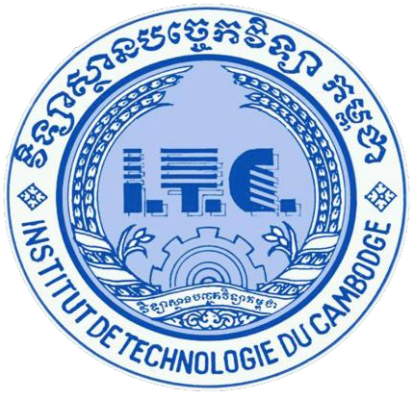

AIMS
Department of Applied Mathematics and Statistics

Institute of Technology of Cambodia

Data Ethics and Privacy

2024-2025

4th Year of Engineering Degree in Data Science

Department of Applied Mathematics and Statistics

Bitcoin Price Prediction Using Machine Learning: RNN and Ensemble

Group members:

Name	ID
Chhon Menghout	e20211474
Dok Dominique	e20210337

Lecturers:

Mr. Sok Kimheng (Course)

Dr. Neang Pheak (TP)

Dr. Phauk Sokkhey (TP)

Bitcoin Price Prediction Using Machine Learning: RNN and Ensemble

Chhon Menghout[1*], Dok Domonique[2*], Neang Pheak[3] Phauk Sokkhey[3] and Sok Kimheng[3]

[1,2,3] Department of Applied Mathematics and Statistics, Institute of Technology of Cambodia, Cambodia

Abstract

This study looks at Bitcoin price prediction using machine learning approaches, specifically comparing the Long Short-Term Memory (LSTM) and XGBoost models. Bitcoin's intrinsic volatility and market complexity pose considerable obstacles to effective forecasting, frequently beyond the capabilities of typical time-series models. To overcome these issues, the study uses advanced deep learning algorithms like LSTM and a robust gradient-boosting algorithm like XGBoost. Historical Bitcoin price data was preprocessed, windowed, and then used to train and test both models. Performance was evaluated utilizing important measures such as MAE, MSE, RMSE, MAPE, and MASE to offer a thorough assessment of prediction accuracy. The findings emphasize LSTM's better capacity to capture temporal dependencies and model non-linear interactions, while XGBoost uses its expertise in feature-based learning to deliver competitive results. This study illustrates the effectiveness of both deep learning and gradient boosting strategies for making sound financial decisions in unpredictable markets.

Table of Contents

List of Figures

List of Tables

Acknowledgement

I would like to extend my heartfelt gratitude to the Department of Applied Mathematics and Statistics at the Institute of Technology of Cambodia for their invaluable support, guidance, and resources throughout this project. Their expertise and encouragement have been instrumental in shaping this research and bringing it to fruition. I am particularly grateful to my professors and mentors, whose insights have guided me at every stage. Furthermore, I wish to express my deepest appreciation to my family for their generous financial support, without which this project would not have been possible. Their commitment to my education and unwavering encouragement have provided me with the stability and motivation necessary to pursue this research with dedication and purpose. This project is a testament to the collaborative spirit and support of both my academic institution and my family, and I am sincerely thankful for their contributions. This research would have never been completely done without these important supporters.

1. Introduction

1.1. Background of project

Bitcoin is an electronic currency that has become increasingly popular since its introduction in 2008. Transactions in the bitcoin system are stored in a public transaction ledger ('the blockchain'), which is stored in a decentralized, peer-to-peer network. Bitcoin provides decentralized currency issuance and transaction clearance. The security of the blockchain depends on a compute-intensive algorithm for bitcoin mining, which prevents double spending of bitcoins and tampering with confirmed transactions.[i]
Since cryptocurrency market, with Bitcoin at its forefront, is known for its significant volatility, which poses a challenge for price prediction for each researcher and investor. While traditional models are often used for time-series forecasting, they struggle with the dynamic nature of Bitcoin prices. So, the aim of this project is to apply machine learning techniques specifically Long Short-Term Memory (LSTM)[ii], and XGBoost[iii] models to predict Bitcoin prices more accurately by capturing patterns that traditional methods might miss.

1.2. Statement of Problem

Existing Bitcoin price prediction models, particularly traditional approaches, often fail to capture the complexity and non-linearity inherent in Bitcoin's price behavior, leading to inaccurate forecasts. This study aims to address these challenges by comparing Long Short-Term Memory (LSTM) and XGBoost models, providing a comprehensive evaluation of their forecasting capabilities. The research explores several key questions: How effective are LSTM and XGBoost models in predicting Bitcoin prices? This question seeks to assess the forecasting accuracy of both models in predicting Bitcoin price movements. What are the key features contributing to the accuracy of each model? This aims to identify the most significant variables or factors influencing the performance of LSTM and XGBoost in Bitcoin price prediction. Finally, how do LSTM and XGBoost compare in terms of performance metrics? This question focuses on comparing the two models based on key performance metrics such as MAE, MSE, RMSE, MAPE, and MASE to evaluate their relative effectiveness in forecasting Bitcoin prices.

1.3. Purpose of study

The study begins with data manipulation, which involves collecting, preprocessing, and analyzing historical Bitcoin price data to ensure it is properly formatted and arranged for training and evaluation. Next, prediction models are developed and trained using LSTM and XGBoost, leveraging their strengths in processing sequential time-series data to effectively capture Bitcoin's complex price trends and high volatility patterns. Finally, model performance is evaluated through a thorough comparative analysis, benchmarking the models' forecasting accuracy by assessing their performance using key metrics and visualizing the results to validate each model's effectiveness.

1.4. Scope and Limitation

1.4.1. Scope

This study focuses on Bitcoin price prediction using machine learning techniques. It involves the development and evaluation of LSTM, and XGBoost models based on historical price data. We've collected the historical price of Bitcoin from 2012 to 2024. This data contains 4707 rows and there are columns such as Start and End which are date, Close, Open, High and Low, which are bitcoin price, Volume, and Market cap. The study also analyzes key features and compares the performance of the models.

1.4.2.Limitation

Data dependency, the accuracy of the model is highly dependent on the quality and availability of historical Bitcoin data. Incomplete datasets, missing values, or external anomalies can negatively impact the model's performance and forecasting accuracy. Market volatility, the unpredictable nature of cryptocurrency markets poses a challenge for the model. Extreme price fluctuations or sudden market shocks, such as regulatory changes or macroeconomic events, can lead to unpredictable shifts in price trends, reducing the model's ability to forecast accurately during such volatile periods. Computational intensity, deep learning models, such as LSTM, require significant computational resources, particularly for training on large datasets. This can make the modeling process resource-intensive and time-consuming, requiring specialized hardware and optimization techniques. Single asset focus, this study focuses exclusively on Bitcoin, which may limit the generalizability of the findings to other cryptocurrencies. Different cryptocurrencies may have unique market dynamics, such as liquidity levels, regulatory considerations, or investor behavior, that could affect their price predictions differently.

2. Literature Review

Vinay Karnati, Lakshmi Dathatreya Kanna, and Trilok Nath Pandey has studied on the performance of the Facebook Prophet, ARIMA, SVM, and LSTM algorithms for time series forecasting and training. Their methods included preprocessing the data, running the algorithms, and evaluating their efficacy with appropriate metrics. The data is collected from the Yahoo finance website dated from year 2013 to 2021. This data contains attributes such as date, highest price, lowest Price, opening price, closing price and market cap on those particular days.[iv] According to their report, LSTM has the best performance following by Prophet, SVM, and ARIMA, respectively.

