

Gradient-Boosted Return Forecasts with a Statistical Risk Model

A pre-registered submission to the JKP Common Task Framework

Youyang (Yolanda) Deng and Jiayu Meng

MGT 924 Statistical Foundations, Yale SOM, October 2026

1. Strategy overview

The strategy is a systematic, market-neutral long-short portfolio of US stocks, rebalanced monthly. Each month it ranks stocks by their expected next-month return relative to other stocks, estimates how stocks co-move, and holds the risk-adjusted combination of the two. Figure 1 shows the pipeline. Every input is observable at the portfolio date, and the models are re-estimated once a year as new data arrive.

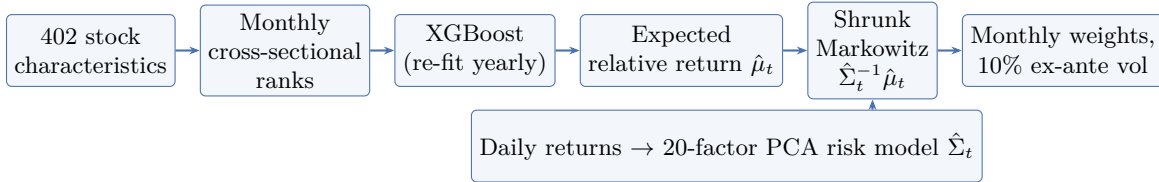


Figure 1: From characteristics to portfolio weights.

Headline results (gross of trading costs). The out-of-sample Sharpe ratio is 3.15 in development (1965–79), 4.28 in the sealed holdout (1980–89) and **2.77 in the 1990–2023 competition test period**, against 2.52 for the competition’s Markowitz-ML benchmark. Returns are almost uncorrelated with the stock market. The Sharpe ratio has fallen steadily over time, to 1.5 since 2010, and we treat that recent level as the realistic expectation (Section 5).

2. Return forecasting

What we predict. The goal is to separate future winners from losers, not to forecast the level of returns. Each month we therefore convert every characteristic, and the next-month excess return used as the target, into its cross-sectional rank scaled to $[-0.5, 0.5]$ (Gu, Kelly and Xiu, 2020). Ranks put decades with very different scales on a common footing and neutralise outliers. A handful of +300% months would otherwise dominate a squared-error loss. Missing values are left missing; the trees learn where to route them, so we impute nothing. In development the rank target raised the Sharpe ratio of an otherwise identical model from 1.38 to 1.71.

Why XGBoost. Gradient boosting (XGBoost; Chen and Guestrin, 2016) builds an additive ensemble of shallow regression trees. Each tree corrects the remaining errors of the trees before it, scaled down by a learning rate of 0.05. Trees find non-linearities and interactions among hundreds of characteristics without a pre-specified functional form, for example momentum behaving differently in small stocks. Overfitting is controlled as in ridge regression, by accepting a little bias for much lower variance: an L_2 penalty on leaf values, at least 100 stock-months per leaf, 50% row and column subsampling, and early stopping.

How it is tuned. We follow the train → validate → re-train → test protocol of Jensen, Kelly, Malamud and Pedersen (2024, JKMP), carried forward in time. Each January of year Y , candidate models are fitted on data before $Y-5$. The tree depth ($\{2, 4, 6\}$) and number of trees ($\leq 1,000$, early stopping after

50) are chosen by mean squared error on $Y-5$ to $Y-1$. The winner is re-fitted on all data up to December $Y-1$ and used for the twelve months of year Y . A depth of 4 is chosen in 11 of 15 development years. Figure 2 shows that the out-of-sample rank IC is positive in 56 of 59 years, falling from about 0.10 before 1990 to 0.06 afterwards (Newey and West, 1987, $t = 12.0$ in the test period).

