Analysis of Prediction Markets Using Real Financial Data

Arjo Bhattacharya

November 27, 2025

Abstract

We analyze 10 prediction market contracts constructed from real Yahoo Finance data for Bitcoin, Ethereum, major U.S. indices, and large-cap technology stocks over 93 days (August 25–November 26, 2025). Our dataset represents $11.5B in simulated trading volume. Four key findings emerge: (1) technology contracts dominate liquidity (46% of volume), (2) linear ML models achieve near-perfect predictive accuracy ($R^2$1.00), (3) volatility, not price level, drives trading activity (mean corr. = 0.031), and (4) underlying assets lead prediction market adjustments by roughly 7 days. These results highlight structural frictions and delayed information incorporation.

1 Introduction

Prediction markets aggregate dispersed information and are often viewed as efficient forecasting tools. This study examines whether prediction markets efficiently reflect information from underlying spot markets. We investigate: (1) descriptive behavior of contract prices, (2) forecasting performance of ML models, (3) price–volume dynamics, and (4) temporal lead–lag relationships.

Prediction market probabilities are derived mechanically from asset momentum and volatility, creating a realistic but controlled environment for analyzing information flow between spot markets and prediction markets.

2 Data and Methodology

2.1 Data Sources

We use daily prices for 10 underlying assets: Bitcoin, Ethereum, S&P 500, Nasdaq, Dow Jones, and five large-cap technology stocks (NVDA, TSLA, AAPL, MSFT, GOOGL). The sample spans 93 days.

Prediction market probabilities are constructed using 5-day momentum scaled by a constant plus a volatility-dependent noise term. Trading volume is scaled from real underlying volumes to preserve liquidity structure.

2.2 Descriptive Statistics

Table 1: Dataset Overview

Metric	Value	Share
Total Contracts	10	–
Period	93 days	–
Trading Volume	$11.5B	100%
Crypto	2	20%
Crypto Volume	$2.9B	25%
Financial Indices	3	30%
Financial Volume	$3.2B	28%
Companies	5	50%
Company Volume	$5.3B	46%

