

Markowitz-ML: An Industry- and Beta-Neutral, Risk-Optimised Machine-Learning Portfolio for US Equities

Aurelia Gong · MGT 924 Statistical Foundations · Problem Set 3 · October 2026

Summary. Each month the model ranks 402 firm characteristics and forecasts each stock's next-month return relative to its industry with an equal-weight ensemble of three learners: a linear ridge, a random-feature ridge and gradient-boosted trees. The forecasts are held in two mean-variance portfolios, one built on a Barra-style factor risk model and one on a shrunk sample covariance of daily returns. Both portfolios have zero net exposure in every 2-digit SIC industry and zero market beta. All design choices were made on 1962-1989; the CTF test sample (Dec 1989 - Nov 2023) was used only to evaluate frozen designs. On it the strategy earns an annualised Sharpe ratio of **3.92** (standard error 0.22; Newey-West $t = 12.6$), against 1.62 for a ridge decile long-short and 0.49 for the 1/N market. Performance has decayed over time (Sharpe 2.56 since 2010), which is how an investor should calibrate expectations (Section 5).

1. Data and the protocol against overfitting

Data. The CTF panel holds 1,369,324 stock-months (Dec 1951 - Nov 2023, 864 months, roughly 500-3,000 mega-, large- and small-cap US stocks per month) with 402 characteristics, the next-month excess return `ret_exc_lead1m`, and 103 million daily excess returns. The characteristics come from the global factor database of Jensen, Kelly and Pedersen (2023), and the competition dataset is the one set up by Hellum, Jensen, Kelly and Pedersen (2025); both were downloaded from WRDS. Nothing else is used.

Split. I use three nested layers of out-of-sample discipline:

1. **Inside the code (every month).** All models are refit every 12 months on an expanding window that contains only returns realised by the portfolio date. Ridge penalties are chosen within each training window: fit on all but the last 60 months, pick the penalty with the highest Sharpe ratio of a prediction-weighted long-short in those 60 months, then refit on the full window. Every other setting is fixed in advance. No hyper-parameter is chosen with future data.
2. **Development sample (portfolio months Jan 1962 - Nov 1989, `ctff_test = 0`).** Every design decision was taken by comparing the out-of-sample Sharpe ratios that layer 1 produces here. All 19 variants I tried are listed in Figure 2, including the rejected ones.
3. **Holdout (Dec 1989 - Nov 2023, `ctff_test = 1`).** The competition's scoring window. It was evaluated twice, each time for a design frozen beforehand: version 1 (Sections 2-3 without the starred additions) and the final version 2. Nothing was tuned on it.

The code passes the competition's own look-ahead test: running `main()` on data truncated at Dec 1970 reproduces every earlier weight exactly (maximum difference 0.0), and two runs give bit-identical output.

2. The model

Features and industries

Each month every characteristic is converted to its cross-sectional rank, scaled to $[-0.5, 0.5]$; missing values are set to 0, the cross-sectional median. Ranks remove outliers and make all 402 inputs comparable. Industries come from SIC codes: Fama-French 12 groups for the risk model and the finer 2-digit SIC groups (about 65 per month) for the neutrality constraint and the prediction target.

Return forecasts: an ensemble of three learners

The target is the next-month excess return minus its month-by-SIC2 average, so the models learn **within-industry** stock selection, the only return source the portfolio can capture.

- **Ridge** on the 402 ranks (the course's workhorse; 1M+ observations make $X^T X$ well estimated).
- **Random-feature ridge**, the “virtue of complexity” approach of Kelly, Malamud and Zhou (2024): 4,096 random Fourier features $z = [\sin(XW), \cos(XW)]$ with $W_{jk} \sim N(0, 12/402)$, followed by ridge.
- **Gradient-boosted trees** (scikit-learn histogram boosting: 300 trees, learning rate 0.05, 255 leaves, at least 2,000 observations per leaf) fitted to $\arctan$ of the month-standardised target, which bounds the influence of extreme returns.

