

Elastic Frontier

Return Forecasts in a Factor-Risk Mean–Variance Portfolio

A submission to the JKP Common Task Framework · MGT 924 Statistical Foundations, Problem Set 3

Dinica Hooda · Yale School of Management · October 2026 · Model code: *main.R* (R 4.4.2), dependencies pinned in *renv.lock*

The strategy holds a dollar-neutral long–short portfolio of an average of 2,171 liquid U.S. stocks, rebalanced monthly. Return forecasts come from an elastic net and gradient-boosted trees, stacked on out-of-fold predictions and trained once on 1952–1989 data. Each month a mean–variance optimizer turns the forecasts into weights, using a ten-factor statistical risk model estimated from the previous year of daily returns. Over the 408 out-of-sample months from January 1990 to December 2023, the portfolio earned a **Sharpe ratio of 1.96** (95% confidence interval 1.62–2.29), 19.6% a year at 10% volatility, and had almost no market exposure (correlation 0.09). The main lesson: construction mattered more than prediction. Holding the forecasts fixed, replacing rank-based rules with the optimizer raised the Sharpe ratio from 1.10 to 1.96, though the advantage is concentrated before 2010. The edge has also decayed: the Sharpe ratio fell from 4.02 in 1990–1999 to 0.78 in 2010–2023, and trading costs absorb most of what remains, so for the model as submitted we expect a forward Sharpe ratio of roughly 0.1–0.4 after costs (Section 6).

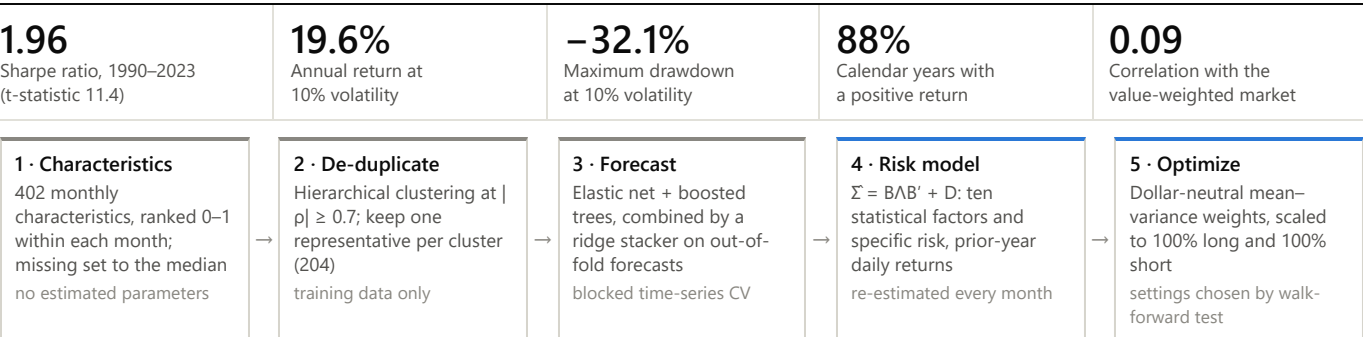


Figure 1. The pipeline. Gray steps have no parameters or are estimated once, on the training sample (returns January 1952–December 1989). Blue steps are re-estimated every test month from data available at that month-end.

1 Data and assumptions

The competition dataset covers 1.37 million U.S. stock-months with 402 characteristics (the JKP set of Jensen, Kelly and Pedersen, 2023), next-month excess returns, and 103 million daily excess returns. The competition’s own flag splits the sample. All estimation uses the 456 training months, whose returns run from January 1952 to December 1989 (483,626 stock-months, median 1,041 stocks a month). Weights are produced for the 408 test months, whose returns run from January 1990 to December 2023 (885,698 stock-months, 1,750–3,034 stocks a month). In the test period the universe has no micro- or nano-caps; every name is in the mega, large, or small size group. We state four assumptions:

- **Return definition.** The target and the realized return are ret_exc_lead1m , the excess return over the month after formation. We checked that it matches compounded daily returns to within 0.01 percentage points.
- **Scoring.** The monthly portfolio return is $\sum_i w_{i,t} r_{i,t+1}$, and the Sharpe ratio is its mean over its standard deviation times $\sqrt{12}$, as on the leaderboard. Results are gross of trading and financing costs; Section 6 adds costs.
- **Timing.** Characteristics dated at month-end t are known at t , the competition’s convention. The risk model uses daily returns only up to the day before month-end, a deliberate one-day buffer; the month-end return is also knowable at t , so including it would be legitimate too.
- **One-time estimation.** The forecasting model is fit once on the training sample and never re-estimated, so every 1990–2023 forecast comes from a model that has seen no data after 1989. This is conservative; a live implementation would refit periodically.

