Machine Learning–Enhanced Pairs Trading in Crypto and Equity Markets

Hand-in date:

9 December 2025

Campus:

BI Oslo

Name:

Dohyeop Lloyd Kim

Contents

Introduction **1**

Method **2**

Data 2

Spread and Z-score Construction 2

Label Construction 3

Feature Engineering 3

Machine Learning Models 3

Trading Rules 3

Results **4**

BTC–IBIT Results 4

S&P 500–IVV Results 6

Conclusion **8**

Reference **8**

GitHub repository
`https://github.com/dlk-4/GRA-41574_1044463`

Introduction

Financial markets move quickly, and short-term traders often struggle to distinguish luck from skill. A profitable trade may feel like pure luck, while a sudden loss raises the question of whether it could have been avoided. This uncertainty creates the need for systematic, data-driven signals that help identify genuine opportunities rather than random fluctuations.

Pairs trading provides one such approach. The idea is intuitive: when two related assets normally move together but suddenly drift apart, the price gap — the spread — often returns to its usual level. Traders can use this temporary divergence to identify potential mean-reversion opportunities that rely less on luck and more on observable structure in the data. Because the rule is simple and grounded in behavior that many markets exhibit, pairs trading has long been viewed as a clear way to study predictable short-horizon returns.

Previous research has built this foundation in several ways. Gatev et al. (2006) introduced a distance-based method for selecting pairs and trading deviations from historical norms. Based on this, Avellaneda et al. (2008) refined the approach using factor-residuals to isolate cleaner mean reversion signals and capture them in U.S. equities. More recently, machine-learning methods have been introduced into statistical arbitrage. Krauss et al. (2017) demonstrated that flexible models can outperform static rules when predicting short-term price movements in U.S. equities.

However, much less is known about whether machine learning performance depends on the characteristics of the market and how it actually adds value across different market environments. Most prior work focuses on stable, highly liquid equity markets. Far fewer studies examine settings where spreads are noisy, volatile, or prone to sudden jumps — conditions typical of the cryptocurrency market. By testing pairs trading with two machine-learning models in both a highly volatile crypto market and a highly efficient equity market, we create a natural testbed for evaluating whether machine learning-based probability estimates can enhance Z-score trading under different market conditions. Comparing these two environments allows us to see how market structure influences the effectiveness of machine-learning signals, offering insight into when ML-based methods are likely to succeed or fail. This question has received little empirical attention, and our study will address it directly.

Therefore, this study allows us to revisit the core question of statistical arbitrage from a new angle: *Can machine learning improve Z–score–based spread trading across different market environments?*

Method

Data

This study uses two related sets of financial data downloaded from Yahoo Finance, covering the period 11 January 2024 to 28 October 2025.

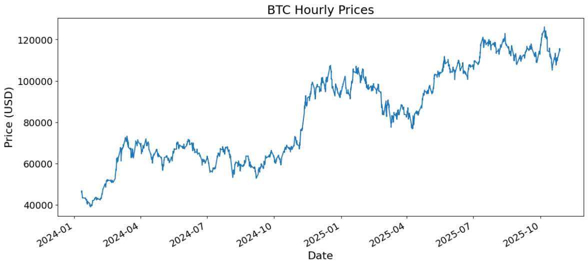

Figure 1: BTC price

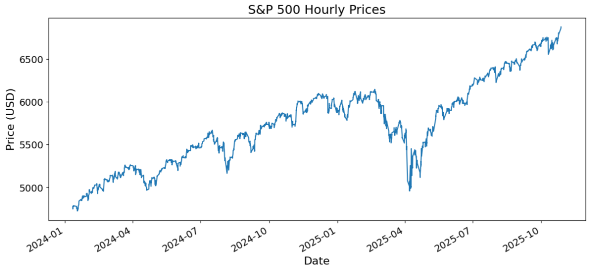

Figure 2: S&P 500 price

The first dataset contains daily closing prices for Bitcoin (BTC-USD), iShares Bitcoin Trust (IBIT), MicroStrategy (MSTR), the S&P 500 Index (GSPC), and the iShares Core S&P 500 ETF (IVV). These assets are included to examine the correlation structure across crypto-linked instruments and equity index products. For the second dataset, two spreads are constructed, BTC–IBIT and S&P 500–IVV, using hourly price data for the pair-trading analysis.

