Cryptocurrency Price Prediction

Jawaad Ahmar
Dept. of Computer Science
Western University
London, On
jahmar@uwo.ca

Mirna Aziz
Dept. of Computer Science
Western University
London, On
maziz60@uwo.ca

Seyed Hirbod Hosseini
Dept. of Computer Science
Western University
London, On
shoss2@uwo.ca

Ryan Hecht
Dept. of Computer Science
Western University
London, On
rhecht@uwo.ca

***Abstract*—This report addresses the task of predicting significant cryptocurrency price movements in a volatile and dynamic market. We trained and back-tested several machine learning models to predict large price increases, starting with a comparison of approaches at 15%+ gains over a 30-day period, then hyperparameter-tuned an XGBoost model to predict price doubles in under 60 days. We included extensive feature engineering across price, volume, volatility, momentum, and technical indicators. We also implemented a novel train-leader-test-follower split strategy to evaluate model generalization. Comparing results from logistic regression, random forest, XGBoost, and LSTM models for the first evaluation showed that deep learning approaches (LSTM) provided the best performance (F1 score: 0.389), while XGBoost (F1-score: 0.383) provided very similar results with significantly reduced training times. In contrast, the optimized XGBoost model with very high probability thresholds led to exceptional financial performance with 1,689.04% portfolio return on test data but provided very low precision (7.97%). We found strong seasonal patterns in cryptocurrency price behavior, with 'month' being the most important predictor variable. Multi-time-frame models, other data sources, and paper trading for real-time validation could be explored in future work.**

I. Introduction

Cryptocurrency markets are one of the most volatile investment landscapes. Unlike mainstream financial markets with established value models, cryptocurrency prices are highly speculative and can be affected drastically by social media sentiment, regulatory changes, and general speculation. This volatility makes cryptocurrencies potentially lucrative targets for high-return predictive investment modeling.

The ability to predict significant price movements is useful for investors and researchers alike. This study attempts to explore conventional financial forecasting through a research-oriented lens. Our research progressed through two consecutive but complementary stages. First, we set up a comparative platform for evaluating different predictive modeling methods to forecast 15%+ price increases within a 30-day horizon. We identify XGBoost as a model that performs almost as well as our deep learning approach (LSTM) and has a reasonable computational demand for training. We then conduct a more targeted optimization study focused on forecasting price doublings within a 60-day horizon using an XGBoost model.

Our objectives across this research were: (1) to develop a full feature engineering pipeline specific to cryptocurrency data; (2) to implement and contrast a number of predictive modeling approaches; (3) to evaluate model performance on realistic splits that test generalization across market segments; (4) to tune the most viable model to financial performance; (5) to examine the real-world influence of model predictions in real trading environments.

The main contribution of this work is a strong methodology for cryptocurrency price forecasting that emphasizes real-world applicability using realistic dataset divisions, extensive feature engineering, and optimization for financial metrics instead of conventional classification metrics. The results have implications for both algorithmic trading and our understanding of the cryptocurrency market.

II. Background & Related Work

A. Cryptocurrency Market Prediction

Predicting cryptocurrency prices has been an area of research since Bitcoin and other altcoins emerged. For instance, research by Mcnally et al. [1] showcased the potential of using machine learning to predict cryptocurrency prices by demonstrating how LSTM models and Bayesian neural networks can predict Bitcoin prices.

In their evaluation of the effectiveness of different machine learning algorithms for managing cryptocurrency portfolios, Alessandretti et al. [2] discovered that gradient boosting techniques performed better than other strategies when short-term trading windows were taken into account. This supported our conclusions about XGBoost's potent performance.

B. Feature Engineering for Cryptocurrency Prediction

Technical indicators have traditionally been used to predict the stock market. Researchers have applied similar approaches to cryptocurrencies. In particular, Ji et al. [3] strongly support the idea that engineered features derived from historical price and blockchain data can improve prediction accuracy, especially in classification tasks. Therefore, technical indicators, such as moving averages, RSI, and MACD provide significant predictive power for cryptocurrency prices.

More recent approaches have incorporated sentiment analysis and on-chain metrics. For example, Abraham et al. [4] used Twitter sentiment and Google Trends data along with technical indicators to achieve improvements in predictive accuracy. Similarly, Chen et al. [6] incorporated on-chain metrics—including transaction volume, mempool activity, and network performance indicators like block size and hash rate—as predictive features.

Our work builds upon these insights by engineering a comprehensive feature set including price momentum, volatility indicators, technical indicators, and temporal features—finding that seasonality plays an unexpectedly important role in cryptocurrency price movements.

