

Boosted Trees on Ranked and Raw Characteristics with an Industry-Neutral Factor Risk Model

JKP Common Task Framework submission, MGT 924 Problem Set 3

Souveer Aggarwal, EK Im, JiWhan Moon, Abdul-Kabiru Nashiru (Yale School of Management)

October 06, 2026

1 Summary

The strategy is a monthly-rebalanced, dollar-neutral US equity portfolio. Each month it predicts next-month relative returns for every stock from the 402 characteristics in the competition data with two gradient-boosted tree models, one on cross-sectional ranks and one on raw values, blended with a ridge regression, and converts the predictions into weights with a mean-variance step that uses a 40-factor risk model built by clustering the same characteristics, with net industry positions removed. On the 1990-2023 test sample the portfolio earns a gross Sharpe ratio of **3.87** with a market beta of 0.08, positive returns in 90% of months, and a maximum drawdown of -5.1% when scaled to 10% annual volatility. The ablation in Section 4 quantifies each step. Every design choice was selected on a 1975-1989 development window, so the test period was evaluated once. Two elements are borrowed, with attribution, from public entries on the competition leaderboard: the raw-characteristic tree model of Paul Geertsema and the industry neutralisation of Abhinav Keshri.

2 Task and data

The dataset covers 1.37 million stock-months of relatively liquid US stocks from December 1951 to November 2023, with 402 firm characteristics per stock-month (JKP characteristics: valuation, profitability, growth, investment, momentum, reversal, seasonality, volatility, liquidity, and so on), next-month excess returns, and daily excess returns. Months from December 1989 onward are flagged as the test sample (408 months). The score is the Sharpe ratio of the portfolio's monthly excess return, gross of trading costs. Two consequences drive the design: only the *ordering* and *relative sizing* of positions matter, not the level of leverage, and turnover is free, so short-horizon signals can be used at full strength.

3 Method

All quantities used to form the portfolio at month-end t are computed from data realised by t . Model refits at t use only stock-months whose next-month return has already been observed by t .

3.1 Pre-processing

Each characteristic is replaced by its cross-sectional rank within the month, mapped to $(-0.5, 0.5)$; missing values are set to 0 (the median). This makes every feature scale-free, robust to outliers, and comparable across decades. The prediction target is the next-month excess return minus its cross-sectional mean, so both models learn *relative* performance. For the tree model on ranks the target is itself rank-transformed.

3.2 Return prediction

Three models are refit on an expanding window and blended.

Ridge regression. Because the design matrix consists of ranks, the per-month cross-products $X_t'X_t$ and $X_t'y_t$ are computed once and summed, so the ridge estimator is available in closed form for any window at negligible cost. It

is refit every month. The penalty is chosen from a grid by the average Spearman correlation between predictions and realised returns over the last 60 months of the training window.

LightGBM on ranks. A gradient-boosted tree ensemble (63 leaves, learning rate 0.05, 50% feature and row subsampling, L2 penalty 10, at least 1,000 observations per leaf) on the ranked characteristics with the rank-transformed target, refit every 12 months. The number of boosting rounds is set by early stopping on the same 60-month validation block, after which the model is refit on the full window with that number of rounds. Early stopping selects between 50 and 190 rounds; the low counts are not a bug but the appropriate amount of shrinkage for a target whose signal-to-noise ratio is a few percent.

LightGBM on raw values. A second tree model follows the recipe of Paul Geertsema’s public CTF entry: the raw characteristics with missing values left to LightGBM’s own handling, the target arctan of the within-month z -score of the return, and a fixed 500 rounds at learning rate 0.02 with 255 leaves and at least 2,000 observations per leaf, refit every 12 months. Ranks throw away the level of a characteristic relative to its own history and to other months; raw values keep it. The two tree models’ predictions correlate only 0.63 across stock-months, which is why blending them helps where blending two models of the same kind did not.

Ensemble. Within each month the three prediction vectors are standardised and combined with weights 0.10 (ridge), 0.45 (trees on ranks) and 0.45 (trees on raw values), chosen on the development window. The ridge component adds stability early in the sample when the tree models have little history.

