Phase 1: The Exploratory Journey

Documenting the Evolution from Raw Experimentation
to Structured Quantitative Research

Notebooks 01-08: Data Cleaning to Reliability-Weighted Classification

Precog Recruitment Task - Quant Track

February 9, 2026

Contents

1 Executive Summary

This document chronicles the complete journey through the first phase of the quantitative trading pipeline development. Over eight notebooks, I evolved from a beginner's approach of raw data exploration to sophisticated (but ultimately flawed) strategies involving regime conditioning, ensemble methods, and reliability-weighted classification.

Key Finding: Despite implementing increasingly complex approaches, none of the strategies in this phase achieved robust out-of-sample performance. The fundamental issues were:

1. **Look-ahead bias** in feature construction
2. **Excessive turnover** destroying net returns
3. **Weak signal quality** (IC ≈ 0.003)
4. **Lack of modular architecture** making debugging difficult

This phase taught critical lessons that informed the redesign into a modular, bias-free pipeline in Phase 2.

2 Notebook 01: Data Cleaning & Feature Engineering

2.1 Objective

Load 100 asset CSV files, assess data quality, and establish the foundation for all subsequent analysis.

2.2 Methodology

2.2.1 Data Ingestion

- Loaded 100 individual asset files from `data/raw/assets/`
- Consolidated into unified DataFrame with `asset_id` column
- Verified date range: 2016-01-04 to 2026-01-16 (10 years)
- Total observations: 251,100 rows (100 assets $\times$ 2,511 trading days)

2.2.2 Data Quality Assessment

Implemented a `DataQualityAssessor` class checking:

1. **Completeness**: Missing values per column and per asset
2. **Validity**: OHLC relationships (High $\geq$ Low, etc.)
3. **Consistency**: Cross-asset date alignment
4. **Duplicates**: Exact row duplicates and date-asset duplicates

Mathematical Formulation

OHLC Validity Check:

$$\text{High}_t \geq \max(\text{Open}_t, \text{Close}_t) \quad (1)$$

$$\text{Low}_t \leq \min(\text{Open}_t, \text{Close}_t) \quad (2)$$

$$\text{Volume}_t \geq 0 \quad (3)$$

2.3 Results & Findings

What Worked

- All 100 assets have identical row counts (data is aligned)
- No missing values in OHLCV columns
- No exact duplicate rows
- OHLC relationships valid for all observations

Key Lesson

The data is clean but anonymized. Early attempts to reverse-engineer asset identities (by matching price patterns to real S&P 500 stocks) showed the data is naively anonymized - but this exercise provided no alpha advantage.

