

Ridge-LightGBM Ensemble with Covariance Shrinkage

Group 6 | Zoey Zhang (yz2635) | Aileen Li (al2945) | Lok Yi Yeung (ly384) | Carol Guo (xg328)
MGT 924 STATISTICAL FOUNDATIONS | PROBLEM SET 3 | JKP COMMON TASK FRAMEWORK

1. Research question and results

Abstract. We test whether linear and nonlinear stock-return signals improve a Ridge portfolio when allocation accounts for correlated risk. The final strategy blends Ridge and LightGBM signals equally and forms monthly dollar-neutral portfolios using a regularized daily covariance estimate. All 402 supplied characteristics enter the models; none are selected by hand.

The official January 1990-December 2023 Sharpe is **3.5572**. Table 1 distinguishes the website display from local exposure conventions.

Evaluation / exposure	Annual return	Annual vol.	Sharpe	Max. drawdown
Initial Ridge: local gross 2	13.39%	6.93%	1.9323	-8.70%
Final ensemble: local gross 2	15.83%	4.45%	3.5572	-4.22%
Final ensemble: official 10% vol.	35.57%	10.00%	3.5572	-9.49%

Table 1. Official score checked October 7, 2026 [2]. Annual return is mean monthly excess return times 12; annual volatility is sample monthly standard deviation times sqrt(12); Sharpe is their ratio. Drawdown compounds monthly excess returns, including initial capital. Returns exclude the illustrative costs on page 5.

CTF multiplies the full monthly return series by one constant to obtain 10% realized annual volatility [2]. Here that constant is approximately 2.247. This leaves Sharpe unchanged but changes returns and compounded drawdown. It is a retrospective comparison convention, not a volatility target used by the trading model.

Why this design

Ridge regularizes correlated features; trees capture thresholds and interactions. Equal blending avoids fitting another set of combination weights. Covariance shrinkage stabilizes noisy, high-dimensional risk estimates. The experiments test their joint usefulness rather than proving each mechanism separately.

Data and scope

Only official WRDS CTF inputs are used: 1,369,324 stock-months, 402 characteristics and 103,382,883 daily-return rows. Formation dates span December 1951-November 2023. The evaluation contains 885,698 stock-month weights and 408 return months. Earlier realized history supplies training and risk estimation; 1980-1989 selects the specification.

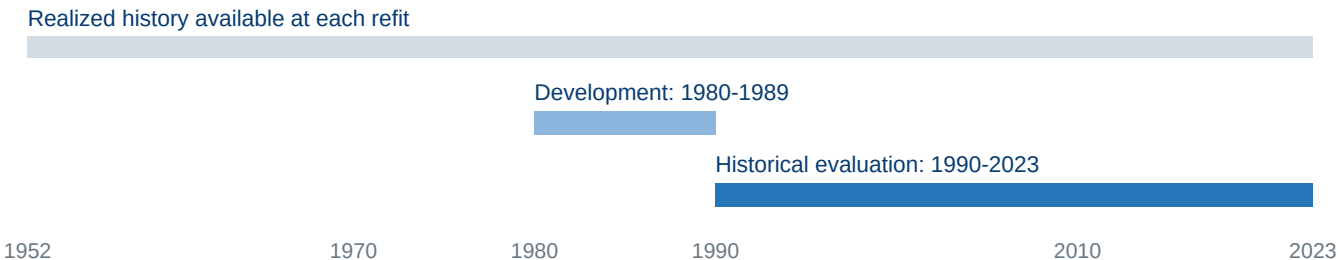


Figure 1. Training incorporates only realized history at each refit. The development window selects the fixed revision; historical evaluation overlaps later expanding training as outcomes become available. Source: project configuration and CTF data.

The 1990-2023 sample had already been viewed during the initial model research and public leaderboard review. Selection among the fixed revision candidates uses pre-1990 returns, but this historical evaluation is not an untouched independent holdout. The historical Sharpe is not a forecast of live performance.

2. From characteristics to portfolio weights



Figure 2. Forecast, risk and portfolio layers of the submitted model. The diagram summarizes the implemented process, not additional models or unrun experiments. Source: submitted project code.