In the study conducted by Junwei Chen, the objective was to develop an algorithmic model with high predictive accuracy for the next-day price of Bitcoin using random forest regression and LSTM. The study also aimed to identify the variables influencing Bitcoin prices. Previous research on Bitcoin price prediction has predominantly utilized time series ARMA models and deep learning LSTM algorithms. Although the Diebold–Mariano test did not conclusively show that random forest regression outperforms LSTM in prediction accuracy, the RMSE and MAPE errors for random forest regression were lower than those for LSTM. The study utilized eight categories (47 variables) as explanatory variables: Bitcoin price variables, specific technical features of Bitcoin, other cryptocurrencies, commodities, market indices, foreign exchange, public attention, and dummy variables of the week. Random forest regression demonstrated better price prediction accuracy than LSTM, despite its limitation in predicting values not present in the training samples. The project was successfully completed, demonstrating that random forest regression can achieve higher prediction accuracy than LSTM for next-day Bitcoin price prediction, despite certain limitations.[v]

In the study conducted by Yaowen Hu, the objective was to perform Bitcoin price prediction using various machine learning models, including Support Vector Machine (SVM), Random Forest, Neural Network, XGBoost, and LightGBM. The Bitcoin price dataset was divided into training and test sets in a ratio of 7:3. The models were trained with the training set and tested with the test set, using stock price change (yield) as the target variable and other variables as input variables. By comparing the MSE, RMSE, MAE, MAPE, and R^2 of the different models, it was found that XGBoost had the best performance and prediction accuracy. The performance of the other four models ranged from good to poor, with LightGBM, Random Forest, SVM, and Neural Network following in descending order of accuracy. The Neural Network model performed the worst, with an MSE significantly higher than the other models. The research results provide valuable reference for future Bitcoin price prediction and for choosing appropriate machine learning models. The source of dataset is CoinDesk.[vi] The dataset has several variables such as Open time, Open Price, Close Price, High Price, Low Price, Volume, Close Time, Quote asset volume, Number of traders, and Taker buy base asset volume.

Yangyu Chen's case study provides an in-depth examination of Bitcoin, exploring its origins, technological foundations, market behavior, trading strategies, and regulatory considerations. The study highlights Bitcoin's

transformative journey from its inception to its current role as a dynamic digital asset, emphasizing the revolutionary impact of blockchain technology. Blockchain Innovation, bitcoin's introduction of blockchain technology has spurred innovations in various industries beyond cryptocurrencies, such as supply chain management and healthcare. Market Behavior, bitcoin exhibits pronounced price volatility and correlations with traditional assets, serving as both a speculative asset and a potential hedge against economic uncertainties. Trading Strategies, the bitcoin market employs diverse trading strategies, including algorithmic trading, sentiment analysis, and high-frequency trading. Cryptocurrency derivatives like futures and options offer new avenues for speculation and risk management. The study also discusses the future of Bitcoin, highlighting the potential for mainstream adoption, increased institutional investment, and broader acceptance. Regulatory clarity and investor protection measures will be crucial in this process. Additionally, blockchain technology is expected to drive further innovations in various sectors.[vii] The project was effectively completed, providing valuable insights into Bitcoin's significance in the digital age and offering a foundation for future research, policymaking, and informed decision-making in the evolving world of cryptocurrencies.

The study conducted by Saber Talazade and Dragan Peraković introduces a novel approach to stock market prediction by integrating sentiment analysis with the traditional Random Forest model. This new methodology, termed "Sentiment-Augmented Random Forest" (SARF), leverages the nuanced understanding of financial sentiments provided by the FinGPT generative AI model to optimize the accuracy of stock price forecasts. The proposed SARF technique incorporates sentiment features into the Random Forest framework. Experiments conducted in the study demonstrate that SARF outperforms conventional Random Forest and LSTM models, with an average accuracy improvement of 9.23% and lower prediction errors in predicting stock market movements. The promising results indicate the potential of this approach for real-world applications in financial forecasting. Key findings from the experiments such as S&P 500: SARF achieved an accuracy of 0.78, compared to 0.67 for traditional Random Forest and 0.58 for LSTM, Nasdaq: SARF achieved an accuracy of 0.85, compared to 0.64 for traditional Random Forest and 0.69 for LSTM, Dow Jones: SARF achieved an accuracy of 0.82, compared to 0.59 for traditional Random Forest and 0.61 for LSTM. Future research will explore the scalability of SARF to handle larger datasets, investigate additional sentiment features using other large language models (LLMs) in the financial domain, and assess its performance in diverse market conditions. Additionally, the integration of real-time sentiment analysis could further enhance the model's responsiveness to dynamic market changes. Machine learning optimization techniques will also be employed to tune the hyperparameters and improve the model's prediction accuracy further.[viii] The project was completed with success, demonstrating the effectiveness of combining sentiment analysis with FinGPT and an optimized Random Forest model for enhancing stock market prediction accuracy.

Yao Yao's study, "Predicting Stock Prices Using The RF-LSTM Combination Model," addresses the challenge of predicting stock prices given the nonlinearity and complexity of stock data. The study proposes a combined Random Forest (RF) and Long-Short Term Memory (LSTM) model to forecast stock closing prices. Stock data was sourced from the Kaggle platform, feature sets were constructed, and data was normalized. Due to the high nonlinearity and information redundancy among multiple features, the optimal feature set was selected using RF to reduce data dimensionality and training complexity. Subsequently, an LSTM network, capable of processing time series data in deep learning, was used to forecast stock prices. The results showed that the RF-LSTM combined model outperformed the single LSTM neural network model, with reductions in root mean squared error (RMSE) and mean squared error (MSE) of 4.14% and 1.00%, respectively. This combined model increased the prediction accuracy of stock prices by incorporating technical indicators linked to stocks, handling dimensionality reduction of the model input data, and simplifying the network structure. Improved Prediction Accuracy, the RF-LSTM model demonstrated higher prediction effectiveness compared to the LSTM model alone, with more stable and accurate prediction findings. Handling Data Complexity, the RF component effectively reduced data dimensionality and training complexity, while the LSTM component processed time series data to forecast stock prices. Technical Indicators, incorporating technical indicators into the RF-LSTM model helped in thoroughly examining the data's latent information and reflecting stock price trends.[ix] The project was successfully finalized,

demonstrating that the RF-LSTM combination model offers a higher prediction accuracy and effectiveness in forecasting stock prices compared to the single LSTM model.

3. Data Preprocessing

Data preparation refers to the act of converting unstructured data into a structured format. It is an initial and very crucial phase in data mining. It is important to evaluate the quality of data before training the models. The initial phases in data preparation include data collection, data cleaning, data reduction, and data transformation.

3.1. Data Collection

The data is collected from the **coincondex** website dated from year 2012 to 2024. This data contains attributes such as date, highest price, lowest Price, opening price, closing price, volume, and market cap on those particular days. [x]

3.2. Data Cleaning

Data cleaning refers to the process of handling missing values and removing erroneous, incomplete, or inaccurate data from time series datasets. This involves ensuring the data is reliable and ready for analysis. When dealing with missing values, one can either remove them entirely or apply imputation techniques, such as replacing them with the mean, median, or mode of the dataset. This step is crucial for maintaining the integrity and consistency of the time series data.

3.3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial initial step in data science projects. It involves analyzing and visualizing data to understand its key characteristics, uncover patterns, and identify relationships between variables refers to the method of studying and exploring record sets to apprehend their predominant traits, discover patterns, locate outliers, and identify relationships between variables. EDA is normally carried out as a preliminary step before undertaking extra formal statistical analyses or modeling.[xi] So, EDA is a crucial step in understanding and interpreting Bitcoin time series data. This phase involves visualizing and summarizing the data to identify underlying patterns, trends, and anomalies.

