My Iowa State University creative component title page:

Predicting Bitcoin Prices Using Machine Learning

by

Chenrui Niu

A creative component submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Information System

Program of Study Committee:

Anthony Townsend, Major Professor

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this creative component. The Graduate College will ensure this creative component is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2025

## Introduction

Bitcoin, as the first and leading cryptocurrency, has garnered widespread attention since its inception in 2009 by Satoshi Nakamoto. Its decentralized nature, which operates independently of traditional banking systems, has made it a revolutionary asset in the digital economy. However, Bitcoin's price volatility poses a significant challenge for both investors and businesses. These frequent and unpredictable price fluctuations make forecasting its value a complex and crucial task especially for those seeing it as an opportunity for portfolio diversification, return enhancement, and protection against inflation.

The growing importance of Bitcoin in the global economy is undeniable. As more companies adopt blockchain-based solutions and cryptocurrencies gain acceptance as a legitimate asset class, understanding Bitcoin's price movements is essential. Accurate predictions can offer investors a strategic advantage, while businesses exploring Bitcoin's potential use cases, such as digital payments or decentralized applications rely on this knowledge to navigate financial risks. Therefore, predicting Bitcoin's price effectively has become a critical issue in both the financial and technological landscapes.

This challenge lies at the intersection of finance, technology, and data science. As businesses integrate more advanced technologies into their operations, the ability to analyze and predict digital asset prices using data-driven approaches is becoming an essential skill for future business.

The primary objective of this research is to compare machine learning models that could predict Bitcoin prices based on historical data and key economic indicators. Specifically, this study aims to:

1. Assess the relationship between Bitcoin prices and macroeconomic indices, including stock market performance, exchange rate, and altcoin data.
2. evaluate predictive models using machine learning techniques, such as linear regression, decision trees, supporter vector machine, and long-short term memory.
3. Compare the performance of different models to identify the most effective approach for forecasting Bitcoin's price.

## Literature Review

The challenge of predicting Bitcoin's price has attracted considerable attention from researchers and financial analysts, with numerous studies exploring various predictive models and techniques. Bitcoin's price volatility, driven by a mix of market sentiment, technological developments, regulatory news, and macroeconomic factors, makes accurate forecasting a difficult yet highly valuable task. The primary goal of Bitcoin price prediction is to develop models that can capture these complex dynamics and provide actionable insights for investors, traders, and businesses.

### Existing Research and Methods

Several studies have attempted to predict Bitcoin's price using a variety of machine learning (ML) and statistical methods. These studies span a range of models, from traditional linear regression to more complex, nonlinear techniques such as decision trees and neural networks. Key approaches and findings from prominent research are discussed below:

Mishra & Kaur (2024) found that various machine learning models, including Support Vector Regression (SVR), Linear Regression, Random Forest Regression, and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, have proven

effective in predicting Bitcoin prices. The analysis and evaluation of these models offer valuable insights into their strengths and weaknesses. The results suggest that LSTM networks may be especially well-suited for long-term predictions due to their capacity to capture sequential patterns in Bitcoin price data.

In the study of Chen (2023), a comprehensive set of explanatory variables was used to predict Bitcoin's price for the following day. These variables were grouped into eight categories: Bitcoin price data, specific technical features of Bitcoin, other cryptocurrencies, commodities, market indices, foreign exchange rates, public attention, and dummy variables for the days of the week. In total, 47 variables from these categories were included in the model. The results indicated that Random Forest Regression provided better prediction accuracy than Long Short-Term Memory (LSTM) networks. Although LSTM has been widely recognized and used in previous research as a high-accuracy model for predicting Bitcoin prices, this paper demonstrates that Random Forest Regression, which has not been as widely applied in this domain, yielded superior results. However, Random Forest Regression does have a limitation: it struggles to predict outcomes outside the range of historical training data. For example, when Bitcoin's price exceeds previous record highs, Random Forest Regression cannot predict prices higher than its training data's historical peak. Nevertheless, as the dataset expands with more transaction history, it is expected that Random Forest Regression will perform better, especially as Bitcoin's price stabilizes.