Bitcoin trades 24/7, while S&P 500 index, IBIT, and IVV trade only during U.S. equity hours. To ensure comparability, BTC-USD is restricted to New York trading hours (10:00-16:00 EST). Therefore, both spreads reflect the same intraday trading window.

Spread and Z-score Construction

For a pair consisting of asset A and asset B, the **spread** at time t is defined as

$$\text{spread}_t = P_t^A - P_t^B,$$

where P_t^A and P_t^B denote the synchronized hourly closing prices.

Using a rolling window of length w, the mean and standard deviation of the spread are computed as:

$$\mu_{t,w} = \frac{1}{w}\sum_{i=0}^{w-1} \text{spread}_{t-i}, \qquad \sigma_{t,w} = \sqrt{\frac{1}{w}\sum_{i=0}^{w-1} \left(\text{spread}_{t-i} - \mu_{t,w}\right)^2}.$$

The Z-score at time t is then computed as

$$Z_t = \frac{\text{spread}_t - \mu_{t,w}}{\sigma_{t,w}}.$$

Label Construction

The target y_t is defined as

$$y_t = \begin{cases} 1, & \text{if } |Z_{t+1}| < |Z_t|, \\ 0, & \text{otherwise.} \end{cases}$$

$y_t = 1$ represents a successful mean-reversion event, while $y_t = 0$ indicates either further divergence or no meaningful correction.

Feature Engineering

The feature set in the machine learning model includes:

- Z_t : the standardized spread
- $|Z_t|$: the absolute magnitude of the deviation
- ΔZ_t : the recent change in the Z-score
- $\texttt{spread\_ret}_t$: the hourly return of the spread
- $\texttt{vol\_spread}_t$: rolling volatility of the spread
- $\texttt{corr}_t$: rolling correlation between the two assets
- $\texttt{beta}_t$: rolling beta of asset A relative to asset B

Machine Learning Models

Logistic Regression

$$P(y_t = 1 \mid X_t) = \sigma(\beta_0 + \beta^\top X_t), \qquad \sigma(z) = \frac{1}{1 + e^{-z}}.$$

HistGradientBoosting

$$P = \sigma\left(\sum_{m=1}^{M} \eta\, h_m(X)\right)$$

Trading Rules

Z-condition

$$\begin{aligned} Z_t &< \text{lower bound} \;\Rightarrow\; \text{long in } P_t^A \text{ , short in } P_t^B \\ Z_t &> \text{upper bound} \;\Rightarrow\; \text{short in } P_t^A \text{ , long in } P_t^B \end{aligned}$$