3.3.1. Summary Statistics

	Open	High	Low	Close	Volume	Market Cap
count	4707.000000	4707.000000	4707.000000	4707.000000	4.707000e+03	4.707000e+03
mean	14791.122668	15125.236917	14448.993594	14812.040377	2.608945e+10	2.806215e+11
std	19774.971069	20211.038246	19339.952351	19808.898442	3.684074e+10	3.841040e+11
min	4.222000	4.222000	4.222000	4.222000	0.000000e+00	3.525001e+07
25%	416.477000	421.687500	410.237500	416.583500	2.616035e+07	5.788002e+09
50%	6328.687636	6434.372794	6232.734067	6330.669843	3.950140e+09	1.099704e+11
75%	23595.945829	24158.389102	23120.778738	23650.821667	4.314177e+10	4.502643e+11
max	91135.390000	93836.940000	90406.060000	92119.470000	2.121958e+11	1.818537e+12

Figure 3.3.1.: Descriptive Statistics

Insights and observations from the data analysis reveal several key trends. First, the large standard deviations in the Open, High, Low, and Close prices highlight Bitcoin's significant price volatility over time, reflecting the unpredictable nature of the cryptocurrency market. Additionally, anomalies such as the minimum values of Volume (0) and Market Cap (~$35.25 million) may point to data entry errors, missing data for specific time periods, or legitimate instances of low activity, especially in Bitcoin's early history. Furthermore, the 25th and 75th percentiles indicate substantial growth over time, with prices shifting from hundreds of dollars ($416) in the early stages to tens of thousands of dollars ($23,650) in later periods, illustrating the cryptocurrency's dramatic rise in value.

3.3.2. Time series analysis of Bitcoin Price

The line plot illustrates the monthly trends in Bitcoin prices over time, providing insights into the cryptocurrency's price dynamics.

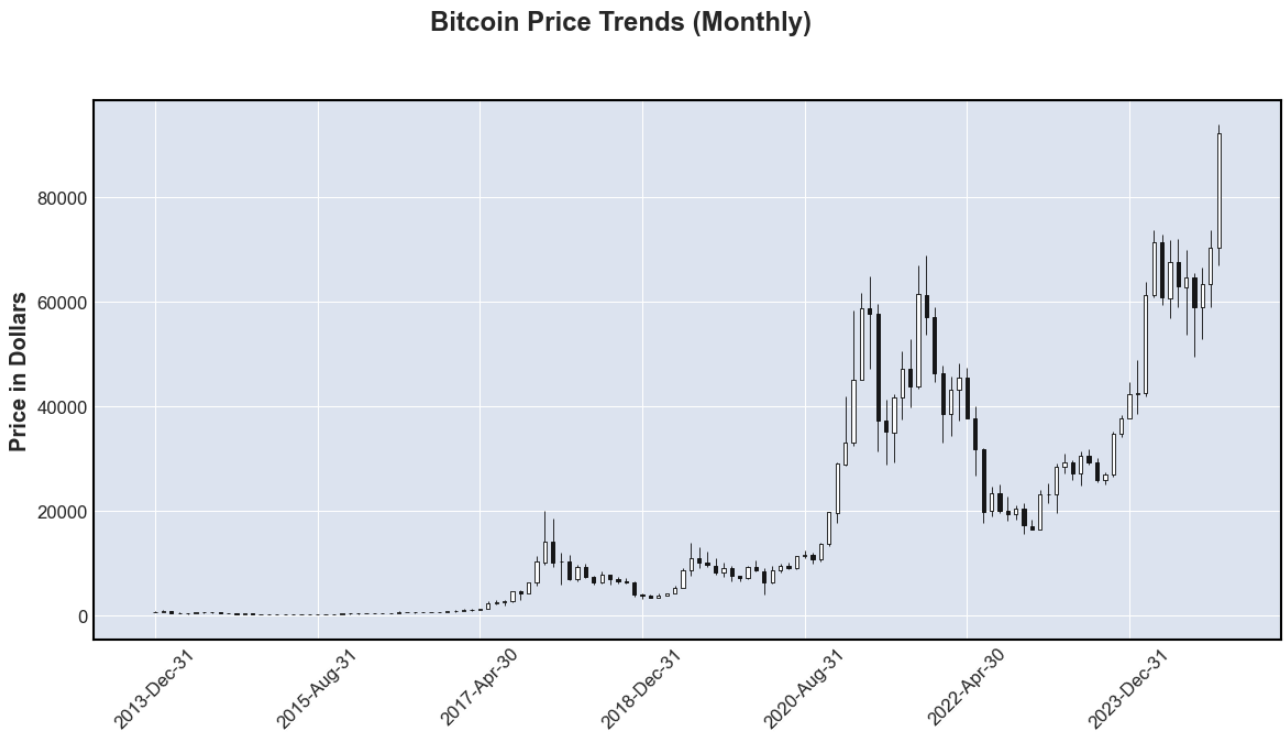

Figure 3.3.2.: Time-series chart

The chart illustrates several key observations regarding Bitcoin's price trends. First, a long-term growth trend is evident, particularly after 2017, suggesting an increase in market adoption, speculative interest, and potential use as an inflation hedge. The chart also highlights periods of significant volatility, characterized by sharp price increases followed by steep declines, such as the rapid surge and subsequent correction between late 2017 and early 2018, and another major price jump from mid-2020 to 2021, followed by a subsequent correction. In addition, recent peaks in Bitcoin's price have seen all-time highs surpassing $90,000, likely influenced by broader market events, institutional investor adoption, and macroeconomic factors. Furthermore, the chart reflects the market's growing maturity over time, as earlier periods feature relatively stable prices, while later periods are marked by extreme fluctuations, indicating increasing investor participation and market evolution. Overall, this visualization underscores the importance of considering market volatility and external influences when developing predictive models for Bitcoin prices.