Beyond the performance of individual models, this research also highlights the effect of incorporating multiple past periods of explanatory variables. Both Random Forest Regression and LSTM declined accuracy as the number of past periods included in the model increased. The model that achieved the highest accuracy was the one that only included explanatory variables

from the immediate previous period. This observation aligns with the Efficient Market Hypothesis, which suggests that past price movements and related variables may not always predict future price changes.

The differing results between the studies by Mishra & Kaur (2024) and Chen (2023) provide valuable insights into the complexities of predicting Bitcoin prices using machine learning models, highlighting both the strengths and limitations of different approaches. And the difference may be caused by several reasons. First, scope of explanatory variables that the extensive set of features likely contributed to the better performance of Random Forest Regression in this study. Random Forest's ability to handle a large number of variables and its robustness to various types of data could have made it more effective than LSTM, which typically requires careful feature engineering and might struggle with incorporating such diverse data types. Second, training data and periods that those studies have been conducted using different training datasets and periods, which could explain variations in model performance. Third, feature selection that LSTM, while powerful for long-term time-series forecasting, is sensitive to data quality and quantity. If the data is noisy or lacks sufficient variability, LSTM may struggle to generalize.

**Role of External Factors**

While Bitcoin price predictions have traditionally relied heavily on historical price data, more recent studies have emphasized the importance of incorporating external factors that influence market behavior. External factors include social media sentiment from social media platforms, especially Twitter, Reddit, and Telegram, which are key spaces for discussions surrounding Bitcoin, often driving rapid price movements. Sentiment analysis, a natural language processing (NLP) technique, is frequently used to analyze public sentiment around Bitcoin. Like

studies by Kraaijeveld & De Smedt (2020) and Valencia, Gómez-Espinosa, & Valdés-Aguirre(2019). Besides, various macroeconomic factors, such as inflation rates, global interest rates, stock market performance, and even geopolitical events, can significantly influence Bitcoin's price. For instance, Bitcoin has often been touted as a "safe haven" asset during times of economic instability or high inflation, which can drive price surges. The relationship between Bitcoin and traditional financial markets, such as stocks and gold, has been the subject of extensive research, particularly in the context of market crises and economic uncertainty. Moreover, regulatory developments, such as government crackdowns, cryptocurrency regulations, or announcements of legal tenders (e.g., El Salvador's adoption of Bitcoin as legal tender), can cause significant price movements. Regulatory uncertainty can either drive Bitcoin's price up (in cases of favorable legislation) or down (in cases of crackdowns), underscoring the need for predictive models to consider these external drivers.

**Machine Learning Techniques in Finance**

Machine learning techniques are becoming increasingly integral to financial modeling. A variety of ML algorithms have been applied in both traditional financial markets and cryptocurrency prediction, each offering unique advantages depending on the data and goals of the prediction. The findings of Coleman, Merkley & Pacelli (2022) show that Robo-Analysts conduct research differently from traditional analysts, with fewer cognitive biases and economic incentives for optimistic reports. They issue more balanced recommendations, revise their forecasts more frequently, and rely on large volumes of complex data.

One of the simplest and most widely used in finance, linear regression seeks to model the relationship between a dependent variable (Bitcoin price) and one or more independent variables (such as historical prices, volume, or economic indicators). While linear regression is effective

for understanding basic trends and relationships, it is limited in capturing the non-linear complexities of cryptocurrency price movements.

Decision trees use a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Random Forest, an ensemble method, creates many decision trees and merges them to improve prediction accuracy and avoid overfitting. These models are widely used in Bitcoin price prediction due to their ability to handle both numerical and categorical data, as well as their effectiveness in uncovering complex relationships between various features (e.g., market sentiment, trading volume, and macroeconomic factors).

As a type of recurrent neural network (RNN), LSTM networks are particularly suited for time-series forecasting, making them a popular choice for predicting Bitcoin's price based on historical data. LSTM models are capable of learning long-term dependencies in sequential data, which allows them to capture patterns over extended periods—critical for forecasting financial time series that exhibit trends and cycles. Studies like Kraaijeveld & De Smedt (2020) have shown that LSTM networks outperform simpler models in predicting Bitcoin prices, especially when trained on large datasets that include both historical prices and external factors.