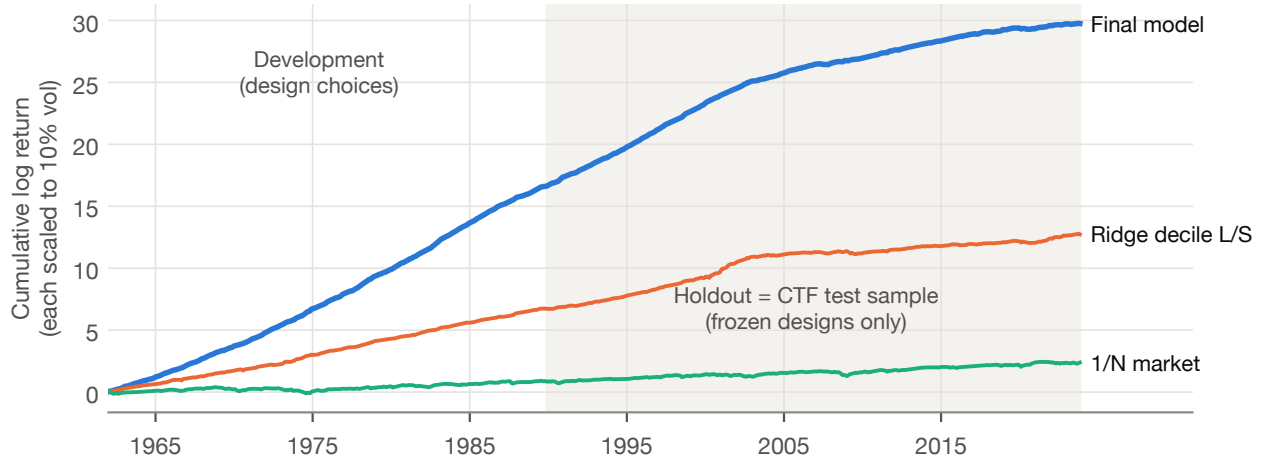


Figure 1: Cumulative log returns, each strategy scaled to 10% annualised volatility over 1962-2023 so slopes are comparable. Shaded: the holdout (= CTF test sample), never used for design choices; it was used only to evaluate the two frozen versions.

Ridge penalties are $\lambda \cdot \bar{e}$ with $\bar{e}$ the average eigenvalue of $X^\top X/n$, so one grid ($\lambda \in 10^{-5}, \dots, 10^3$) serves both ridge models. Forecasts are standardised each month and averaged with equal weights.

Two risk models

Factor model. Following the Barra approach from class 2-3 and Jensen et al. (2024), $\Sigma_t = B_t F_t B_t^\top + D_t$ with 12 industry dummies and 13 style factors. The style factors come from hierarchical clustering (average linkage, distance $1 - |\rho|$) of the 402 characteristics, re-estimated every year on the preceding 10 years (13 clusters, as in Jensen, Kelly and Pedersen 2023); a stock's exposure is the standardised average of its sign-aligned ranks in the cluster. Daily cross-sectional WLS regressions on the previous month-end exposures give 18,000 days of factor returns and residuals. F_t is an exponentially weighted covariance (half-lives 504 days for correlations, 84 for variances) with a Newey-West adjustment (5 lags); specific variances are EWMA of squared residuals. On the development sample, realised/predicted volatility is 0.93 for random long-short portfolios and 1.23 for the market (1.58 before the Newey-West adjustment).

Sample covariance. S_t is the Ledoit-Wolf (2004) shrinkage estimate from each stock's trailing 756 daily returns (at least 250), regularised as $S_t + \bar{e}I$. It captures co-movement that 25 factors miss, while the factor model is stabler for stocks with short histories.

Portfolio construction

Under each risk model $\Sigma \in \{BFB^\top + D, S + \bar{e}I\}$ the weights solve

$$\max_w \mu_t^\top w - \frac{\gamma}{2} w^\top \Sigma w \quad \text{s.t.} \quad H_t^\top w = 0 \quad \Rightarrow \quad w \propto \Sigma^{-1}(\mu_t - H_t \lambda_t), \quad \lambda_t = (H_t^\top \Sigma^{-1} H_t)^{-1} H_t^\top \Sigma^{-1} \mu_t,$$

where H_t holds the SIC2 dummies and each stock's trailing market beta (EWMA of daily returns against the equal-weighted market, half-life 252 days), so the book is dollar-, industry- and beta-neutral. The two books are scaled to equal ex-ante risk, averaged 50/50, and scaled to 10% ex-ante annualised volatility under the factor model; the scale does not affect the Sharpe ratio.

3. How the design was chosen

Version 1 (no stars in Figure 2). Three lessons shaped it. Portfolio construction matters as much as prediction: the same ridge forecasts earn 3.1 as an equal-weighted decile sort and 4.9 once a risk model decides the weights. Finer neutrality pays: moving constraint and target from FF12 to SIC2 adds 0.8. Nonlinearity helps modestly: random features add 0.3-0.4. Several reasonable ideas failed: volatility-scaled forecasts, shrinking specific variances, a rank-transformed target, and **Portfolio-ML** (class 7-9) on the 402 characteristic-managed portfolios (0.41; a 402-dimensional tangency portfolio cannot be estimated from 100-300 months).

