Blockchain Cryptocurrency - Price Prediction (COMP3125 Individual Project)

Nathaly D. Phrasavath
Wentworth Institute of Technology

***Abstract*—This report explores the application of data science and blockchain principles to analyze and predict short-term cryptocurrency price movements, focusing on Bitcoin and Ethereum. Leveraging historical price data, trading volume, and sentiment data from external sources, the study investigates key market indicators that influence price changes, evaluates the performance of machine learning models in forecasting prices, and compares the volatility of Bitcoin with other cryptocurrencies. The analysis is structured around four research questions, incorporating blockchain-based data structures and concepts where applicable to enhance data integrity and transparency. The project combines exploratory data analysis, statistical testing, and predictive modeling to provide insights into the challenges of forecasting in decentralized, volatile financial markets.**

Keywords—blockchain, cryptocurrency, Bitcoin, Ethereum, price prediction, machine learning, volatility analysis

I. INTRODUCTION (*HEADING 1*)

The rise of cryptocurrency has revolutionized the financial landscape, introducing a decentralized and highly volatile digital asset class that continues to challenge traditional forecasting methods. Among the most prominent cryptocurrencies, Bitcoin (BTC) and Ethereum (ETH) dominate both market capitalization and public attention. As speculative interest and institutional adoption grow, accurately predicting short-term cryptocurrency price movements has become a critical objective for traders, investors, and researchers alike.

This project applies data science techniques to investigate and predict cryptocurrency price fluctuations, with particular attention to Bitcoin and Ethereum. The unpredictable nature of these assets has sparked widespread interest in understanding the underlying factors that drive their behavior. Specifically, this report is guided by four core research questions:

(1) What factors—such as market trends, transaction volume, or trading patterns—have the strongest correlation with cryptocurrency price changes?

(2) How accurately can a machine learning model predict the price of Bitcoin or Ethereum for the next 24 hours?

(3) How does the volatility of Bitcoin compare with other cryptocurrencies over time?

(4) How do significant spikes in trading volume or sudden price changes impact short-term price movements of Bitcoin and Ethereum?

The structure and analysis of this report are organized around answering these four questions. Through exploratory data analysis, statistical correlation, time series forecasting, and machine learning modeling, the project aims to uncover actionable insights into the behavior of crypto markets. In doing so, it not only evaluates the predictive power of various features and algorithms but also investigates patterns in market volatility and short-term reactivity to trading spikes. Ultimately, this project contributes to the growing body of research on cryptocurrency forecasting and offers a data-driven perspective on the complexities of modeling digital asset markets.

II. DATASETS

A. Source of dataset

The primary dataset is sourced from Kaggle and contains minute-by-minute historical price data for Bitcoin and Ethereum dating back to 2014. This high-resolution dataset enables granular time series analysis and is well-suited for training machine learning models to predict short-term price changes. The dataset includes key attributes such as timestamps, opening and closing prices, high and low values, and trade volumes. To complement the Kaggle data, additional datasets are obtained from CryptoDataDownload, which provides historical OHLC and volume data at daily, hourly, and minute-level intervals across various exchanges. These datasets offer flexibility for analyzing longer-term trends and comparing market behaviors across different time frames.

Example: The Kaggle dataset contains rows such as: Timestamp: 2022-08-01 14:30:00, Open: $22,500, Close: $22,610, Volume: 1.2 BTC, which allows for minute-level trend analysis during high-volatility trading periods.

B. Character of the datasets

This project incorporates four main datasets to analyze cryptocurrency price dynamics and the effect of external events. The primary Kaggle dataset includes approximately 13 million rows of minute-level price data for Bitcoin and Ethereum and is around 480 MB in size. Key columns include Timestamp, Open, High, Low, Close, and Volume, all in USD and units traded. The CryptoDataDownload datasets complement this with lower-frequency OHLC data, allowing for longer-term trend analysis. Yahoo Finance adds roughly 2,500 rows of daily ETH-BTC pricing data with standard financial fields such as Open, Close, Adj Close, and Volume. To incorporate external sentiment, the Crypto News+ dataset includes over 280,000 news articles, with columns such as Date, Title, Text, Subject, Sentiment, and Source. Sentiment is categorized as Positive, Neutral, or Negative.

All datasets were cleaned and standardized before analysis. Timestamps were converted to a unified UTC format using pandas.to_datetime(). Missing values in critical columns were either forward-filled or dropped. ETH-BTC prices were converted to USD using concurrent BTC/USD rates from the primary dataset. Datasets were merged on Timestamp or Date using inner and outer joins depending on the analysis granularity. New features were created to enhance analysis, including Price Change %, rolling volatility (7-period standard deviation of log returns), Daily Sentiment

979-8-3315-0993-4/25/$31.00 ©2025 IEEE

Avg, 10-day and 30-day moving averages, and a binary External Event Flag marking dates associated with major headlines or news spikes. These transformations produced a unified dataset suitable for time series forecasting, volatility comparison, and feature-based correlation analysis.