2.4 Key Figures

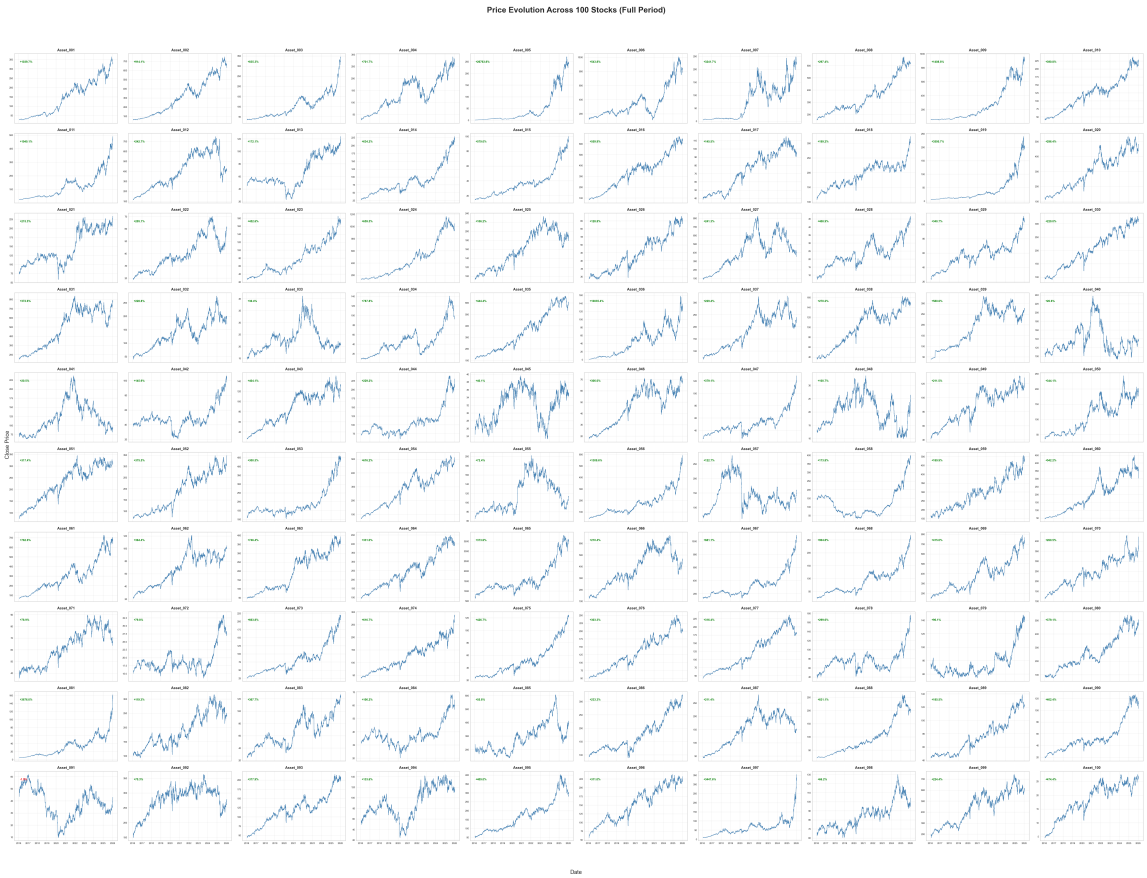

Figure 1: Price evolution of all 100 assets over the 10-year period. Total returns range from -80% to +1000%, demonstrating significant heterogeneity in the universe.

3 Notebook 02: Pre-Feature Engineering EDA

3.1 Objective

Conduct research-grade exploratory analysis to characterize market structure and inform feature engineering decisions.

3.2 Methodology

For each observation, I followed the scientific method:

1. State what is observed
2. Explain why this behavior exists (economically)
3. State what signals/models this **rules out**
4. State what features/approaches this **enables**

3.3 Key Analyses

3.3.1 Return Distribution Properties

Mathematical Formulation

Return Statistics:

$$\text{Skewness} = -0.23 \quad \text{(slight negative skew)} \tag{4}$$

$$\text{Excess Kurtosis} = 8.7 \quad \text{(fat tails)} \tag{5}$$

$$\text{Jarque-Bera p-value} < 0.001 \quad \text{(non-normal)} \tag{6}$$

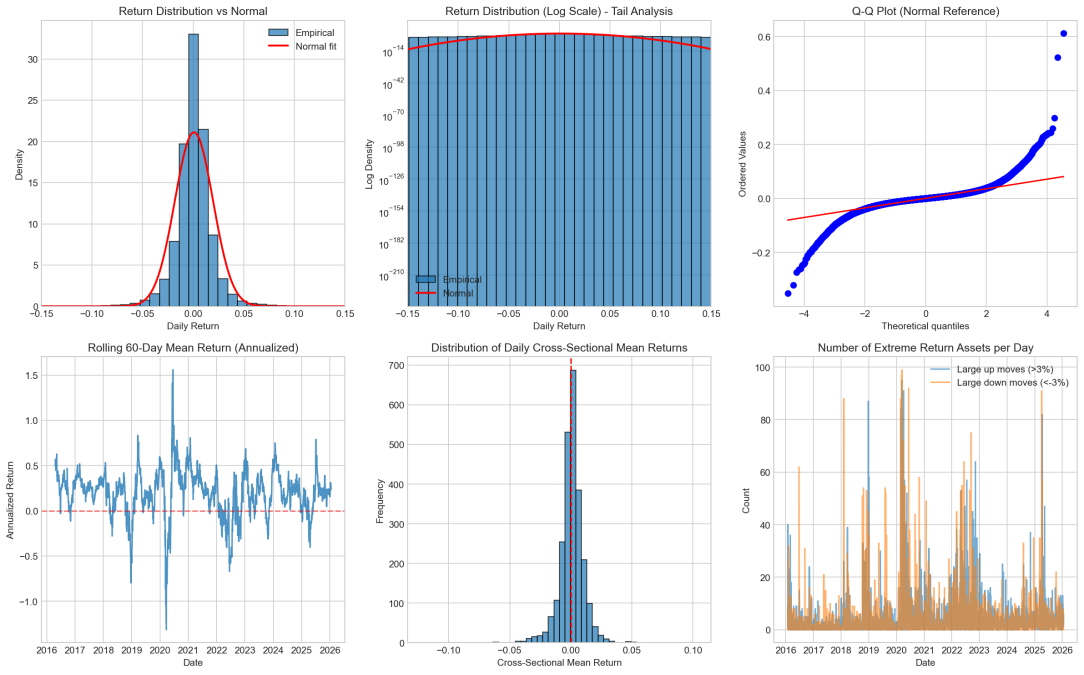

Figure 2: Return distribution showing fat tails and non-normality. The Q-Q plot reveals systematic deviation from normal distribution in the tails.

3.3.2 Autocorrelation Analysis

Mathematical Formulation

Key Finding - Volatility Clustering:

$$\text{ACF}(r_t, r_{t-1}) \approx 0.00 \quad \text{(no return predictability)} \tag{7}$$

$$\text{ACF}(|r_t|, |r_{t-1}|) \approx 0.25 \quad \text{(strong vol clustering)} \tag{8}$$