Linear signal: Ridge

Each characteristic is converted to a within-month percentile rank, with missing ranks set to zero, and the cross-section is demeaned. The return label is winsorized at the monthly 1st/99th percentiles and demeaned. Annually refitted regressions use expanding realized history and the latest 120 months; each month receives equal weight. The penalty is 0.10 times the mean diagonal of the training second-moment matrix. The two prediction series are standardized, averaged and standardized again to form the revised Ridge signal.

Nonlinear signal: LightGBM

LightGBM uses all raw float32 characteristics and handles missing values through its learned split routing. Within each formation month, finite next-month excess returns are demeaned and divided by their sample standard deviation, then passed through arctan. This limits the influence of extreme labels. Predictions represent relative attractiveness rather than calibrated dollar returns. Pooled stock-month training gives larger cross-sections more total weight.

Fixed tree settings are 500 boosting rounds, learning rate 0.02, 63 leaves, 63 bins and L2 penalty 1.0. Minimum leaf size is $\min(2,000, \max(20, \text{training rows} / 50))$, with integer division. Training requires 36 realized months, uses seed 42 and deterministic column-wise fitting, and refits annually at January month-end. Only labels realized by that cutoff enter the fit. Each signal is centered, standardized and clipped to ± 3 before the fixed 50/50 blend; combination weights are not optimized on test returns.

Risk allocation

At each formation month, the strategy uses up to 756 prior trading days ending at that date. Stocks need at least 250 finite daily returns to enter the daily-risk allocation. Missing daily returns are replaced with zero and each stock series is demeaned. Let S be the daily sample covariance and m its mean eigenvalue. The risk estimate is $(1-\delta)S + (\delta+1)mI$: δ is the estimated Ledoit-Wolf shrinkage intensity, and mI is an additional ridge penalty. The identity terms stabilize the estimate when the number of stocks is large relative to daily observations.

The standardized signal is multiplied by the inverse regularized covariance, with a projection imposing zero net-dollar and zero estimated market-beta exposure. Beta is estimated against the equal-weight daily return of the current eligible stocks. The portfolio is scaled so absolute weights sum to 2, corresponding to approximately 100% long and 100% short. A Woodbury identity solves in the smaller time dimension; it changes computation, not the allocation formula.

Stocks excluded from the daily-risk allocation receive zero weight. If the allocator has fewer than five eligible stocks or inadequate daily variation, a 60-month causal risk model is used. With n observed months, beta is shrunk toward one with data weight $n/(n+24)$. Residual variance is shrunk halfway toward its cross-sectional median and floored at one-quarter of that median; fewer than 12 observations use beta one and the median variance. It enforces the same neutrality and gross-exposure constraints. Neither allocator imposes hard industry, stock-position or turnover limits.

3. Model development and validation

Baseline and fixed alternatives

The initial strategy averages centered prediction ranks from expanding and recent Ridge regressions and allocates using monthly beta and residual-risk estimates. Five fixed configurations are tested across two 60-month development blocks. Figure 3 compares annualized Sharpe before and after illustrative costs.

