Intraday VIX Term-Structure Signal on ES

A Quantitative Research Study

1. Motivation

The VIX index reflects the market's expectation of 30-day implied volatility, while VIX3M measures 90-day implied volatility. The difference between the two represents the term structure of implied volatility, which may contain information about short-term risk sentiment.

The goal of this project is to test whether the VIX–VIX3M slope exhibits predictive power for future intraday ES returns, specifically the next 60-minute log return.

2. Data and Preprocessing

I use:

- ES 1-minute OHLCV data (Regular Trading Hours, 09:30–16:00 US/Eastern);
- Daily VIX and VIX3M data, forward-filled within each day (no lookahead);
- Aligned timestamps using timezone-aware conversion.

A key limitation is that I only get access to VIX3M data begins in late 2007. As a result, most of the ES sample (2005–2007) cannot be used for feature construction. Only November–December 2007 form a clean training window.

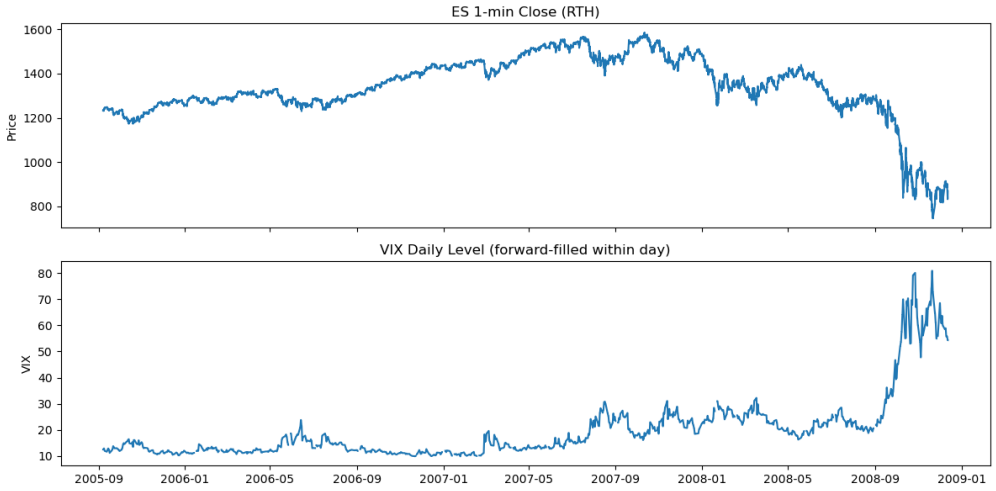

Figure 1: ES 1-minute Close (top) and VIX Daily (bottom).

3. Feature Engineering

3.1 Log Returns

$$r_{1m}(t) = \log\left(\frac{P_t}{P_{t-1}}\right), \qquad r_{60m}(t) = \log\left(\frac{P_{t+60}}{P_t}\right)$$

3.2 VIX Term Structure

$$\text{Slope}(t) = \text{VIX}(t) - \text{VIX3M}(t)$$

A rolling Z-score is computed using a 7700-minute window ($\approx$ 20 trading days):

$$\text{ZSlope}(t) = \frac{\text{Slope}(t) - \mu_t}{\sigma_t}.$$

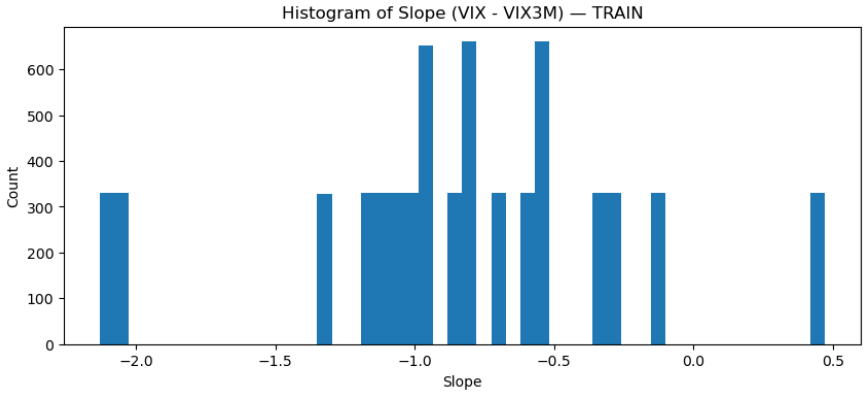

Figure 2: Distribution of Slope (VIX–VIX3M) on the training set.

