# Bitcoin Price Prediction Using Regression and LSTM

**Junhao Qu, Wendeng Wang, Yiwei Huang, Haoran Liu**

## Abstract

This study investigates efficient machine learning approaches for Bitcoin price prediction, focusing on their effectiveness in addressing the challenges posed by the cryptocurrency's high volatility and complex time-series patterns. By utilizing state-of-the-art methods, such as polynomial regression and Long Short-Term Memory (LSTM) networks, the research explores the trade-offs between interpretability, computational efficiency, and predictive accuracy. Through rigorous preprocessing techniques, including log transformations and feature normalization, the study prepares a robust dataset to test various predictive models. The results reveal insights into short-term price movements, feature importance, and the limitations of current methodologies in capturing Bitcoin's intricate dynamics.

## 1 Introduction

Bitcoin's extreme price volatility and its role as a leading cryptocurrency present unique challenges for time-series forecasting. This study leverages a rich dataset of daily Bitcoin prices, spanning from November 2019 to November 2024, to predict future trends using machine learning. A thorough exploratory data analysis highlights significant variability in price and trading volume, necessitating advanced preprocessing methods such as log transformation, z-score normalization, and Box-Cox adjustments. These ensure the dataset's suitability for predictive modeling by reducing skewness and stabilizing feature distributions.

Building upon state-of-the-art techniques, the study employs polynomial regression to establish foundational models and incorporates advanced LSTM networks to address the temporal dependencies inherent in Bitcoin price data. The comparative analysis of these methodologies underscores the strengths and limitations of each approach, offering valuable perspectives for researchers and practitioners in cryptocurrency forecasting.

## 2 Dataset Description and Exploratory Analysis

The dataset used for this study was obtained from the historical Bitcoin price data available at CoinCodex, Bitcoin Historical Data (CoinCodex, 2014). It spans daily data from November 29, 2019, to November 27, 2024, and includes 1825 records. Each record contains the following attributes:

- **Start, End**: The start and end dates of the trading period.
- **Open, High, Low, Close**: Daily Bitcoin price statistics in USD.
- **Volume**: Total daily trading volume in USD.
- **Market Cap**: Daily Bitcoin market capitalization in USD.

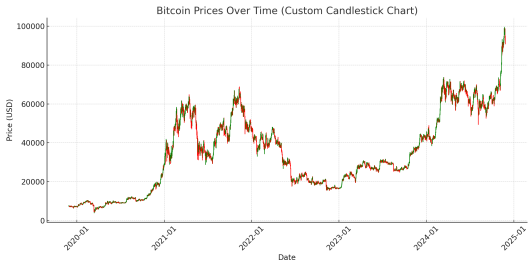


Figure 1: Bitcoin price over time

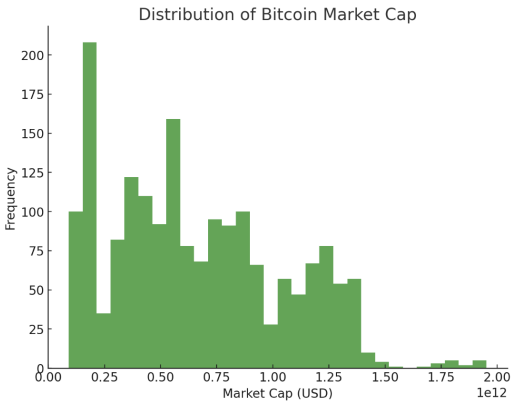


Figure 2: Distribution of Bitcoin Market Cap

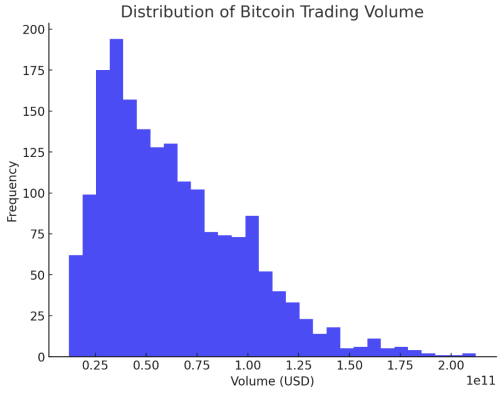


Figure 3: Distribution of Bitcoin Trading Volume

- Green Bars: The price closed higher than it opened (bullish day).
- Red Bars: The price closed lower than it opened (bearish day).
- Vertical lines represent the full range (High-Low) for each day.
- Volume Distribution: A skewed distribution indicates days with exceptionally high trading activities.
- Market Cap Distribution: The market cap also exhibits skewness, reflecting the influence of rapid market movements.