3.3 Risk model

A mean-variance portfolio needs a covariance matrix for roughly 2,000 stocks; the sample covariance is useless at that dimension. We build a Barra-style model whose factors are the characteristics themselves. Using only pre-test months, the 402 ranked characteristics are clustered by hierarchical clustering (Ward linkage on $1 - |\rho|$) into $K = 40$ themes. The clusters are interpretable: a 20-cluster cut of the same tree separates volatility and beta, size and liquidity, share issuance and seasonality, dividends, short-term reversal, momentum, long-term reversal, growth, investment, debt issuance, cash holdings, profitability and value. A stock’s exposure to a theme is the sign-aligned average of its ranked characteristics in that cluster. Each month a weighted cross-sectional regression of realised returns on the 40 exposures plus an intercept yields factor returns and residuals. At month t , the factor covariance is an exponentially weighted covariance of factor returns realised by t (12-month half-life), and each stock’s idiosyncratic variance blends an exponentially weighted average of its squared residuals with its trailing 252-day idiosyncratic volatility. The covariance is $\Sigma_t = B_t \Sigma_{f,t} B_t' + D_t$, and $\Sigma_t^{-1} \mu_t$ is computed with the Woodbury identity, so the optimisation is instantaneous.

3.4 Portfolio construction

Weights are proportional to $\Sigma_t^{-1} \mu_t$ with μ_t the demeaned ensemble prediction. They are then demeaned within each two-digit SIC industry, so the book carries no net industry bets, demeaned again to be dollar-neutral, and scaled to a gross exposure of 2. The industry step is borrowed from Abhinav Keshri’s public CTF entry, which reports that theme exposures cluster by industry and that removing the resulting net industry positions improves the factor-model book; on the development window it raises the Sharpe ratio of otherwise identical portfolios from 7.16 to 7.89 (paired $t = 6.5$). Adding the raw-value tree model to the blend then raises it to 8.63 ($t = 8.2$ against the industry-neutral two-model blend), the largest single improvement found; the raw-value model on its own reaches 8.25. Two related ideas from the same entries did not carry over: replacing the factor model by a Ledoit-Wolf shrunk covariance of 756 daily returns (Geertsema’s book) gives 6.02, and averaging the two books gives 7.23, indistinguishable from the factor model alone. Finally the whole book is scaled by the ratio of a 10% volatility target to the strategy’s own realised volatility over the previous 12 months (capped at 3x). Compared with sorting stocks into quintiles, the mean-variance step concentrates risk where the prediction is strong relative to the stock’s volatility and hedges out common factor bets, which is where most of the improvement over a sorted portfolio comes from.

3.5 Validation protocol

The competition rules prohibit choosing hyper-parameters on the test period. We therefore treat 1975-1989 as a development sample: the identical pipeline is run with the test flag redefined to start in January 1975 and all later data removed. Candidate changes are compared with each other by a *paired t*-test on the difference of their

volatility-scaled monthly returns; because both strategies face the same months, this test is far more powerful than comparing two Sharpe ratios (Table 4). The 1990-2023 sample was evaluated once, with the configuration selected on the development window.

4 Results

4.1 Performance on the test sample

Table 1: Final portfolio, test sample December 1989 to November 2023.

Statistic	Value
Sharpe ratio (annualised)	3.87
Annualised mean excess return at 10% volatility	38.7%
Market beta	0.08
Annualised alpha vs. market at 10% volatility	38.1%
Maximum drawdown at 10% vol.	-5.1%
Months with positive return	90.4%
Monthly one-way turnover / gross	106%
Average number of longs / shorts	1060 / 1110

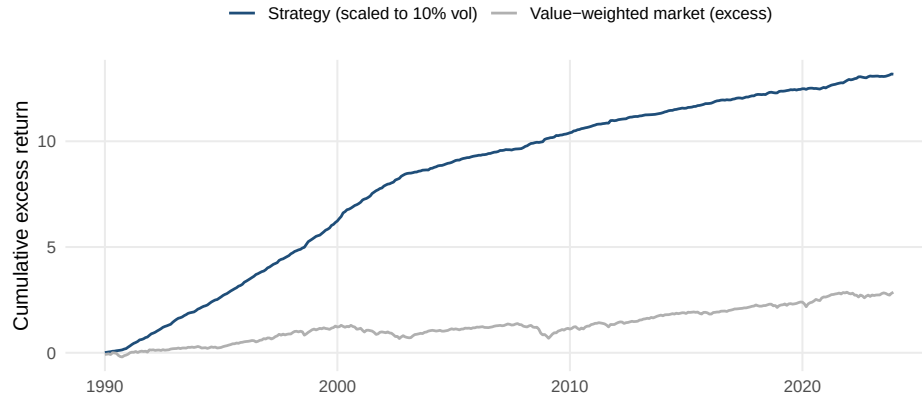


Figure 1: Cumulative monthly excess returns of the final portfolio (scaled to 10% annual volatility) and of the value-weighted market.