C. Classification vs. Regression in Financial Prediction

Financial forecasting research typically follows either regression approaches (predicting exact price values) or classification approaches (predicting directional movements). Regression models often struggle with the inherent randomness in financial markets, while classification models have shown promise in predicting directional movements. For instance, Campisi et al. [7] compared classification and regression approaches for stock market prediction, finding that classification models provided more actionable investment insights.

Our research extends this direction by focusing specifically on the binary classification of significant price movements rather than minor directional changes or exact price values.

D. Model Evaluation for Financial Applications

A critical issue in financial prediction models is the evaluation methodology. Conventional random splits often lead to optimistic performance metrics that fail to generalize to real-world scenarios. Lopez de Prado [5] introduced the concept of purged cross-validation to address look-ahead bias in financial time series prediction.

Traditional classification metrics like accuracy, precision, and recall may not align with financial performance objectives. Patel et al. [8] proposed using financial metrics such as returns and the Sharpe ratio for evaluation. Our work follows this guidance by optimizing for portfolio returns.

E. Research Gap

Despite extensive research in cryptocurrency price prediction, several gaps remain:

- Most studies focus on major cryptocurrencies but overlook altcoins
- Few studies specifically target significant price movements that represent substantial investment opportunities
- Limited research explores the trade-off between classification metrics and financial performance
- Seasonality effects in cryptocurrency markets remain underexplored

Our research attempts to address these gaps by developing models specifically designed to identify high-potential cryptocurrencies, optimizing for financial returns rather than classification metrics, and investigating seasonal patterns in cryptocurrency price movements.

III. Methods

A. Research Objectives

Our research progressed through two phases with complementary objectives:

1) Phase 1: Model Comparison:

1) Develop a comprehensive feature engineering framework tailored to cryptocurrency time series data that captures price trends, volatility patterns, volume dynamics, and technical indicators
2) Implement and evaluate multiple machine learning approaches for cryptocurrency price prediction, including traditional models (logistic regression, random forest), gradient boosting (XGBoost), and deep learning (LSTM)
3) Design and implement a train-leader-test-follower split methodology to evaluate model generalization from higher market cap cryptocurrencies to lower cap alternatives. This works by taking test data from coins with lower market caps under the expectation that they will follow larger market cap coins.

2) Phase 2: XGBoost Optimization:

1) Develop an optimized XGBoost model targeting cryptocurrency price doublings within the next 60 days
2) Identify optimal features, parameters, and probability thresholds for maximizing financial returnse

B. Research Methodology

1) Data Collection and Preprocessing: We utilized the CoinGecko API to collect historical price, volume, and market capitalization data for over 2,000 cryptocurrencies. The data was recorded daily for 365 days.

Key preprocessing steps included:

- Filtering out stablecoins by analyzing price stability patterns and removing coins that maintained prices within a 3% threshold for over 90% of observations
- Converting date strings to timestamp format
- Sorting data chronologically for time-based splitting

2) API Rate Limiting and Collection Strategy: A significant technical challenge in our research was handling the rate limitations imposed by the CoinGecko API. To solve this, we implemented an internal rate-limiting strategy in our `CoinGeckoAPIScraper` that waited and retried requests when we hit rate limiting:

- Detecting 429 status codes (rate limit exceeded) in API responses
- Implementing dynamic waiting periods based on the `retry-after` header value returned by the API
- Base 1-second delays between consecutive requests to mitigate triggering rate limits
- Batch processing with pagination for collecting large datasets efficiently

This approach allowed us to collect data for 2,000 cryptocurrencies without encountering errors that would have to be manually dealt with. The implementation included error handling and logging to ensure data integrity.

3) Research Evolution: From Sentiment Analysis to Technical Indicators: It is worth noting that our initial research proposal envisioned a different approach focused on sentiment analysis of social media content (particularly Reddit and

Twitter) to predict price movements of low market cap cryptocurrencies (colloquially termed "crap coins"). The original plan included:

- Using BERT-based models for sentiment analysis of cryptocurrency-related social media posts
- Implementing LSTM models to predict price trends based on both market data and sentiment indicators
- Focusing specifically on predicting dramatic price surges ("mooning" events) in highly volatile, low-cap cryptocurrencies

However, practical constraints necessitated a pivot in our research direction:

- The manual daily collection process for low-cap cryptocurrency data was impractical given our project timeline as it only gave us a few weeks of data for each coin
- We wanted to prioritize developing a deeper understanding of core predictive models before expanding to incorporate sentiment analysis

This pivot allowed us to focus on developing predictive models based on technical indicators and price patterns, establishing a foundation for future work that could incorporate sentiment analysis and other alternative data sources.

