
Problem Set 3

MGT 924: Statistical Foundations

Boosted Stock Rankings in a Two-Book Mean-Variance Portfolio: LightGBM
Forecasts with Factor and Shrinkage Risk Models

Marco Chen
Maggie Ma
Cecilia Xin
Simon Zhang

Course Details:

Instructor: Prof. Theis Ingerslev Jensen
Date: October 7, 2026

Yale School of Management

Contents

1	Summary	1
2	Data and Assumptions	1
3	Return Forecasts	2
4	Portfolio Construction	2
5	How We Arrived at the Model	3
6	Reproducibility and Computational Environment	3
7	Validation and Compliance with Competition Rules	3
8	Results	4
9	What an Investor Should Expect	5
10	References and Disclosures	5

1 Summary

We rank stocks by predicted next-month return with two gradient-boosted tree models trained on all 402 characteristics in the competition data and refit every year on an expanding window. The forecast is demeaned within industry and fed into two mean-variance portfolios that differ only in their risk model, a Barra-style factor covariance and a shrunk sample covariance. Both are dollar- and beta-neutral, and the submitted weights are their 50/50 average. Over the 408 evaluation months (1990–2023) the strategy earns a gross Sharpe ratio of 3.89 (95% interval 3.26–4.53), against 0.85 for the ridge baseline and 0.54 for the value-weighted market (Figure 1). Most of the gain over the baseline comes from portfolio construction rather than from the predictor. Performance has declined over time and the evaluation window was used more than once during development, so we plan around a gross Sharpe ratio of 1.5–2.5 before costs.

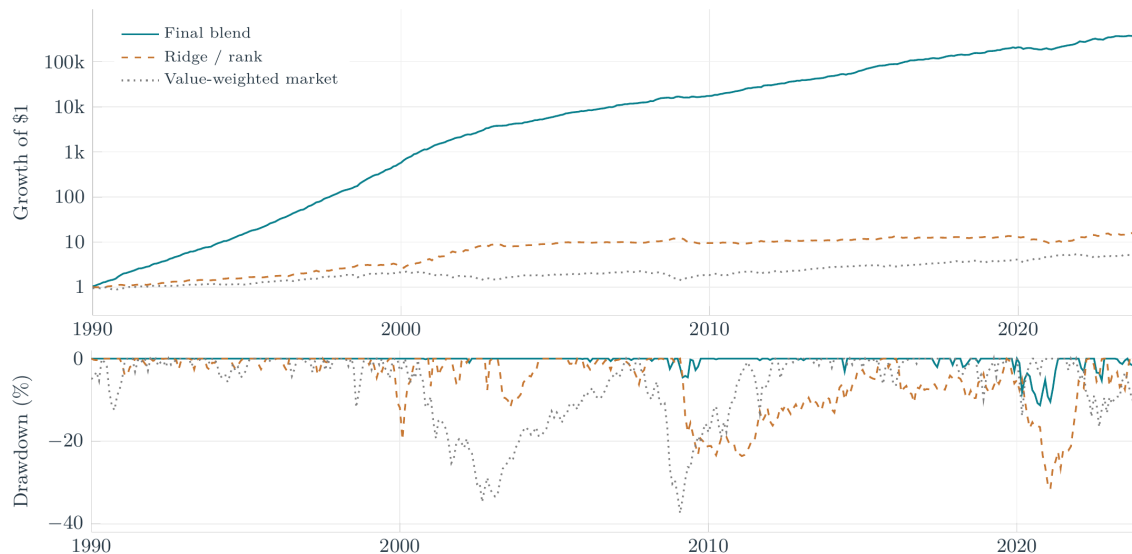


Figure 1: Growth of \$1 (top, log scale) and drawdown from the running peak (bottom), 1990–2023, each series scaled to 10% annual volatility. Teal: final blend; dashed orange: ridge baseline; dotted gray: value-weighted market. Returns are gross of all costs.