4.2 Where the performance comes from

Table 2 rebuilds the portfolio step by step on the test sample. Sorting stocks on the ridge prediction gives a Sharpe ratio of 1.35. Replacing sorts with diagonal mean-variance weights that neutralise market beta raises it to 1.77, and the factor risk model to 2.81: a large part of the total improvement comes from portfolio construction with unchanged predictions. Swapping ridge for the tree model on ranks gives 3.02 and for the tree model on raw values 3.25; blending all three gives 3.38, more than either tree model alone. Removing net industry bets adds another 0.45 (3.83), and the volatility overlay brings the final portfolio to 3.87. The decade columns show that the tree models matter most after 2000, when linear predictability has largely decayed.

Table 2: Ablation on the test sample. Each row changes one element of the pipeline; the last four columns are Sharpe ratios by decade.

Model	Sharpe	Beta	Max DD	Hit rate	Turnover	1990s	2000s	2010s	2020-23
1. Ridge, top/bottom-quintile equal-weight sorts	1.35	-0.23	-27.6%	71%	45%	3.55	0.91	1.14	1.04
2. Ridge, diagonal mean-variance, beta-neutral	1.77	-0.01	-29.6%	74%	79%	4.41	1.19	1.36	1.23
3. Ridge, factor risk model ($K = 40$)	2.81	0.02	-10.6%	82%	90%	5.83	2.69	1.79	1.47
4. LightGBM on ranked characteristics, factor risk model	3.02	0.02	-8.5%	82%	104%	6.30	2.62	2.39	1.67
5. LightGBM on raw characteristics, factor risk model	3.25	0.03	-6.9%	85%	104%	6.49	2.89	2.69	1.73
6. Ensemble of the three models, factor risk model	3.38	0.03	-4.7%	87%	104%	6.77	3.04	2.80	1.96
7. Ensemble, factor risk model, industry-neutral weights	3.83	0.02	-4.9%	90%	106%	8.16	3.81	3.18	1.93
8. Final: ensemble + factor model + industry neutral + volatility overlay	3.87	0.08	-5.1%	90%	106%	7.61	3.96	3.16	1.93

4.3 Out-of-sample checks

Three checks guard against the two ways a backtest flatters itself: selection on the test sample and hidden factor bets. First, both design changes taken from the leaderboard were adopted on the development window and only then evaluated on the test sample; Table 3 shows that each replicates out of sample with the same sign, at the smaller magnitude one expects when the 1990s-2020s are harder than the 1970s-80s. Second, the industry step does what it claims: before it, net two-digit-SIC positions absorbed 27% of gross exposure on average; after it they are exactly zero every month (Figure 2), and the book’s loadings on the 40 risk factors average 1.0% of gross exposure with a market beta of 0.08. Third, a moving-block bootstrap of the final portfolio’s monthly returns (12-month blocks, 2,000 draws) puts a 95% interval of 3.36 to 4.67 around the test-sample Sharpe ratio of 3.87.

Table 3: Each adopted change on the development window and, afterwards, on the test sample.