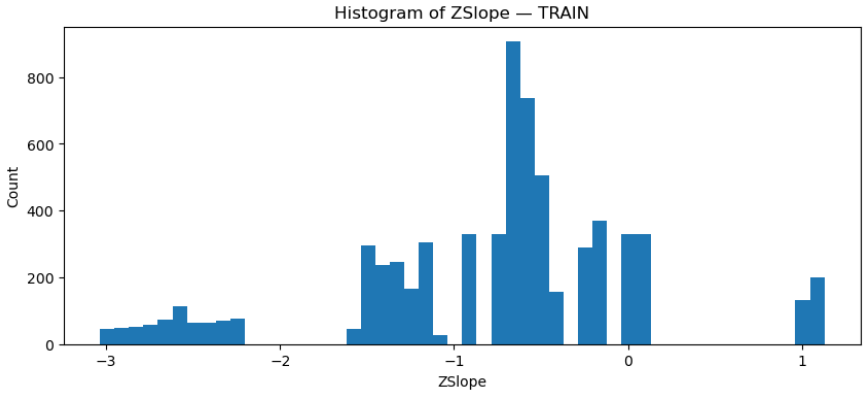

Figure 3: Distribution of the Z-scored slope (ZSlope) on the training set.

4. Exploratory Signal Analysis

4.1 Correlation

The Pearson correlation between ZSlope and future 60-minute return is:

$$\rho \approx -0.039,$$

indicating almost no linear predictive power.

4.2 Scatter

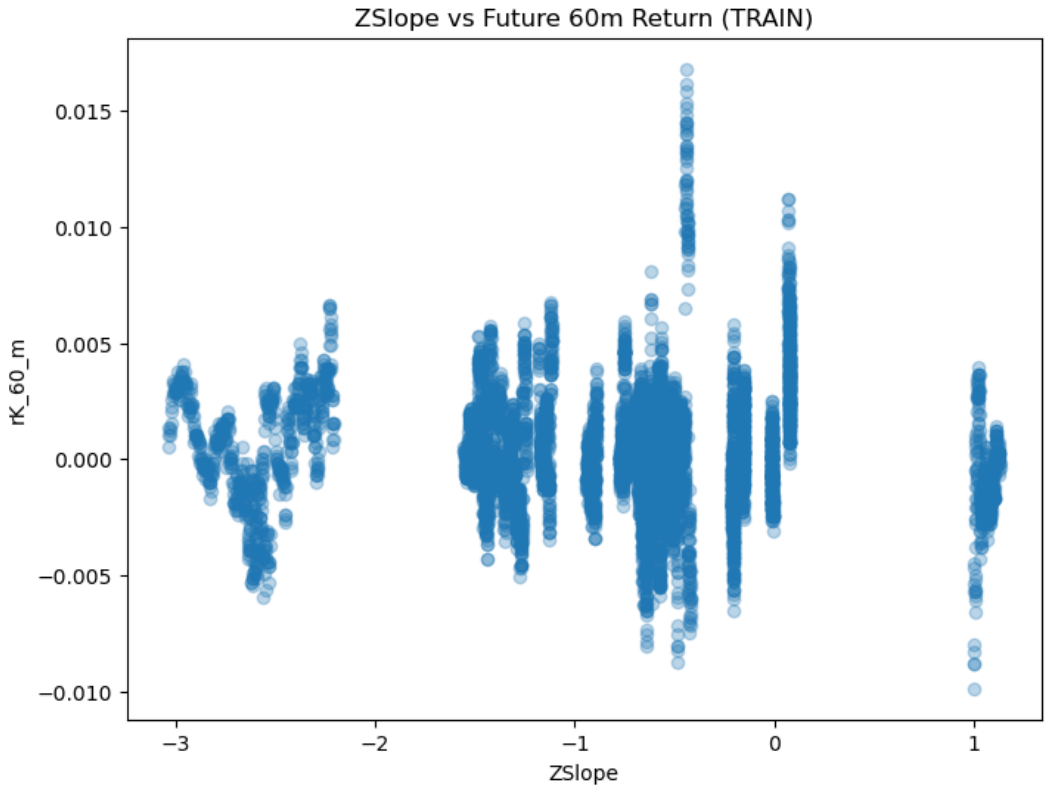

Figure 4: ZSlope vs future 60-minute return (TRAIN).