2 The model

2.1 Characteristics and feature reduction

Each month, every characteristic is replaced by its cross-sectional percentile rank, with ties averaged. Missing values are set to 0.5, the cross-sectional median. Exact zeros, which usually flag an absence such as no dividend, are pinned at 0. Ranks make characteristics comparable, neutralize outliers, and need no estimated scaling. Many of the 402 characteristics are

near-duplicates. On training data only, we cluster them hierarchically (average linkage, distance $1 - |\rho|$) and cut the tree at $|\rho| = 0.7$. From each of the resulting 204 clusters we keep the member most correlated with next-month returns. The code recomputes this selection from whatever data it receives, so no characteristic is chosen by hand (CTF Rule 2).

2.2 Return forecasts

Two learners use the same 204 inputs and target. All tuning uses **blocked** three-fold cross-validation: each fold is a contiguous block of months, so no fold is ever predicted from its own future. The **elastic net** (Zou and Hastie, 2005) searches the mixing weight $\alpha \in \{0.1, 0.3, 0.5, 0.7, 1\}$ and the full penalty path. It selects $\alpha = 0.1$ and $\lambda = 0.0026$, which leaves 30 of 204 coefficients nonzero. The **boosted trees** (XGBoost; Chen and Guestrin, 2016) search depth $\in \{2, 3, 4\}$ and learning rate $\in \{0.05, 0.10\}$, use 80% row and column subsampling, and set the number of trees by early stopping. They select depth 2 and seven trees, so the data support only a shallow nonlinear model. A **ridge stacker**, fit on out-of-fold forecasts, combines the two: $\hat{\mu} = 0.52 \cdot \text{EN} - 0.25 \cdot \text{XGB} + \text{const}$. The negative tree weight is a warning sign. Breiman (1996) constrains stacking weights to be non-negative to rule this out, and seven depth-2 trees are close to a step function of a few characteristics, so the weight most likely reflects collinearity with the elastic net rather than signal. Under Breiman's constraint the trees would get zero weight and the portfolio would be built on the elastic net alone. We have not dropped them because that change has not yet passed the pre-1990 walk-forward test that every other change had to pass; Section 7 lists it as the first test to run.

2.3 Risk model

For each test month we fit a statistical factor model (Connor and Korajczyk, 1986) to the previous 370 calendar days of daily excess returns. It covers stocks with at least 80% of trading days and uses $k = 10$ principal components: $\hat{\Sigma} = BAB' + D$, scaled from daily to monthly variance ($\times 21$). This is the market-model covariance from class with ten estimated factors in place of one. Stocks without enough history get zero factor loadings and the average specific variance. Specific variances are floored at 5% of the month's median. The floor matters because the smallest eigenvalue of $\hat{\Sigma}$ is bounded below by $\min D$. One stock whose daily returns happened to be almost fully explained by the factors made a portfolio direction look riskless, and an unconstrained optimizer bet heavily on it. We traced this in nine months spanning calm and crisis periods; each time it was a different, fully traded stock, not a data error.