SVM is a supervised learning model that works well for classification and regression tasks. In cryptocurrency prediction, SVMs are often used to classify price movements (up or down) or to predict future price ranges. SVMs are particularly effective when the data is not linearly separable, which is common in financial markets.

## Methodology

Machine learning methods can be broadly categorized into supervised learning, unsupervised learning, and reinforcement learning, depending on the presence or absence of a target variable. Since the goal of this study is to predict future Bitcoin prices, a supervised learning approach is employed. The general process in machine learning involves defining an algorithm, training a model on the available data, and refining the model's performance through repeated training and validation. Once the model has been adequately trained, it is evaluated using test data to assess its accuracy and applicability.

In this study, both Random Forest Regression and the LSTM model are implemented using open-source Python libraries. The linear regression model, Random Forest Regression model, and support vector machine model utilize the scikit-learn library, while the LSTM model is built using Keras. Data preprocessing and preparation are carried out with the help of the pandas library.

**Data Collection and Preprocessing**

Historical Bitcoin price data will be obtained from Yahoo Finance, which provides reliable daily closing prices over an extended period. This time series data will serve as the primary variable for predicting Bitcoin price fluctuations. To enhance prediction accuracy, additional features will be incorporated, including macroeconomic indicators like Treasury bond yields, which reflect investor sentiment and broader market conditions. Exchange rates for major fiat currencies (USD, EUR, etc.) will also be considered, as fluctuations in these currencies can influence Bitcoin prices. Additionally, the prices of other cryptocurrencies, such as Ethereum, will be included, given their correlation with Bitcoin's price movements.

For data preprocessing, missing values will be addressed using techniques such as forward/backward filling or interpolation. In cases of substantial missing data, rows may be removed or imputed with an average value. To ensure that all features are on a comparable scale, normalization or scaling methods (e.g., Min-Max Scaling or Standardization) will be applied, particularly for models sensitive to feature magnitude, such as Support Vector Machines (SVM) or neural networks. Exploratory Data Analysis (EDA) will be conducted to uncover trends, patterns, and anomalies in the data. This will involve visualizing time series data for Bitcoin and other variables, calculating correlations to understand relationships.

**Model Selection and Approach**

To predict Bitcoin price movements, several machine learning models will be employed, starting with Linear Regression, which provides a simple, interpretable approach to understand linear relationships between Bitcoin prices and the selected features. Although basic, this model offers a useful baseline. Decision Trees will be used to capture non-linear relationships and identify complex decision boundaries between features that influence Bitcoin prices. Building on this, Random Forests will combine multiple decision trees to enhance performance, reduce overfitting, and provide more robust predictions. Additionally, Support Vector Machines (SVM) will be tested to handle high-dimensional data and non-linear relationships, utilizing different kernel functions to identify the most effective model for capturing Bitcoin price patterns. LSTMs are extremely effective for time series forecasting because they can model sequential dependencies in data. Thus, this model is used to predict stock prices, weather patterns, sales data, and other time-dependent phenomena such as Bitcoin prices.