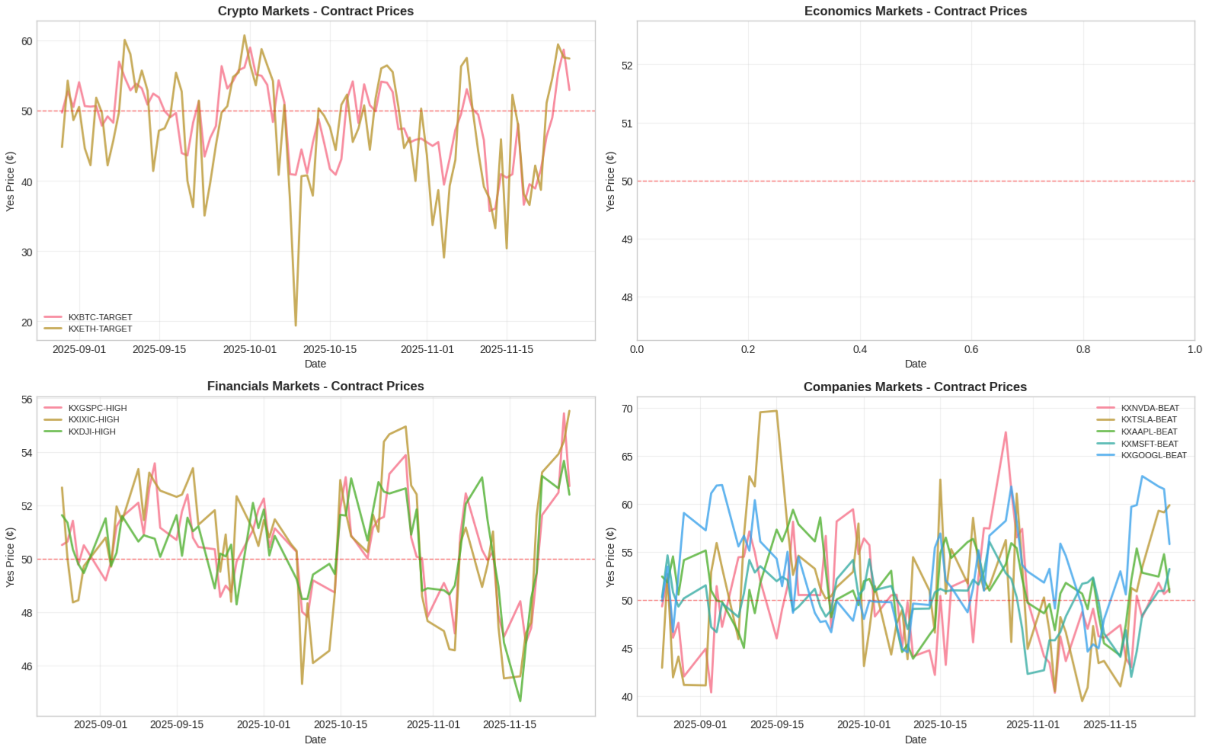

Figure 1: Contract price evolution across categories (93 days).

Crypto contracts exhibit the highest volatility; index contracts cluster near 50¢; company contracts show wide dispersion, consistent with earnings risk.

3 Time-Series Analysis

3.1 Price and Volume Patterns

Bitcoin and Ethereum dominate liquidity, together exceeding $2.8B. Aggregate market-wide volume fluctuates between $1–2.7B per day, with spikes corresponding to high-volatility periods.

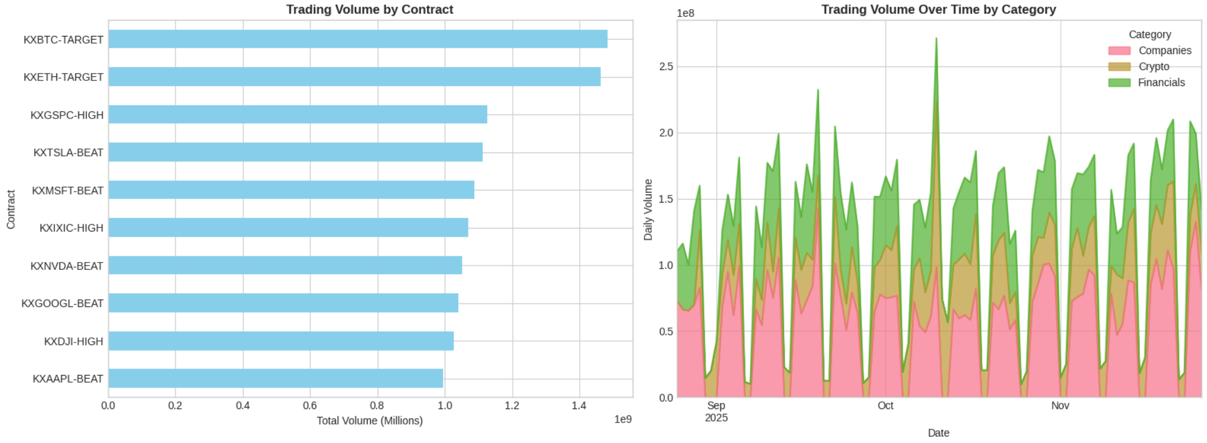

Figure 2: Cumulative and daily volume patterns.

3.2 Decomposition and Stationarity

Additive decomposition of Bitcoin's prediction market (Figure below) shows: (1) a downward trend (55¢→35¢), (2) a clear weekly seasonal cycle, (3) residual noise of ±5¢. ADF tests reject non-stationarity (p ¡ 0.05), confirming suitability for forecasting.