ML probability filter

$$P(y_t = 1) > p_{\text{long}},$$
$$P(y_t = 1) > p_{\text{short}}.$$

Results

BTC–IBIT Results

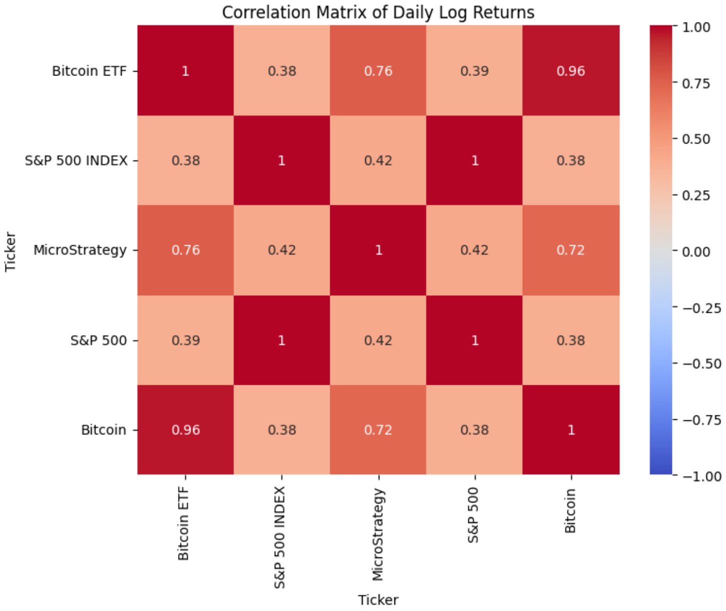

Figure 3: Correlation Matrix

Correlation is used to verify co-movement between assets, which is essential for constructing spreads in pairs trading. To obtain stable covariance estimates, log returns are used for the first dataset. Assets with strong historical relationships tend to generate spreads with clearer mean-reversion signals. As shown in Figure 3, IBIT and MicroStrategy closely track Bitcoin, while IVV closely tracks the S&P 500 index. This motivates the selection of BTC–IBIT and S&P 500–IVV as trading pairs. These pairs also represent two distinct market environments. BTC–IBIT reflects a volatile, 24/7 crypto market with frequent short-term imbalances, whereas S&P 500–IVV represents a highly efficient and tightly arbitraged equity market. This contrast shows how mean-reversion signals behave across markets with very different levels of volatility, efficiency, and structure.

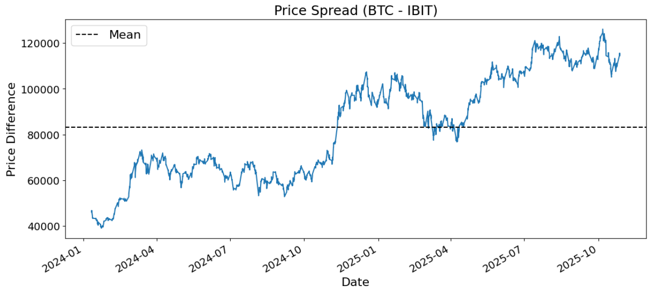

Figure 4: Price spread (BTC - IBIT)

Figure 4 displays the BTC–IBIT spread, which fluctuates around a slowly shifting mean but exhibits occasional sharp, short-lived deviations. Examining the raw spread helps identify whether deviations occur frequently enough to generate meaningful trading signals. These temporary divergences highlight the presence of mean-reversion dynamics. Since the magnitude of deviations changes across the sample, a rolling Z-score is used to standardize the spread and ensure that signals are comparable over time.

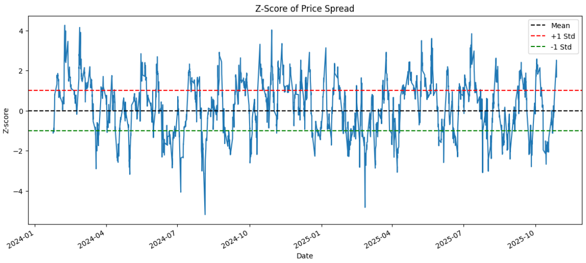

Figure 5: Rolling Z score of price spread

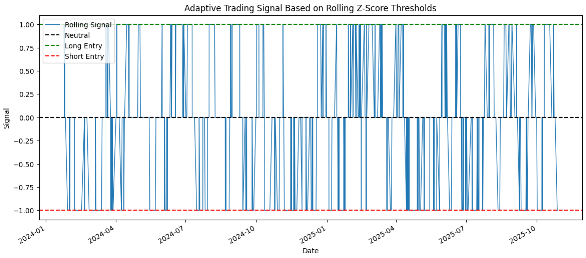

Figure 6: Trading signal

Figure 5, Figure 6 show the rolling Z-score of the BTC–IBIT spread and the corresponding long/short signals generated from fixed ±1 thresholds. Following standard practice in statistical arbitrage (Avellaneda et al. 2008), a rolling window $w = 60$ is used for the BTC–IBIT pair. In Figure 5, the rolling Z-score standardizes the spread and reveals mean-reversion opportunities. Figure 6 visualizes the long/short signals triggered by the ±1 thresholds, indicating how often the strategy identifies potential reversion points.