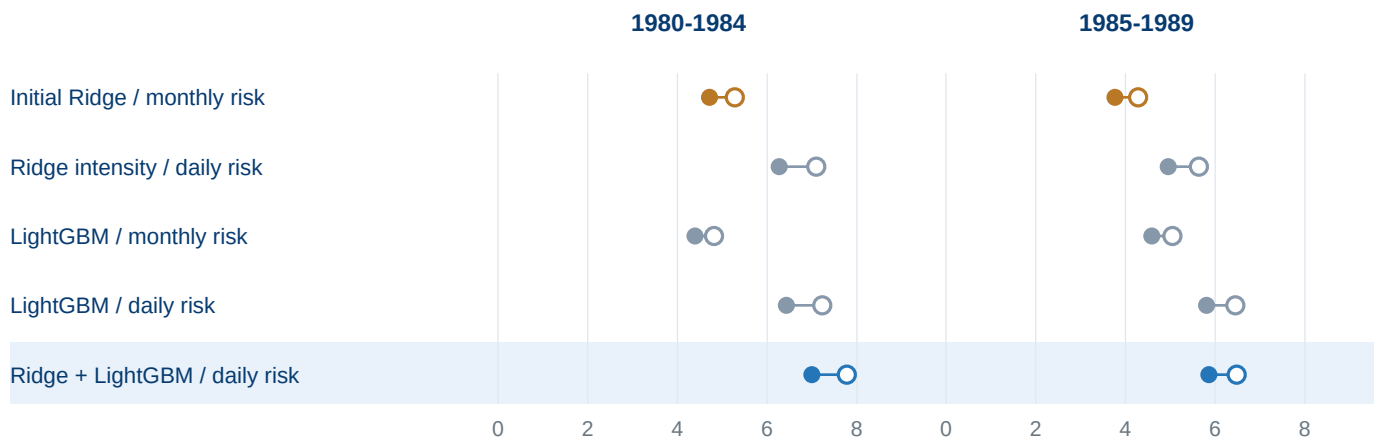


Figure 3. Hollow dots: gross Sharpe; filled dots: after 10 basis points per unit of absolute target-weight change. Each block has 60 months. The final row is the selected ensemble. These are development results, not official test-period scores. Source: saved development experiment results.

The predeclared rule selects the highest average cost-adjusted Sharpe, provided gross Sharpe improves in both blocks and average cost-adjusted Sharpe exceeds the initial model; otherwise it retains the initial model. The ensemble scores 7.78 / 7.00 and 6.48 / 5.86 (gross / cost), averaging 6.43 after costs versus 4.24 for the initial model. All candidates retain the supplied feature list.

What the comparison establishes

The strongest development candidates combine nonlinear signals with daily risk allocation, and the ensemble improves on either signal alone in both blocks. However, the revised Ridge signal also changes magnitude handling, so the baseline comparison is not a clean experiment isolating covariance estimation. Nor do five candidates exhaust the possible model choices. The selection rule limits direct test-period tuning but cannot remove all research-selection uncertainty.

Chronological safeguards

A forward-return label enters training and monthly risk history only after its recorded realization date. Daily-risk windows stop at the formation date. Preprocessing is local to the relevant cross-section; feature definitions, missing-value rules and fixed model settings do not use future observations. Test flags identify scoring observations rather than authorizing their future labels for training. After 1990, past test-period outcomes can enter later refits once realized; future outcomes cannot.

Reproducibility and checks

The submitted function accepts the three official input tables and returns integer security IDs, formation month-ends and finite weights. Dependencies are pinned to NumPy 2.3.5, pandas 2.2.3, SciPy 1.16.3 and LightGBM 4.6.0. Local checks cover deterministic output, input ordering, future truncation, future feature/daily-return perturbations and unavailable-label perturbations. Actual entry-point calls on 123-month panels include a 46,020-row real-data panel. The time-dimension covariance solve is checked against a direct stock-covariance solve.

All 885,698 final weight keys match the official test observations; recalculated returns match within $4.8e-14$. A full portfolio-core check reproduces the CSV using saved causal annual tree fits and verified Ridge coefficients, not a second fresh full-data entry-point refit. The published official score matches local Sharpe to four decimals.

4. Historical performance and stability

Full-period local Sharpe improves from 1.93 to 3.56, while annual volatility falls from 6.93% to 4.45%. Figure 4 uses the same gross-two exposure convention for both versions, without the official 10% rescaling.