2.4 Portfolio construction

Each month we solve, with the OSQP solver (Stellato et al., 2020),

$$\max_w w' \hat{\mu} - \gamma w' \hat{\Sigma} w - \kappa \|w - w_{t-1}\|_1 \quad \text{subject to} \quad 1' w = 0.$$

The walk-forward test in Section 3 selects $\kappa = 0$. The solution is then the classic mean–variance portfolio, the classic formula $\pi^* = \gamma^{-1} \Sigma^{-1} \mu$ with a zero-net-investment constraint: $w^* \propto \hat{\Sigma}^{-1}(\hat{\mu} - \eta \mathbf{1})$, where η makes the book dollar-neutral. Weights are rescaled to 100% long and 100% short, which makes γ irrelevant. There are no position limits; the largest single position in 34 years was 5.2% of capital. A simple fallback, with weights proportional to demeaned forecasts, covers months without enough daily history. It is needed for the competition's 123-month validation file and was never triggered in the test period, where the optimizer ran in 408 of 408 months.

Table 1. Every modeling choice and how it was made. “CV” is blocked cross-validation inside the training sample. Every parameter was set on pre-1990 data; the choice of model family was informed by test-period results (Section 4).

Component	Setting in the submission	How it was chosen
Characteristic transform	Monthly percentile rank; missing $\rightarrow 0.5$; exact zeros $\rightarrow 0$	Fixed ex ante; no parameters
Feature clustering	Average linkage, cut at $ \rho = 0.7$: 402 $\rightarrow$ 204 characteristics	Threshold fixed ex ante; representatives chosen on training data
Elastic net	$\alpha = 0.1$, $\lambda = 0.0026$; 30 nonzero coefficients	CV over α grid and λ path
Boosted trees	Depth 2, learning rate 0.10, 7 trees	CV over 6-point grid; early stopping
Stacking weights	+0.52 elastic net, −0.25 trees	Ridge on out-of-fold forecasts; penalty by CV
Risk model	10 PCA factors; 370-day daily window; $\geq 80\%$ coverage	Fixed ex ante (standard range)
Specific-variance floor	5% of the month's median specific variance	Structural, after diagnosing nine real months
Turnover penalty κ	0	Walk-forward test, 75 backtests before 1990 (Figure 2)
Risk aversion γ ; exposure	Irrelevant after scaling; 100% long, 100% short	Convention

3 Validation

Validation has two layers, both inside the training sample. Hyperparameters of the learners are set by blocked CV within each fit (Table 1). Construction choices are set by a **walk-forward test** that reruns the whole pipeline (clustering, both learners, the stacker, the risk model, and the optimizer) on expanding windows. Each of five folds fits on all months before a

24-month evaluation block (Figure 2a). We evaluated $\gamma \in \{0.5, 1, 2, 5, 10\} \times \kappa \in \{0, 0.001, 0.01\}$, 75 backtests in all, and chose by the **median** Sharpe ratio across folds, which no single unusual window can drive.

$\kappa = 0$ is the only setting that never fails (Figure 2b): the penalty helped in 1980–81 and 1986–87 but collapsed in 1984–85 and 1988–89. We traced the failure to a bug, not to turnover control as such: the penalty compared the optimizer's unscaled solution with last month's rescaled weights. We therefore set $\kappa = 0$, under which each month's portfolio is independent of the last and every γ gives the same scaled weights, and tested a correctly scaled alternative, smoothing weights after the optimizer (Table 2, first row). The walk-forward Sharpe ratios of 3–5 are far above what the strategy earned after 1990. They come from short windows in an era when many anomalies were stronger and not yet published (McLean and Pontiff, 2016). We use them only to rank settings, never as performance estimates. Table 2 lists the refinements screened the same way. None cleared the bar.