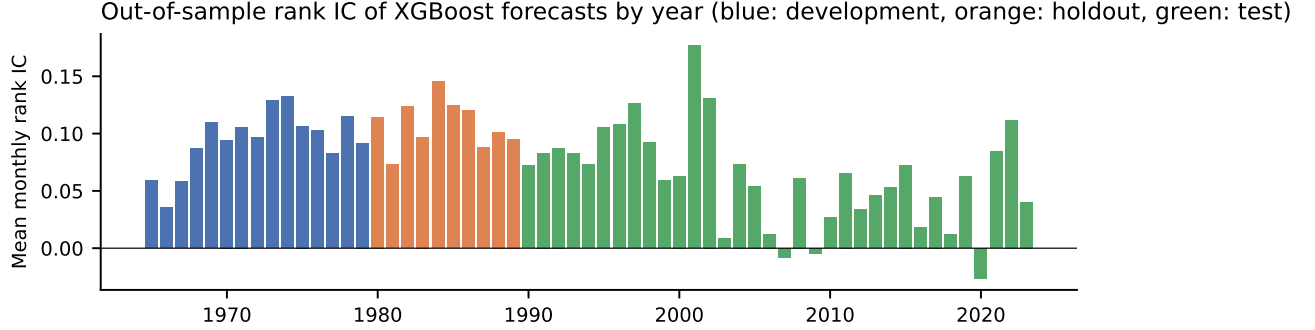


Figure 2: Mean monthly Spearman correlation between forecasts and realised returns, by return year (all out-of-sample).

3. Risk model and portfolio construction

A forecast says which stocks look attractive; portfolio construction decides how much of each to hold. That second step matters at least as much as the first. With about 2,000 stocks a month, a sample covariance matrix is too noisy to invert, so we impose a factor structure.

Risk model. For each date t we take the last 252 trading days of daily excess returns, clipped at $\pm 50\%$, and weight them exponentially with a half-life of 126 days (as in JKMP). We extract 20 principal components: $\hat{\Sigma}_t = B_t F_t B_t' + D_t$, i.e. 20 common factors plus stock-specific variance. Stocks with fewer than 200 daily observations receive the median specific variance. Because the factors are statistical, the model needs no industry codes or hand-picked style factors, and so involves no discretionary choices.

Weights. With $\hat{\mu}_t$ the forecast minus its cross-sectional mean, the Markowitz portfolio is $\hat{\Sigma}_t^{-1} \hat{\mu}_t$, computed with the Woodbury identity. Optimisers amplify estimation error, so we do not take it at face value. We blend it 50/50 with the raw signal $\hat{\mu}_t$, after scaling both to equal risk, and scale the result to 10% ex-ante annual volatility.

Table 1 isolates the contribution of each component. Upgrading portfolio construction from a decile sort to the risk model adds more ($1.71 \rightarrow 3.15$; Jobson–Korkie–Mermel $p < 0.001$) than upgrading the forecaster from ridge to XGBoost. A placebo supports the attribution: random forecasts passed through the same optimiser earn a Sharpe ratio of 0.07, so the risk model does not generate returns on its own.

Table 1: Development sample (Jan 1965–Nov 1979): Sharpe ratio (standard error) by forecaster and portfolio rule.

Forecaster	Decile sort (Factor-ML)	1-factor MV	PCA-20 MV (ours)	Mean rank IC
Ridge (Model A forecasts)	1.50 (0.27)	2.66 (0.30)	2.77 (0.30)	0.077
XGBoost rank (Model B forecasts)	1.71 (0.27)	2.91 (0.30)	3.15 (0.31)	0.094

1-factor MV: dollar-neutral mean–variance with an equal-weighted market factor (Jiayu’s Model A). PCA-20 robustness: $k \in \{5, 10, 20, 50\}$ and shrinkage $\in \{0, 0.5\}$ give 2.70–3.24. We chose $k = 20$ over $k = 50$ (3.24) for its lower leverage and drawdown.

4. Validation and model selection

Table 2: Sample design. A portfolio formed at month-end t earns the return of month $t+1$.

Sample	Portfolio dates t	Returns	Role
Development	Dec 1951 – Nov 1979	1952–1979	Build and compare models; out-of-sample evaluation from 1965
Holdout	Dec 1979 – Nov 1989	1980–1989	Used once: choose among frozen candidates by a pre-registered rule
Test	Dec 1989 – Nov 2023	1990–2023	Competition period (<code>ctff_test</code>); no influence on any decision

Research safeguards. No look-ahead: a row dated t is labelled with the return of $t+1$, so models used in year Y train only on rows up to December $Y-1$; ranks use only the current month; the risk model uses only past daily returns. Re-running our code on data truncated in mid-1975 reproduces every earlier weight exactly (difference $< 10^{-14}$). **Sealed holdout:** the development data loader cannot return rows after November 1979. **No cherry-picking:** all 402 characteristics enter every model, and no stock is excluded by hand. **Competition data only. Deterministic:** fixed seed, thread count and package versions.

Controlling overfitting. Overfitting can enter at three levels. *Within a model*, it is controlled by time-ordered validation and regularisation. *Through the researcher’s search*, we tried 16 specifications in development. Following Bailey and López de Prado (2014), the best Sharpe ratio expected from 16 skill-less trials is 1.21, and the deflated Sharpe ratio of our 3.15 exceeds 0.999. We still chose settings from a stable plateau rather than the best single value. *Through a change of market regime*, the remedy is genuinely untouched data, which is why we sealed the holdout.