3.3.3. Distribution of Bitcoin Close Price

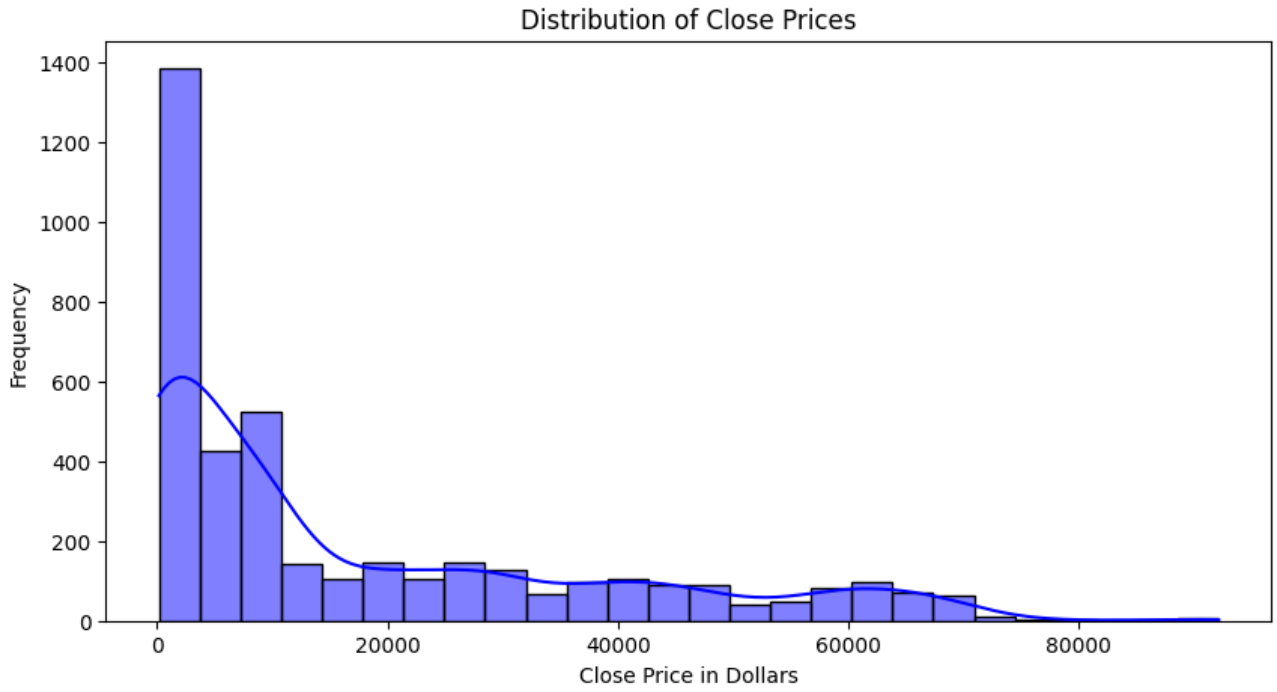

Figure 3.3.3.: Close Price Distribution

The histogram provides insights into the distribution of Bitcoin's closing prices during the analyzed period. The data shows a highly skewed distribution with a strong positive skew, where the majority of closing prices are concentrated in the lower price range, particularly below $20,000. This suggests that Bitcoin's price has predominantly remained at lower levels throughout the period. As the closing prices rise, the frequency of occurrences decreases significantly, indicating that periods of higher Bitcoin prices, such as above $40,000, were less frequent. Additionally, the long tail of the distribution reflects occasional price spikes, with rare instances reaching $60,000 or higher, likely driven by major market events or speculative trends. These occasional high prices might represent potential outliers, highlighting periods of exceptional market activity that warrant further exploration.

3.3.4. Correlation Matrix

This heatmap illustrates the correlation matrix of Bitcoin-related variables, highlighting the strength of linear relationships between them.

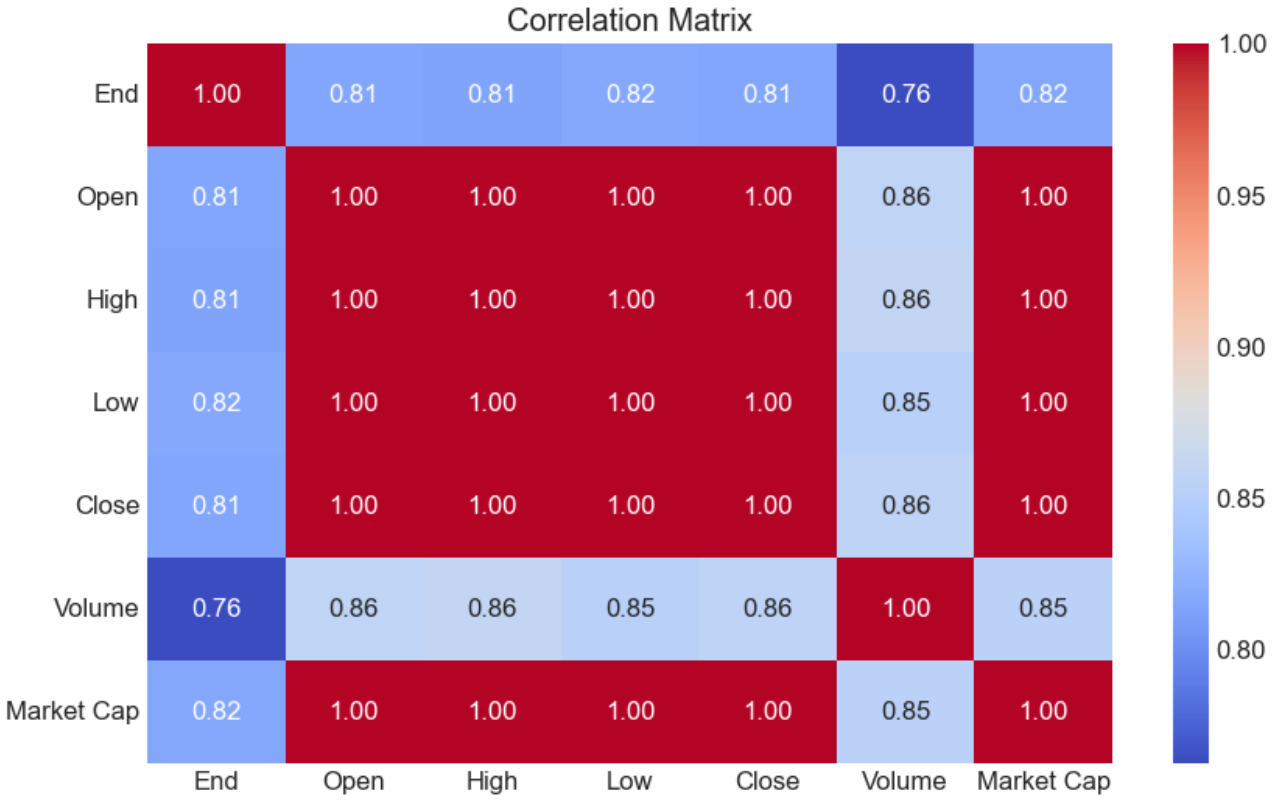

Figure 3.3.4.: Heatmap

The heatmap reveals important insights into the relationships between Bitcoin's price and market metrics. It shows strong correlations among the Open, High, Low, and Close prices, which are perfectly correlated (correlation coefficient = 1.00), as these values typically move together within a given time frame. Market Capitalization also demonstrates a perfect correlation (1.00) with the price variables, reflecting that it is inherently derived from these metrics. On the other hand, Volume has a moderate correlation with the End price (0.76), suggesting that trading volume does not always align perfectly with price movements. The color gradient in the heatmap helps interpret these relationships, with darker red areas indicating higher correlations and blue areas, such as those related to Volume, representing weaker associations. This visualization emphasizes the need for careful feature selection in predictive modeling to avoid multicollinearity, especially with highly correlated variables like Open, High, Low, and Close.

4. Model Implementation Time Series Forecasting

Time series forecasting involves predicting future values based on historical data. In this project, we employed a Long Short-Term Memory (LSTM) and XGBoost models to forecast values in a univariate time series dataset. The approach included windowing the dataset, building and training both models, and evaluating their performance on test data.

4.1.1. Data Preparation

4.1.1.1. Windowing Dataset

Windowing transforms a time series dataset into a supervised learning problem by creating labeled windows of historical data to predict future values. **Horizon:** The number of future steps to predict (set to 1) and **Window Size:** The number of past timesteps used for prediction (set to 7). For example: Input: [1, 2, 3, 4, 5,

6] and Output: ([1, 2, 3, 4, 5], [6]). A function, get_labelled_windows, was created to label the windowed dataset.

4.1.1.2. Creating Windows

The make_windows function was implemented to convert a 1D array of time series data into sequential windows of size WINDOW_SIZE. The output included both windows (inputs) and their corresponding labels (targets). (Total Windows: 4,700) Example Output: **First Window:** [1, 2, 3, 4, 5, 6, 7] → Label: [8] and **Last Window:** [n-7, n-6, ..., n] → Label: [n+1]

4.1.1.3. Train-Test Split

The dataset was split into training and test sets using an 80-20 split: Training Windows and Test Windows are 3,760 and 940, respectively.