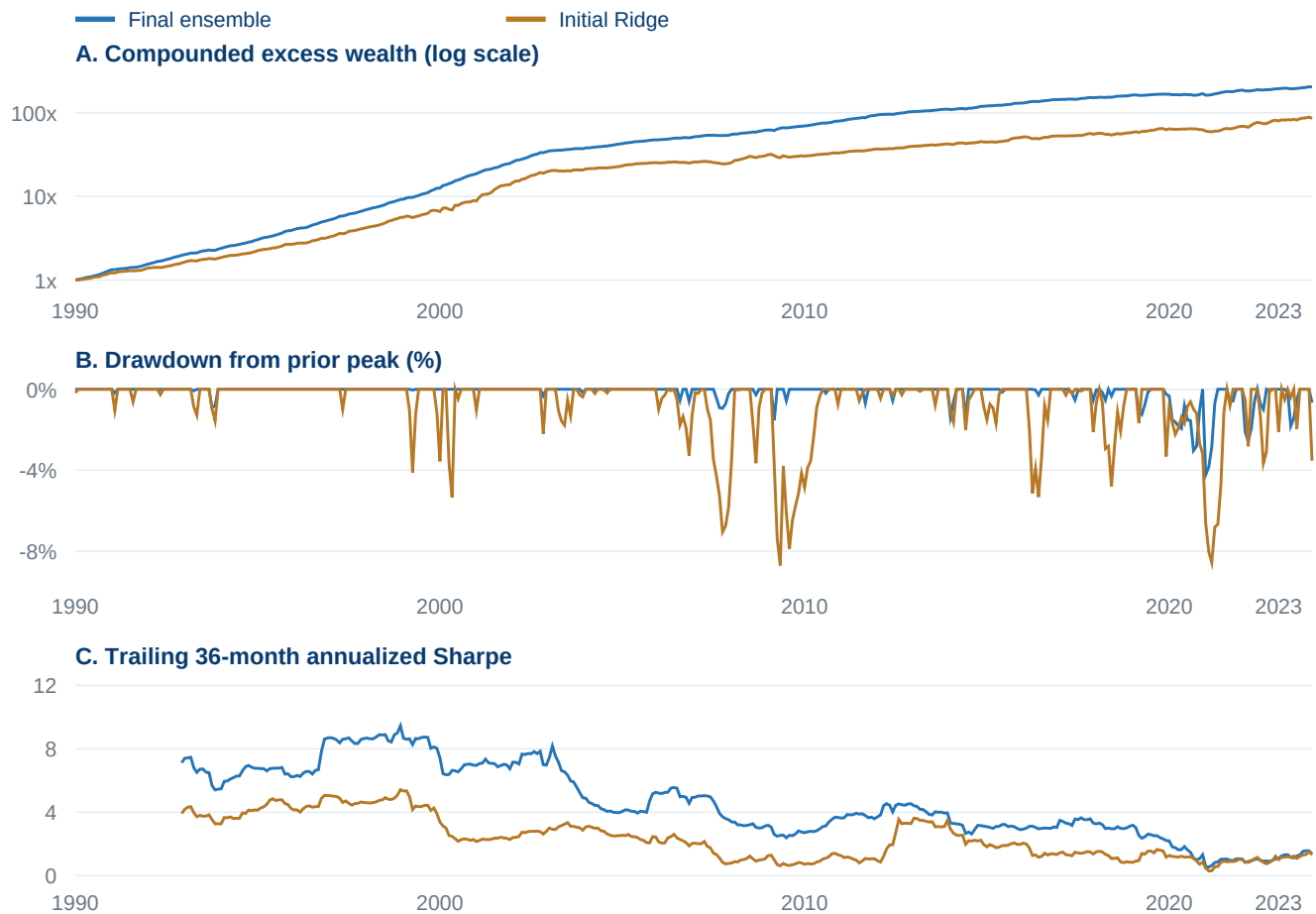


Figure 4. Wealth compounds monthly excess returns from capital one; drawdown includes initial capital. Rolling Sharpe uses the previous 36 realized returns (sample standard deviation, annualized); its first value is December 1992. These descriptive charts omit trading costs, cash interest and financing. Source: saved 408-month return series.

Realization period	Annual return	Annual vol.	Sharpe
1990-1999	25.62%	3.63%	7.05
2000-2009	17.28%	4.81%	3.59
2010-2019	8.88%	2.76%	3.22
2020-2023	5.07%	4.40%	1.15

Table 2. Final strategy at local gross-two exposure. The last block has 48 months; each preceding block has 120. Source: saved monthly returns.

What an investor should take from this

Annual mean excess return falls from 25.62% in the 1990s to 5.07% in 2020-2023; Sharpe falls from 7.05 to 1.15. The recent block contains the largest local drawdown, 4.22%. Full-period averages therefore give an incomplete current expectation. The cause of the decline is not identified; live evaluation and regime/exposure diagnostics are needed before allocating capital.