2 Data and Assumptions

The dataset has 1,369,324 stock-months (December 1951–November 2023) with 402 characteristics, the next-month excess return and about 103 million daily excess returns. Scored portfolios are formed December 1989–November 2023. Earlier months are used only for training and design decisions, and no external data is used. We assume (i) a characteristic dated at month-end t is observable at t ; (ii) the score is the gross Sharpe ratio of $\sum_i w_{i,t} r_{i,t+1}$, so costs and leverage limits are not modelled; and (iii) because returns are rescaled to 10% volatility, only the relative scale of the weights across months matters.

3 Return Forecasts

Within each month every characteristic is converted to its percentile rank on $[-0.5, 0.5]$, with missing values set to 0 (the median). The target is the within-month rank of next-month return, which removes the market component and limits the effect of extreme returns. All transformations are cross-sectional.

We fit two LightGBM models of different size (Table 1), z-score each model's predictions within the month and average them. The combined signal is demeaned within SIC industry and z-scored again. Models are refit every December from 1989 to 2022 (34 fits) on all observations whose label is realised by that date, and each fit forecasts the next twelve months. The baseline is a ridge regression on the same features, with the penalty chosen on the last 60 months of each training window; it gets zero weight in the final model.

Table 1: LightGBM settings. Both models use 400 rounds, learning rate 0.05 and 63 feature bins.

	Small	Large
Maximum leaves / minimum observations per leaf	31 / 2,000	127 / 5,000
Feature fraction / row fraction	50% / 50%	30% / 50%
L_2 penalty	10	50

4 Portfolio Construction

Book A (factor risk). Following Jensen, Kelly and Pedersen (2023), we cluster the 402 characteristics with Ward linkage on $1 - |\rho|$, using the 120 months before 1990, into 12 groups (value, size/beta, investment, leverage, reversal, issuance, profitability and others). Each group's exposure is the sign-aligned average of its members, z-scored monthly; with 14 SIC industries this gives 26 factors. As in Jensen, Kelly, Malamud and Pedersen (2024), we run daily weighted regressions of returns on previous month-end exposures (square-root market-cap weights). Factor and specific variances use an 84-day EWMA half-life and factor correlations 504 days. The monthly covariance is $\Sigma_A = 21(XFX' + D)$.

Book B (sample risk). We use 756 trailing trading days for stocks with at least 250 observations, zero-fill missing days and demean. With $S = R'R/T$ and $\nu = \text{tr}(S)/N$, the Ledoit–Wolf estimator is $\Sigma_B = (1 - \delta)S + \delta\nu I$, and the allocation uses $H_B = \Sigma_B + \nu I$ to stabilise the inverse.

Weights. With forecast s , market-implied betas $\beta = \Sigma m / (m' \Sigma m)$ and $C = [\mathbf{1}, \beta]$, each book solves mean-variance subject to dollar and beta neutrality:

$$w_0 = H^{-1}s - H^{-1}C(C'H^{-1}C)^{-1}C'H^{-1}s, \quad w_{\text{final}} = \frac{1}{2}w_A + \frac{1}{2}w_B.$$

Each book is scaled to 10% predicted volatility before averaging. Both solves use the Woodbury identity, so no $N \times N$ matrix is inverted.

5 How We Arrived at the Model

We started from the suggested baseline (ridge forecasts, rank weights) and changed one component at a time, comparing choices on a pseudo out-of-sample window: training from 1951 and evaluating 1975–1989. The evaluation window was run three times (Table 2). The gap in v1 led us to two faults: average linkage had put 375 of 402 characteristics into one cluster, and early stopping had cut recent models to 50 trees. Both fixes were chosen on 1975–1989 only. For v3 we fixed a promotion rule in advance (+0.10 design Sharpe, no decline in either half, no higher volatility). Book B alone failed it, but the blend passed. Rejected alternatives (1,000 trees, adding ridge, an arctan target, shrinking the factor model) all scored lower in the design window. Because evaluation results prompted revisions, the final score is not a clean holdout.

Table 2: Development history: Sharpe ratios in the design and evaluation windows.

