

Cryptocurrency Price Analytics and Prediction Using a Layered Data Pipeline and Machine Learning

Sheena Patel

Abstract—Cryptocurrency markets generate highly volatile and non-linear time-series data, making price prediction a challenging task. This paper presents an end-to-end cryptocurrency price analytics and prediction pipeline implemented in Python, following a layered Bronze–Silver–Gold architecture. Real-world market data are ingested via a public API, cleaned and analyzed through exploratory data analysis, enriched with time-series feature engineering, and modeled using machine learning techniques. Experimental results demonstrate that feature-engineered datasets significantly improve predictive performance, with linear and ensemble-based models effectively capturing short-term price dynamics.

Index Terms—Cryptocurrency, Data Pipeline, Time-Series Analysis, Feature Engineering, Machine Learning

I. INTRODUCTION

Cryptocurrencies have emerged as a major component of modern financial markets, exhibiting high volatility and complex temporal behavior. Accurate analysis and prediction of cryptocurrency prices require robust data pipelines in addition to predictive models. This work presents a complete analytics pipeline integrating data engineering and machine learning for cryptocurrency price prediction.

II. RELATED WORK

Prior studies have applied statistical and machine learning techniques, including ARIMA, support vector machines, random forests, and deep learning models such as LSTM networks, to cryptocurrency price prediction. While these methods demonstrate predictive potential, many assume pre-cleaned datasets and overlook data pipeline design. This paper emphasizes reproducible pipeline architecture alongside modeling.

III. SYSTEM ARCHITECTURE

The proposed system follows a layered Bronze–Silver–Gold architecture. The Bronze layer stores raw API data, the Silver layer contains cleaned and preprocessed time-series data, and the Gold layer includes feature-engineered datasets optimized for machine learning tasks.

IV. EXPLORATORY DATA ANALYSIS

Exploratory data analysis was conducted to examine historical price trends, return distributions, and volatility behavior. Visual analysis revealed volatility clustering and heavy-tailed return distributions, consistent with financial time-series characteristics.

V. FEATURE ENGINEERING

Feature engineering included lagged price features, simple and exponential moving averages, percentage returns, and rolling volatility measures. These features capture temporal dependencies and trend information essential for effective prediction.

VI. MATHEMATICAL FORMULATION

The mathematical definitions of the key features and evaluation metrics used in this study are presented below.

The percentage return is defined as:

$$r_t = \frac{P_t - P_{t-1}}{P_{t-1}} \quad (1)$$

where P_t denotes the cryptocurrency price at time t .

The simple moving average (SMA) over a window of size n is defined as:

$$SMA_t = \frac{1}{n} \sum_{i=0}^{n-1} P_{t-i} \quad (2)$$

Model performance was evaluated using the Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (3)$$

where y_i and \hat{y}_i represent the actual and predicted prices, respectively.

VII. EXPERIMENTAL RESULTS

Experiments were conducted using historical Bitcoin price data obtained via a public cryptocurrency market API. The final Gold-layer dataset consisted of N observations after preprocessing and feature engineering. An 80/20 chronological train-test split was used to avoid data leakage.

Linear Regression achieved an RMSE of 18.66 USD and an R^2 score of 0.9999, while the Random Forest model achieved an RMSE of 117.56 USD and an R^2 score of 0.9963. RMSE values are reported in U.S. dollars, reflecting absolute price prediction error.

The exceptionally high R^2 score of the Linear Regression model is attributable to strong temporal autocorrelation in cryptocurrency prices and the inclusion of lag-based and trend-oriented features, which enable effective modeling of short-term price dynamics.

TABLE I
MODEL PERFORMANCE COMPARISON

Model	RMSE (USD)	R^2
Linear Regression	18.66	0.9999
Random Forest	117.56	0.9963

VIII. DISCUSSION

Results indicate that feature engineering significantly improves predictive accuracy. While linear models perform well for short-term prediction, ensemble models provide robustness to non-linear patterns but are sensitive to volatility.

IX. CONCLUSION AND FUTURE WORK

This paper presented a complete end-to-end cryptocurrency analytics pipeline integrating data engineering and machine learning. Future work includes multi-asset modeling, deep learning approaches, and real-time streaming pipelines.

AI USE DISCLOSURE

Generative artificial intelligence tools were used to assist with language refinement, structural organization, and clarity of presentation. All experimental design, data collection, feature engineering, model implementation, evaluation, and interpretation of results were performed and verified by the author.

REFERENCES

- [1] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," 2008.
- [2] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, 2009.
- [3] A. Géron, *Hands-On Machine Learning*, O'Reilly, 2019.
- [4] M. Armbrust et al., "Delta Lake: High-Performance ACID Table Storage over Cloud Object Stores," PVLDB, 2020.