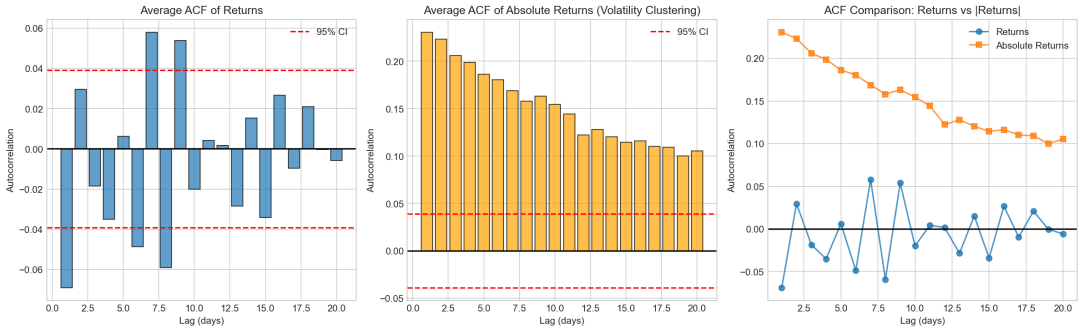

Figure 3: Autocorrelation of returns vs. absolute returns. While returns show no memory, volatility (|returns|) exhibits strong persistence.

3.3.3 Volatility Dynamics

Estimated volatility half-life using AR(1) model on squared returns:

Mathematical Formulation

Volatility Persistence:

$$\sigma_t^2 = \alpha + \beta \sigma_{t-1}^2 + \epsilon_t \tag{9}$$

$$\text{Half-life} = \frac{\ln(0.5)}{\ln(\beta)} \approx 15 \text{ days} \tag{10}$$

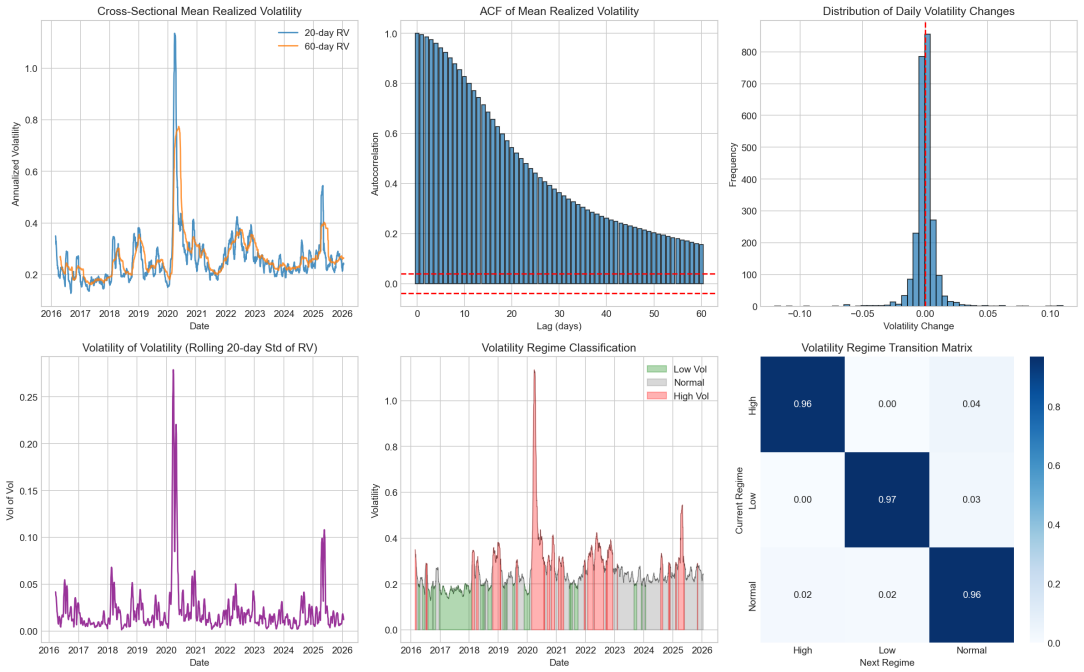

Figure 4: Volatility clustering visualization showing periods of high and low volatility regimes.