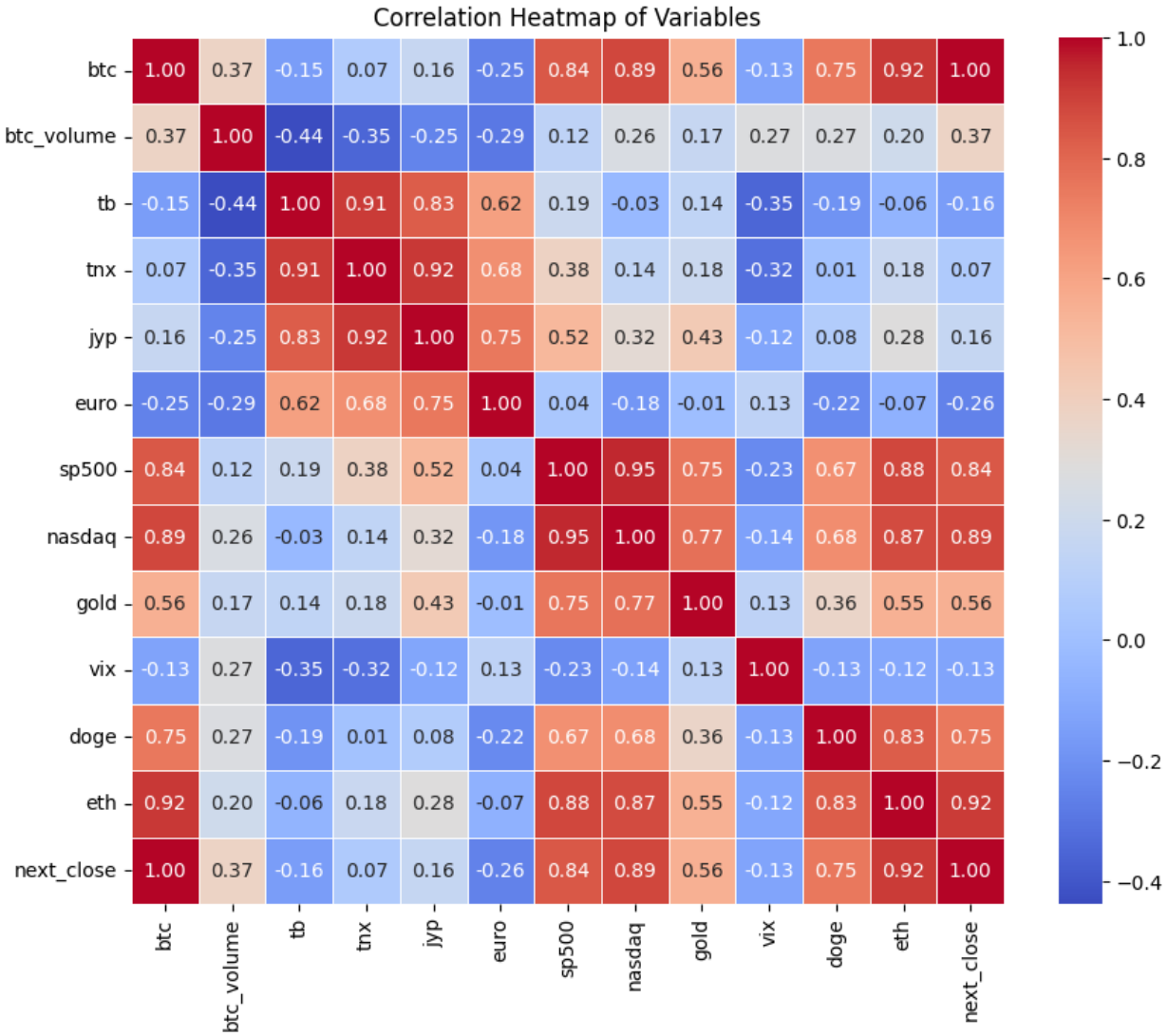
**Training and Testing**

The dataset will be divided into training and testing sets with a 70/30 split. The training set will be used to fit the models, while the testing set will assess model performance on unseen data. To optimize model performance, hyperparameters will be tuned using grid search or random search methods. For example, in Random Forests, parameters like the number of trees, maximum depth, and minimum samples per leaf will be fine-tuned, while in SVM, kernel types and regularization parameters will be adjusted.

## Model Evaluation and Results

This section will present the performance metrics for each machine learning model used to predict Bitcoin prices. The evaluation focuses on metrics such as Root Mean Squared Error (RMSE), R-squared ($R^2$), and Mean Absolute Error (MAE), which are commonly used in regression tasks to assess the model's predictive accuracy and robustness. Since the result of the random forest outperforms all aspects of the result of the decision tree, thus, only the result of the random forest is recorded.