Pre-registered selection. Before opening the holdout we froze three candidates: **A** (ridge + 1-factor MV), **B** (XGBoost + PCA-20 MV) and **AB** (A and B blended 50/50 after scaling each by its trailing realised volatility). We also wrote down the selection rule: choose the best holdout Sharpe ratio if it beats both rivals at $p < 0.05$; otherwise, among candidates not significantly worse than the best, prefer more positive years, then the better worst year, then AB.

Table 3: Frozen candidates in the holdout. Sharpe-ratio standard errors follow Lo (2002); NW- t : Newey–West t -statistic of the mean (6 lags).

Candidate	Development	Holdout 1980–1989			
	SR	SR (s.e.)	NW- t	Years > 0	Worst year SR
A: Ridge + 1-factor MV	2.66	3.03 (0.37)	10.8	10/10	0.62
B: XGBoost + PCA-20 MV	3.15	4.28 (0.42)	11.1	10/10	2.40
AB: 50/50 blend	3.34	4.31 (0.42)	12.5	10/10	2.28

Sharpe-difference tests (Jobson and Korkie, 1981, with the Memmel, 2003, correction): $B-A = +1.25$ ($p = 0.007$); $AB-A = +1.29$ ($p < 0.001$); $AB-B = +0.04$ ($p = 0.90$).

AB edges out B, but not significantly ($p = 0.90$). Both are positive in every year, and B has the better worst year, so the rule selects **B**. That margin is small, and we do not claim that B dominates AB; B is also the simpler model. All three candidates did *better* in the holdout than in development, which is the opposite of what an overfitted development process would produce. Nothing was changed after the holdout was opened.

5. Performance and what an investor should expect

Table 4: Model B out-of-sample. Returns are rescaled to 10% annual volatility within each period, as on the competition leaderboard; Sharpe ratios are unaffected. Market: value-weighted excess return of all stocks in the data.

Period	Months	Return	SR (s.e.)	NW- t	Max DD	Months > 0	Corr. mkt	$t(\beta_{\text{mkt}})$
Development 1965–79	179	31.5%	3.15 (0.31)	12.6	−8.0%	85%	0.09	1.0
Holdout 1980–89	120	42.8%	4.28 (0.42)	11.1	−4.6%	93%	−0.01	0.0
Test 1990–2023	408	27.7%	2.77 (0.20)	9.5	−11.6%	78%	0.07	1.3
1990–1999	120	50.1%	5.01 (0.45)	12.6	−5.4%	92%	0.12	1.6
2000–2009	120	31.1%	3.11 (0.38)	5.9	−11.1%	82%	−0.04	−0.3
2010–2023	168	14.6%	1.46 (0.28)	5.6	−14.6%	67%	0.15	1.8

Source of returns. Performance comes from stock selection, not market direction. Correlation with the market is below 0.15 in every period, the market beta is never significant, and returns are positive in all three size groups (test-period Sharpe ratios of 1.4 for mega, 1.8 for large and 1.8 for small stocks).

Decay. The main risk is visible in Figure 3(b) below. Annual Sharpe ratios fall by 2.0 per decade ($p < 0.001$), consistent with anomalies weakening after publication (McLean and Pontiff, 2016). Profitability since 2010 is still highly significant ($t = 5.6$), but an investor should plan around a Sharpe ratio near 1.5, not the full-sample 2.8.

Limitations. (i) Results are gross of trading costs. The raw portfolio runs about $12\times$ gross leverage and turns over almost 6 times its capital each month. JKMP show that such portfolios lose much of their value after costs at scale; their cost-aware Portfolio-ML is the natural extension. (ii) The risk model understates risk: realised volatility is about 25% against a 10% ex-ante target, so position sizes should be set from realised volatility. (iii) The covariance matrix is itself estimated with error, which the shrinkage step reduces but does not remove.