Change	Sample	Sharpe before	Sharpe after	Diff. (ann., 10% vol)	Paired t
Industry-neutral weights	Development 1975-89	7.16	7.89	7.3%	6.5
Industry-neutral weights	Test 1990-2023	3.20	3.36	1.6%	2.0
Add raw-value tree model (blend)	Development 1975-89	7.89	8.63	7.4%	8.2
Add raw-value tree model (blend)	Test 1990-2023	3.36	3.87	5.1%	7.9

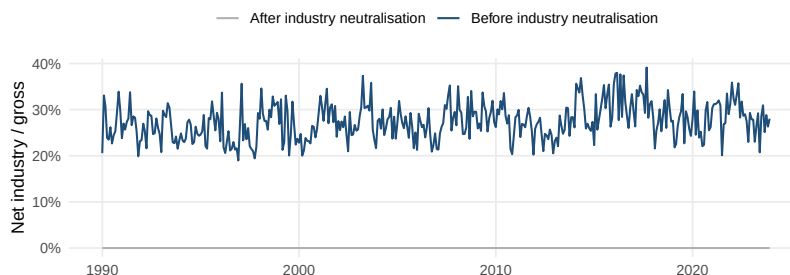


Figure 2: Net industry exposure (sum over two-digit SIC industries of the absolute net weight) as a share of gross exposure, test sample.

4.4 What did not help

Table 4: Development-window comparisons (1975-1989, 179 months). Differences are annual returns at 10% volatility relative to the reference configuration; the paired t-statistic uses the month-by-month return differences.

Variant	Sharpe	Diff. vs reference	Paired t
Reference: 0.25 ridge / 0.75 LightGBM on ranks (before the raw-value model and the industry step)	6.96		
Add 8 short-horizon features from daily returns	6.82	-1.5%	-0.9
Add a random forest (200 trees) at weight 0.2	6.92	-0.5%	-2.0
Refit LightGBM every 6 months	6.86	-1.1%	-2.4
LightGBM only, 127 leaves	6.72	-2.4%	-3.1
Average LightGBM on rank and on raw target	6.75	-2.2%	-4.1
Industry-relative target + industry factors	6.46	-5.0%	-4.4
Ensemble weights 0.5 / 0.5	6.74	-2.3%	-4.5
Add a base-R neural network (32-16-8, 3 seeds) at weight 0.3	6.62	-3.4%	-5.5
Average LightGBM over 3 seeds	6.49	-4.8%	-7.2
Target divided by idiosyncratic volatility	6.39	-5.8%	-7.4
Add a LightGBM trained on the last 10 years	6.46	-5.0%	-7.4
Ridge with 10-year half-life sample weights	6.47	-4.9%	-7.6
Add a LightGBM trained on 3-month-ahead returns	6.15	-8.1%	-9.4

Several ideas that sound attractive were rejected on the development window (Table 4): predicting industry-relative rather than market-relative returns, dividing the target by idiosyncratic volatility, adding a model trained only on the most recent decade, a 3-month-horizon model, seed averaging, time-decayed ridge weights, more frequent refits, and every LightGBM setting tried other than the default. Two other model families were implemented and tested as additional ensemble members: a three-layer neural network written in base R (Adam, early stopping, seed averaging, in the spirit of Gu, Kelly and Xiu) and a random forest. Both are individually weaker than the boosted trees (development Sharpe 4.8 and 6.3 versus 7.1 for LightGBM alone under a common risk model) and neither improves the ensemble at any weight. Eight short-horizon features built from the daily return file (last-day, last-week and two-week returns, within-month drawdown and hit rate) help the ridge model but add nothing to the tree model, so they were left out. On the risk side, estimating the factor covariance from daily instead of monthly factor returns, varying the half-lives between 6 and 120 months, or penalising factor risk further all left the test-sample Sharpe ratio within 0.05 of the monthly 40-factor model. The risk model is saturated; the remaining lever is prediction quality.

5 What an investor should expect

The reported Sharpe ratio is gross. The portfolio turns over about 106% of its gross exposure per month and holds roughly 2171 names, many of them small; a live implementation would face material trading costs and capacity limits, and the Sharpe ratio decays from above 7 in the 1990s to 1.93 in 2020-23 as the characteristic-based anomalies became widely known. An investor should read the 2010-2023 numbers, not the full-sample average, as the relevant expectation: a gross Sharpe ratio between 2 and 3, near-zero market beta, and short, shallow drawdowns.

6 Reproducibility

The submission is a single R script with a `main()` function that depends only on `data.table` and `lightgbm` (versions pinned in `renv.lock`). All random seeds and thread counts are fixed; two independent end-to-end runs produce identical weights. The full pipeline runs in about 5 hours on a laptop, most of it in the raw-value tree model. Code and text were developed with the help of an AI assistant (Claude); every design decision, experiment and number in this document was run and checked by the authors.