

A Stacked Return Forecast with a Forecast-Aligned Factor Risk Model

MGT 924 Statistical Foundations, Problem Set 3: submission to the JKP Common Task Framework
Thomas Robichaud (tmr62) and Abhay Kumar (ak3299), Yale School of Management, October 2026

1. Summary

We build a dollar-neutral US equity portfolio from the 402 characteristics in the competition dataset. Three learners forecast next-month returns, a stacking step weights them by their own past out-of-sample accuracy, and a 93-factor risk model turns the forecast into mean-variance weights. Over the competition test window (December 1989 to November 2023, 408 months) the portfolio earns an **annualized Sharpe ratio of 2.87** (standard error 0.20), computed locally with the competition's code structure.

- **Portfolio construction is worth as much as prediction.** The same forecasts earn a Sharpe ratio of 1.50 when rank-weighted and 2.87 through the risk model (shown in Figure 1).
- **The risk model is calibrated where it matters.** Adding the forecast itself as a risk factor cut the optimized portfolio's realized-to-forecast volatility from 2.1 to 1.0 in our design sample.
- **Performance decays.** The Sharpe ratio is 6.6 in the 1990s, 2.6 in the 2000s and 1.6 in 2010 to 2023. An investor should plan on the last number, not the full-sample one.

2. Data and protocol

The data are the three competition tables: 1,369,324 stock-months of characteristics for 12,325 US stocks (December 1951 to November 2023, about 2,170 stocks per month in the test window), their next-month excess returns, and 103 million daily excess returns. This is the only data in scope for this exercise. Industry groups are derived from the SIC code supplied in the characteristics table using fixed Fama-French-12-style ranges.

Every quantity behind the portfolio formed at month t uses data dated t or earlier. A month's return enters training only once it is realized; models are refit each December on an expanding window (the calendar, not the sample end, sets the schedule); and every tuning step scores candidates on months before t . We verified this by deleting all data after a cut-off date left every earlier weight unchanged to machine precision at three cut-off dates on synthetic data with the competition schema. The code is deterministic (fixed seeds, identical output across repeated runs and row orderings) and runs on a 123-month dataset.

The first portfolio is formed in December 1956, after 60 months of training data. That leaves **396 pre-test months (December 1956 to November 1989)** of genuine out-of-sample portfolios, which we used as the design sample for available choices.

3. Model

Features and target. Each characteristic is ranked within the month and mapped to $[-0.5, 0.5]$; missing values (about 8% of the entire dataset) are set to 0, the cross-sectional median. The target is the next-month excess return, standardized within the month and capped at three standard deviations to control the outliers.

Three learners. (i) Ridge regression on the 402 ranks. (ii) Ridge on 1,024 random Fourier features of the ranks (512 Gaussian frequencies at three bandwidths), a smooth non-linear model. (iii) LightGBM (400 trees, learning rate 0.03, 63 leaves, half the features and 70% of rows per tree, minimum leaf size 0.1% of the training rows), which captures interactions. Each ridge penalty is chosen from 13 values at every refit by squared error on the most recent 60 training months, then the model is refit on all of them. The LightGBM settings were fixed in advance and not tuned.

Stacking. Each month we store the three forecasts actually used. Every December, non-negative least squares on the trailing 120 months of those out-of-sample forecasts against realized returns gives the ensemble weights; before 36 months of history exist the weights are equal. The weight on LightGBM averages 0.71 in the test window and varies between 0.3 and 1.0 (Figure 3b).

Risk model. Following the Barra approach from class and Jensen, Kelly, Malamud and Pedersen (2024), the covariance matrix is $B\Omega B' + D$ with 93 factors: 12 industries, 80 style factors and the model's own forecast. The style factors come from average-linkage hierarchical clustering of the characteristics on one minus their rank correlation, re-estimated at each refit (Jensen, Kelly and Pedersen, 2023); a stock's exposure is its standardized mean rank within the cluster. Daily cross-sectional regressions of returns on the previous month-end exposures give factor returns and residuals. Ω is an

exponentially weighted covariance of daily factor returns and D holds exponentially weighted residual variances, clipped at the 10th and 99th percentiles.