4.1.2. Model Development

4.1.2.1. LSTM

Neural networks are useful for bitcoin price prediction because of their capacity to capture complicated patterns and nonlinear correlations in data, allowing them to learn from previous trends and create accurate projections based on many factors impacting bitcoin pricing [xii]. Long Short-Term Memory is a type of Recurrent neural network architecture which is effective in handling the issue of Vanishing gradients in traditional RNN's [xiii]. It is considered as an upgraded version of RNN model [xiv]. LSTM layers are categorized into 3 main layers. They are input layer, LSTM layer and output layer [xv]. Input layer basically takes the input data which can be in any format such as text, audio or any other time series data and the input data is preprocessed and encoded. [xvi] LSTM layer is responsible for long-term dependencies in the input data and LSTM layer consists of cells where each cell consists of 3 gates such as Input gate, this gate has the ability to control information. It tells whether to let new information enter the cell or not, forget gate according to the requirement this gate forgets the information in the previous state, and Output gate, it tells which part of the cell state should be output as the LSTM's hidden state. The output layer takes the final output and provides us with predicted results. The output layer consists of SoftMax activation function which is used to solve classification problems and for regression problems we use linear activation function [xvii]. Training Approach, LSTM models were trained using historical Bitcoin price data as input sequences and the corresponding target values. The objective was to capture long-term dependencies and temporal patterns in the data. Optimization Algorithm, LSTM models employ optimization algorithms like Stochastic Gradient Descent (SGD) or Adam to update the network weights and minimize the loss function. Hyperparameter Tuning, Hyperparameters such as the number of LSTM layers, the number of hidden units, the learning rate, and the batch size were tuned to optimize the model's performance.

4.1.2.1.1. LSTM Model Architecture

The LSTM model was designed to handle sequential data with the architecture such as Input Layer, it accepts data with a shape of (7, 1), LSTM Layer is A single LSTM layer with 128 units and ReLU activation, and Dense Layer is a fully connected layer to produce a single output.

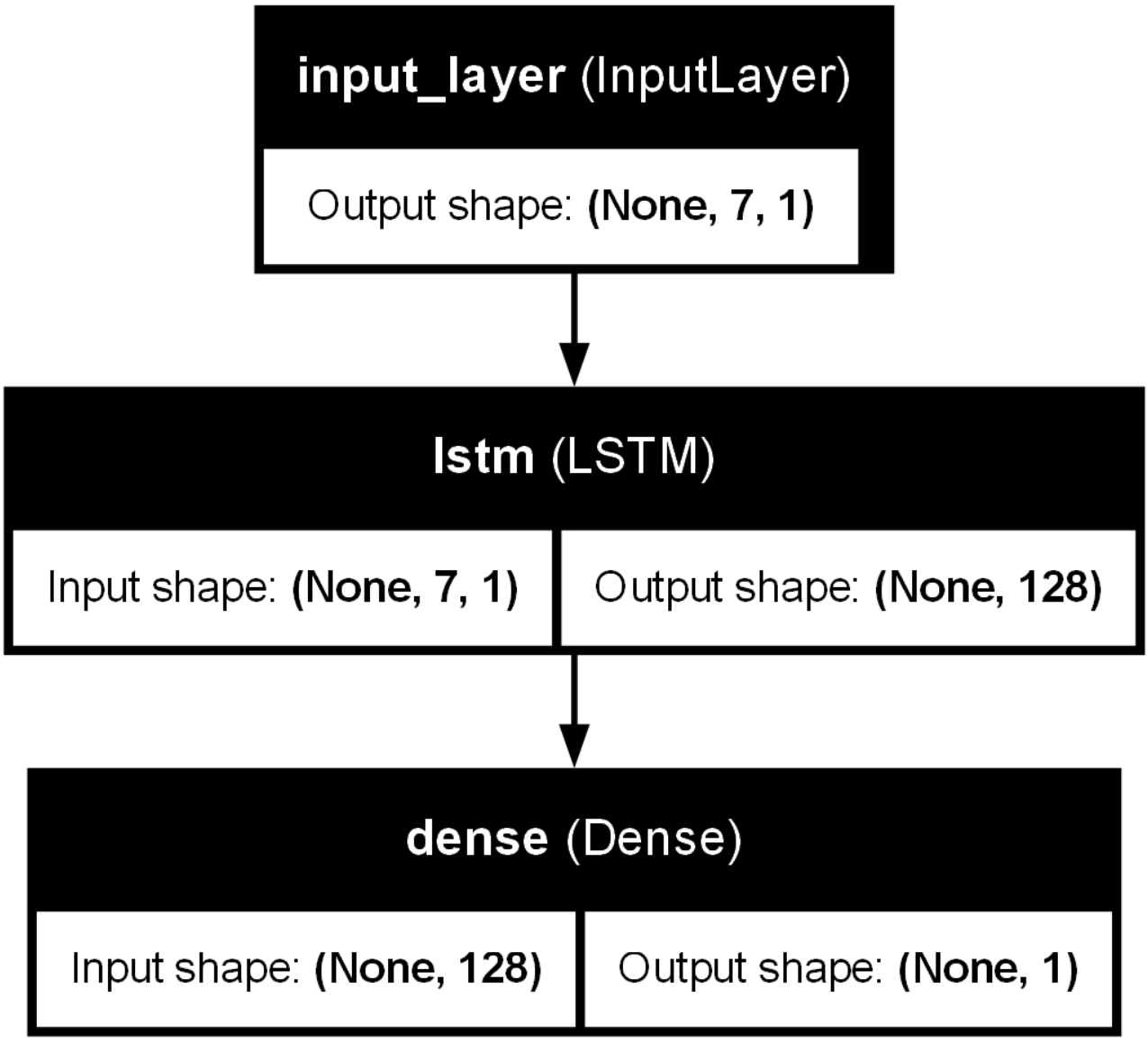

Figure 4.1.2.1.2: Long Short-term Memory Model's Architecture

4.1.3. XGBoost

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way. The same code runs on major distributed environment (Hadoop, SGE, MPI) and can solve problems beyond billions of examples.[xviii] The XGBoost model, a tree-based ensemble method, was also applied to the Bitcoin price prediction task. While it does not inherently model temporal dependencies like LSTM, its gradient-boosted framework excels at capturing relationships between features.

4.1.3.1. XGBoost Model's Architecture

The XGBoost model was implemented as an alternative to the LSTM model for time series forecasting with specific configurations to optimize its performance. The dataset was first converted into the DMatrix format for efficient processing. The objective function used was 'reg:squarederror', which is appropriate for regression tasks. Key hyperparameters included a maximum tree depth of 6, which controls the complexity of

the individual trees, and a learning rate (eta) of 0.05, slightly increased to enhance model convergence. To improve diversity and reduce overfitting, a subsample of 0.8 was applied to the rows and a colsample_bytree of 0.8 was used for the features. Regularization was incorporated with an L2 regularization term (lambda) of 1.0 and an L1 regularization term (alpha) of 0.1 to prevent overfitting. Lastly, a random seed of 42 was set for reproducibility of the results. These settings were chosen to balance model complexity, prevent overfitting, and optimize performance in predicting Bitcoin prices.