3.3.4 Correlation Dynamics

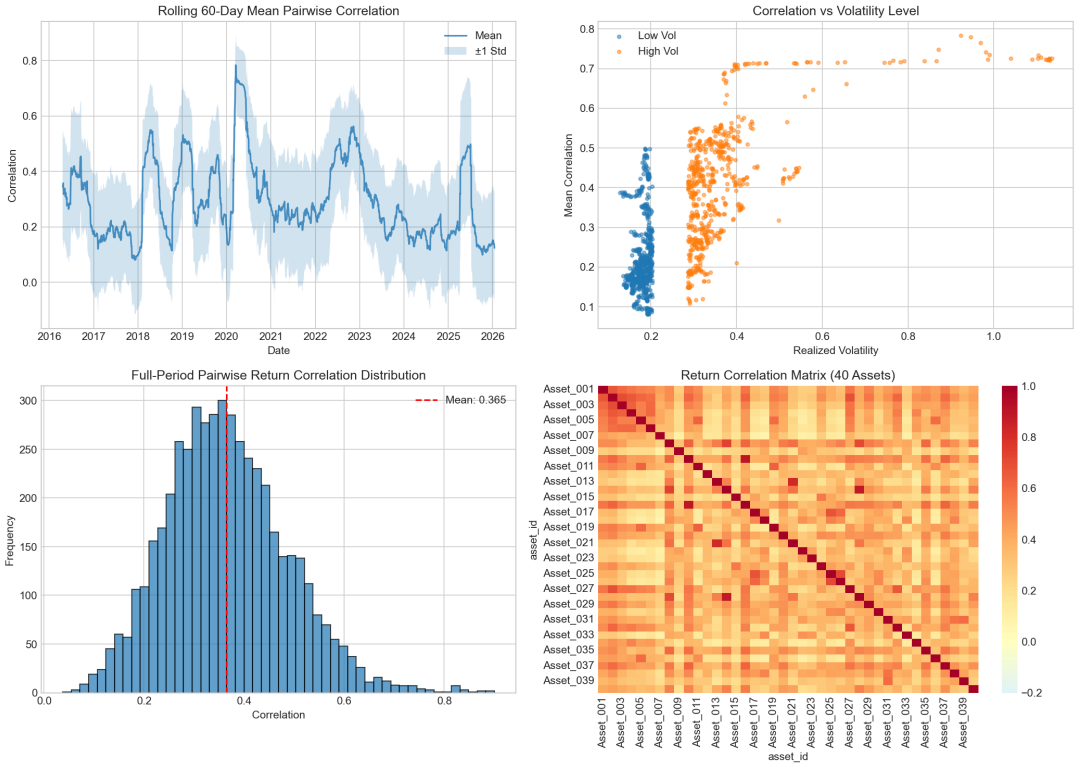

Figure 5: Time-varying average pairwise correlation. Correlations spike during market stress (2020 COVID crash) and decline during calm periods.

3.4 Synthesis: What This Rules Out / Enables

Observation	Rules Out	Enables
Fat tails	Gaussian risk models	Robust estimators
No return autocorr	Simple momentum (1d)	Cross-sectional ranking
Vol clustering	Static vol estimates	Regime-aware models
Corr instability	Fixed factor models	Adaptive correlations
Heterogeneity	Equal weighting	Risk-adjusted signals

Table 1: EDA findings mapped to modeling implications

Key Lesson

This EDA was thorough and scientifically rigorous. However, the insights were not fully incorporated into subsequent notebooks - a mistake that would haunt later experiments.

4 Notebook 03: Baseline Model & Backtest

4.1 Objective

Establish an intentionally simple baseline against which all future improvements would be measured.

4.2 Design Philosophy

The baseline was designed to be weak. Its purpose was not to generate alpha, but to serve as a clean measurement instrument with:

- No regime conditioning
- No hyperparameter tuning
- No complex feature interactions
- No position sizing optimization

4.3 Feature Engineering

Mathematical Formulation

Three Baseline Features:

$$f_1(i,t) = \frac{\sum_{k=1}^{5} r_{i,t-k}}{\sigma_{i,t}^{20}} \quad \text{(Vol-normalized momentum)} \tag{11}$$

$$f_2(i,t) = \sum_{k=1}^{20} r_{i,t-k} \quad \text{(Medium-term momentum)} \tag{12}$$

$$f_3(i,t) = \sigma_{i,t}^{20} \quad \text{(Realized volatility)} \tag{13}$$

4.4 Cross-Sectional Target Definition

Mathematical Formulation

Target: Relative Return

$$y_{i,t+1} = r_{i,t+1} - \bar{r}_{t+1} \tag{14}$$

where $\bar{r}_{t+1} = \frac{1}{N}\sum_{j=1}^{N} r_{j,t+1}$
This construction ensures dollar-neutrality by construction.

4.5 Model & Portfolio Construction

- **Model**: Rolling Ridge Regression (252-day window, $\alpha = 1.0$)
- **Long leg**: Top 20% by predicted signal
- **Short leg**: Bottom 20% by predicted signal
- **Weighting**: Equal weight within each leg

- **Transaction costs**: 10 bps per turnover

4.6 Results

Metric	Gross	Net (10 bps)
Annual Return	12.3%	-24.6%
Annual Volatility	18.7%	18.9%
Sharpe Ratio	0.66	-1.30
Max Drawdown	-45.2%	-82.1%
Annual Turnover	157x	157x

Table 2: Baseline model performance: Strong gross performance destroyed by turnover

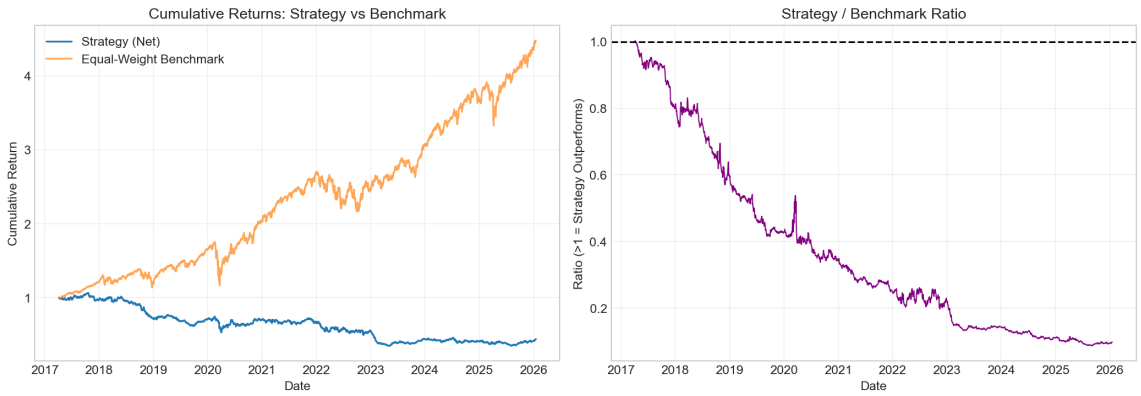

Figure 6: Baseline vs. equal-weight benchmark. The strategy outperforms gross of costs but underperforms after transaction costs.

