

# Markowitz-ML with a Clustered Barra Risk Model

**Author:** Xiaoxuan Shang

**Summary.** A gradient-boosted tree model forecasts each stock’s next-month return relative to the cross-section from 816 inputs: the 402 CTF characteristics plus features generated by fixed rules. The forecasts feed a mean-variance portfolio that is neutral to beta and to industries, built from two complementary risk models: a Barra-style factor model with 40 clustered style factors, and a shrunk daily sample covariance. Over the test period the strategy earns an annualised Sharpe ratio of 4.39 gross of costs, with a maximum drawdown of  $-7.7\%$  at  $10\%$  volatility.

## 1. Design principles and validation protocol

**Data and timing.** The panel has 1.37 million stock-months (1952–2023, about 1,500–2,900 liquid U.S. stocks per month) with 402 characteristics and the next-month excess return `ret_exc_lead1m`, plus 103 million daily returns. A portfolio formed at month-end  $t$  uses characteristics dated  $t$ , daily returns up to  $t$ , and training labels from months before  $t$  only, which are realised by  $t$ . Every feature transformation is within-month, so no information crosses dates.

**Validation protocol.** Choosing among model variants by their 1990–2023 performance would itself be look-ahead bias: in a simulation of 20 skill-less strategies, the best one’s 34-year Sharpe ratio averages  $+0.32$  by luck alone. I therefore made every design decision on a *validation* window (Jan 1975 – Nov 1989, 179 months), with each candidate run walk-forward exactly as it would be live. The *test* period (Dec 1989 – Nov 2023, 408 months) was evaluated once per adopted building block, for reporting only. In total about 30 variants were compared on validation (Tables 1–2); none was revisited after its test value was seen. Validation Sharpe ratios exceed test values throughout because short-horizon anomalies were stronger before 1990; only rankings within the validation sample were used.

**Building blocks.** Following the course, the model was built in steps, each kept only if it improved validation performance: (i) Factor-ML (tree model, decile long-short), (ii) feature engineering, (iii) a multivariate Barra risk model, (iv) Markowitz-ML portfolio construction with neutrality constraints and a second risk model.

## 2. Return forecasts

**Characteristics.** Each characteristic is converted every month to its cross-sectional percentile rank on  $[-0.5, 0.5]$ , with missing values set to 0 (the median). Ranking removes 70 years of scale drift (the median market cap rose from \$161M

in 1960 to \$3.3B in 2023), caps outliers and puts all variables on one scale. On validation, ranks beat raw values with LightGBM’s native missing-value handling (IC 0.108 vs 0.105).

**Rule-based feature engineering.** To comply with Rule 2, engineered features are generated by fixed rules applied to all inputs rather than chosen as specific anomalies. 816 inputs are used in the LightGBM model:

- 402 provided characteristics
- 5 short-horizon returns: the stock’s excess return over the last  $h \in \{1, 2, 3, 5, 10\}$  trading days
- 1 data coverage indicator: the fraction of missing characteristics for the stock
- 408 industry-relative versions: the stock’s rank minus the average rank of all stocks in its industry that month of all inputs above

**Target and model.** The model predicts  $\arctan(z_{i,t+1})$ , where  $z$  is the within-month z-score of next-month excess return: demeaning removes the market (which the portfolio hedges) and the arctan bounds extreme returns (up to +1,600%). The learner is LightGBM with 1,000 trees, learning rate 0.01, 255 leaves and at least 2,000 observations per leaf, the main regulariser. Each tree sees 50% of rows and features. It is refit every 12 months on an expanding window of all prior stock-months; predictions are standardised within month.