Version	Configuration	1975–89	1990–2023
v1	Early stopping; average-linkage clusters	5.90	2.26
v2	400 trees; Ward clusters; factor book	6.66	3.28
v3	50/50 blend; industry-demeaned forecasts	8.05	3.89

6 Reproducibility and Computational Environment

- **Deterministic execution.** All randomness (row and feature subsampling in LightGBM) is driven by one fixed seed, 20261006. LightGBM runs in deterministic mode with a fixed thread count of 16, and the input panel is sorted by month and stock before fitting. Repeated runs on a test sample produced identical weights.
- **Dependency pinning.** Package versions are pinned in `renv.lock`: R 4.4.2, `data.table` 1.18.6.1 and `lightgbm` 4.7.0. The model was developed and run under R 4.6.1, so identical results under the pinned R 4.4.2 runtime are expected but have not yet been verified.
- **Consistency of the submitted weights.** The submitted weights are the direct output of the submission script on the full dataset. Recomputing the portfolio return from these weights reproduces the reported Sharpe ratio of 3.89. The look-ahead check (identical weights after truncating the data in December 2005) is reported in Section 7.

7 Validation and Compliance with Competition Rules

- **No Look-Ahead Bias.** All portfolio weights at month-end t use strictly information knowable at t : models consume labels realised by each refit date, risk models use prior month-end exposures, and clustering relies entirely on pre-1990 data. Truncating the dataset at December 2005 reproduces all 439,387 overlapping weights exactly ($\max |\Delta w| = 0.00$).

- **No Cherry-Picking.** All 402 characteristics are utilized; risk factors are derived algorithmically via hierarchical clustering rather than manual feature selection.
- **Competition Data Only.** Inputs are strictly restricted to the supplied JKP CTF dataset without external data sources.
- **Output Schema Compliance.** The submission file (`output.csv`) matches all format constraints: exactly 885,698 rows covering the scored evaluation period (1989-12-31 to 2023-11-30), structured precisely as `id`, `eom`, `w` with zero missing values or duplicate keys.
- **Placebo Test.** Shuffling return forecasts across stocks within each month yields a mean Sharpe ratio of 0.02 (maximum 0.36), confirming that strategy alpha is driven by the predictive signal rather than optimizer artifacts.

8 Results

Table 3 reports evaluation-window performance at 10% ex-post volatility. Intervals use a Newey–West estimator with six lags (Lo, 2002) and do not account for model selection. On the public leaderboard on 7 October 2026 the model would rank fifth of 56, above the Markowitz-ML (2.52) and IPCA (1.95) benchmarks.

Table 3: Performance, 1990–2023 (408 months). Mean and drawdown in percent.

Portfolio	Sharpe	95% CI	Mean	Max DD	Corr. mkt	$t(\alpha)$
Final blend	3.89	[3.26, 4.53]	38.9	−11.3	0.14	12.2
Ridge baseline	0.85	[0.46, 1.25]	8.5	−31.6	−0.51	6.2
Market	0.54	[0.17, 0.92]	5.4	−37.4	1.00	—

Where the performance comes from. Holding forecasts fixed (Figure 2), trees raise the Sharpe ratio from 0.85 to 1.53 under rank weights. The same tree forecasts reach 3.28 with the factor book and 3.69 with the sample book, the blend 3.83, and industry demeaning adds 0.06. Portfolio construction is the largest step.

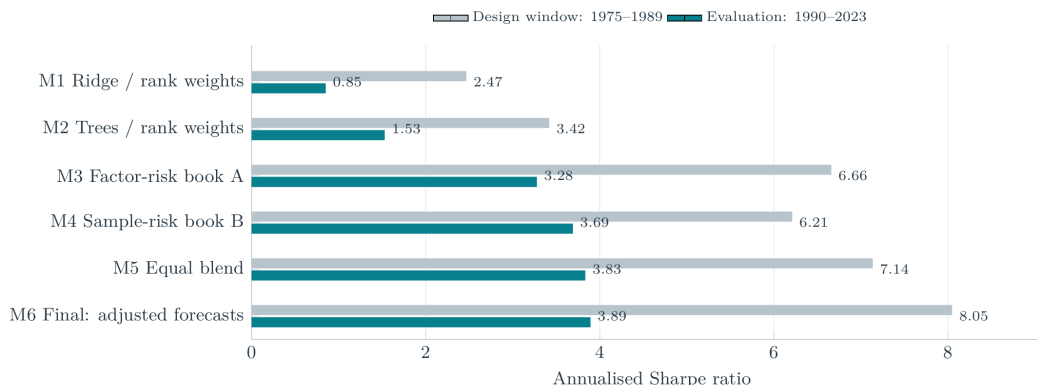


Figure 2: Sharpe ratios of matched variants: design window (gray) and evaluation window (teal).