Points form several clusters corresponding to intra-regime behavior within the short 2007 window, but no clear linear or nonlinear structure is visible.

4.3 Quantile Buckets

ZSlope was split into deciles, and the mean future return was computed for each.

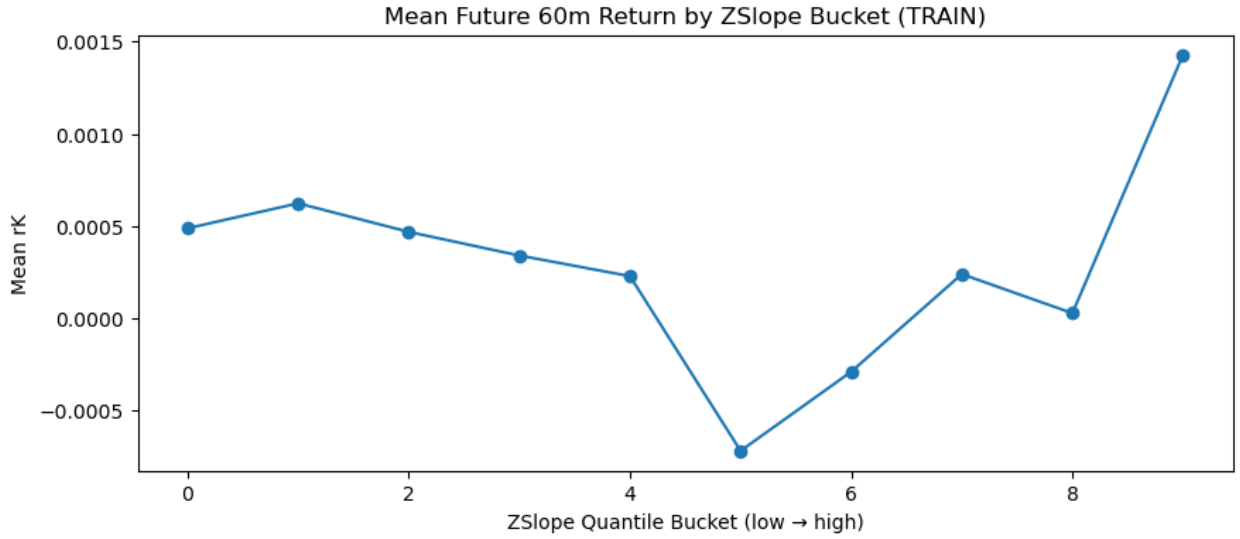

Figure 5: Mean future return by ZSlope bucket.

The lowest decile shows the highest positive average future returns, and the highest decile also shows modestly positive returns. The pattern is weak and non-monotonic.

4.4 Intraday Robustness

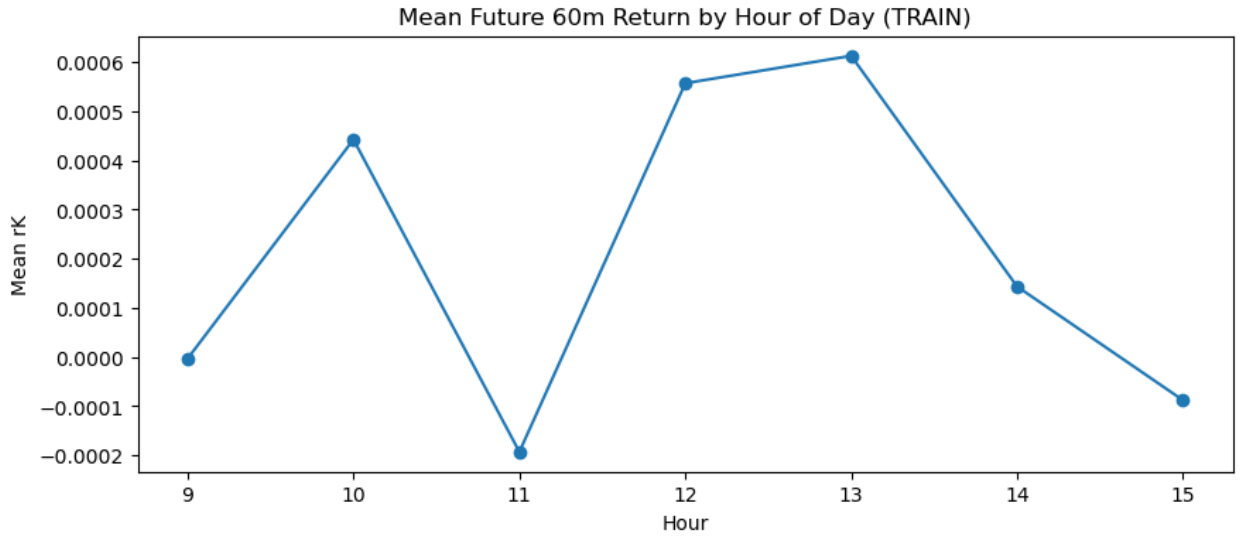

Figure 6: Mean future return by hour of day.