### 2.1 Interesting findings

The dataset's descriptive statistics reveal significant volatility in Bitcoin prices. For example, the mean opening price is $34,550.22 with a standard deviation of $19,595.44. Similarly, trading volume averages $63.55 billion daily, while the market cap averages $663.70 billion. These statistics highlight Bitcoin's highly dynamic nature.

Key observations from the analysis include:

- **Volatility**: High variability in prices and trading volume over time.
- **Trends**: Long-term growth periods interspersed with sharp declines.
- **Temporal**: Dependencies: Significant short-term correlations between trading volume, market capitalization, and price suggest strong temporal relationships that are valuable for predictive modeling.

These findings underscore the need for robust models capable of handling extreme variability while capturing underlying temporal dynamics. By identifying these patterns and trends, the study lays the foundation for designing effective machine learning approaches to Bitcoin price prediction.

## 3 predictive task on this dataset

The primary predictive task for this dataset is forecasting the next day's Bitcoin closing price using features derived from daily market data. These features include market volume, market capitalization, and price volatility (calculated as the difference between the highest and lowest prices each day). This task is chosen due to the strong temporal dependencies identified in the dataset and the need to model Bitcoin's high volatility effectively.

### 3.1 Validity of Predictions and Significance Assessment

Metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and $R^2$ (coefficient of determination) will quantify the model's performance. A lower MSE and higher $R^2$ on the validation set will confirm prediction accuracy.

### 3.2 Data Preprocessing

- Log Transformation: Applied to features like market volume and market capitalization to stabilize variance and mitigate the effect of extreme values.
- Feature Engineering: Created lagged features, such as the previous day's price, market volume, and volatility, to capture temporal dependencies. Applied transformations like $ln(x) + x$ to enhance feature stability and predictive power.

### 3.3 Justification of Features Based on Exploratory Analysis

Volatility and Short-Term Dependencies: High variability in daily Bitcoin prices and trading volumes suggested that features like daily price ranges (High-Low) and lagged values of trading volume would be predictive.
Normalization Needs: Skewness in the distributions of market volume and market capitalization underscored the necessity of transformations (e.g., log and $ln(x) + x$) to reduce outlier effects and stabilize variance.

## 4 Regression - Using previous day's features to predict the current day's price

### 4.1 Input features used in regression and dataset splitting method

This study considers the previous day's Bitcoin **market volume**, **market capitalization**, and **volatility** (highest price minus lowest price) as candidate input features. This study will test the fitting performance of different feature combinations after applying the ln(x) + x transformation to the input features, in order to select the best-performing regression model.

For dataset splitting, use the dataset from 11/29/2019 to 11/26/2023 as the training set, and from 11/27/2023 to 11/27/2024 as the validation set, i.e., split the training and validation sets in a **4:1** ratio.

### 4.2 Try and select the best feature combination

Since the trend of Bitcoin does not exhibit a clear periodicity and is evidently non-linear, the approach is to apply the ln(x) + x transformation to the input features before fitting.

For the three transformed input features, test the fitting performance of all seven possible feature combinations on the validation set, with the results shown in Table 1. It can be concluded that using market volume, market capitalization, and volatility as input features simultaneously yields the best fitting performance.

Based on this feature combination, the specific fitting formula used is as follows:
$x_1' = \ln(x_1) + x_1 \quad x_2' = \ln(x_2) + x_2 \quad x_3' = \ln(x_3) + x_3$
$y = \beta_0 + \beta_1 x_1' + \beta_2 x_2' + \beta_3 x_3' + \epsilon$

### 4.3 Regression results

Use the model trained on the training set to predict the Bitcoin closing price for each day in the validation set, resulting in the training curve shown in the Figure 4 below. The MSE on the validation set is 6,922,968.29, and after standardization, it is 2.13.

| Volume | Market Cap | Volatility | $R^2$ |
|---|---|---|---|
| Yes | No | No | -2.16 |
| No | Yes | No | 0.88 |
| No | No | Yes | -1.69 |
| Yes | Yes | No | 0.92 |
| Yes | No | Yes | -1.73 |
| No | Yes | Yes | 0.91 |
| Yes | Yes | Yes | 0.94 |

Table 1: $R^2$ Scores for Different Feature Combinations

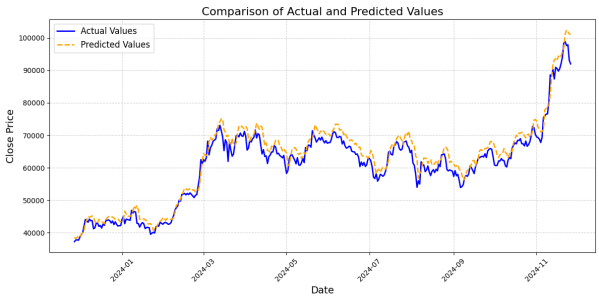


Figure 4: Comparison of Actual and Predicted Values

Although using a regression method to predict Bitcoin prices can yield certain results on the validation set, it has a significant drawback— it can only effectively predict the price for the next day and struggles to predict prices over a longer period. Therefore, this study will use a time series model in an attempt to address this limitation.