However, since the spread volatility changes over time, a fixed ±1 threshold may not consistently capture the strongest signals. This motivates optimizing the upper and lower thresholds.

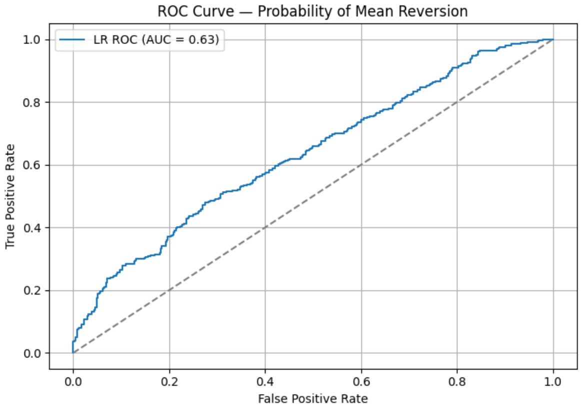

Figure 7: ROC Curve - Logistic Regression

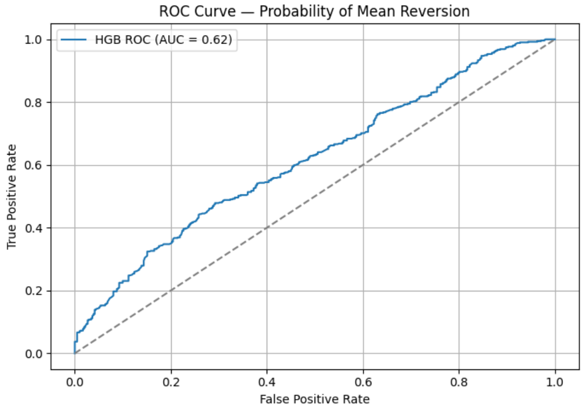

Figure 8: ROC Curve - HistGradientBoosting

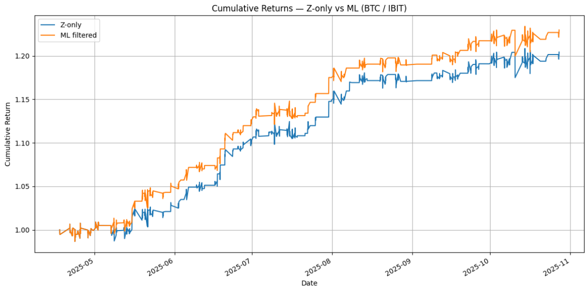

Figure 9: Cumulative Return - Logistic Regression vs Z

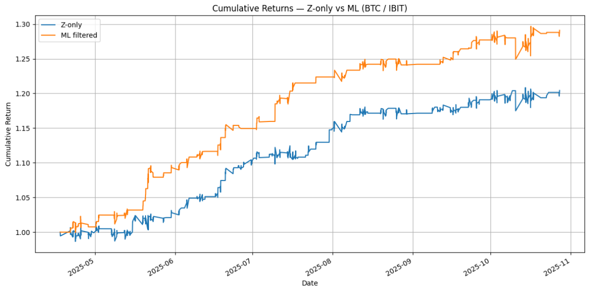

Figure 10: Cumulative Return - HistGradientBoosting vs Z

The Z-score thresholds are first optimized on the training set (70%) by selecting the bounds that maximize the Sharpe ratio. The Sharpe ratio is appropriate for this setting because the trading strategy is executed frequently and has varying volatility. Using this approach, the BTC–IBIT pair produces frequent long/short signals due to its high volatility. The optimized thresholds (upper: 0.63, lower: -1.06) are then applied to the 30% test set. Probability cutoffs for the ML models are chosen using the same procedure.