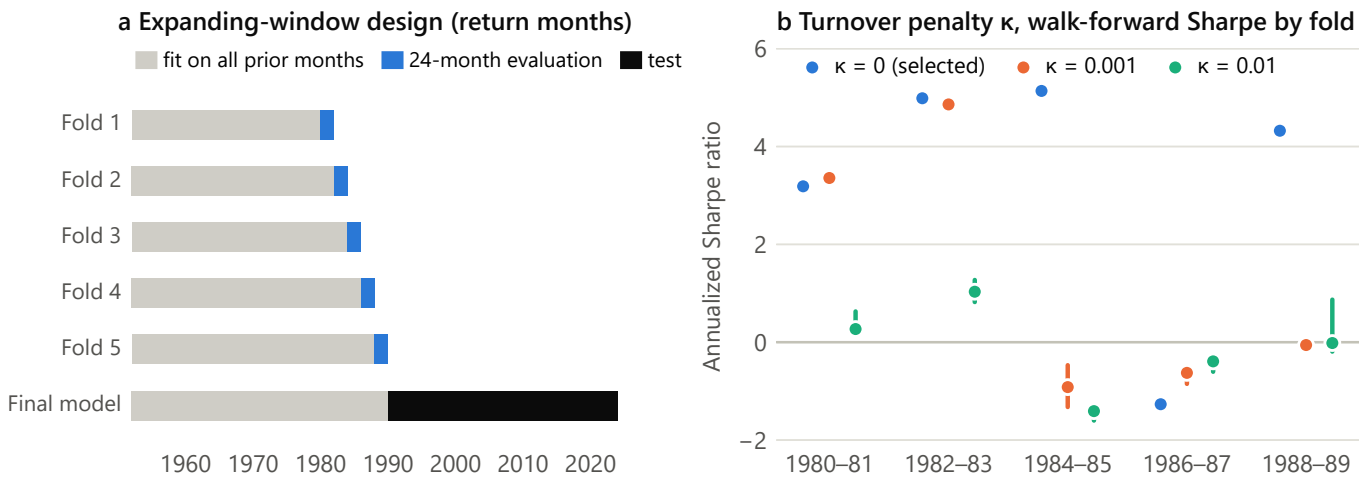


Figure 2. Walk-forward validation, entirely before 1990. (a) Every fold refits the full pipeline on data before its evaluation block. (b) Dots are medians over the five risk-aversion values $\gamma \in \{0.5, 1, 2, 5, 10\}$; vertical lines span their range. $\kappa = 0$ has a median Sharpe ratio of 4.32 across folds, against 0.36 for the best $\kappa > 0$. With $\kappa = 0$, γ does not change the scaled portfolio, so its range collapses to a point.

Table 2. Refinements screened by the same walk-forward test. Median Sharpe ratio across the five pre-1990 folds, variant versus the reference pipeline.

Refinement	Variant	Reference	Outcome
Smooth turnover after optimizing (90% new, 10% old weights)	4.46	4.32	Won 2 of 5 folds; 1.93 vs 1.96 on 1990–2023. Not adopted
Debias factor eigenvalues (spiked-covariance correction)	4.32	4.32	Difference 0.1%. Not adopted
Skip feature clustering (use all 402)	4.31	4.32	Mean favored skipping (3.37 vs 3.27); mixed. Kept
Shrink specific variances toward their mean	—	—	Estimated optimal shrinkage was 1–1.8% in 20 months; nothing to fix
Third learner: one-layer neural network	—	—	CV chose the smallest network; runtime tripled. Dropped

Engineering checks: on synthetic data shaped like the competition files, repeated runs give identical output, and truncating the data leaves all earlier weights unchanged (the competition's look-ahead test). Every month is exactly dollar-neutral with 200% gross exposure. A full run takes 2–3.5 hours on 8 cores, depending on CPU speed, and fits in a 62 GB machine, within the CTF's 300 GB and 24-hour limits; no GPU is needed. A rerun on a different server with newer package versions reproduced every fitted parameter in the log ($\lambda = 0.0026267$, 30 nonzero coefficients, 7 trees, stacking weights 0.5173 and -0.2499) and the Sharpe ratio to four decimals. The walk-forward drivers are separate development scripts, available on request.

4 How we arrived at this design

We followed the course's advice to start simple (Figure 3). The course recipe earned 1.38: ranked characteristics, an elastic net, long the top and short the bottom. About twenty variants of rule-based, risk-aware construction (risk-adjusted ranks, top and bottom slices, position caps, turnover blending), with boosted trees added, never beat 1.41. The optimizer of Section 2.4, fed *identical* forecasts, raised the Sharpe ratio from 1.10 to 1.96. Over the full period the binding constraint was construction, not prediction, although Section 6 shows the advantage is concentrated before 2010. The optimizer first needed two fixes. An early version fed it ranks instead of return forecasts, and daily instead of monthly variances. With mismatched units, the solver returned positions of millions of dollars per dollar of capital. The second fix was the specific-variance floor of Section 2.3.