## 5 LSTM

### 5.1 Leveraging LSTM for Short-Term Bitcoin Price Prediction

LSTM (Long Short-Term Memory) is a specialized type of Recurrent Neural Network (RNN) that excels in time series forecasting. It is particularly well-suited for analyzing financial data with high volatility and time dependency, such as Bitcoin. Bitcoin prices often exhibit short-term trends, where the price and trading volume of the past few days are strongly correlated with the price of the following day. To leverage this relationship, we utilize data from the past 7 days to predict the price for the next day.

In our approach, we combine **market volume** and **market capitalization** as market-

related features, and incorporate **open, close, high, and low prices** as price-related features. The dataset is divided into training and test sets with a ratio of 90:10, ensuring that the model is trained effectively while leaving enough data for robust evaluation.

### 5.2 Result

The performance results of the LSTM model are presented in Table2 2. A visual representation of the model's predictions compared to actual values can be seen in Figure 5, 6, 7

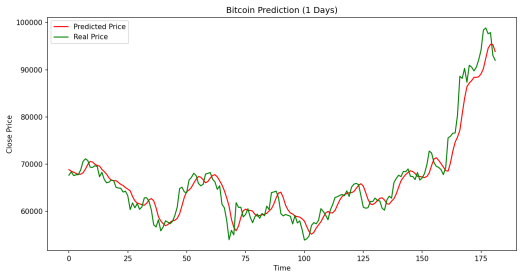


Figure 5: Comparison of Actual and Predicted Values (Price Features)

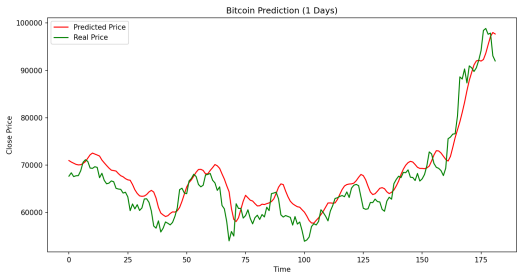


Figure 6: Comparison of Actual and Predicted Values (Market Features)

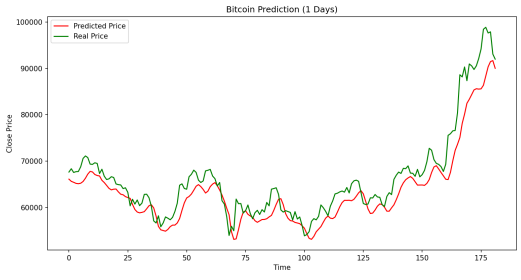


Figure 7: Comparison of Actual and Predicted Values (Price + Market Features)

Based on the above statistical results, we can conclude that the LSTM models trained with different feature sets are generally able to predict the overall trend of Bitcoin prices. However, the model using only the price features performs better both visually and statistically. We speculate that both trading volume and Bitcoin market capitalization are much larger data points compared to Bitcoin prices. Additionally, the volatility of trading volume is more severe, and even though we applied normalization, this feature still has a negative impact on the predictions.

| | **Price** | **Market** | **Price + Market** |
|---|---|---|---|
| MSE | 6717998 | 11478697 | 14398558 |
| MAE | 1915 | 2790 | 2982 |
| $R^2$ | 0.9254 | 0.8726 | 0.8402 |

Table 2: Statistics Result For Different Features

In summary, for the three different feature combinations, we are still unable to predict Bitcoin price movements in advance, nor can we forecast the precise trend of Bitcoin within a short time frame. We can only predict a rough, lagging trend.

## 6 Related literature

### 6.1 Datasets Used in This Study and Prior Work

For this study, we use Bitcoin historical data from CoinCodex, which provides daily records of open, high, low, and close prices, trading volume, and market capitalization. The dataset was preprocessed to create lagged features, such as the previous day's trading volume, market capitalization, and high-low price difference, to capture temporal dependencies in the data. These transformed features were then used as predictors to model the Close price of Bitcoin for the subsequent day.