5. Implementation limits and disclosures

Trading-cost sensitivity

The following charges are illustrative assumptions, not observed execution costs or a required competition fee. Each is charged on all absolute target-weight changes, counting purchases and sales. Weights remain fixed across scenarios, at local gross-two exposure.

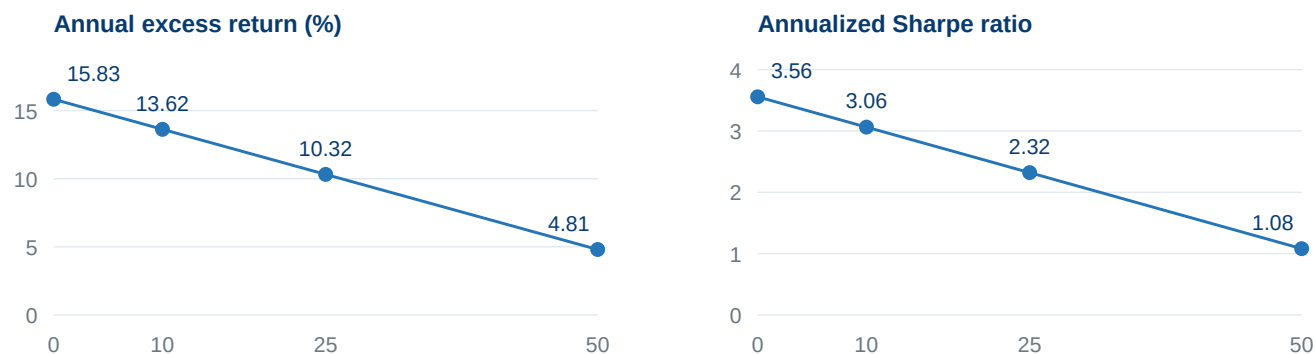


Figure 5. Horizontal axes: basis points per unit of absolute target-weight change. Annual volatility stays near 4.44%-4.45%. Mean monthly trade volume is 1.837 units of capital. A 10-basis-point charge gives 13.62% annual excess return and Sharpe 3.06. Initial entry is included; weight drift, market impact, borrow fees and financing are excluded. Source: saved cost-sensitivity calculations.

Assumptions and remaining risks

The supplied characteristics are assumed point-in-time; their publication lags are not independently reconstructed. Zero-filled missing daily returns may understate risk. Estimated beta neutrality does not guarantee realized beta neutrality or industry neutrality. The maximum observed absolute stock weight is 0.69%, but there is no hard cap. Liquidity, short availability and actual execution costs are not tested. Fixed gross exposure also leaves realized volatility variable.

The final signals are relative scores, not expected returns in dollars. Their standardized magnitudes can affect risk allocation despite clipping. Five candidate trials, prior inspection of historical test results and learning from public entries create selection uncertainty. The official score establishes historical competition performance; it does not establish live profitability or remove these limitations.

Borrowed ideas and AI assistance

Paul Geertsema's public entry informed the raw-feature LightGBM, bounded-label and shrunk daily-covariance choices [3]. Abhinav Keshri's entry informed beta-neutral allocation and model diversification [4]. Implementations were written for this project, including the time-dimension covariance calculation; they are not newly invented methods. MAXSER-PR was reviewed but is not implemented. OpenAI Codex provided extensive assistance: drafting code, running training and validation, preparing figures and drafting the report. The submitting students remain responsible for reviewing the methods, checking the results and explaining the final submission.

Visual presentation borrows the academic layout and diagram approach of Wen et al. [6]; its model designs, placebo results and performance figures are not used in this study.

References

- [1] Jensen, T. I. *MGT 924 Statistical Foundations: Problem Set 3*, pp. 1-4.
- [2] Global Factor Data. [CTF leaderboard](#) (October 7, 2026); [CTF rules](#) (updated September 29, 2026).
- [3] Geertsema, P. [LightGBM with penalised VCV \(MV - GMV\), raw features](#), public CTF code.
- [4] Keshri, A. [XGB+MLP + two risk models \(beta neutral\)](#), public CTF code.
- [5] LightGBM. [Parameter documentation](#).
- [6] Wen, X., Huang, Y., Wang, B. and Wu, E. *Placebo-Audited LightGBM with Size-Block Covariance*, supplied report (visual layout reference only).