Stability and risk. The Sharpe ratio falls from 8.01 (1990–2000) to 3.79 (2001–2011) and 2.53 (2012–2023), and is 1.49 over 2020–2023 (Figure 3). Each book targets 10% volatility, but the blend realises 25.4%, with average gross exposure of 14.2 times capital.

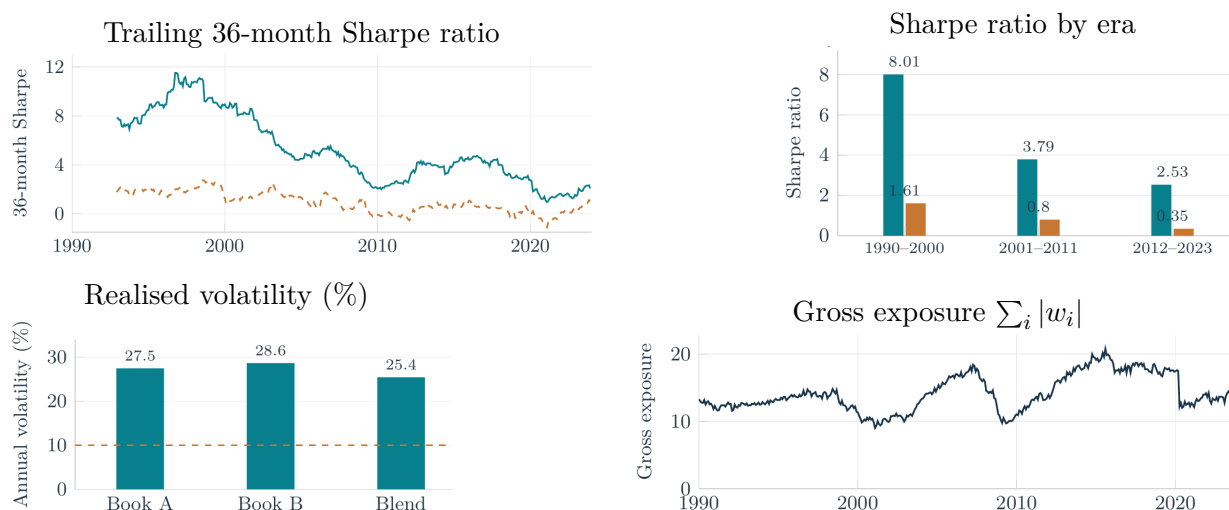


Figure 3: Performance over time (top) and realised risk and leverage (bottom). Teal: final blend; orange: ridge. The dashed line marks each book’s 10% volatility target.

9 What an Investor Should Expect

- **Lower Sharpe ratio.** Given the decline across eras and the repeated use of the evaluation window, we plan around a gross Sharpe ratio of 1.5–2.5, a judgement rather than a forecast.
- **Costs.** Weights turn over about 103% of gross exposure per month (59% for the baseline), and leverage is high, so net returns would be materially lower.
- **Risk.** Realised volatility is about 2.5 times the target, so the risk target should be calibrated on realised outcomes. The returns keep a 0.14 correlation with the market.

10 References and Disclosures

Jensen, Kelly and Pedersen (2023), *Is There a Replication Crisis in Finance?*

Jensen, Kelly, Malamud and Pedersen (2024), *Machine Learning and the Implementable Efficient Frontier*

Ledoit and Wolf (2004), *A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices*

Lo (2002), *The Statistics of Sharpe Ratios*.

Disclosures. Ideas from public leaderboard entries, without copied code: penalized sample covariance (P. Geertsema) and two risk books (A. Keshri). We used AI tools (Claude, Codex) for help with code, checks and figures; all choices and results were reviewed by the authors.