Project Report

Title: Predicting Short-Horizon Cryptocurrency Returns Using Time-Series Features and Gradient Boosting

Abstract

This project investigates the feasibility of predicting short-horizon cryptocurrency returns using historical price data and engineered time-series features. A structured data science workflow is followed, including data exploration, feature engineering, model development, and evaluation. An expanding-window cross-validation strategy is used to prevent look-ahead bias. Gradient boosting models are compared against naive benchmark strategies to assess the presence of predictive signal. The results highlight the challenges of noisy financial time series and the importance of appropriate target design and baseline selection.

1. Introduction and Problem Definition

Cryptocurrency markets operate continuously and exhibit high volatility, making short-horizon price prediction a challenging task. Despite the noisy nature of financial time series, prior research suggests that carefully engineered features may capture weak but exploitable patterns.

The objective of this project is to evaluate whether short-term price movements can be predicted using historical market data. Specifically, the task is formulated as a regression problem where the goal is to predict the future return over a fixed horizon based on past price behavior.

Objectives

- Explore and understand the structure of cryptocurrency price data
- Engineer time-series and cross-sectional features
- Train and evaluate a machine learning model for return prediction
- Compare model performance against naive benchmarks
- Discuss limitations and potential improvements

2. Data Exploration and Understanding

2.1 Dataset Description

The dataset consists of minute-level price data for multiple cryptocurrency trading pairs. Each observation includes:

- Timestamp
- Asset identifier
- Closing price
- Target variable (future return)

The target variable is defined as the percentage return over a fixed future horizon.

2.2 Sampling and Structure

An initial inspection of the data reveals millions of observations across multiple assets. A sample of the dataset is shown in Table 1.

Table 1: Sample observations

Timestamp	Asset_ID	Close	Target
t_0	BTCUSDT	93125.2	0.0012
t_1	BTCUSDT	93130.7	-0.0008

2.3 Summary Statistics

Key numerical features were summarized using descriptive statistics, including mean, standard deviation, and percentiles. The target variable exhibits:

- Mean close to zero
- Heavy-tailed distribution
- Significant outliers

2.4 Data Abnormalities

Several characteristics requiring attention were identified:

- Missing values due to asynchronous trading
- Non-stationarity across time
- Large price jumps during volatile periods

These issues motivate careful preprocessing and robust evaluation.