4) Feature Engineering: We implemented a feature engineering framework that created features across five categories:

a) Price Features:

- Moving averages (7, 14, 30, 60, 90 days)
- Price/MA ratios
- MA crossovers
- Distance from historical highs/lows
- Price momentum (1-day, 7-day, 14-day, 30-day, 90-day percentage changes)

b) Volume Features:

- Volume moving averages
- Volume changes
- Price-volume correlations
- On-Balance Volume (OBV) and its momentum

c) Volatility Features:

- Standard deviation of returns
- Price ranges
- Exponential volatility measures
- Bollinger Bands width

d) Momentum Features:

- Relative Strength Index (RSI) across multiple timeframes (7, 14, 30 days)
- Momentum indicators
- Rate of change metrics
- Moving Average Convergence Divergence (MACD)

e) Time-Based Features:

- Day of week
- Month
- Quarter
- Weekend indicator

We defined two target variables at different phases of our research:

- **Phase 1**: Binary indicator of whether a cryptocurrency's price would increase by at least 15% over a 30-day future horizon
- **Phase 2**: Binary indicator of whether a cryptocurrency's price would double within 60 days

5) Train-Leader-Test-Follower Split: We developed a novel dataset split methodology to better approximate real-world conditions:

1) Cryptocurrencies were ranked by market capitalization
2) The top 50% ("leaders") were allocated to the training set
3) A random sample from the bottom 50% ("followers") was selected for testing

This approach evaluates the model's ability to generalize from patterns observed in established cryptocurrencies to emerging ones. For our XGBoost optimization phase, we supplemented this with a time-based split to prevent look-ahead bias.

6) Handling Class Imbalance: Both of our target events (15% price increase and price doubling) represented minority classes in our dataset, with doubling events constituting less than 1% of observations. To address this severe class imbalance, we implemented a two-step resampling approach:

1) **Undersampling**: Reducing the majority class to a manageable proportion
2) **Oversampling**: Applying Synthetic Minority Oversampling Technique (SMOTE) to create synthetic examples of the minority class

7) Model Development: During our first research phase, we implemented four distinct model architectures:

1) **Logistic Regression**: Baseline linear model with L2 regularization and class balancing
2) **Random Forest**: Ensemble of decision trees with controlled depth to prevent overfitting
3) **XGBoost**: Gradient boosting implementation with hyperparameters tuned for imbalanced classification
4) **LSTM (Long Short-Term Memory)**: Deep learning approach with sequence data preparation, dual LSTM layers, and dropout regularization

For our second phase, we focused on optimizing XGBoost.

8) Threshold Optimization: A key insight from our research was that the default probability threshold of 0.5 was suboptimal for financial performance. We conducted a threshold sensitivity analysis and found that higher thresholds (up to 0.95) maximized portfolio returns despite reducing the number of positive predictions.

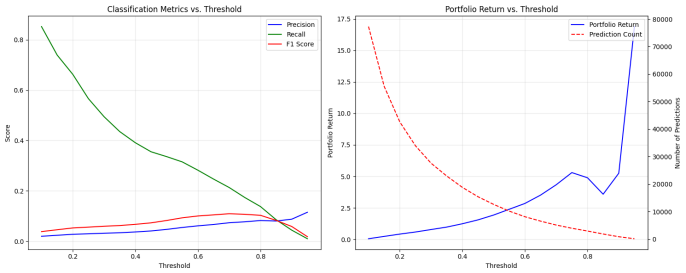

Fig. 1. Left: Precision, recall, and F1 score across decision thresholds. Right: Estimated portfolio return vs. threshold, alongside the number of predictions. A peak return is observed at the optimal threshold, balancing selectivity and gain.

9) Evaluation Framework: We evaluated our models on multiple dimensions:

- **Classification Metrics**:
  - Accuracy, precision, recall, F1-score
  - Confusion matrix
- **Financial Performance Metrics**:
  - Portfolio return (mean return of predicted positives)
  - True positive and false positive returns
  - Win rate and loss rate
  - Profit factor (ratio of gains to losses)
  - Return volatility (standard deviation)
- **Feature Importance Analysis**:
  - Feature importance scores
  - Analysis of top contributors to prediction

This multi-faceted evaluation approach ensured that models were assessed not just on statistical measures but on practical financial outcomes.

IV. Results

A. Model Comparison

Our initial model comparison revealed significant differences in predictive performance across model architectures:

TABLE I
Initial Model Comparison

Model Type	Train F1-score	Test F1-score	Train-Test Gap
LSTM	0.475	0.389	0.086
XGBoost	0.510	0.383	0.127
Random Forest	0.453	0.370	0.083
Logistic	0.346	0.312	0.034

The LSTM model achieved the highest test F1-score (0.389), slightly outperforming XGBoost (0.383); however, XGBoost showed strong performance while requiring significantly less computational resources and training time, making it our choice for further optimization.