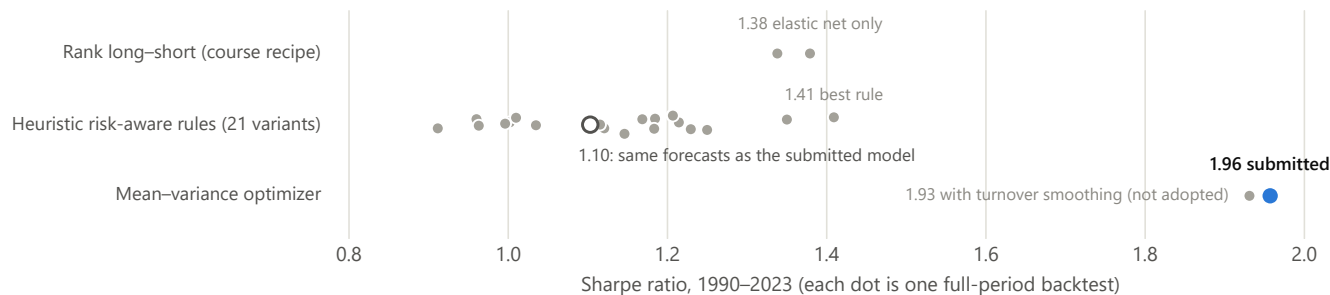


Figure 3. Every full-period backtest we ran (25). The open circle and the submitted model use the same forecasts and differ only in portfolio construction.

Disclosure. All 25 backtests in Figure 3 were scored on the 1990–2023 test period. From the optimizer on, a change was run on that period only if it first won the pre-1990 walk-forward test. One change, turnover smoothing, won narrowly, lost on the full period, and was not adopted. Because the test sample was viewed repeatedly, Section 6 deflates the Sharpe ratio for 25 trials.