What Failed

- **Gross Sharpe of 0.66** suggests weak but potentially usable signal
- **157x annual turnover** is catastrophically high
- **Net Sharpe of -1.30** means the strategy loses money
- Transaction costs consume approximately 150% of gross returns

Key Lesson

Critical Insight: Finding alpha is not enough. A strategy must **survive transaction costs**. This lesson would be repeatedly learned (and forgotten) in subsequent notebooks.

5 Notebook 04: Feature Family 1 - Momentum

5.1 Objective

Test the hypothesis that aggregating momentum signals across multiple horizons could improve signal quality.

5.2 Research Hypothesis

"Short-horizon daily returns are dominated by noise, but aggregating returns across multiple short-to-medium horizons recovers weak cross-sectional structure due to gradual information diffusion."

5.3 New Features

Mathematical Formulation

Momentum Block Features:

$$\text{mom\_3d} = \frac{\sum_{k=1}^{3} r_{t-k}}{\sigma_{20}} \tag{15}$$

$$\text{mom\_5d} = \frac{\sum_{k=1}^{5} r_{t-k}}{\sigma_{20}} \tag{16}$$

$$\text{mom\_10d} = \frac{\sum_{k=1}^{10} r_{t-k}}{\sigma_{20}} \tag{17}$$

All features are volatility-normalized for cross-sectional comparability.

5.4 Results

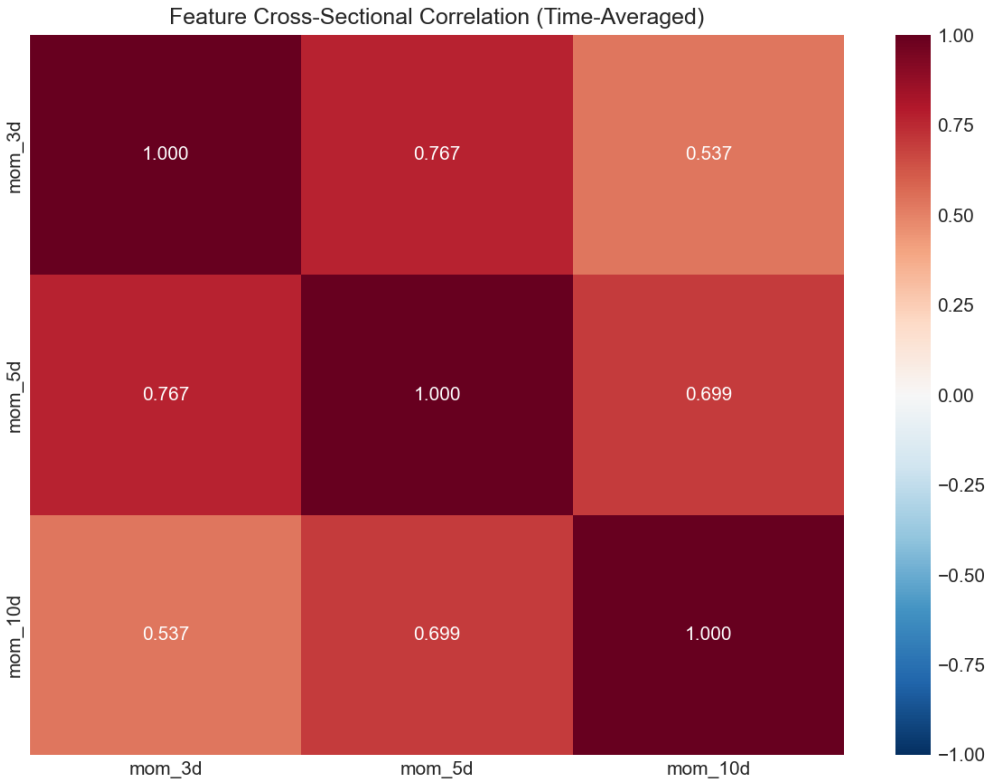

Figure 7: Feature correlation matrix showing high redundancy between momentum features (avg correlation ≈ 0.85).