Portfolio. Weights are proportional to $\Sigma^{-1}\mu$ with the minimum-variance portfolio subtracted so they sum to zero. A ridge term κ times the mean residual variance is added to D; κ is picked each December from five values by the realized Sharpe ratio of past out-of-sample portfolios. Weights are scaled to 10% forecast annual volatility, giving gross leverage of about 5 on average.

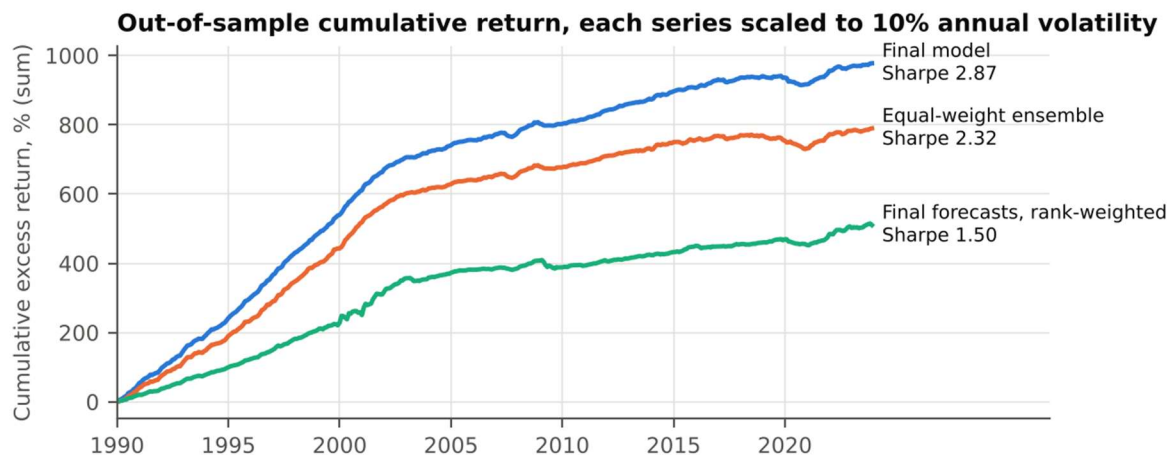


Figure 1. Test-window cumulative excess return (sum of monthly returns), each series rescaled to 10% realized annual volatility. The gap between the top and bottom lines is the risk model; the gap between the top two is stacking.

4. Validation of design choices

We compared seven designs on the full history before running LightGBM, using the ridge learners with equal weights, and chose on the pre-test Sharpe ratio (Table 1). There were three considerations that shaped the model:

- **Forecast as a risk factor.** With industry and style factors only, the risk model was accurate for random long-short portfolios (realized-to-forecast volatility 0.99) but the optimized portfolio realized 2.1 times its forecast. The optimizer was loading heavily on the forecast. Adding it as a factor brought the ratio to 1.01 and raised the Sharpe ratio.
- **Granularity.** Moving from 20 to 80 style clusters raised the pre-test Sharpe ratio from 4.09 to 4.49. Replacing the 12 industries with one market factor lowered it to 3.37.
- **What we dropped.** A step that rescaled the portfolio by its trailing realized-to-forecast volatility became redundant once the forecast was a factor (4.09 with it, 4.26 without), so we removed it.

Design (ridge learners, equal weights)	Pre-test Sharpe	Test Sharpe	Realized / forecast vol., pre-test
A. Industries + 20 style clusters	3.95	2.25	2.14
B. A + forecast as a risk factor	4.09	2.35	1.01
C. B with 40 clusters	4.31	2.35	0.99
D. B with 80 clusters	4.49	2.38	0.96
E. B without volatility rescaling	4.26	2.33	1.01
F. B with a lower κ grid	4.14	2.36	1.00
G. B with a market factor, no industries	3.37	2.23	1.09

Table 1. Design comparison. Volatility ratios are for the optimized portfolio before any rescaling. The final model combines B, D, E and F; that combination was not tested separately before the full run.