5 Out-of-sample performance, 1990–2023

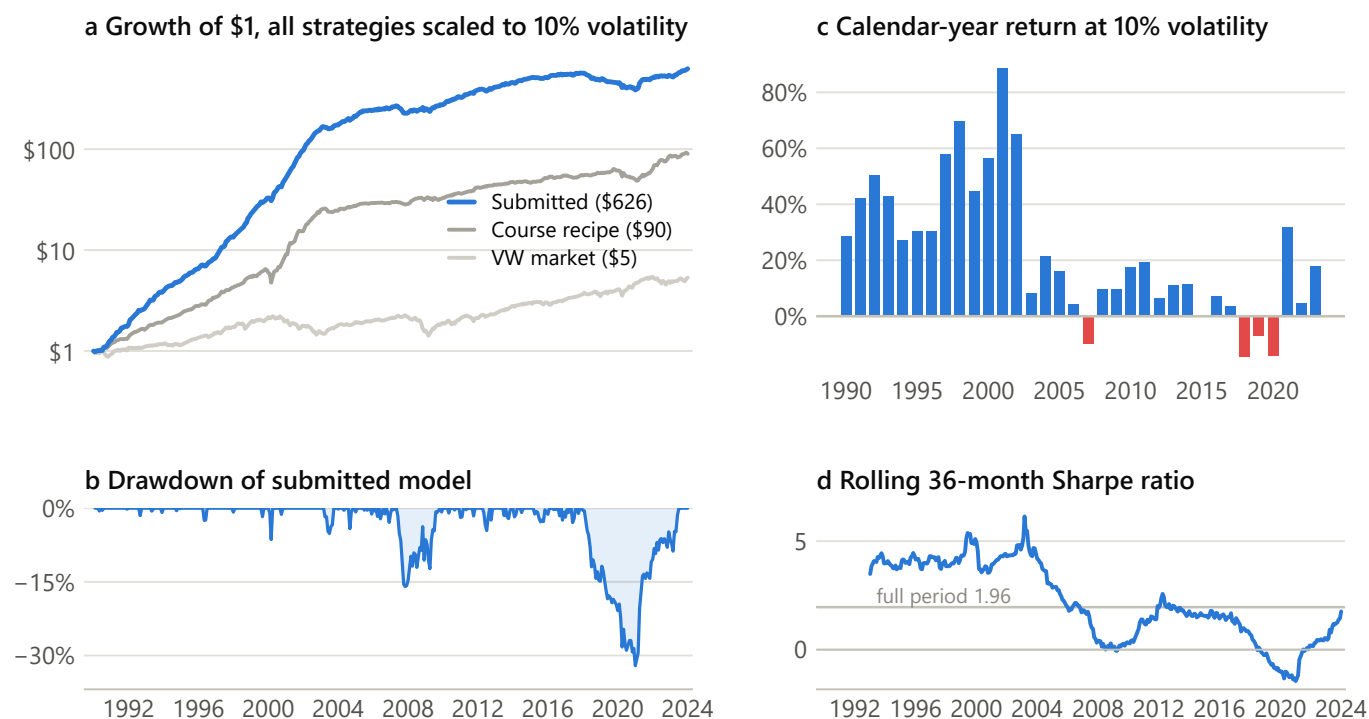


Figure 4. Performance of the submitted weights. All series are scaled to 10% annualized volatility so that cumulative returns compare like with like. “Course recipe” is the elastic-net rank long–short baseline; “market” is the value-weighted excess return of the same universe.

Figure 4 and Table 3 summarize the out-of-sample record. At 10% volatility the model compounds faster than both the course recipe and the market; 30 of 34 calendar years are positive, and the worst year, 2018, lost 14.5%. The deepest peak-to-trough drawdown, –32.1% at 10% volatility (–15.1% as built), ran from October 2017 to December 2020; the portfolio regained its previous peak in June 2023. Returns are essentially independent of the market: correlation 0.09 and beta 0.03 with the value-weighted market excess return. The edge weakens over time (Figure 4d): the Sharpe ratio falls from 4.02 in 1990–1999 to 0.79 in 2010–2019, and over 2010–2023 the simpler course recipe did better (0.92 against 0.78).

On the public leaderboard (7 October 2026), 1.96 would rank about 24th of 57: above the organizers’ instrumented-PCA, rank-weighted-factor, minimum-variance and equal-weight benchmarks, but below their Markowitz-ML benchmark (2.52) and the leading entries (4.42 and 4.11), and with a deeper drawdown than any entry above it. Eight of the top ten name gradient-boosted trees (LightGBM or XGBoost) as their forecaster, whereas our cross-validated trees are nearly inactive (seven depth-2 trees), so the gap most likely lies in nonlinear forecasting rather than in portfolio construction.

Table 3. Performance summary, gross of costs, January 1990–December 2023 (408 months). “At 10% vol” applies a constant leverage factor.

	Submitted	Course recipe		Submitted	Course recipe
Sharpe ratio	1.96	1.38	Sharpe ratio, 1990–1999	4.02	2.99
95% confidence interval	1.62–2.29	1.00–1.76	Sharpe ratio, 2000–2009	1.93	1.27
Volatility as built (100/100)	4.3%	9.5%	Sharpe ratio, 2010–2019	0.79	1.23
Leverage for 10% volatility	2.3×	1.0×	Sharpe ratio, 2020–2023	0.82	0.81
Return at 10% vol	19.6%	13.8%	Correlation with market	0.09	−0.16
Maximum drawdown at 10% vol	−32.1%	−26.8%	CAPM alpha t-statistic	11.2	8.7
Worst month at 10% vol	−7.7%	−18.1%	Monthly turnover, $\Sigma \Delta w $	191%	157%
Positive months	71%	70%	Skewness / kurtosis	0.48 / 4.4	0.56 / 13.0

6 What an investor should expect

The model's parameters used no post-1989 data, but five sources of uncertainty surround the 1.96. Sampling error and serial correlation widen the range; multiple testing, costs, and decay push the expectation down.