Notably, all models exhibited a performance gap between training and testing, reflecting the challenge of generalizing from higher to lower market cap cryptocurrencies.

TABLE II
Initial Financial Performance Metrics

Model Type	ROI (%)	Profit Factor	Win Rate
LSTM	0.51	1.14	0.28
XGBoost	0.42	1.12	0.27
Random Forest	0.19	1.05	0.26
Logistic	-0.61	0.84	0.22

B. Financial Performance Assessment

The initial financial performance metrics for our model comparison phase revealed:

During our threshold optimization experiments with XGBoost, we observed that higher prediction thresholds (e.g., 50%, 85%, 100%) showed better ROI figures (up to 4.16% for the 100% threshold model). While these higher-threshold models showed promising theoretical ROI, their extremely low win rates (below 10%) and increased volatility means that they require much deeper analysis for practical use.

This finding led to our second research phase focused on optimizing XGBoost specifically for financial performance.

C. Optimized XGBoost Performance

Our fully optimized XGBoost model with a probability threshold of 0.95 achieved the following classification metrics on the test dataset when predicting price doublings within 60 days:

- **Accuracy**: 0.9836
- **Precision**: 0.0797
- **Recall**: 0.0864
- **F1 Score**: 0.0829

The confusion matrix revealed:

```
[[201347   1754]
 [  1608    152]]
```

At first glance, these metrics might appear disappointing, particularly the low precision and recall; however, the financial performance tells a dramatically different story:

- **Portfolio Composition**:
  - Total predictions: 1,906
  - True doublings: 152 (7.97%)
  - False doublings: 1,754 (92.03%)
- **True Positive Returns**:
  - Mean Return: 1,044.05%
  - Median Return: 422.01%
  - Min Return: 101.12%
  - Max Return: 7,048.51%
- **False Positive Returns**:
  - Mean Return: -31.88%
  - Median Return: -44.41%
  - Min Return: -96.33%
  - Max Return: 99.34%
- **Overall Portfolio Performance**:
  - Portfolio Return: 1,689.04%
  - Win Rate: 13.12%

– Profit Factor: 9.59

It is extremely important to note that these financial performance metrics must be interpreted with caution because both training and testing data share the same timespan. This means market-wide trends affecting all cryptocurrencies during a specific time period are embedded in both datasets. As a consequence, these results may not necessarily generalize to future market conditions that differ substantially from those observed in our testing period. More research needs to be done to validate this.

D. Feature Importance Analysis

Our feature importance analysis revealed interesting insights into cryptocurrency price movements across both research phases.

In our initial model comparison, momentum indicators (RSI, price momentum) and recent price trends (7-day and 30-day moving averages) provided the strongest predictive signals. Our optimized XGBoost model reveals a different but complementary picture:

Top 10 Features by Importance:

1) month (0.204259) - Indicates strong seasonality in cryptocurrency markets
2) ema_26 (0.064084) - 26-day exponential moving average
3) ema_12 (0.063423) - 12-day exponential moving average
4) market_cap (0.063270) - Market capitalization
5) bollinger_width (0.055262) - Width of Bollinger Bands (volatility indicator)
6) ma_7 (0.046612) - 7-day moving average
7) ath_ratio (0.045261) - Ratio of current price to all-time high
8) drawdown_30d (0.039911) - 30-day drawdown from local high
9) price (0.039163) - Current price
10) rolling_max (0.035950) - Rolling maximum price

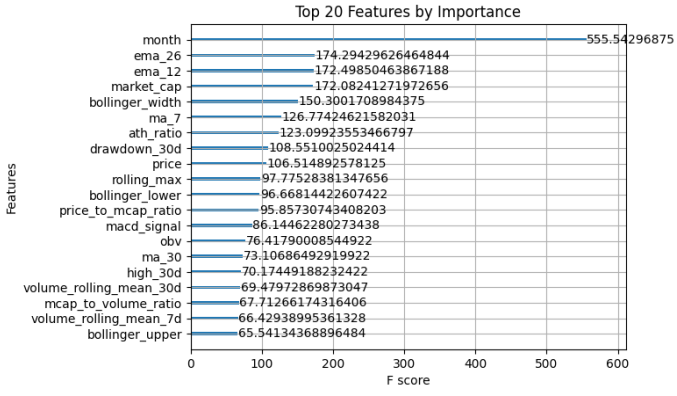

Fig. 2. Top 20 features ranked by XGBoost model importance (F score). Month, EMA, and market metrics are among the most influential.