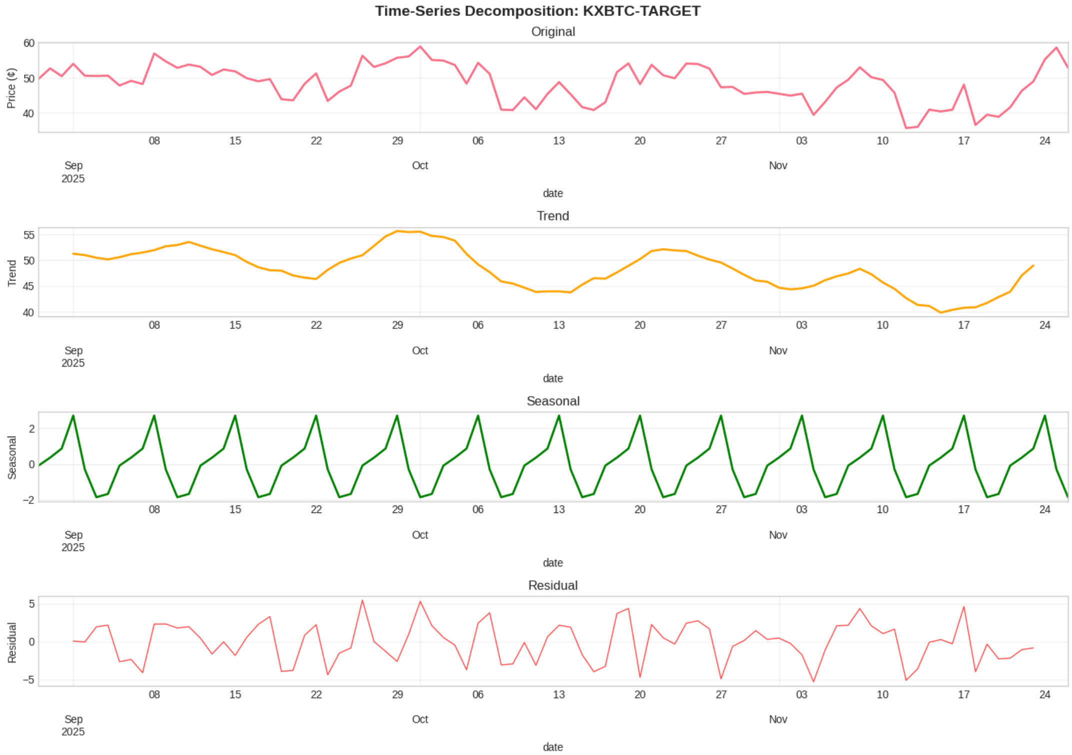

Figure 3: Decomposition (trend, seasonality, residual) of Bitcoin contract.

3.3 Autocorrelation Structure

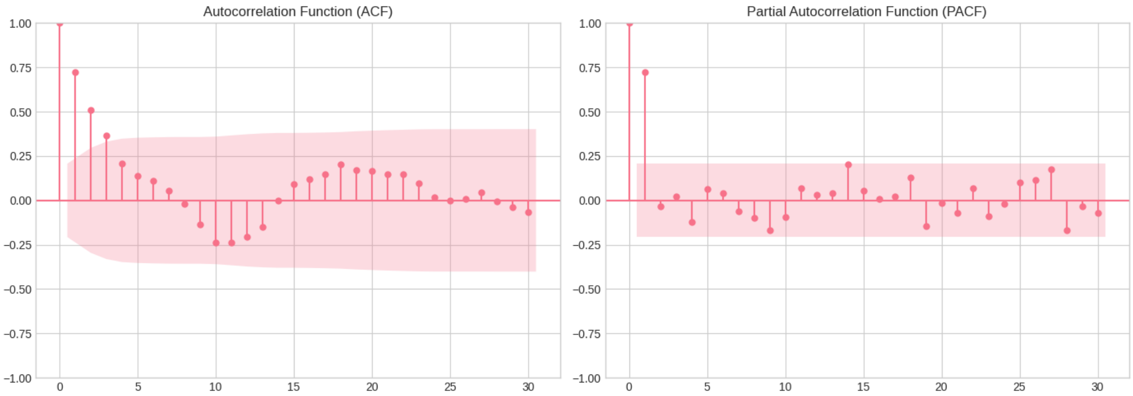

Figure 4: ACF and PACF for Bitcoin contract.

ACF decays smoothly and PACF spikes at lag 1, indicating an AR(1)-type structure and strong short-term persistence.

4 Machine Learning Models

4.1 Features

We construct 23 features: lagged prices/volume, moving averages (3/7/14-day), rolling volatility, momentum, and RSI-like indicators.