Similarly, the referenced article (Velankar et al., 2018) utilizes datasets from Quandl and CoinMarketCap, which include time-series data recorded daily for Bitcoin over several years. The data features, such as transaction volume, block sizes, and price movements, were normalized using various techniques like log transformations, z-score normalization, and Box-Cox normalization to ensure data consistency and reduce volatility. These preprocessing steps align closely with our approach, where we also applied logarithmic transformations combined with scaling $\ln(x) + x$ to stabilize variances in our features.

### 6.2 Other Similar Datasets and Their Usage

In addition to CoinCodex, other research has frequently employed datasets from platforms like Yahoo Finance, and social sentiment analysis tools. These datasets often integrate traditional financial metrics with external variables such as macroeconomic indicators, news sentiment, and social media data. For instance, studies combining blockchain-specific features with sentiment analysis have demonstrated improvements in predictive accuracy by capturing both market behavior and public sentiment trends. However, our study focuses exclusively on internal blockchain and market attributes, avoiding external data to streamline the modeling process and reduce noise.

### 6.3 State-of-the-Art Methods and Their Relevance

The referenced article explores various predictive methods, including **Bayesian regression**, **Generalized Linear Models (GLMs)**, **Random Forests**, and **artificial neural networks (ANNs)**. Bayesian regression was applied to identify probabilistic relationships between features and outcomes, while GLMs and Random Forests were used for feature selection and time-series predictions. Neural networks, such as Multi-Layer Perceptrons (MLPs), were employed to model nonlinear relationships in price movements. Additionally, hybrid approaches combining statistical techniques with machine learning algorithms demonstrated improved performance in capturing both short-term and long-term trends.

Our methodology aligns with some aspects of these state-of-the-art approaches. Specifically, we adopted polynomial regression using lagged features to model temporal dependencies in Bitcoin prices. Unlike the article, which uses neural networks and more complex ensemble techniques, we opted for a simpler linear model enhanced with feature transformations ln(x) + x to stabilize the input data. This decision reflects a trade-off between interpretability and predictive power, focusing on computational efficiency and robustness over capturing higher-order nonlinear patterns.

### 6.4 Contributions and Comparisons

Our findings align with the referenced article in emphasizing the importance of robust preprocessing and advanced modeling techniques for Bitcoin price prediction. However, our approach differs in scope and methodology. While the article integrates external factors such as macroeconomic indicators and social sentiment into its predictions, our study focuses exclusively on internal features like price, volume, and market capitalization, offering a streamlined framework.

The article employs a broader range of machine learning techniques, including Bayesian regression, Random Forests, and neural networks, to address Bitcoin's volatility and capture nonlinear relationships. In contrast, our study demonstrates that regression models, when combined with carefully engineered features and transformations such as ln(x) + x, can achieve high performance ($R^2 = 0.94$) while maintaining simplicity and interpretability.

One key distinction lies in the handling of external data. The article's inclusion of external features broadens its analysis but introduces additional complexity in data integration. Our work suggests incorporating such factors as a direction for future research, aiming to build upon the stable foundation established by our internal feature-based models.

## 7 Conclusion

This study demonstrates the utility of machine learning models in predicting short-term Bitcoin price trends. The following key conclusions were drawn:

- **Model Effectiveness:** Regression models with appropriately engineered features performed well in capturing price movements, achieving an $R^2$ of 0.94. LSTM models added the ability to account for temporal dependencies, but their performance degraded slightly when noisy features (e.g., trading volume) were included.
- **Feature Impact:** Price-related features (open, high, low, close) significantly outperformed market-related features (volume, market capitalization) due to their stability and direct correlation with closing prices. Transformations such as $\ln(x) + x$ were critical for improving the models' robustness.
- **Limitations:** The inability to predict long-term trends and reliance on past prices for short-term forecasting highlight the challenges posed by Bitcoin's volatility.

Furthermore, the exclusion of external factors, such as macroeconomic indicators or social sentiment, constrained the model's scope.

In summary, while the developed models successfully predict daily Bitcoin prices within the dataset's constraints, future research should explore incorporating external data, including regional trends and macroeconomic factors, to enhance predictive accuracy and provide more comprehensive insights.

## References

CoinCodex. 2014. Bitcoin (BTC) historical data. Online; accessed 27-Nov-2024.

Siddhi Velankar, Sakshi Valecha, and Shreya Maji. 2018. Bitcoin price prediction using machine learning, feb. Online; accessed 27-Nov-2024.