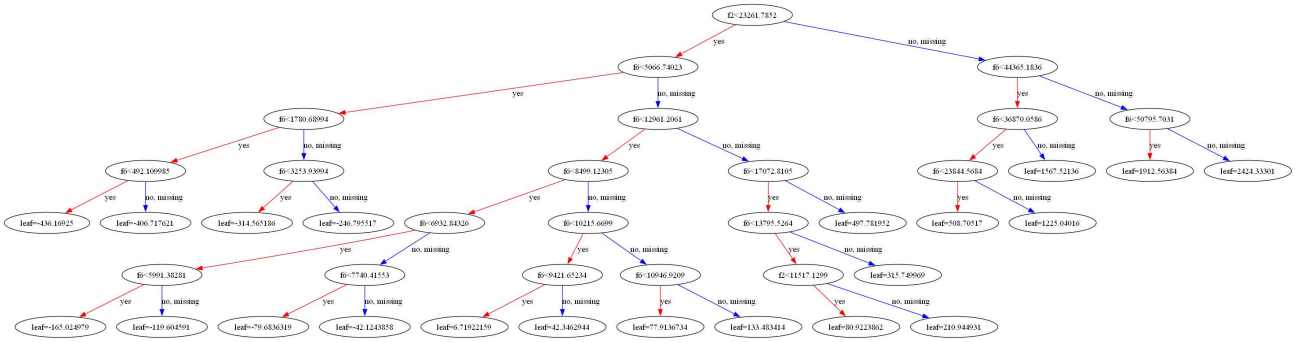

Figure 4.1.3.3: XGBoost Model's Architecture

4.1.3.2. Training Process

4.1.3.2.1. LSTM Training Process

The LSTM training process utilizes several key components to optimize its performance. The loss function used is Mean Absolute Error (MAE), which measures the average of the absolute errors between predicted and actual values. MAE is commonly employed in regression tasks and is less sensitive to outliers compared to other loss functions like Mean Squared Error (MSE). The Adam optimizer is chosen for adjusting the model's weights during training, as it combines the benefits of AdaGrad and RMSProp, providing an efficient method for optimizing the model's parameters. The model is trained over 300 epochs, where an epoch represents a full pass through the entire training dataset. A batch size of 32 is used, meaning the dataset is divided into smaller batches of 32 samples, with the model's weights updated after each batch to improve computational efficiency and convergence speed. Additionally, the ModelCheckpoint callback is employed to save the model either after every epoch or at the best point in training, preventing the loss of the best-performing model and facilitating later testing or deployment. A custom callback, PrintEpochPerformance, is used to log or print the model's performance after each epoch, offering insights into the training process and helping to determine when to stop training or make adjustments to hyperparameters.

4.1.3.2.2. XGBoost Training Process

The XGBoost training process follows a custom training loop rather than relying on a built-in function, offering greater control over the training steps. This loop provides flexibility in determining when to apply early stopping, the number of boosting rounds to execute, and when to save the best model based on performance. The model is trained for 300 boosting rounds, in which a new weak learner is added to target the errors made by previous rounds, aligning with the LSTM model's 300 epochs to ensure consistency between the two models. Early stopping is integrated with a patience of 10 rounds, meaning training will stop if performance (measured by loss or accuracy) does not improve for 10 consecutive rounds, which helps prevent overfitting. Additionally, a custom callback logs the model's performance after each boosting round, tracking metrics like training and validation loss, allowing for insights into the training process. This information is crucial for monitoring model behavior and deciding whether to manually halt the training process based on the progression of the performance metrics.

5. Result and Analysis

Time-series forecasting prioritizes the accuracy of predictions relative to actual values over traditional classification metrics like accuracy. In this study, various error metrics are employed to assess the performance of the models in predicting Bitcoin prices such as Root Mean Squared Error (RMSE) has been used to assess the average difference between the predicted and actual Bitcoin prices, with larger errors penalized more heavily due to squaring. It provides a robust measure of how well the model minimizes significant deviations, Mean Absolute Error (MAE) evaluates the average absolute difference between predicted and actual values, offering a straightforward and interpretable metric for assessing prediction accuracy, Mean Squared Error (MSE) calculates the mean of squared differences between predicted and actual values, emphasizing larger errors due to squaring. This metric helps identify the model's effectiveness in minimizing substantial deviations, Mean Absolute Percentage Error (MAPE) measures prediction errors as a percentage of the actual values, providing a scale-independent metric that is particularly useful for comparing performance across different datasets or models, Mean Absolute Scaled Error (MASE) compares the prediction accuracy of the model against a naïve baseline, offering insights into the relative improvement achieved by the model over simple forecasting methods. These metrics collectively provide a comprehensive evaluation of the model's performance in predicting Bitcoin prices.

5.1. LSTM Results

The application of Long Short-Term Memory (LSTM) to Bitcoin price prediction from 2012 to 2024 yielded high accuracy. The model effectively captured temporal dependencies and patterns within the data, leveraging its recurrent structure and memory cells. This demonstrates LSTM's capability to model complex, non-linear relationships in financial time series data.

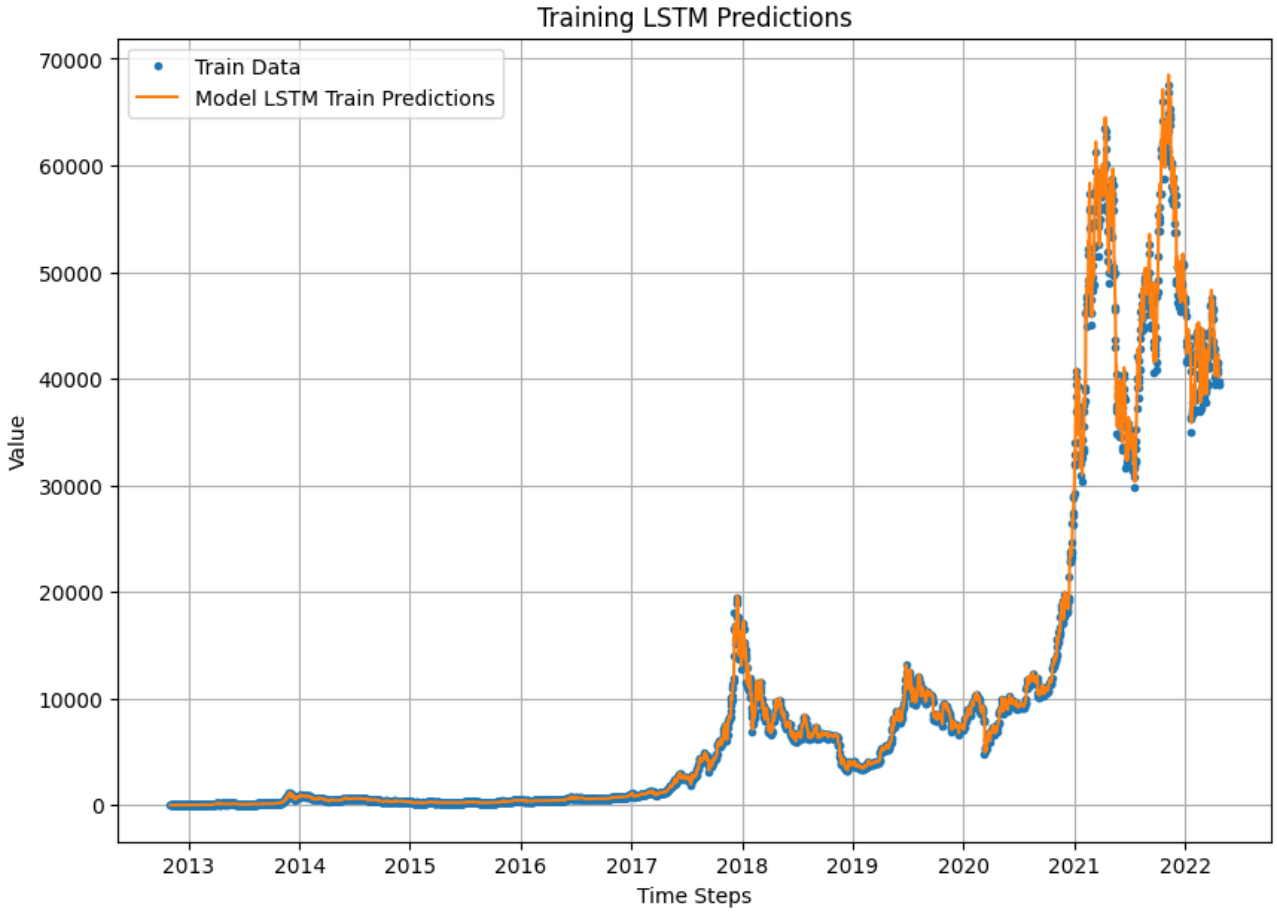

Figure 5.4.: LSTM Training Performance

The orange line (model predictions) closely follows the blue line (actual training data) across the 11 years. This indicates the model's strong ability to capture trends and minimize errors during training.