**Table 1.** Average monthly Spearman correlation of forecasts with realised returns, validation 1975–1989 between models

| Input                   | Model                                                        | Rank IC         |
|-------------------------|--------------------------------------------------------------|-----------------|
| Raw 402 features        | LightGBM                                                     | 0.105           |
| Ranked 402 features     | Ridge/ LightGBM                                              | 0.095 / 0.108   |
| 816 rule-based features | Ridge with targets as                                        | 0.114 / 0.118 / |
|                         | raw / z-score / rank /<br>arctan z                           | 0.122 / 0.121   |
| 816 rule-based features | LightGBM                                                     | 0.134           |
| 816 rule-based features | LightGBM                                                     | 0.126 / 0.131 / |
|                         | recency-weighted<br>training, half-life 5 / 10<br>/ 20 years | 0.135           |

When selecting targets, Ridge is used rather than LightGBM because it is quicker to run. Recency weighting does not help in this window, so all training months are weighted equally.

### 3. Risk model

**Clustered Barra factor model.** Covariance is modelled as  $\Sigma_t = B_t F_t B_t^\top + D_t$ .

- Exposures  $B_t$ : 12 FF12 industry dummies and 40 style factors. Each style factor is a cluster of similar characteristics (hierarchical clustering, re-estimated every five years from the prior ten); exposure is the stock's standardised average rank in the cluster.
- Factor returns: daily cross-sectional regressions of returns on prior month-end exposures ( $\sqrt{\text{market cap weights}}$ ).
- $F_t, D_t$ : exponentially weighted factor covariance over 10 years (half-life 84 days for volatilities, 504 for correlations) and 84-day residual variance per stock.

On validation, Sharpe rises with the number of style factors up to 40, then flattens (6.8, 7.2, 7.5, 7.6 for 10–80); evenly sized Ward clusters did slightly worse.

**Second risk model.** The factor model only captures covariance that runs through characteristics. As a complement I use the 756-day daily sample covariance with Ledoit-Wolf shrinkage plus a ridge equal to its mean eigenvalue. The two books' validation returns are only 0.59 correlated.

### 4. Portfolio construction

**Book 1** solves  $\max_w w^\top \mu - \frac{\gamma}{2} w^\top \Sigma w$  s.t.  $A^\top w = 0$ , where  $A$  holds each stock's beta to the value-weighted market and the industry dummies;  $w \propto \Sigma^{-1}(\mu - A\lambda)$  is computed via the Woodbury identity and is also dollar-neutral. **Book 2** is tangency minus global-minimum-variance under the shrunk sample covariance (dollar-neutral). Each book is scaled to 10% ex-ante volatility, averaged 50/50, and rescaled to 10%.

**Table 2.** The same forecasts roughly double their Sharpe ratio through portfolio construction.

| Portfolio rule (same forecasts)                             | Validation SR               | Test SR |
|-------------------------------------------------------------|-----------------------------|---------|
| Decile long-short, equal-weighted                           | 3.74                        | 1.55    |
| Barra mean-variance: unconstrained / dollar- / beta-neutral | 7.39 / 7.42 / 7.43          | 3.59    |
| Weight on Ledoit-Wolf book: 0 / 25 / 50 / 75 / 100%         | 8.4 / 9.2 / 9.5 / 9.3 / 8.6 |         |

## 5. Guarding against overfitting and look-ahead bias

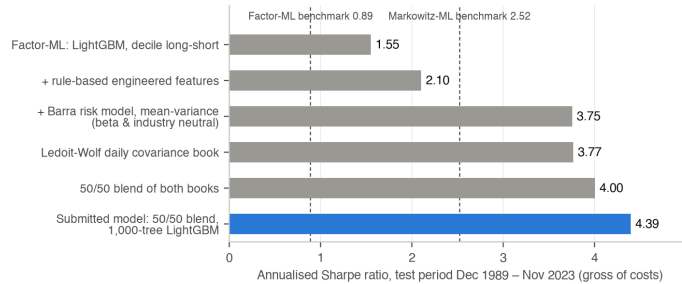
**Simulation study.** To check that the pipeline finds real signal and nothing else, I kept the real panel (stock-months, the 402 ranked characteristics, size groups) and replaced returns with synthetic ones of known predictability:  $r_{i,t+1} = s \cdot g(x_{it}) + \beta_i f_{t+1} + \varepsilon_{i,t+1}$ . The full pipeline was then run unchanged on two worlds:

- Null world ( $s = 0$ ), nothing to find. Every pipeline earns a Sharpe ratio within two standard errors of zero ( $-0.38$  to  $-0.22$ ), so the walk-forward machinery does not manufacture performance.
- Signal world  $g$  combines a linear, an interaction and a threshold term in four randomly chosen characteristics and explains 0.5% of the cross-section. LightGBM recovers 80% of the Sharpe ratio of an oracle that knows  $g$ , and beats ridge, as expected for a nonlinear signal.

Because the synthetic noise has no realistic covariance structure, this test isolates the forecasting step; it does not test the risk model.

**Look-ahead and reproducibility tests.** Mirroring the competition’s automated check, `main()` was run on a 123-month panel and again with the last two years removed. All 99,090 overlapping weights were bit-identical, so no weight depends on later data. A second run reproduced the first exactly, and all dependencies are pinned with Python 3.13 wheels verified. These tests rule out look-ahead in the code.

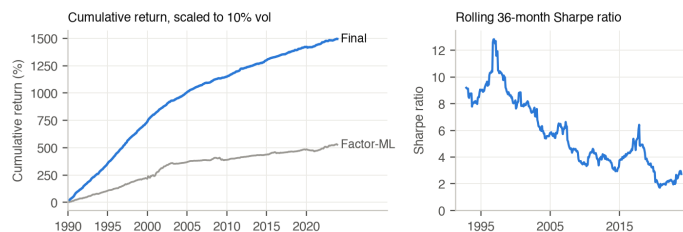
## 6. Results



**Figure 1.** Test-period Sharpe ratio of each building block

Feature engineering lifts the decile portfolio; replacing the decile sort with a risk-based Markowitz portfolio adds the most, because it avoids loading on volatile, unpriced factor exposures; blending two imperfectly correlated risk models adds a further 0.25. The submitted model’s test Sharpe ratio of **4.39** ( $t = 7.8$ ) comes with a maximum drawdown of **−7.7%** at 10% volatility (Figure 2).

**Figure 2.** Submitted model in the test period



### What an investor should expect.

1. **Decay.** Predictability fell sharply around 2005 and the Sharpe ratio falls from 9.2 in the 1990s to 4.7 in the 2000s and 3.2 since 2010, consistent with anomaly publication, decimalisation and the growth of quantitative trading. The model adapts only slowly: annual refits on an expanding window, where post-1990 data reach 64% of the sample by 2023. Recency weighting did not help on validation, but that window contains no comparable regime shift, so this remains a known risk.
2. **Costs and capacity.** Monthly turnover is 1.13 times gross exposure and 43% of gross exposure is in small stocks; under moderate costs the Sharpe ratio falls to 3.10, and capacity is limited.
3. **Risk forecasts.** Realised volatility was 2.1 times the ex-ante target, most likely because the optimiser concentrates in positions whose risk is underestimated. A Barra-style volatility regime adjustment, scaling positions by the trailing ratio of realised to predicted risk, corrects the level (it shrinks positions to about one third in the validation window) but does not improve the validation Sharpe ratio (9.23–9.31 vs 9.36), so it is not part of the model.

## 7. Disclosures

**Borrowed ideas.** The arctan-z target, the LightGBM hyper-parameters and the Ledoit-Wolf tangency-minus-GMV book follow the public submission “LightGBM with penalised VCV” (P. Geertsema); blending two risk books was inspired by A. Keshri’s submission. These settings were re-checked on my validation window, but their authors may have chosen them with knowledge of 1990–2023 performance, so they are the one place where test-period information could have entered indirectly. The Barra methodology follows Jensen, Kelly, Malamud and Pedersen (2024) and the course; factor clustering follows Jensen, Kelly and Pedersen (2023).

**AI assistance.** Code and drafting were developed with an AI assistant (Claude), as permitted by the course. The author remains responsible for everything submitted.