The feature importance analysis reveals that variable Bitcoin price from the previous day had the most significant impact on Bitcoin price predictions. Thus, this variable is taken out to avoid the model potentially failing to learn the true underlying relationships in the data because it relies too much on one variable that is too predictive of the target. The heatmap shows the correlations between all variables:




**Performance Metrics**

RMSE measures the average magnitude of the prediction error. It gives an idea of how far the model's predictions are from the actual values. A lower RMSE indicates better performance. RMSE values for each model:

Linear Regression: 0.2734

Random Forest Regression: 0.1053

Support Vector Machine (SVM): 0.1360

LSTM (Long Short-Term Memory): 0.6052

In this case, Random Forest demonstrated the lowest RMSE, indicating that it produced predictions closer to the actual Bitcoin prices compared to the other models.

The R-squared value represents the proportion of variance in the target variable (Bitcoin price) that the model explains. An $R^2$ value closer to 1 indicates a better fit. $R^2$ values for each model:

Linear Regression: 0.9231

Random Forest Regression: 0.9886

Support Vector Machine (SVM): 0.9809

LSTM (Long Short-Term Memory): -0.6052

The Random Forest model again demonstrated the highest $R^2$ value, explaining a significant portion of the variance in Bitcoin prices.

MAE is the average of the absolute differences between predicted and actual values. It is another important metric that provides a direct measure of prediction error, with lower values being preferable. MAE values for each model:

Linear Regression: 0.1875

Random Forest Regression: 0.0561

Support Vector Machine (SVM): 0.0919

LSTM (Long Short-Term Memory): 0.3335

The Random Forest had the lowest MAE, again outperforming all other models. This suggests that the random forest model not only fits the data well but also minimizes errors in its predictions.