Interestingly, the most important feature is 'month', accounting for over 20% of the model's predictive power. This suggests a strong seasonal component to cryptocurrency price movements.

E. Market Capitalization and Predictability

We observed a clear relationship between market capitalization and prediction accuracy. The models generally performed better on higher market cap cryptocurrencies, with performance declining as market cap decreased. Smaller cryptocurrencies may exhibit more erratic behavior that is influenced by factors beyond what the technical indicators in our features can capture.

V. Conclusions & Future Work

This research demonstrates that machine learning models can predict significant cryptocurrency price movements with potentially very high financial returns and yields several important insights:

- Deep learning approaches (LSTM) and gradient boosting methods (XGBoost) outperform traditional models for cryptocurrency price prediction
- XGBoost offers a great balance between performance and computational efficiency
- Optimizing for financial returns rather than classification metrics can lead to dramatically different model configurations
- Cryptocurrency markets may exhibit strong seasonality patterns that can be exploited for prediction
- Models likely demonstrate better predictive performance on higher market cap cryptocurrencies

The most striking finding is the financial performance achieved by our optimized XGBoost model, which generated a 1,689.04% portfolio return. However, we must emphasize that these returns were observed in backtesting on a specific market period; therefore, performance levels should not necessarily be expected in future deployments due to changing market dynamics and the shared temporal context between our training and testing data.

A. Limitations

Several limitations of our approach should be acknowledged:

1) **Temporal Validation**: While we implemented a market cap-based split to test generalization across different cryptocurrency segments, our training and testing datasets share the same timespan. This means the model may have captured market trends specific to particular time periods rather than generalizable patterns.
2) **Backtested Performance**: The extraordinary portfolio returns observed (1,689.04%) represent backtested performance in a specific market context and should not be interpreted as returns for future deployments.
3) **Stop-Loss Assumptions**: Our financial evaluations assumed stop-loss orders at constant levels, which may not be achievable in highly volatile market conditions.

B. Future Work

Several promising directions for future research emerge from this work:

- **Incorporate Additional Data Sources**:
  - On-chain metrics (transaction volume, active addresses)
  - Social media sentiment analysis
  - Developer activity metrics (GitHub commits)
- **Implement Real-Time Validation**:
  - Develop a paper trading system to validate model predictions in real-time
  - Compare actual market performance with backtested results
  - Implement automatic model retraining as new data becomes available
- **Explore Risk Management Strategies**:
  - Dynamic stop-loss based on volatility
  - Position sizing optimized for cryptocurrency volatility
  - Portfolio construction techniques to balance high-potential opportunities

Lessons learned include the importance of realistic evaluation methodologies that approximate real-world conditions, the value of comprehensive feature engineering tailored to the unique characteristics of cryptocurrency markets, and the critical need to assess financial performance when developing prediction models for investment applications.

REFERENCES

[1] S. Mcnally, J. Roche, and S. Caton, "Predicting the Price of Bitcoin Using Machine Learning," in *Proceedings of the 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP)*, 2018, pp. 339–343.

[2] L. Alessandretti, A. ElBahrawy, L. M. Aiello, and A. Baronchelli, "Anticipating Cryptocurrency Prices Using Machine Learning," *Complexity*, vol. 2018, pp. 1–16, 2018.

[3] S. Ji, J. Kim, and H. Im, "A Comparative Study of Bitcoin Price Prediction Using Deep Learning," *Mathematics*, vol. 7, no. 10, p. 898, 2019.

[4] J. Abraham, D. Higdon, J. Nelson, and J. Ibarra, "Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis," *SMU Data Science Review*, vol. 1, no. 3, 2018.

[5] M. Lopez de Prado, "The 10 Reasons Most Machine Learning Funds Fail," *The Journal of Portfolio Management*, vol. 44, no. 6, pp. 120–133, 2018.

[6] Z. Chen, C. Li, and W. Sun, "Bitcoin price prediction using machine learning: An approach to sample dimension engineering," *Journal of Computational and Applied Mathematics*, vol. 365, 2020.

[7] G. Campisi, S. Muzzioli, and B. De Baets, "A comparison of machine learning methods for predicting the direction of the US stock market on the basis of volatility indices," International Journal of Forecasting, vol. 40, no. 3, pp. 869–880, Jul. 2024. doi:10.1016/j.ijforecast.2023.07.002

[8] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, "Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques," *Expert Systems with Applications*, vol. 42, no. 1, pp. 259–268, 2015.