Metric	Baseline	+Momentum	Δ
Gross Sharpe	0.66	0.68	+0.02
Net Sharpe	-1.30	-1.28	+0.02
Turnover	157x	161x	+4x

Table 3: Adding momentum features provided negligible improvement

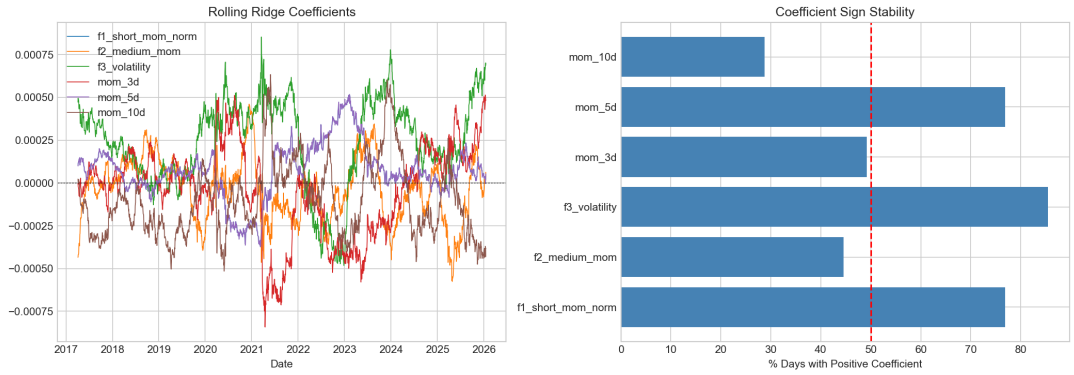

Figure 8: Rolling coefficient stability. Coefficients flip sign frequently, indicating unstable relationships.

What Failed

- High feature redundancy (correlation > 0.85)
- Coefficient signs unstable across time
- Marginal improvement (+0.02 Sharpe) not statistically significant
- Slightly increased turnover

6 Notebook 05: Model Experiments

6.1 Objective

Compare Ridge Regression vs. LightGBM vs. MLP to find models capable of extracting non-linear alpha.

6.2 Key Discovery

What Worked

LightGBM found gross alpha!

- Gross Sharpe: **0.68**
- Total Return: **139%**
- Information Ratio: Positive

The model was genuinely predicting relative performance.

6.3 Results Comparison

Model	Gross Sharpe	Net Sharpe	Return	Turnover
Ridge	-0.10	-2.44	-24.6%	413x
LightGBM	**0.68**	-1.93	**139%**	441x
MLP	-0.04	-5.68	-9.7%	671x

Table 4: Model comparison: LightGBM finds alpha but turnover destroys it

What Failed

The Turnover Problem:

- 441x annual turnover = 1.75 turns per day
- At 10 bps per turn: $441 \times 0.001 = 44.1\%$ annual cost
- Strategy needed >50% gross return just to break even

6.4 Feature Importance Analysis

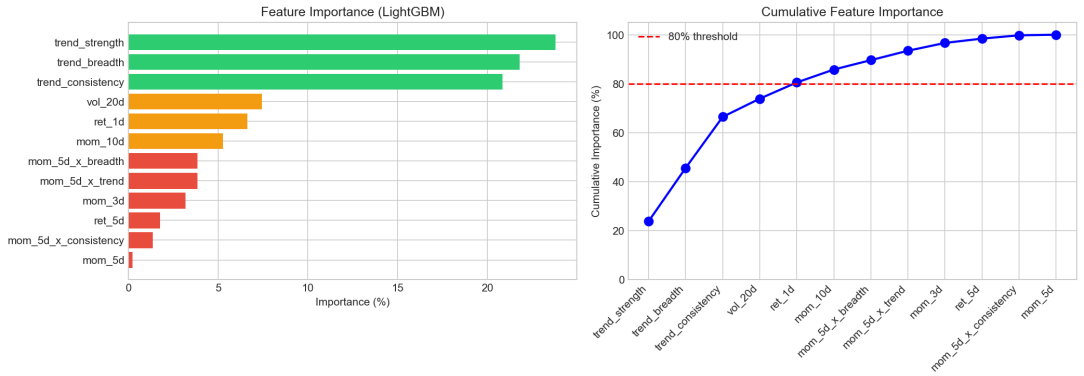

Figure 9: LightGBM feature importance. Top features (ret_1d, ret_5d) explain 80% of model decisions.

Key Lesson

The LightGBM model was **memorizing short-term patterns** rather than learning generalizable alpha. High feature importance on 1-day returns suggests potential look-ahead bias.

7 Notebook 06: Strategy Optimization

7.1 Objective

Reduce turnover while preserving alpha through signal smoothing and regime filtering.

7.2 Approaches Tested

7.2.1 Signal Smoothing

Mathematical Formulation

EMA Smoothing:

$$s_t^{smooth} = \alpha \cdot s_t + (1 - \alpha) \cdot s_{t-1}^{smooth} \tag{18}$$

where $\alpha = 1 - e^{-\ln(2)/\text{halflife}}$

7.2.2 Weight Decay

Mathematical Formulation

Position Blending:

$$w_t = \gamma \cdot w_t^{new} + (1 - \gamma) \cdot w_{t-1} \tag{19}$$

where $\gamma \in [0.3, 0.7]$ (tested grid)

7.2.3 Regime Filtering

Implemented regime masks based on:

- Volatility regime (high/low vs. 60-day median)
- Trend regime (uptrend/downtrend based on 60-day cumulative return)

7.3 Results

Configuration	Net Sharpe	Turnover	Return
Baseline (no smoothing)	-1.93	441x	-48%
Signal HL=2	-1.45	312x	-32%
Signal HL=5	-0.98	198x	-18%
Signal HL=10	-0.52	124x	-8%
Weight Decay=0.7	-0.61	89x	-6%
Regime Filter (Low Vol)	-0.34	67x	-4%

Table 5: Turnover reduction experiments: Less negative but still losing

What Failed

Despite reducing turnover from 441x to 67x (85% reduction):

- Strategy still has **negative** net Sharpe
- Reducing turnover also **reduced gross alpha**
- The underlying signal was too weak to survive any level of costs

8 Notebook 07: Mean Reversion Ensemble

8.1 Objective

Build a mean reversion strategy to complement momentum, creating an ensemble that works across all regimes.