Version 2 (starred). I then studied the public leaderboard entries with holdout Sharpe ratios above 3 and took four recurring ideas to the development sample, fixing their settings in advance. Three survived. A second, Ledoit-Wolf risk-model book (from Geertsema's entry) and holding one forecast in two books (Keshri) add 0.6: the two

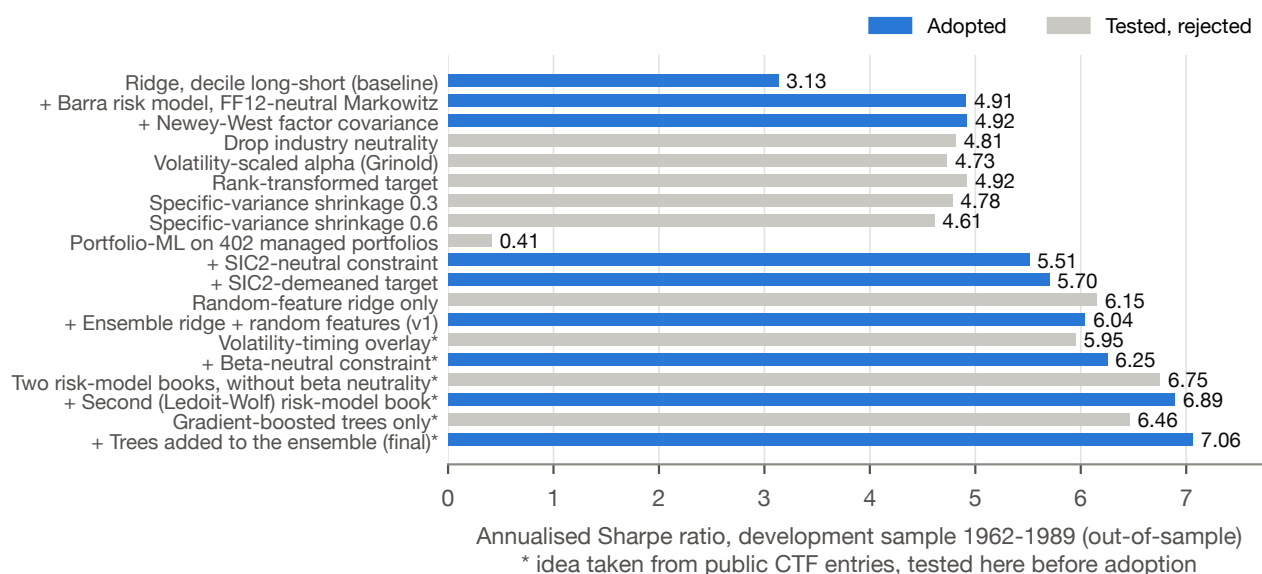


Figure 2: Every variant tried, in the order tested, with its out-of-sample Sharpe ratio on the development sample. Blue steps were adopted.

books' returns are only 0.82 correlated. Beta neutrality (Keshri) adds 0.2. Gradient-boosted trees add 0.2 to the ensemble. A volatility-timing overlay (Moreira and Muir 2017), which several entries credit with much of their performance, **lowered** the development Sharpe ratio and was dropped. Because these entries were themselves selected for their holdout performance, I adopted an idea only on development-sample evidence.

4. Out-of-sample performance (holdout, Dec 1989 - Nov 2023)

Strategy	Sharpe	s.e.	NW t	Skew	Hit rate	Max DD ^a
Final model (v2, submitted)	3.92	0.22	12.6	0.18	88%	10.9%
v1 forecasts in the v2 portfolio	3.58	0.21	11.9	0.07	87%	11.7%
Version 1 (ridge + RF, factor book)	2.86	0.20	11.5	0.19	81%	13.2%
Ridge decile long-short (baseline)	1.62	0.18	7.2	1.28	69%	21.9%
Equal-weight market (1/N)	0.49	0.17	2.9	-0.49	61%	39.4%

Table 1: Holdout statistics, monthly returns, annualised. The first row is scored from the submitted weights file. s.e.: Lo (2002) standard error of the Sharpe ratio; NW t : Newey-West (6 lags) t -statistic of the mean. ^aMaximum drawdown of cumulative returns, scaled to 10% annualised volatility.

Version 2 raises the holdout Sharpe ratio from 2.86 to 3.92, in line with the development evidence: the new portfolio alone, with the version 1 forecasts, gives 3.58, and the tree model accounts for the rest. The holdout Sharpe ratio is about 55% of the development one (7.06). The same decay appears for the plain baseline (3.12 to 1.62), consistent with the decline of anomaly returns after publication and as markets became more liquid, so it is not a sign of overfitting to the development sample.