Logistic Regression is selected as the main model because it is fast, interpretable, and well-suited for high-frequency financial data. To capture nonlinear and interaction-driven reversion patterns that Logistic Regression misses, a HistGradientBoosting model is also fitted. Figure 7 and Figure 8 show the ROC curves for the two models. On the test set, Logistic Regression achieves an AUC of 0.63, while HistGradientBoosting achieves a similar AUC of 0.62, indicating modest but meaningful predictive ability. They are trained on the 70/30 split using features derived from the spread, with feature $|Z_t|$ receiving the highest importance because large deviations are more likely to revert.

Figure 9 shows that Logistic Regression filters out noisy trades and delivers higher cumulative returns than the Z-only baseline. On the test set, the cumulative return increases from 1.204 to 1.2297 (+2.13%), and the Sharpe ratio rises from 1.92 to 2.16. Figure 10 shows that HistGradientBoosting further improves returns to 1.291 (+7.2%) and Sharpe to 3.101, reflecting its ability to capture nonlinear patterns and asymmetric movements in the spread. Overall, ML-filtered trades outperform the optimized Z-rule on BTC–IBIT, with HistGradientBoosting providing the strongest risk-adjusted performance.

S&P 500–IVV Results

The same methodology applied to the BTC–IBIT pair is now evaluated on the S&P 500–IVV spread. The S&P 500–IVV spread, however, is substantially more stable. To quantify the difference, a spread variance ratio is computed:

$$\frac{\mathrm{Var}(\text{BTC–IBIT spread})}{\mathrm{Var}(\text{S\&P 500–IVV spread})} = 2595.36.$$

This variance ratio highlights how differently the two spreads behave and explains why the S&P 500–IVV pair produces far fewer and smaller mean-reversion deviations. The

S&P 500–IVV exhibits much lower variance than BTC–IBIT spread, confirming that the equity pair is far more stable, tightly arbitraged, and less prone to large deviations. Since deviations in S&P 500–IVV are very small and short-lived, a long window such as $w = 60$ obscures nearly all meaningful Z-score deviations. A shorter window of $w = 8$ is therefore used to capture the microstructure-level fluctuations, as it provides the best balance between generating sufficient signals and limiting noise across the window lengths tested.

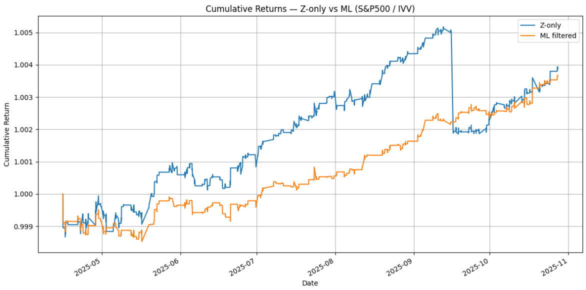

Figure 11: Cumulative Return - Logistic Regression

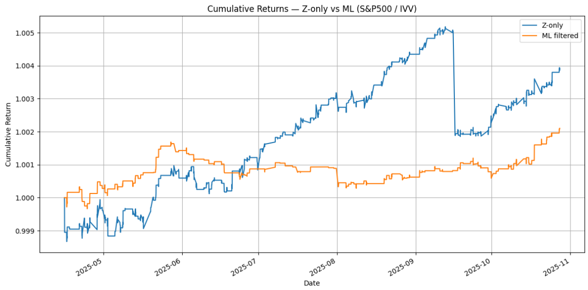

Figure 12: Cumulative Return - HistGradientBoosting

Figure 11 shows that the Z-score strategy produces a smooth and steadily rising equity curve with limited day-to-day variation, reflecting the stability and limited number of trading opportunities in this market. However, a notable exception occurs in late September 2025, when the Z-score rule triggers a sharp and abrupt drop. After this isolated event, the curve stabilizes and resumes its slow upward trend, indicating that the drop was driven by a single misfiring signal rather than a broader change in market conditions.

