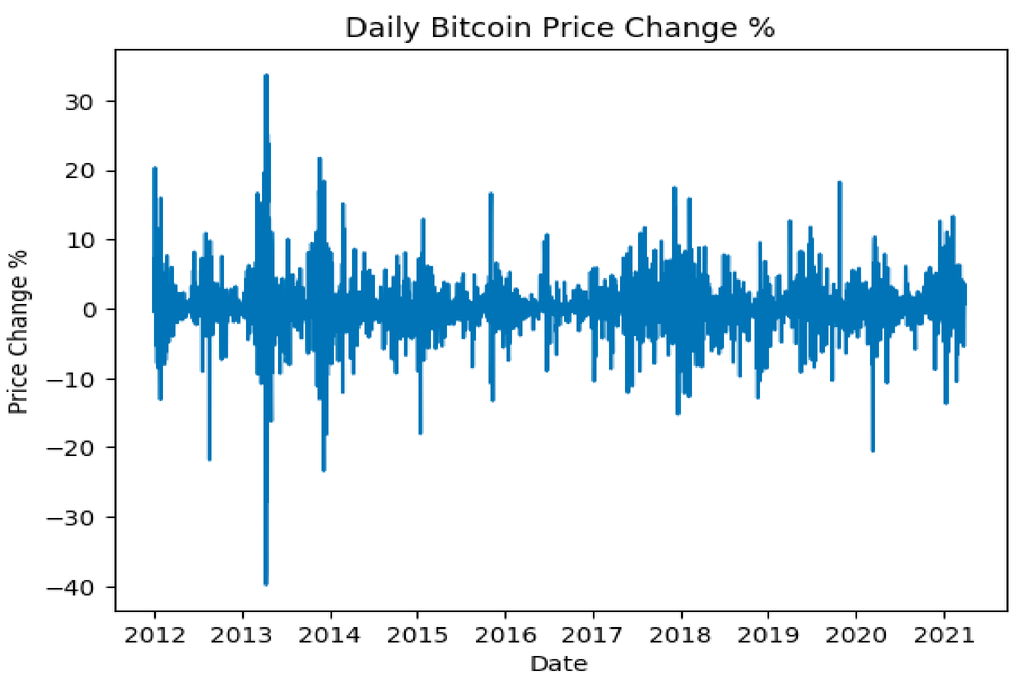
Overall, the LSTM model emerged as the most recommended model due to its lower MAPE and ability to generate more accurate predictions. Its suitability for predicting time series data, coupled with its capability to capture patterns and handle long-term dependencies, further strengthens its appeal. For resolving business challenges pertaining to Bitcoin and making daily price predictions, we highly recommend utilizing our LSTM model. With its impressive accuracy, the model serves as a valuable tool for providing business insights into market trends, facilitating informed decision-making, and adjusting investment strategies for our users. Furthermore, integrating this model with other tools can enhance decision-making capabilities by offering users access to a wider range of resources and information.

While our study focuses on exploring the application of deep learning algorithms, specifically RNN and LSTM, in cryptocurrency price prediction, it highlights the validity and potential of these techniques in the market. However, it is crucial to recognize that predicting Bitcoin's daily prices is a complex task that involves numerous factors beyond the scope of any single model. Therefore, while our findings demonstrate promising results, it is essential to approach Bitcoin price prediction with caution and consider additional factors and analysis techniques to make informed decisions in the dynamic cryptocurrency market.

Appendix



Graph 1



Graph 2

References

Tealab, A. (2018, November 15). *Time series forecasting using Artificial Neural Networks Methodologies: A systematic review.* ScienceDirect. https://www.sciencedirect.com/science/article/pii/S2314728817300715

ÖZTEMIZ, M. (2021, November 14). *Bitcoin price prediction with RNN*. Kaggle. https://www.kaggle.com/code/mustiztemiz/bitcoin-price-prediction-with-rnn

Panchal, D., Kapur, I., & Prakash, S. (2021, January 10). *BTC-price-prediction-ml-project.* GitHub. https://github.com/dushyant18033/BTC-Price-Prediction-ML-Project