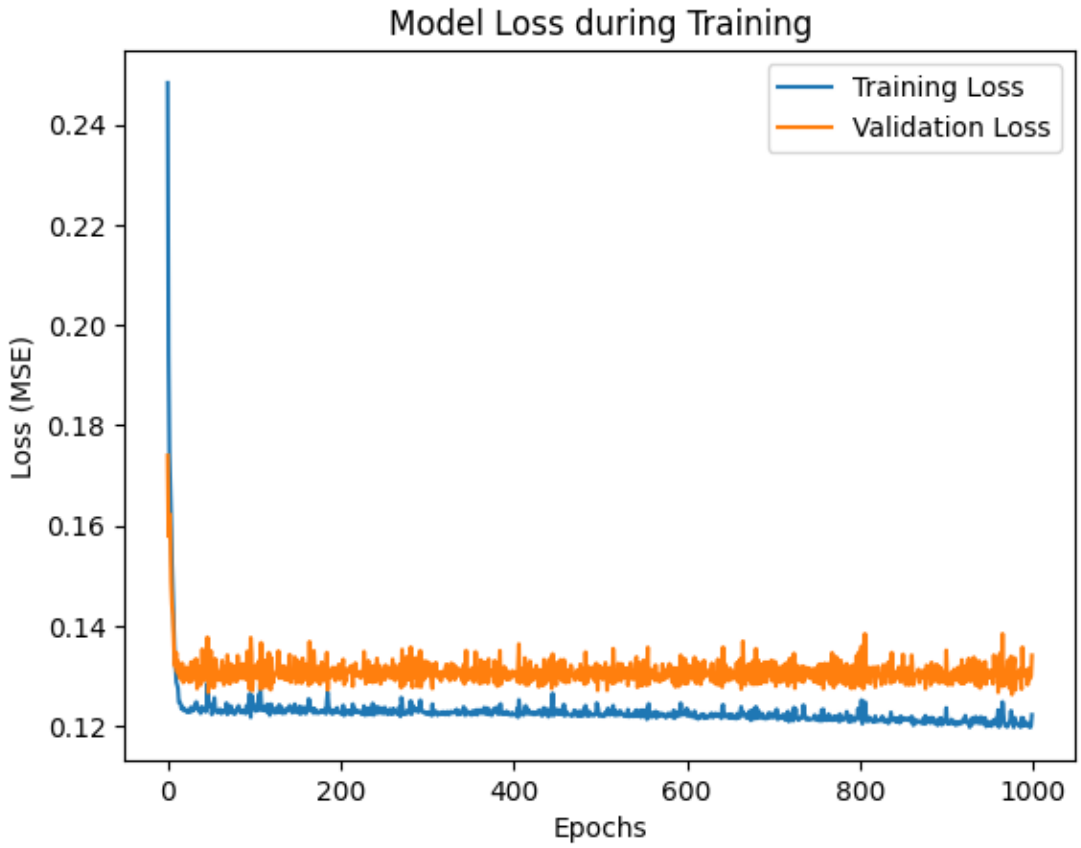
**Model Comparison**

In terms of performance, the random forest model emerged as the most robust and accurate predictor of Bitcoin prices, outperforming traditional machine learning models like Linear Regression and Support Vector Machines.

Random Forest Regression offered strong predictive performance with high accuracy and robustness. It is an ensemble method that performs well with non-linear data and complex interactions, which is particularly useful for financial data that often contains noisy or hidden relationships. Support Vector Machines (SVM) showed moderate performance. SVMs tend to be more sensitive to the choice of kernel and hyperparameters, and while they can work well in high-dimensional spaces. Linear Regression performed reasonably well but showed relatively lower accuracy. As expected, linear models struggle to capture the complex, non-linear patterns inherent in cryptocurrency price movements, making them less suitable for time-series forecasting of volatile assets like Bitcoin.

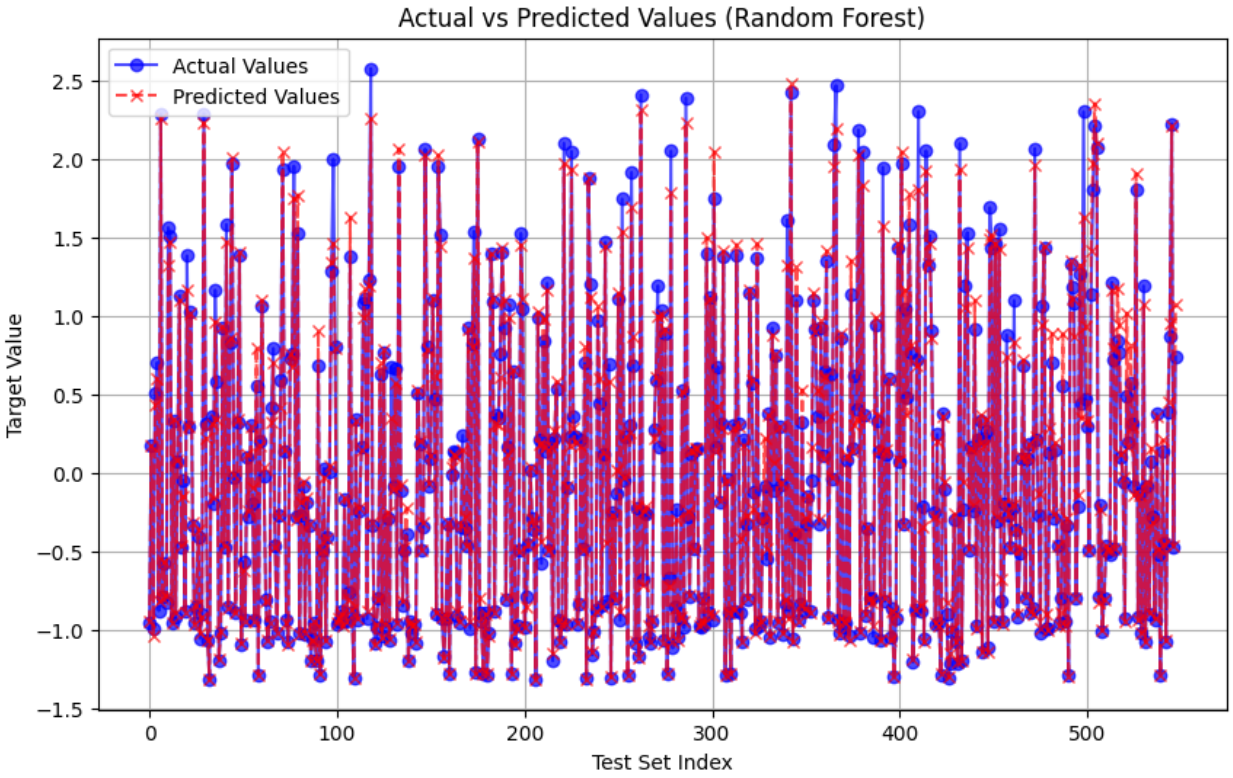
LSTM, on the other hand, is specifically designed to handle time-series data by learning long-term dependencies. However, the performance of LSTM might have been limited by factors such as the dataset might not have enough temporal patterns or seasonal trends to fully exploit LSTM's strengths. Also, The LSTM model could be more sensitive to hyperparameter settings such as the number of layers, units in each layer, and learning rates. If not tuned properly,