Hourly grouping reveals no consistent intraday structure.

5. Rule Construction

Rules were derived programmatically based on deciles with the largest spread between mean returns. Due to the noisy and unstable bucket structure from a limited training window, resulting thresholds favored *higher ZSlope for long entries* and *lower ZSlope for shorts*.

However, the bucket means show the opposite tendency (low ZSlope had higher returns). The training window was too small to make the directional mapping reliable.

6. Backtest Results

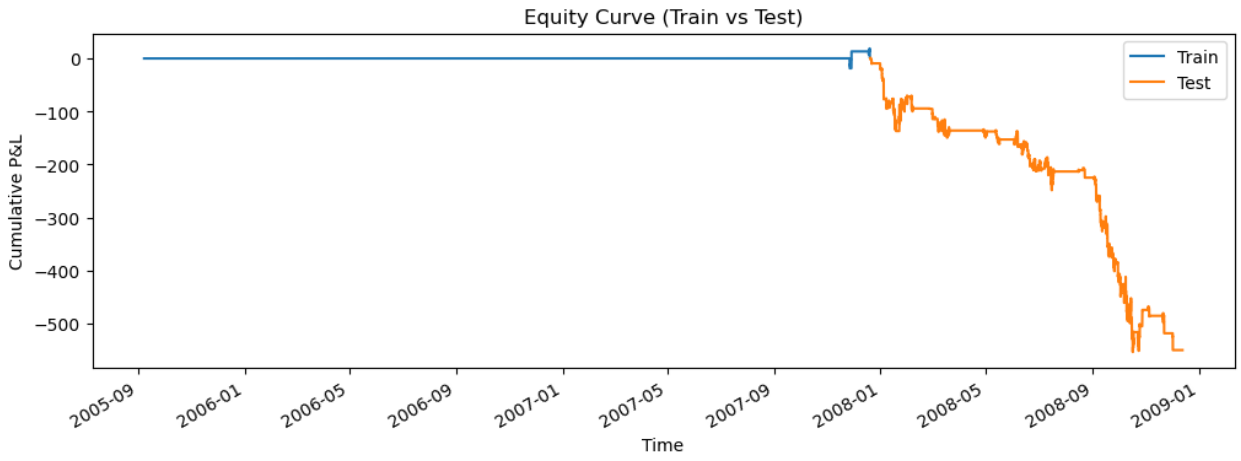

Figure 7: Equity curve for train vs test period.

Train Performance (Nov–Dec 2007)

- Trades: 26
- Win rate: 50%
- Total PnL: +2.95 ES points

Small sample size prevents meaningful interpretation.

Test Performance (2008)

- Trades: 843
- Win rate: 46%
- Total PnL: –550 ES points
- Profit factor: 0.73
- Max drawdown: –555 ES points
- Long-biased (778 longs, 65 shorts)

The strategy collapses out-of-sample, largely due to:

- misaligned directional logic from unstable bucket statistics;
- enormous regime shift into the 2008 crisis;
- insufficient volume of training data (2 months).

7. Conclusions

Key Findings

- The ES+VIX+VIX3M alignment pipeline functions correctly.
- The VIX term structure slope exhibits **no stable predictive power** in the available window.
- The usable dataset is severely restricted because VIX3M history begins in late 2007.
- Bucket-based rule discovery is unstable with limited data.

Overall Assessment

The VIX–VIX3M slope does not represent a reliable intraday signal for ES during 2007–2008. The weak correlation, inconsistent bucket behavior, and poor out-of-sample performance all support the conclusion that the feature lacks predictive power in this context.

8. Improvements

- Restrict all analysis to the VIX3M era and acquire more post-2007 data;
- Flip directional rules to match bucket sign structure;
- Use smoother volatility term structure features (e.g., VIX/VIX3M ratios);
- Reduce label overlap and sampling frequency;
- Add regime filters based on realized volatility or VIX level.