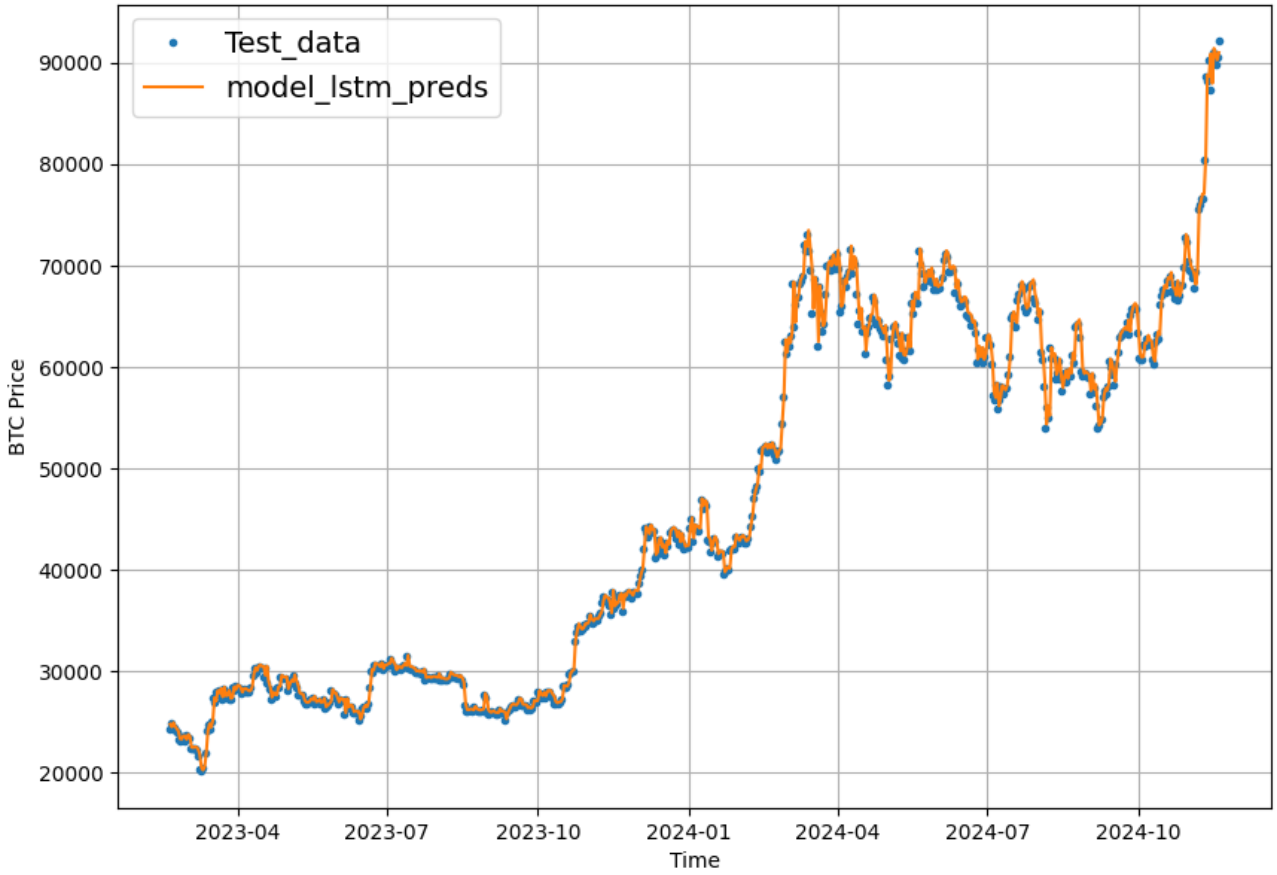

Figure 5.5.: LSTM Testing Performance

The predicted values (orange line) align well with the test data (blue points) from 2023 to 2024. The model successfully generalizes to unseen data, indicating robust performance. Strengths of LSTM: Recurrent architecture allows it to model temporal dependencies. High accuracy in capturing trends and sudden changes in volatile data, as evident in the testing phase.

5.2. XGBoost Result

The model performs well on the training data, with predictions closely matching actual values, indicating strong training performance. On the test data, the predictions align reasonably well with the actual values, though there are some deviations, suggesting room for improvement in generalization. Overall, the model demonstrates good predictive capabilities but may benefit from further refinement, such as feature engineering or hyperparameter tuning, to improve accuracy on unseen data. (Figure 9)

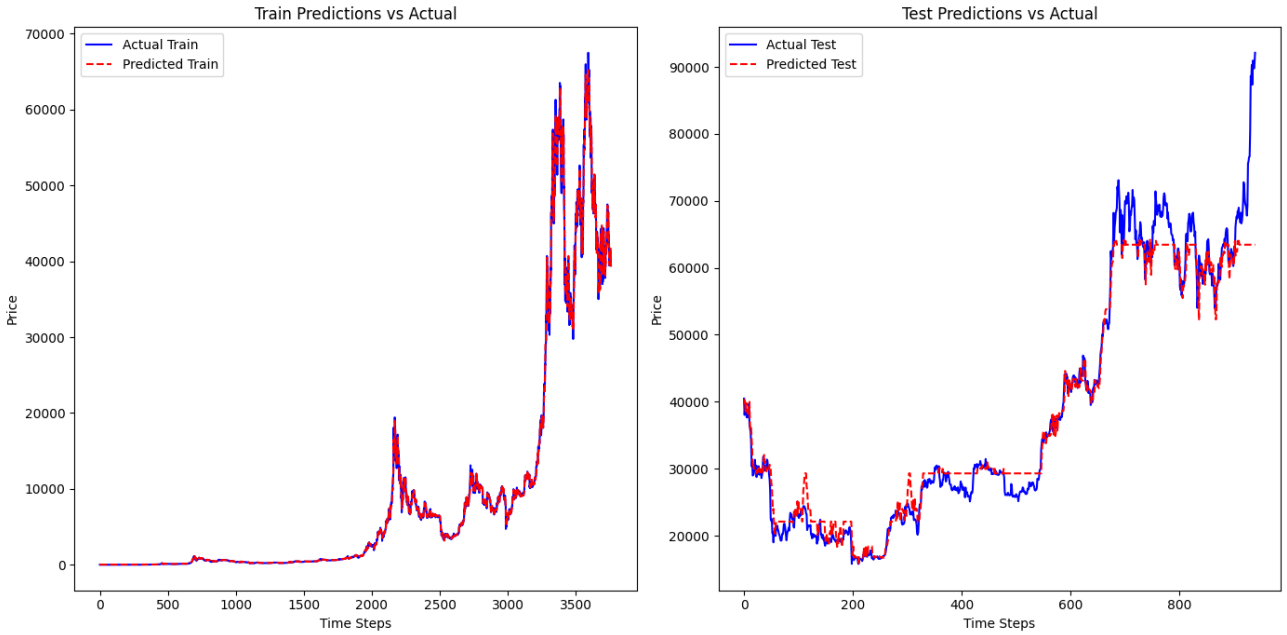

Figure 5.2.: XGBoost Training vs. Testing Predictions

Training vs. Testing Predictions:

Left Panel (Training), the red dashed line (predicted prices) aligns closely with the blue line (actual prices), indicating accurate fitting, while the Right Panel (Testing), he red dashed line (predictions) follows the overall trend of the blue line but struggles slightly during periods of high volatility, showing minor deviations.

5.3. Model Performance Comparison: LSTM vs. XGBoost

This section provides a performance comparison of the Long Short-Term Memory (LSTM) model and the XGBoost model on the test dataset. The evaluation metrics used include Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and Mean Absolute Scaled Error (MASE).