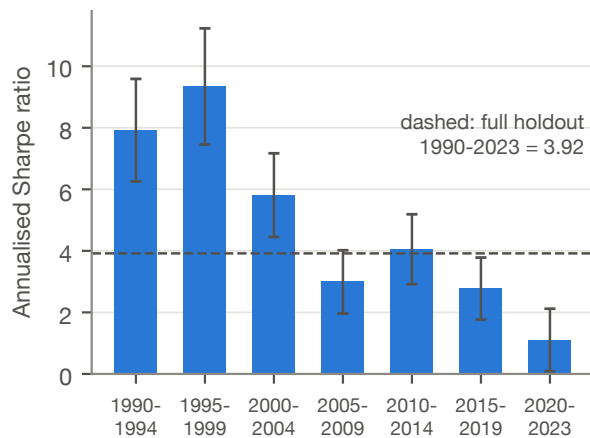


Figure 3: Holdout Sharpe ratio of the final model by five-year period, with 95% confidence intervals.

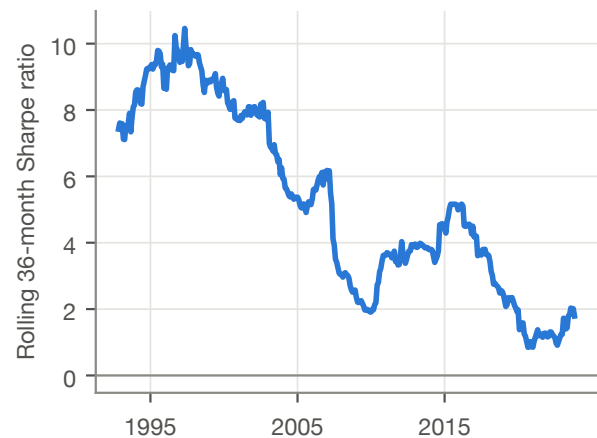


Figure 4: Rolling 36-month Sharpe ratio of the final model over the holdout.

5. What an investor should expect, and the risks

Expected performance. The decline in Figures 3-4 matters more than the full-sample average. The Sharpe ratio was 2.56 over 2010-2023 and 1.10 over 2020-2023, the latter with a wide confidence band. A forward-looking expectation somewhat below the post-2010 figure, **before trading costs**, is more honest than the full-sample 3.92.

Risks and limitations. (i) **Trading costs are ignored**, as in the competition; the portfolio turns over about 101% of its gross exposure each month, so a real implementation needs a turnover penalty (e.g. the Portfolio-ML trading-cost model from class). (ii) **Leverage**: a 10% ex-ante volatility target requires gross exposure of about $11\times$ capital because the hedged book has little predicted risk, and realised volatility runs about 2.0 times the forecast because errors in μ are not part of Σ . (iii) **Residual market exposure**: after beta neutralisation the book's beta to the equal-weighted market is 0.20 (correlation 0.20). (iv) **Worst episode**: the largest drawdown, 10.9% at 10% volatility, ran from 2019-10 to 2020-09, through the COVID-19 turmoil, and the Sharpe ratio was lowest in 2020-2023.

6. Reproducibility

The submission is a single Python 3.13 file (`ctf_model.py`) with `main(chars, features, daily_ret)`. All data handling runs in `ibis` on DuckDB and the numerics in `numpy`, `scipy` and `scikit-learn`; `pandas` appears only where the harness hands over and takes back its DataFrames. Dependencies are pinned in `requirements.txt`. The only randomness (random-feature weights, the clustering subsample, the tree learner) uses fixed seeds. A full run took about 11 hours on an 8-core, 16 GB laptop, partly while other jobs ran; the tree refits dominate and parallelise, so the competition's 32 cores bring it well inside the 24-hour limit. **Borrowed ideas**: the competition's rank-transform example; the clustering of characteristics from Jensen, Kelly and Pedersen (2023); random features from Kelly, Malamud and Zhou (2024); the Barra structure and the mean-variance framework from class; and, from public CTF entries, the Ledoit-Wolf daily-covariance book with Tikhonov ridge and the arctan target (Geertsema) and the two-book and beta-neutral construction (Keshri). Their ideas were re-implemented, not copied. AI assistance (Claude) was used for coding; all results were produced by the submitted code.

References. Hellum, Jensen, Kelly, Pedersen (2025), Common Task Framework competition, SSRN working paper 5242901. · Jensen, Kelly, Pedersen (2023), Is There a Replication Crisis in Finance? *Journal of Finance* (data source). · Jensen, Kelly, Malamud, Pedersen (2024), Machine Learning and the Implementable Efficient Frontier. · Kelly, Malamud, Zhou (2024), The Virtue of Complexity in Return Prediction, *Journal of Finance*. · Ledoit, Wolf (2004), A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices, *Journal of Multivariate Analysis*. · Lo (2002), The Statistics of Sharpe Ratios, *Financial Analysts Journal*. · Moreira, Muir (2017), Volatility-Managed Portfolios, *Journal of Finance*. · Public CTF entries by P. Geertsema and A. Keshri, jkpfactors.com leaderboard (2026).