Our first full run averaged the three learners equally and earned a test Sharpe ratio of 2.51, against 3.25 for LightGBM alone in that run. Rather than select LightGBM on test results, we added the stacking step, which uses only past data. It also improves the pre-test Sharpe ratio (4.04 with equal weights, 4.41 stacked), but we decided to add it after seeing

test-period numbers, so the 2.87 is mildly optimistic. The pre-test sample was also consulted repeatedly, so its 4.41 is an upper bound.

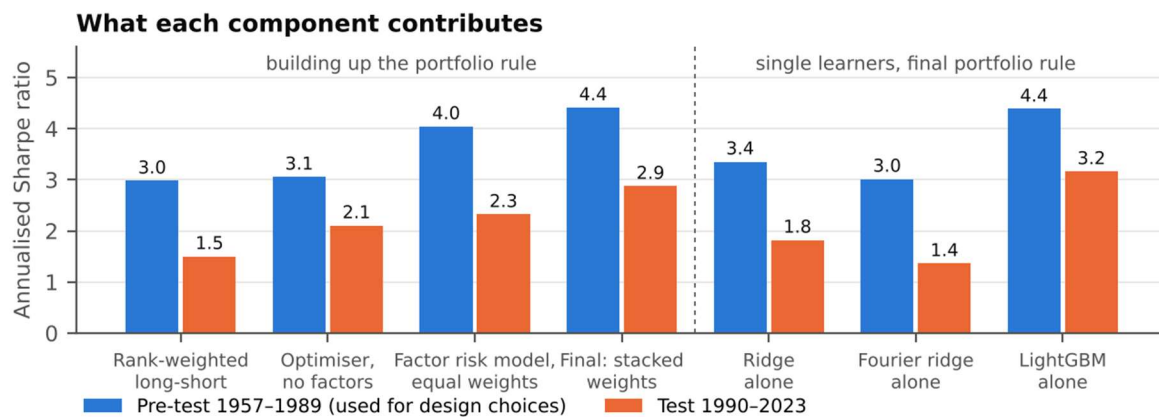


Figure 2. Annualized Sharpe ratio by component, in the final run. Left group: the same stacked forecast under progressively richer portfolio rules. Right group: each learner's forecast alone under the final rule. LightGBM alone (3.16 in the test window) still beats the stacked model (2.87).

5. Results

Period	Months	Sharpe (s.e.)	Mean IC	Positive months	Realised / forecast vol.	LightGBM weight
Pre-test, 1957–1989	396	4.41 (0.23)	0.101	91%	0.92	0.55
Test, 1990–2023	408	2.87 (0.20)	0.067	81%	1.07	0.71
1990–1999	120	6.57 (0.53)	0.096	97%	0.88	0.80
2000–2009	120	2.62 (0.36)	0.056	80%	1.07	0.51
2010–2023	167	1.56 (0.28)	0.054	71%	0.86	0.79

Table 2. Out-of-sample performance of the final model. IC is the monthly cross-sectional correlation between the stacked forecast and the standardized return. Sharpe standard errors use the Lo (2002) approximation for independent returns. The test row covers December 1989 to November 2023.

Forecasts. The stacked forecast has a mean monthly information coefficient of 0.067 in the test window (t-statistic 12.2). LightGBM is the most accurate learner (0.067, against 0.051 for each ridge model) and the only one whose accuracy holds after 2010 (0.057 against 0.035). The ridge portfolios are highly correlated with each other (0.84) and less so with LightGBM (0.45 to 0.57), which is why the stacking weights move toward LightGBM over time.