- **Sampling error.** Allowing for skewness and fat tails (Lo, 2002), the Sharpe ratio's standard error is 0.17, so the 95% interval is 1.62–2.29. It describes the full-period average, not a stable future Sharpe ratio (see Decay).
- **Multiple testing.** Given the dispersion of our 25 full-period backtests (Figure 3), the best of 25 skill-free strategies would show a Sharpe ratio of about 0.51. The deflated Sharpe ratio (Bailey and López de Prado, 2014), the probability that the true Sharpe ratio exceeds that bar, is above 0.999, so the result is not an artifact of search.
- **Serial correlation.** Monthly returns are autocorrelated: lag 1 is 0.24 (band ± 0.10), lags 1–11 sum to 2.78, and the Ljung–Box test over 12 lags gives $p < 0.001$. Taken at face value, Lo's (2002) adjustment cuts the annualized Sharpe ratio to 0.99. But the dependence comes almost entirely from the slowly declining mean: after removing a centered five-year moving average, lag 1 is 0.00 and lags 1–11 sum to 0.04. The real risk is decay, not short-run dependence.
- **Trading costs.** With no turnover penalty, the portfolio trades 191% of capital a month ($\Sigma|\Delta w|$), and 39% of gross exposure sits in small caps. At 10 basis points per dollar traded the full-period Sharpe ratio falls to 1.42, and at 20 basis points to 0.89; the break-even cost is 37 basis points.
- **Decay.** The Sharpe ratio was 4.02 in 1990–1999, 1.93 in 2000–2009, 0.79 in 2010–2019 and 0.82 in 2020–2023. The course recipe, also fit once on pre-1990 data, declined less (2.99, 1.27, 1.23, 0.81), so this is more than the post-publication decay documented by McLean and Pontiff (2016): the optimizer's advantage is concentrated before 2010.

Bottom line. For the model as submitted, the post-2010 record is the right guide, not the full period: a gross Sharpe ratio of 0.78 over 2010–2023, or 0.45 and 0.12 after costs of 5 and 10 basis points per dollar traded. We therefore expect a forward Sharpe ratio of roughly 0.1–0.4 after costs, with returns largely independent of the market. That is positive but modest, and uncertain; the 1.96 headline mostly reflects the 1990s. Refitting, a stronger nonlinear forecaster and a turnover penalty are the changes most likely to raise it (Section 7).

7 Limitations and next steps

Four tests come first, each on the pre-1990 walk-forward folds: periodic re-estimation, since the submitted model never sees post-1989 data and its edge decayed faster than the course recipe's; non-negative stacking, equivalent here to elastic-net forecasts alone; a correctly scaled turnover penalty, since costs consume most of the post-2010 edge; and, as the leaderboard suggests, a stronger nonlinear forecaster with deeper and many more trees, still under non-negative stacking weights. A more aggressive specific-variance floor raised the first walk-forward fold from 3.19 to 3.64; testing it on all five folds is the next step.

Acknowledgments. The weight convention started from the CTF rules' example; the code was developed with an AI coding assistant (Claude, Anthropic). No leaderboard code was reused; the author is responsible for all content.

Breiman, L. (1996). Stacked regressions. *Machine Learning* 24, 49–64.

Bailey, D. H., and M. López de Prado (2014). The deflated Sharpe ratio. *Journal of Portfolio Management* 40(5), 94–107.

Chen, T., and C. Guestrin (2016). XGBoost: A scalable tree boosting system. *Proceedings of KDD*, 785–794.

Connor, G., and R. Korajczyk (1986). Performance measurement with the arbitrage pricing theory. *Journal of Financial Economics* 15, 373–394.

Jensen, T. I., B. Kelly, and L. H. Pedersen (2023). Is there a replication crisis in finance? *Journal of Finance* 78(5), 2465–2518.

Jensen, T. I., B. Kelly, S. Malamud, and L. H. Pedersen (2024). Machine learning and the implementable efficient frontier. Working paper.

Lo, A. W. (2002). The statistics of Sharpe ratios. *Financial Analysts Journal* 58(4), 36–52.

McLean, R. D., and J. Pontiff (2016). Does academic research destroy stock return predictability? *Journal of Finance* 71(1), 5–32.

Stellato, B., G. Banjac, P. Goulart, A. Bemporad, and S. Boyd (2020). OSQP: An operator splitting solver for quadratic programs. *Mathematical Programming Computation* 12, 637–672.

Zou, H., and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B* 67(2), 301–320.