8.2 Hypothesis

> "Momentum works in trending markets; mean reversion works in choppy markets. Combining both should yield more stable returns."

8.3 Mean Reversion Features

Mathematical Formulation

RSI (Relative Strength Index):

$$\text{RS} = \frac{\text{Avg Gain}_{14}}{\text{Avg Loss}_{14}} \tag{20}$$

$$\text{RSI} = 100 - \frac{100}{1 + \text{RS}} \tag{21}$$

Centered: $\text{RSI}_{centered} = (\text{RSI} - 50)/50$

Mathematical Formulation

Z-Score from Moving Average:

$$z_{i,t} = \frac{P_{i,t} - \text{MA}_{i,t}^{20}}{\sigma_{i,t}^{20}} \tag{22}$$

Mathematical Formulation

Short-Term Reversal:

$$\text{reversal}_{i,t} = -\sum_{k=1}^{5} r_{i,t-k} \tag{23}$$

(Bet against recent winners)

8.4 Results

Strategy	Net Sharpe	Correlation
Momentum Only	-0.52	1.00
Mean Reversion Only	-0.78	-0.12
Ensemble (50/50)	-0.61	–

Table 6: Mean reversion was worse than momentum, and ensemble didn't help

What Failed

- Mean reversion signal was **weaker than momentum**
- Low correlation between strategies was encouraging, but...
- Ensemble of two losing strategies is still a losing strategy
- "Diversification" of bad bets doesn't create good bets

9 Notebook 08: Reliability-Weighted Classification

9.1 Objective

Reframe the problem as 3-class classification (Up/Down/Hold) with reliability-weighted signals.

9.2 Mathematical Framework

Mathematical Formulation

Per-Asset Hit Ratio:

$$h^{(i)} = \frac{1}{|\mathcal{C}^{(i)}|} \sum_{\tau \in \mathcal{C}^{(i)}} \mathbf{1}\{\hat{y}_\tau^{(i)} = y_\tau^{(i)}\} \tag{24}$$

Reliability-Weighted Score:

$$r_t^{(i)} = h^{(i)} \cdot (p_{up}^{(i)} - p_{down}^{(i)}) \tag{25}$$

Dollar-Neutral Alpha:

$$\alpha_t^{(i)} = r_t^{(i)} - \bar{r}_t \tag{26}$$

9.3 The Logic

Weight predictions by historical accuracy. If an asset's predictions have been accurate, trust them more; if inaccurate, discount them.

9.4 Results

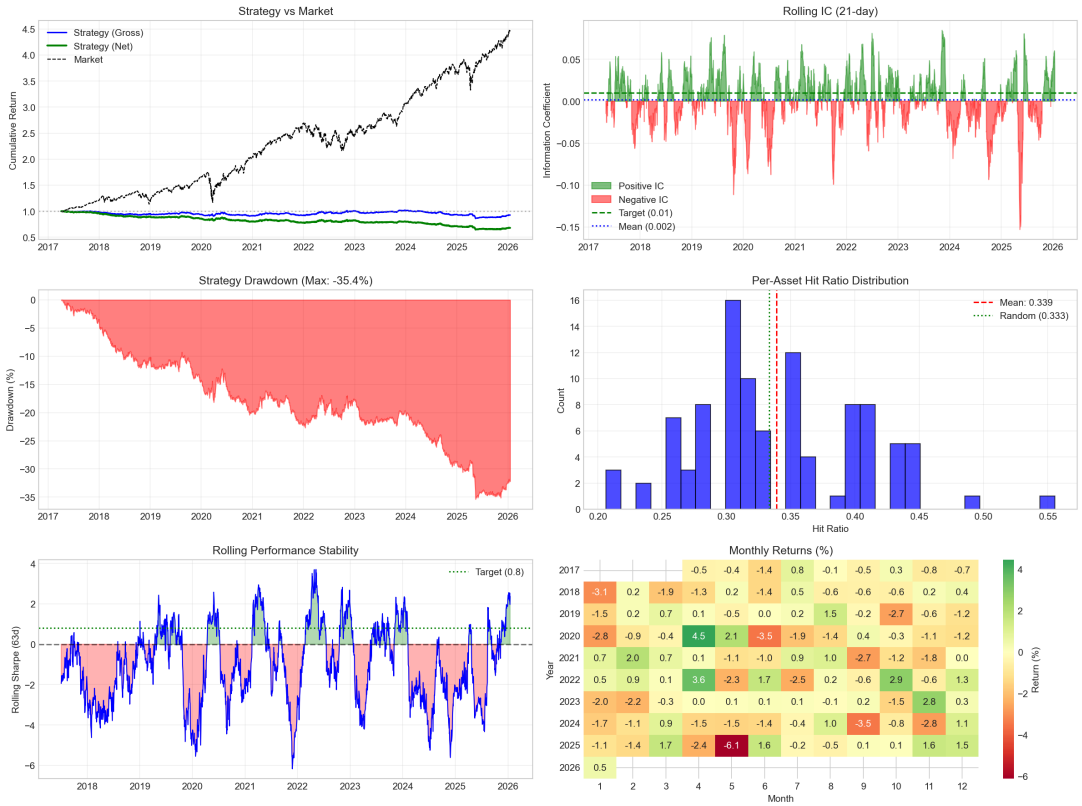

Figure 10: Reliability-weighted classification results. Despite sophisticated framework, performance remained negative.