In Figure 11, the cumulative-return curves for the Logistic Regression–filtered strategy demonstrate clear efficiency gains. Logistic Regression reduces the number of trades from 327 to 268 while preserving an almost identical overall return (Z-only: 1.0039 vs. ML: 1.0037) and increasing the Sharpe ratio significantly (1.12 → 2.042). This indicates that the probabilistic filter effectively removes marginal, noise-driven trades.

However, Figure 12 shows that HistGradientBoosting performs significantly worse than both the Z-only rule and Logistic Regression. The ML-filtered equity curve fails to follow the upward movement of the Z-score baseline and remains consistently lower throughout the period. Although the model reduces the number of trades from 327 to 227 and produces a slightly higher Sharpe ratio (1.1235 → 1.6422), this improvement stems mainly from its lower volatility rather than from better trading performance. In practice, HGB filters out too many potentially profitable signals, causing the strategy to miss most of the limited opportunities available in this highly efficient spread. As a result, HGB offers strong noise reduction but substantially weaker returns, making it far less effective than Logistic Regression for this market.

The ML models behave differently in the S&P 500–IVV pair. Logistic Regression delivers better performance by reducing unnecessary trades and avoiding major Z-score misfires. HistGradientBoosting becomes overly conservative and fails to follow the Z-score baseline, resulting in weaker cumulative returns in this highly efficient market.

Conclusion

This study set out to evaluate whether a combination of interpretable and flexible machine-learning models can enhance classical Z-score–based spread trading by predicting short-term mean reversion. By examining two fundamentally different environments, this study provides new empirical insight into classical pairs trading.

For BTC–IBIT, both Logistic Regression and HistGradientBoosting outperformed the optimized Z-score rule by removing noisy trades, raising Sharpe ratios, and producing smoother returns—showing that ML adds value when the spread contains short-lived but informative deviations. In contrast, the S&P500–IVV results show that ML uncovers limited predictability in this highly efficient market: Logistic Regression improves stability by reducing fragile trades, whereas HistGradientBoosting becomes overly conservative and fails to capture the few available opportunities.

There are important limitations to acknowledge. The models' predictive accuracy for BTC–IBIT is modest ($AUC \approx 0.63$), meaning they should not be used as a standalone predictor of mean reversion. Even so, this level of performance is still effective for filtering out low-quality trades in markets with informative deviations. In highly efficient markets, ML contributes mainly through risk control via Logistic Regression, while HistGradientBoosting becomes overly conservative and misses the limited opportunities. Therefore, ML-based filters should be applied cautiously in efficient equity markets, as their conservativeness can suppress the few trading opportunities that exist.

Returning to the main research question—*Can machine learning improve Z–score–based spread trading across different market environments?*—the results provide a balanced answer. **Yes**, machine learning offers meaningful improvements when the market generates informative deviations. **However**, in highly efficient markets, ML adds limited value: Logistic Regression provides only modest risk control, while HistGradientBoosting becomes overly conservative and weakens performance. Overall, this study highlights both the promise and the limits of ML-enhanced statistical arbitrage and reveals that ML's effectiveness in pairs trading is **market-dependent rather than universal**. Understanding this distinction can help traders apply ML tools more appropriately, improving the practical application of machine learning within market microstructure.

References

Avellaneda, Marco and Jeong-Hyun Lee (2008). "Statistical arbitrage in the US equities market". In: *Quantitative Finance* 10.7.

Gatev, Evan, William N. Goetzmann, and K. Geert Rouwenhorst (2006). "Pairs Trading: Performance of a Relative-Value Arbitrage Rule". In: *Review of Financial Studies.*

Krauss, Christopher, Xuan A. Do, and Nicolas Huck (2017). "Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500". In: *European Journal of Operational Research* 259.2.