LSTM's performance can degrade, leading to suboptimal results. Besides, LSTM models are more computationally intensive and require careful training, which could have influenced their performance in this analysis. The following graph shows the model loss during training:

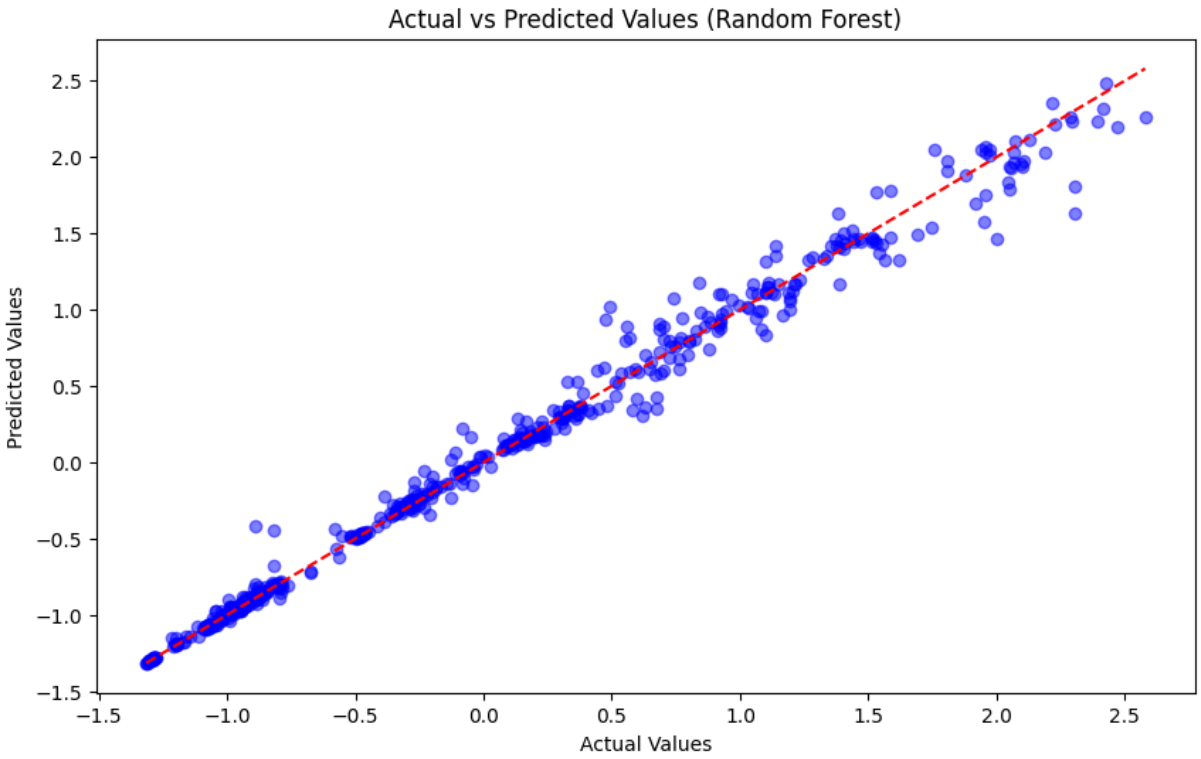



**Visualizing Predictions vs Actual Prices**

A line plot of actual vs predicted Bitcoin prices for the LSTM model is shown below. The close alignment of predicted and actual values reflects the model's performance.