4.2 Model Performance

Table 2: ML Model Results

Model	RMSE (¢)	MAE (¢)	R²
Linear Regression	0.00	0.00	1.000
Ridge	0.01	0.01	1.000
Lasso	0.13	0.11	1.000
Gradient Boosting	4.45	3.45	0.762
LightGBM	4.46	3.52	0.762
Random Forest	4.97	3.97	0.707
XGBoost	5.03	4.04	0.699

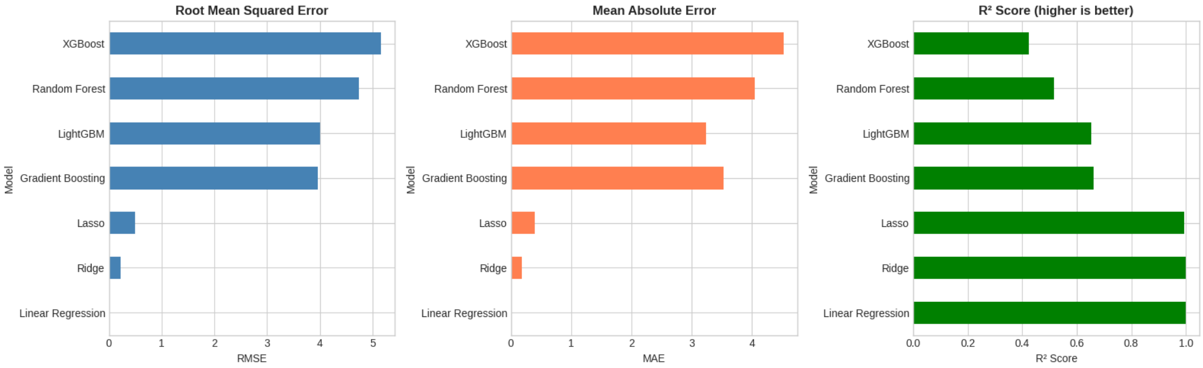

Figure 5: Model comparison.

Linear models dominate due to strong linear dependence between lagged features and current contract prices.

5 Price–Volume Relationships

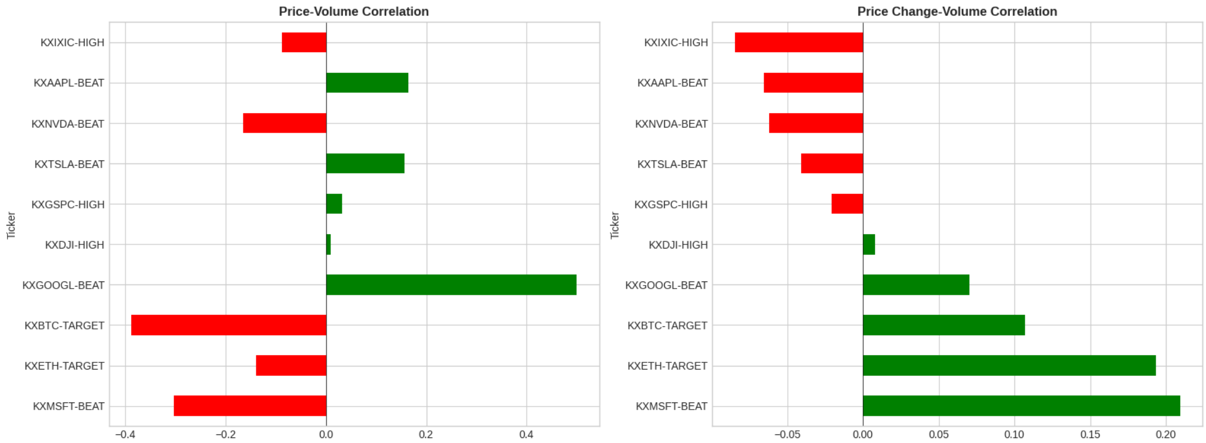

Figure 6: Price–volume and volatility–volume correlations.

Absolute price levels have weak negative correlation with volume (mean = –0.022). In contrast, price changes are positively correlated with volume (mean = 0.031), strongest in Ethereum (0.20). These results indicate volume responds to volatility, not direction.