Risk. Realized volatility is within 14% of forecast in every sub-period. Figure 3a shows the 36-month ratio staying between 0.5 and 1.3; the model tends to overstate the optimized portfolio's risk in calm periods, which costs some leverage but not accuracy. The κ choice matters little: the five candidates earn test Sharpe ratios between 2.57 and 2.92, and the rule picked 1.0 in 348 months and 4.0 in 60.

Drawdown. At 10% volatility the worst drawdown is 26.6%, from October 2019 to September 2020, and the worst month is February 2020 at -10.3%, more than three forecast standard deviations. Calendar-year returns are negative in 2009 (-0.1%), 2019 (-4.3%) and 2020 (-19.4%), followed by +39.3% in 2021.

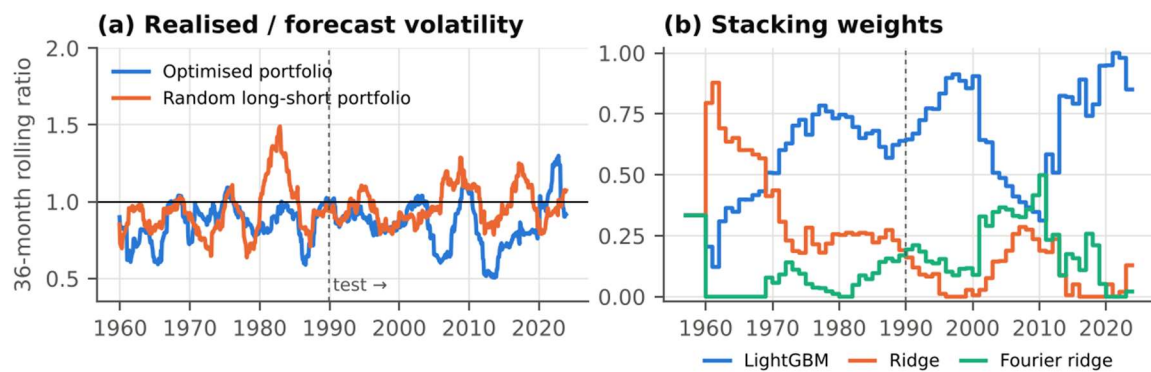


Figure 3. (a) Rolling 36-month realized volatility divided by the 10% forecast, for the final portfolio and for a random dollar-neutral portfolio scaled by the same risk model. (b) Stacking weights, updated each December.

6. What an investor should expect in practice

- **A Sharpe ratio nearer 1.5 than 2.9.** The most recent 14 years are the relevant guide for a realistic Sharpe expectation, and the decline is consistent with the published evidence that characteristic-based predictability weakens after discovery.
- **No trading costs are included.** The competition scores gross returns. The portfolio runs at about five times gross leverage across roughly 2,100 names and is rebuilt monthly; we did not measure turnover, and costs would reduce the figures above.
- **Crowding risk.** The 2019 to 2020 drawdown shows the portfolio can lose a quarter of its value at 10% volatility when many characteristic strategies fail together.
- **Reproducibility.** Results were produced locally on Python 3.14; the competition runs Python 3.13, so its score may differ slightly from ours.

7. Sources and disclosure

Using LightGBM on this dataset and imposing dollar neutrality as the tangency portfolio minus the minimum-variance portfolio follow Paul Geertsema's public leaderboard submissions. The factor risk model follows Jensen, Kelly, Malamud and Pedersen, *Machine Learning and the Implementable Efficient Frontier*; the clustering follows Jensen, Kelly and Pedersen, *Is There a Replication Crisis in Finance?*; random Fourier features follow Kelly, Malamud and Zhou, *The Virtue of Complexity in Return Prediction*. The stacking step, the forecast-as-factor adjustment and the pre-test design protocol were developed for this submission. AI assistance (Anthropic's Claude) was used to write the code and to generate the first draft and format of this document.