3. Exploratory Visualization

3.1 Target Distribution

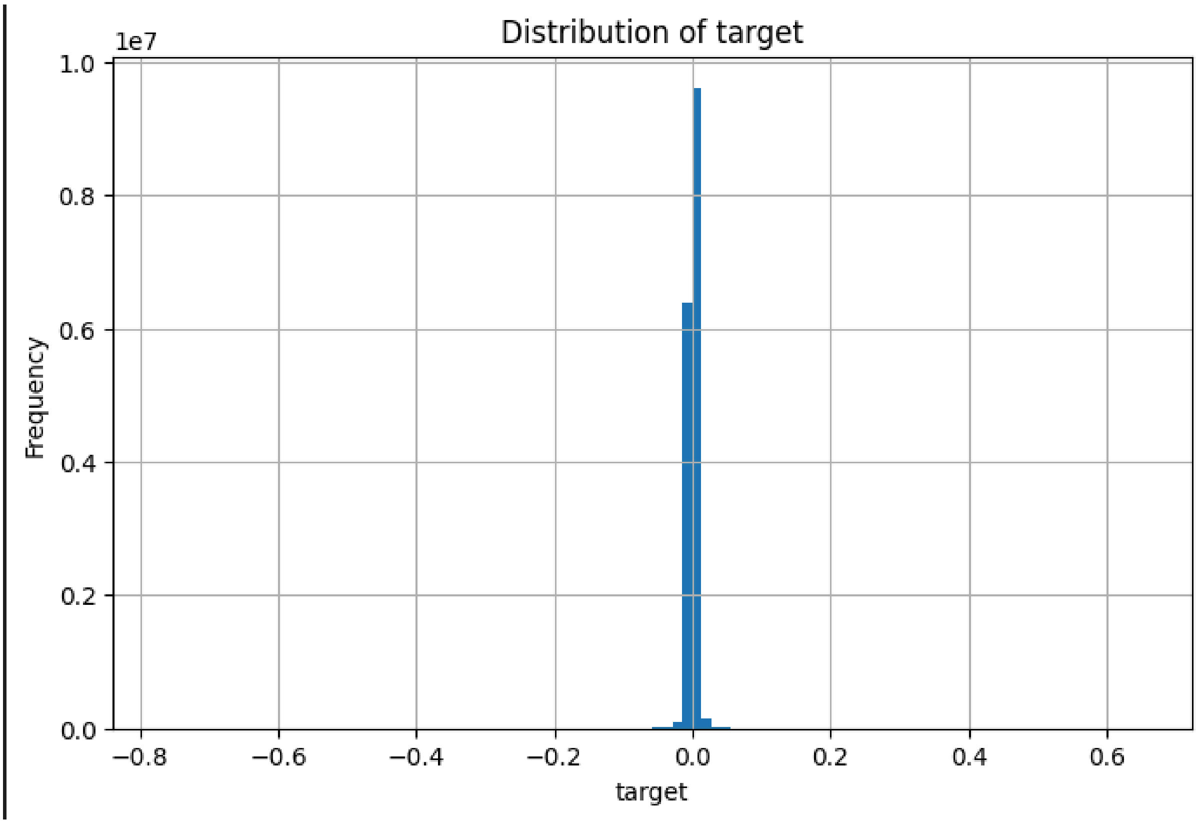

Figure 1: Distribution of future returns

The distribution is centered near zero with heavy tails, indicating that extreme returns occur more frequently than under a normal distribution. This characteristic is typical of financial return series and suggests that prediction will be noisy.

---

4. Feature Engineering and Methodology

4.1 Feature Design

Several categories of features were engineered:

- **Lagged log returns:** capture short-term momentum and mean-reversion
- **Rolling averages and trends:** capture local price dynamics
- **Cross-sectional features:** isolate asset-specific behavior relative to the market

4.2 Target Definition

The prediction target is defined as the future return over a fixed horizon. This formulation allows the model to learn directional and magnitude-based signals but introduces overlapping windows, which must be considered during evaluation.

4.3 Data Alignment

To ensure temporal consistency:

- Each asset is sorted by timestamp
- Missing prices are forward-filled within a limited window
- Feature values are computed using only past information

---

5. Model Development and Training

5.1 Model Selection

A gradient boosting regression model (LightGBM) is used due to its:

- Ability to model non-linear relationships
- Robustness to feature scaling
- Strong performance on tabular data

5.2 Training Strategy

An expanding-window cross-validation strategy is employed:

- Training data includes all observations prior to the validation window
- Validation windows progress forward in time
- This approach prevents look-ahead bias

5.3 Benchmark Models

To contextualize model performance, naive benchmarks are used:

- Copy-last-return baseline

These benchmarks establish a minimum performance threshold.

---

6. Evaluation and Results

6.1 Evaluation Metric

Model performance is evaluated using the Pearson correlation between predicted and realized returns. This metric is closely related to the information coefficient commonly used in quantitative finance.

6.2 Cross-Validation Results

The trained model achieves a small but consistently positive average correlation across validation folds. This indicates the presence of weak predictive signal, though the magnitude remains modest.

6.3 Benchmark Comparison

The model outperforms naive return-based baselines, suggesting that engineered features provide incremental information beyond simple heuristics.

---

7. Discussion

The results demonstrate that while short-time return prediction is extremely challenging, machine learning models can extract limited predictive signal from historical price data. However, the magnitude of this signal is small and sensitive to target design and evaluation methodology.

---

8. Limitations and Future Work

Several limitations are identified:

- High noise-to-signal ratio in returns
- Lack of order book features

Future work could explore:

- Incorporation of volume, funding, or order flow data

---

9. Conclusion

This project presents a structured approach to short-horizon cryptocurrency return prediction using time-series feature engineering and gradient boosting. While predictive performance remains limited, the methodology provides a solid foundation for further exploration and highlights critical considerations in financial modeling, including data leakage, target design, and benchmark selection.

References

1.G-Research Crypto Forecasting

https://www.kaggle.com/competitions/g-research-crypto-forecasting