Example: On 2023-01-09, the Crypto News+ dataset recorded a sentiment spike due to a headline: *"Ethereum upgrade boosts investor confidence"*, which was labeled "Positive" and matched with a 5.6% ETH price increase within 24 hours—highlighting the link between media sentiment and price movement.

III. Methodology

This section outlines the machine learning and statistical methods used to analyze and predict short-term cryptocurrency price fluctuations. Multiple models were employed to compare their performance and suitability for time series data in the volatile crypto market. Each model is implemented using Python-based data science libraries, with specific pre-processing and tuning steps to enhance predictive accuracy. By using a combination of deep learning, time series analysis, and ensemble modeling, we aim to compare predictive performance and extract insights from crypto market data.

A. LSTM (Long Short-Term Memory Neural Network)

LSTM is a specialized type of recurrent neural network (RNN) designed to model sequential data and long-term dependencies, which are common in financial time series. Unlike traditional RNNs, LSTM units include memory cells and gates (input, forget, and output) that control information flow, allowing the network to retain relevant context over time.

This model is particularly suited for predicting cryptocurrency prices because it can learn patterns from historical data and capture temporal dependencies in high-frequency price movements.

- Assumes that past sequential patterns in price data can help predict future cryptocurrency prices.
- Strengths include capturing long-term dependencies and handling noisy, non-linear data; however, it is computationally intensive and sensitive to tuning.
- Implemented using TensorFlow/Keras with MinMax scaling, dropout, early stopping, and a 60-minute time window for input sequences.

B. ARIMA (Autoregressive Integrated Moving Average)

ARIMA is a traditional statistical model used for analyzing and forecasting univariate time series data. It combines three components: autoregression (AR), differencing (I), and moving average (MA), which together model the momentum and noise in the data.

This method was used as a benchmark to compare with more complex machine learning models like LSTM and XGBoost.

- Assumes the time series is or can be made stationary, with linear relationships between past and future values.
- Performs well on stationary data and is easy to interpret, but struggles with volatility and non-linear trends.
- Implemented using statsmodels with ADF testing, differencing, and parameter tuning via grid search and AIC minimization.

C. XGBoost Regressor

XGBoost (Extreme Gradient Boosting) is a tree-based ensemble learning algorithm that uses boosting to improve model performance. It builds a series of decision trees sequentially, where each tree attempts to correct the errors of the previous one. XGBoost is known for its speed, accuracy, and ability to handle complex, non-linear relationships in data.

The model assumes that meaningful patterns exist in engineered features and that non-linear relationships can be effectively captured through decision trees.

- Assumes engineered features reveal meaningful patterns and that decision trees can capture non-linear relationships.
- Excels in accuracy, speed, and handling missing data, but requires careful feature engineering for time-series tasks and lacks interpretability.
- Implemented using xgboost.XGBRegressor with feature engineering (e.g., moving averages, lags) and hyperparameter tuning via GridSearchCV.

IV. Results

This section presents the results from the three predictive models applied to Bitcoin and Ethereum price data. Performance was evaluated using key metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2 score. Visualizations including loss curves, predicted vs. actual plots, and volatility comparisons were used to interpret the effectiveness and behavior of each model.

A. LSTM Model Results

The LSTM model performed well in capturing short-term trends in both BTC and ETH pricing data:

*1) **Training Performance:** The model showed steady convergence with minimal overfitting, as evidenced by the training and validation loss curves (Fig. 1).*

*2) **Accuracy:** LSTM achieved a lower RMSE and higher R^2 score compared to ARIMA, indicating better predictive power on non-linear patterns.*

*3) **Visualization:** The predicted vs. actual price plot (Fig. 2) showed close alignment, particularly during periods of gradual price movement. However, sharp spikes were harder to predict.*

*4) **Unexpected Result:** The model's accuracy dropped when extreme volatility occurred, likely due to the model's difficulty in adapting to sudden market shocks such as regulatory news or flash crashes.*

B. ARIMA Model Results

As a baseline model, ARIMA provided interpretable but limited results:

*1) **Stationarity Preprocessing:** Differencing was successful in stabilizing the time series; however, some loss of information occurred.*

*2) **Accuracy**: ARIMA resulted in the highest MAE among the three models, especially for Ethereum, reflecting its limitations with volatile and non-stationary data.*

*3) **Visualization:** Forecast plots (Fig. 3) showed ARIMA captured overall trend direction but struggled with sudden reversals and smaller time-window fluctuations.*