Table 5.3: Metrics Comparison

Metric	LSTM	XGBoost	Comparison
MAE	963.55	2107.53	LSTM demonstrates a significantly lower MAE, indicating better accuracy in capturing absolute differences.
MSE	1,884,897.2	13,777,630.0	LSTM achieves a much lower MSE, showing that it is better at penalizing larger errors compared to XGBoost.
RMSE	1,372.92	3,711.82	The RMSE of LSTM is substantially lower, highlighting its superior ability to reduce large deviations.
MAPE	0.0253 (2.53%)	0.0548 (5.48%)	LSTM outperforms XGBoost in terms of MAPE, offering more reliable percentage-based predictions.
MASE	1.297	2.837	LSTM has a lower MASE, signifying better forecasting accuracy relative to a naïve baseline.

6. Conclusion and Future work

6.1. Conclusion

In this study, we compared two machine learning models, Long Short-Term Memory (LSTM) and XGBoost, for univariate time series forecasting using historical Bitcoin price data. The results highlighted several key insights regarding model performance. The LSTM model significantly outperformed XGBoost across all evaluation metrics, including MAE, MSE, RMSE, MAPE, and MASE. LSTM's ability to model temporal dependencies made it especially effective in capturing trends and sudden changes in volatile data, which is critical for financial time series forecasting. While XGBoost performed well on the training data, its predictions on test data showed slightly higher errors and deviations, suggesting challenges in generalizing to unseen data without temporal awareness. In terms of model strengths, LSTM excelled at modeling non-linear patterns and long-term dependencies due to its recurrent architecture and memory cells. On the other hand, XGBoost, although lacking inherent temporal modeling capabilities, provided reasonable performance by effectively leveraging feature interactions. From a practical standpoint, the high accuracy of LSTM makes it a more suitable choice for Bitcoin price forecasting or similar financial time series tasks. However, XGBoost may still be a competitive option in cases where temporal relationships are less critical, or where feature engineering can be applied to explicitly encode temporal dependencies.

6.2. Future work

To further enhance and expand upon this research, several directions are proposed. First, feature engineering could be explored by incorporating additional variables such as trading volume, sentiment analysis from social media, macroeconomic indicators, and market sentiment data to enrich model inputs and potentially improve forecasting accuracy. Additionally, hybrid models combining the strengths of LSTM and XGBoost could be investigated to leverage both temporal modeling capabilities and feature interaction strengths. Advanced hyperparameter optimization techniques, such as Bayesian optimization or genetic algorithms, could be applied to fine-tune the parameters of both models for better performance. Another potential direction is extending the study to multivariate time series forecasting by including external variables, like financial indices or macroeconomic factors, to create a more comprehensive prediction framework. Alternative deep learning architectures, including Transformer-based models or Convolutional Neural Networks (CNNs), could also be explored as they have shown promise in time series forecasting. Ensemble methods, combining multiple models, could be employed to reduce overfitting and improve robustness in predictions. Robustness testing under varying market conditions, such as periods of high volatility, could help evaluate the models' reliability and scalability. Finally, a real-time forecasting pipeline could be developed to provide actionable insights for traders and stakeholders, including automated alerts and dashboard visualizations. This study lays the foundation for accurate and effective Bitcoin price forecasting, and future work can build upon these findings to create even more powerful and robust predictive systems, ensuring their relevance and applicability in dynamic and volatile financial markets.

References

[i] Sustainability of bitcoin and blockchains - ScienceDirect
ii tf.keras.layers.LSTM | TensorFlow v2.16.1
iii XGBoost Python Package — xgboost 2.1.3 documentation
iv Prediction and Analysis of Bitcoin Price using Machine learning and Deep learning models | EAI Endorsed Transactions on Internet of Things
v Analysis of Bitcoin Price Prediction Using Machine Learning
vi
https://www.researchgate.net/publication/380180701_Bitcoin_Price_Prediction_Based_on_Multiple_Machine_Learning_Algorithms
viihttps://www.researchgate.net/publication/376887854_Understanding_Bitcoin_A_Case_Study_Method_to_Understand_Market_Dynamics_Strategies_and_Risks_of_Cryptocurrency
viii
https://www.researchgate.net/publication/384811602_SARF_Enhancing_Stock_Market_Prediction_with_Sentiment-Augmented_Random_Forest
ix https://www.researchgate.net/publication/383230619_Predicting_Stock_Prices_Using_The_RF-LSTM_Combination_Model
x https://coincodex.com/crypto/bitcoin/historical-data/
xi What is Exploratory Data Analysis? - GeeksforGeeks
xiibing.com/ck/a?!&&p=80d0683244ffb2cc572f7d59046905f47e34c9c95700e9ef396ae83685c8bf26JmltdHM9MTczNTk0ODgwMA&ptn=3&ver=2&hsh=4&fclid=17b6659c-9435-6b9a-13fd-7174956f6ad9&psq=Pandey%2c+T.N.%2c+Priya%2c+T.%2c+Jena%2c+K.%3a+Prediction+of+Exchange+rate+in+a+cloud+computing+environment+using+machine+learning+tools.+Intell.+Cloud+Comput.+137–146+(2021).&u=a1aHR0cHM6Ly9saW5rLnNwcmluZ2VyLmNvbS9jaGFwdGVyLzEwLjEwMDcvOTc4LTk4MS0xNS02MjAyLTBfMTU&ntb=1
xiii https://doi.org/10.1007/978-981-16-8739-6_41
xiv Bitcoin price prediction using ARIMA and LSTM | E3S Web of Conferences
xvbing.com/ck/a?!&&p=80d0683244ffb2cc572f7d59046905f47e34c9c95700e9ef396ae83685c8bf26JmltdHM9MTczNTk0ODgwMA&ptn=3&ver=2&hsh=4&fclid=17b6659c-9435-6b9a-13fd-7174956f6ad9&psq=Pandey%2c+T.N.%2c+Priya%2c+T.%2c+Jena%2c+K.%3a+Prediction+of+Exchange+rate+in+a+cloud+computing+environment+using+machine+learning+tools.+Intell.+Cloud+Comput.+137–146+(2021).&u=a1aHR0cHM6Ly9saW5rLnNwcmluZ2VyLmNvbS9jaGFwdGVyLzEwLjEwMDcvOTc4LTk4MS0xNS02MjAyLTBfMTU&ntb=1
xvibing.com/ck/a?!&&p=1a031c4475e3f097c9ee94a08ede235198b53193cc2f7404689811cbd0e2ae6eJmltdHM9MTczNTk0ODgwMA&ptn=3&ver=2&hsh=4&fclid=17b6659c-9435-6b9a-13fd-7174956f6ad9&psq=Samiksha+Marne%2c+Shweta+Churi%2c+Delisa+Correia%2c+Joanne+Gomes%2c+2021%2c+Predicting+Price+of+Cryptocurrency+–+A+Deep+Learning+Approach%2c+INTERNATIONAL+JOURNAL+OF+ENGINEERING+RESEARCH+%26+TECHNOLOGY+(IJERT)+NTASU+–+2020+(Volume+09+–+Issue+03)%2c%27&u=a1aHR0cHM6Ly93d3cuaWplcnQub3JnL3Jlc2VhcmNoL3ByZWRpY3RpbmctcHJpY2Utb2YtY3J5cHRvY3VycmVuY3ktYS1kZWVwLWxlYXJuaW5nLWFwcHJvYWNoLUlKRVJUUQ09OVjlJUzAzMDgzLnBkZg&ntb=1
xvii Prediction of Bitcoin Price Based on LSTM | Semantic Scholar
xviii XGBoost Documentation — xgboost 2.1.3 documentation

Appendix

https://drive.google.com/file/d/1fvH3WkpuEr7C5cQdwJLnpRUag2OkJ-Pb/view?usp=sharing EDA

https://drive.google.com/file/d/1J6b0n_EmEiMMza5korShGwq_zi2BdxMW/view?usp=sharing Models