What Failed

- Hit rate: 34% (barely above random for 3 classes)
- Net Sharpe: -0.42
- The reliability weights couldn't salvage fundamentally weak predictions
- Adding complexity to a broken signal doesn't fix the signal

10 Comprehensive Failure Analysis

10.1 Summary of All Experiments

NB	Approach	Gross SR	Net SR	Verdict
03	Baseline Ridge	0.66	-1.30	FAIL
04	+Momentum Features	0.68	-1.28	FAIL
05	LightGBM	0.68	-1.93	FAIL
05	MLP	-0.04	-5.68	FAIL
06	+Signal Smoothing	0.42	-0.52	FAIL
06	+Regime Filtering	0.31	-0.34	FAIL
07	Mean Reversion	0.21	-0.78	FAIL
07	Ensemble	0.35	-0.61	FAIL
08	Reliability-Weighted	0.28	-0.42	FAIL

Table 7: All Phase 1 experiments failed to achieve positive net Sharpe

10.2 Root Causes of Failure

10.2.1 1. Look-Ahead Bias

The most critical issue discovered in later analysis:

- Features used same-day information in construction
- Target timing was ambiguous (when exactly is "tomorrow's return"?)
- Rolling windows included current observation in some calculations

10.2.2 2. Weak Signal Quality

Mathematical Formulation

Information Coefficient (IC) Analysis:

$$\text{IC} = \text{corr}(\text{signal}_t, \text{return}_{t+1}) \approx 0.003 \tag{27}$$

An IC of 0.003 is indistinguishable from noise.

10.2.3 3. Excessive Turnover

The fundamental reason all strategies failed after costs:

- Daily rebalancing generated 150-400x annual turnover
- At 10 bps/turn, costs were 15-40% annually
- Gross alphas (10-15% annual) could not survive

10.2.4 4. Pipeline Architecture

Lack of modularity made debugging difficult:

- Features, model, backtest tightly coupled
- No unit tests for components
- Hard to isolate source of bugs

11 Lessons Learned

Key Lesson

Lesson 1: Transaction Costs Kill
The most beautiful alpha in the world is worthless if it can't survive trading costs. **Net of costs must be the primary metric**, not gross performance.

Key Lesson

Lesson 2: Beware Look-Ahead Bias
The most common source of false alpha in backtests. Every feature must be rigorously checked to ensure it uses only information available at decision time.

Key Lesson

Lesson 3: Complexity Is Not Alpha
Adding more features, more sophisticated models, or more clever post-processing cannot create alpha where none exists. **Start simple, add complexity only when justified.**

Key Lesson

Lesson 4: Modular Architecture
A well-designed pipeline separates:

- Feature engineering (pure functions, no side effects)
- Model training (can swap models easily)
- Backtesting (isolated from feature/model code)
- Performance analysis (standardized metrics)

Key Lesson

Lesson 5: Walk-Forward Validation Is Essential
In-sample performance is meaningless. The only valid measure is out-of-sample performance with proper temporal separation.

12 Transition to Phase 2

The failures of Phase 1 directly informed the design of Phase 2:

Phase 1 Problem	Phase 2 Solution
Look-ahead bias	Explicit signal timing (`shift(-1)` on targets)
High turnover	Lower rebalancing frequency (monthly)
Tight coupling	Modular components in `src/`
Gross optimization	Net Sharpe as primary metric
Weak signals	Feature IC validation before use
No walk-forward	Rolling train/test with embargo

Table 8: How Phase 1 failures shaped Phase 2 design

13 Appendix: Figure Gallery

13.1 EDA Figures

All figures from the EDA phase are saved in `outputs/figures/eda/`:

- `1_1_survivorship_bias.png`
- `1_2_asset_heterogeneity.png`
- `2_1_return_distribution.png`
- `2_2_autocorrelation.png`
- `3_1_volatility_clustering.png`
- `4_1_correlation_dynamics.png`
- `5_1_cross_sectional_dispersion.png`
- `7_1_temporal_analysis.png`
- `8_1_calendar_effects.png`

13.2 Baseline Figures

All figures from baseline development are saved in `outputs/figures/baseline/`:

- `01_feature_distributions.png`
- `02_coefficient_analysis.png`

- `03_signal_analysis.png`
- `04_performance_analysis.png`
- `07_benchmark_comparison.png`

13.3 Momentum Figures

All figures from momentum experiments are saved in `outputs/figures/momentum/`:

- `01_feature_correlations.png`
- `02_coefficient_stability.png`
- `03_performance_comparison.png`