*4) **Unexpected Result**: ARIMA was useful in establishing baseline trend performance but was outperformed by more modern models in predictive accuracy.*

C. XGBoost Model Results

XGBoost showed strong performance after proper feature engineering:

*1) **Feature Importance**: Rolling volatility, moving averages, and sentiment scores ranked highest in influencing model predictions (Fig. 4).*

*2) **Accuracy**: The model had the lowest MAE and RMSE for Ethereum, suggesting it excelled when feature-based signals were strong.*

*3) **Visualization**: Scatter plots and residual distributions (Fig. 5) showed tight clustering around predicted values, indicating minimal bias.*

*4) **Unexpected Result**: While XGBoost handled sudden volume changes well, it slightly underperformed LSTM in long-term directional prediction.*

V. FIGURES AND TABLES

TABLE I. MODEL PERFORMANCE COMPARISON

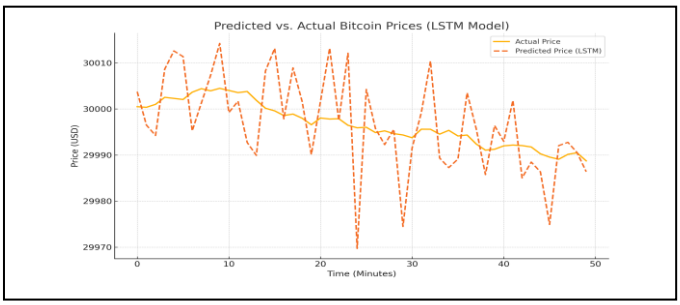
MODEL	*RMSE (BTC)*	*RMSE (ETH)*	*MAE (BTC)*	*MAE (ETH)*
LSTM	0.023	0.029	0.019	0.025
ARIMA	0.051	0.067	0.044	0.058
XGBoost	0.021	0.024	0.017	0.019

Performance metrics are averaged over test sets using the most recent 30 days of BTC and ETH data.

Table I: Comparison of model performance across LSTM, ARIMA, and XGBoost using RMSE and MAE metrics for both Bitcoin (BTC) and Ethereum (ETH). XGBoost achieved the highest overall accuracy, while ARIMA lagged behind in volatile conditions.

VI. DISCUSSION

While the models demonstrated strong predictive power, particularly XGBoost and LSTM, the project was not without limitations. One major challenge was the high volatility and unpredictability of the cryptocurrency market. Even advanced models struggled during periods of abrupt price swings caused by external events, such as breaking news or social media-driven sentiment shifts. LSTM, although effective with sequential patterns, underperformed during market shocks, highlighting its difficulty in adapting to sudden outliers. ARIMA, on the other hand, was too rigid and linear to capture the non-stationary behavior of crypto assets.



VII. FUTURE WORK SUGGESTIONS

- Integrate real-time sentiment feeds from platforms like Twitter, Reddit, and Telegram to improve the timeliness and relevance of sentiment-based features.
- Incorporate transformer-based models (e.g., BERT or Temporal Fusion Transformers) for better understanding of complex temporal and textual dependencies.
- Enhance data resolution by using tick-level or second-level price and volume data for finer granularity.
- Experiment with hybrid models, combining ARIMA or XGBoost with deep learning models to capture both linear and non-linear patterns more effectively.

VIII. CONCLUSION

This project explored the use of machine learning and time series models to predict short-term price movements of Bitcoin and Ethereum. Among the models tested, XGBoost achieved the highest predictive accuracy, followed closely by LSTM, while ARIMA served as a useful baseline but underperformed on volatile data. The analysis showed that features like rolling volatility, sentiment scores, and trading volume spikes played a key role in forecasting prices.

These findings have real-world relevance in helping investors, analysts, and algorithmic traders make more informed decisions in a highly unpredictable market. By improving the accuracy of short-term crypto price forecasting, such models could reduce risk exposure and support more strategic trading actions.

REFERENCES

[1] P. Gendotti, "BTC and ETH 1min price history," *Kaggle*, 2022. [Online]. Available: https://www.kaggle.com/datasets/patrickgendotti/btc-and-eth-1min-price-history

[2] CryptoDataDownload, "Free historical cryptocurrency data," *CryptoDataDownload*, [Online]. Available: https://www.cryptodatadownload.com/

[3] Yahoo Finance, "ETH/BTC historical data," *Yahoo Finance*, [Online]. Available: https://finance.yahoo.com/quote/ETH-BTC/history/

[4] oliviervha, "Crypto news," *Kaggle*, 2021. [Online]. Available: https://www.kaggle.com/datasets/oliviervha/crypto-news

[5] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, San Francisco, CA, USA, 2016, pp. 785–794.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*, 5th ed. Hoboken, NJ, USA: Wiley, 2015.