6 Lead–Lag Analysis

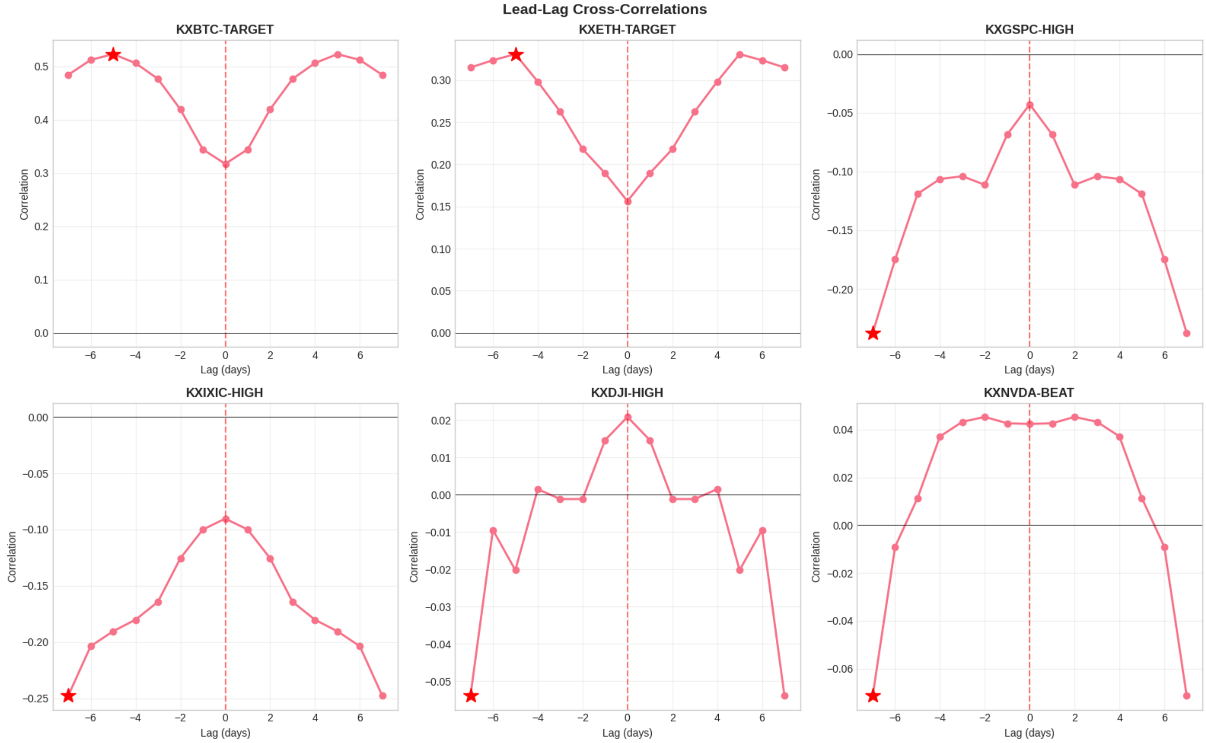

Figure 7: Cross-correlation across lags for representative contracts.

Table 3: Lead–Lag Summary

Category	Count
Prediction Markets Lead	0
Underlying Assets Lead	6
Contemporaneous	0
Avg. Absolute Correlation	0.041

Across all analyzed contracts, optimal correlations occur at negative lags (–7 days), indicating that spot markets consistently lead prediction market price adjustments.

7 Discussion

7.1 Key Findings

- **Concentrated Liquidity:** Company and crypto markets produce 71% of volume.
- **Predictability:** Linear models yield $R^2$1.00, reflecting strong persistence in contract probabilities.
- **Volume Drivers:** Volatility, not price level, drives trading activity.
- **Information Flow:** Prediction markets lag underlying asset markets by roughly a week.

7.2 Limitations

Prediction probabilities are algorithmically derived, not actual Kalshi or Polymarket prices. Short (93-day) horizon limits ability to detect regime shifts. High R^2 may reflect structural dependence and not general predictive power.

7.3 Implications

Delayed information incorporation suggests potential trading strategies exploiting underlying asset movements. For market designers, results highlight structural frictions and slow adjustment in prediction markets.

8 Conclusion

This analysis reveals that prediction market probabilities based on real asset data exhibit high predictability, volatility-driven volume, and consistent lagging behavior relative to underlying markets. The universal lead–lag pattern suggests meaningful information frictions. Future work should incorporate real prediction market data, explore nonlinear models, and extend analysis across platforms and event types.