A scatter plot of predicted vs actual values helps visualize the accuracy of the model. Ideally, the points should lie close to the 45-degree line, indicating good model performance.




## Conclusion

In this study, several machine learning models are applied to predict Bitcoin prices using a variety of explanatory variables, including macroeconomic indicators, financial market data, and cryptocurrency prices. The models we evaluated included Random Forest, LSTM (Long Short-Term Memory), Support Vector Machines (SVM), and Linear Regression. While LSTM was expected to perform best due to its ability to capture temporal dependencies in time-series data, the results showed that Random Forest outperformed LSTM in terms of predictive accuracy and robustness.

The Random Forest model demonstrated the best performance across all evaluation metrics, including Root Mean Squared Error (RMSE), R-squared ($R^2$), and Mean Absolute Error (MAE). It was able to accurately predict Bitcoin prices, outperforming LSTM, which is typically strong at modeling sequential data. The success of Random Forest can be attributed to its ability

to handle complex, non-linear relationships and interactions between features, as well as its robustness to noisy data.

Despite LSTM's ability to capture long-term dependencies in time-series data, its performance was not superior in this case. This suggests that for Bitcoin price prediction, where volatility and short-term market movements are critical, ensemble methods like Random Forest may be more effective than deep learning models like LSTM. Additionally, LSTM models require careful hyperparameter tuning and substantial computational resources, which can sometimes hinder their effectiveness, especially when the data does not exhibit clear, long-term patterns.

Moreover, external factors, such as geopolitical events, regulatory changes, or market sentiment shifts, are not captured in the data, which could significantly affect Bitcoin price movements but were not included in the analysis. These factors, while difficult to quantify and model, could substantially influence the cryptocurrency market, and future studies should attempt to account for them. Thus, future studies could incorporate sentiment analysis techniques, particularly from social media platforms, to enhance cryptocurrency prediction models. Given that cryptocurrencies are often heavily influenced by public sentiment, advanced natural language processing (NLP) methods could more effectively capture the subtleties of online discussions. By analyzing sentiment signals from platforms like Twitter, Reddit, and other forums, these methods could offer valuable insights. Moreover, integrating multimodal sentiment analysis, which combines text, images, and video content, could provide a more holistic understanding of market sentiment, further improving predictive accuracy.

## Reference

Chen, J. (2023). Analysis of Bitcoin Price Prediction Using Machine Learning. *Risk Financial Manag.* 2023, *16*, 51. https://doi.org/10.3390/jrfm16010051

Mishra, A., & Kaur, P. (2024). Cryptocurrency price prediction analysis using machine learning algorithms. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4490209

Coleman, B., Merkley, K., & Pacelli, J. (2022). Human versus machine: A comparison of robo-analyst and Traditional Research Analyst Investment Recommendations. *The Accounting Review*, *97*(5), 221–244. https://doi.org/10.2308/tar-2020-0096

Kim, H., Bock, G., & Lee, G. (2021). Predicting Ethereum prices with machine learning based on Blockchain information. *Expert Systems with Applications*, *184*, 115480. https://doi.org/10.1016/j.eswa.2021.115480

Abraham, J., Higdon, D., Nelson, J., & Ibarra, J. (2018). Cryptocurrency price prediction using tweet volumes and sentiment analysis. *SMU Data Science Review*, *1*(3), 1. https://scholar.smu.edu/cgi/viewcontent.cgi?article=1039&context=datasciencereview

Kraaijeveld, O., & De Smedt, J. (2020). The predictive power of public Twitter sentiment for forecasting cryptocurrency prices. *Journal of International Financial Markets Institutions and Money*, *65*, 101188. https://doi.org/10.1016/j.intfin.2020.101188

Valencia, F., Gómez-Espinosa, A., & Valdés-Aguirre, B. (2019). Price movement prediction of cryptocurrencies using sentiment analysis and machine learning. *Entropy*, *21*(6), 589. https://doi.org/10.3390/e21060589