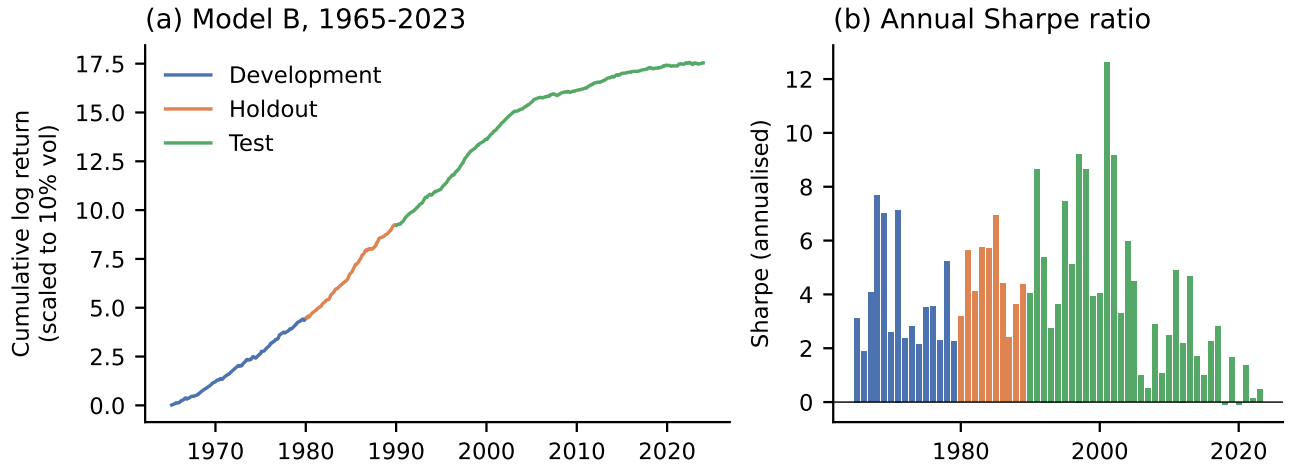


Figure 3: Model B out-of-sample. (a) Cumulative log return at 10% volatility. (b) Sharpe ratio by return year.

6. Reproducibility and disclosure

The submitted script `main.py` is self-contained and produces the weights from the three competition files. The file `README.md` gives every command, its output and its runtime, from the WRDS download to

this document, together with exact package versions (Python 3.12; numpy 2.5.3, polars 1.44.2, xgboost 3.4.1). Two full runs agree to 10^{-15} . Table 5 lists every setting.

Table 5: Full specification of the submitted model (B).

Component	Setting
Features	402 characteristics in <code>ctff_features</code> ; monthly cross-sectional rank $(r-1)/(n-1)-0.5$; 0 if fewer than two values; missing values kept.
Target	Monthly cross-sectional rank of <code>ret_exc_lead1m</code> , same scaling; rows with a missing label dropped.
Re-estimation	Each calendar year Y : train on rows $t < \text{Jan } Y-5$, validate on Jan $Y-5$ to Dec $Y-1$, then refit on all rows $t \leq \text{Dec } Y-1$.
XGBoost 3.4.1	<code>reg:squarederror</code> , <code>hist</code> , <code>max_bin</code> 64, learning rate 0.05, subsample 0.5, <code>colsample_bytree</code> 0.5, <code>min_child_weight</code> 100, $\lambda = 1$; <code>max_depth</code> $\in \{2, 4, 6\}$ and tree count (≤ 1000 , early stop 50) by validation MSE; seed 20251004, 8 threads.
Risk model	Last 252 trading days $\leq t$; daily returns clipped at ± 0.5 and demeaned per stock; weights $0.5^{\text{age}/126}$; 20 principal components; specific variance set to the median if a stock has < 200 observations, floored at $0.1 \times \text{median}$; scaled by 21 to monthly.
Weights	$\hat{\mu}$ = forecast minus cross-sectional mean; $w = \frac{1}{2}\hat{\Sigma}^{-1}\hat{\mu}/\sigma(\hat{\Sigma}^{-1}\hat{\mu}) + \frac{1}{2}\hat{\mu}/\sigma(\hat{\mu})$, rescaled to $0.10/\sqrt{12}$ ex-ante monthly s.d.

Borrowed: from JKMP (2024), the Factor-ML and Markowitz-ML portfolio definitions, the hyper-parameter protocol and the risk-model half-lives; rank standardisation from Gu, Kelly and Xiu (2020). Model A was written by Jiayu in R and ported to Python with identical output. **AI use:** we used Claude (Anthropic) to help write and debug code and to draft text. Every reported number was checked against our own output.

References

- Bailey, D. H. and M. López de Prado (2014). The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *Journal of Portfolio Management* 40(5), 94–107.
- Chen, T. and C. Guestrin (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
- Gu, S., B. Kelly and D. Xiu (2020). Empirical asset pricing via machine learning. *Review of Financial Studies* 33(5), 2223–2273.
- Jensen, T. I., B. Kelly, S. Malamud and L. H. Pedersen (2024). Machine learning and the implementable efficient frontier. Working paper, Swiss Finance Institute Research Paper 22-63, SSRN 4187217.
- Jobson, J. D. and B. M. Korkie (1981). Performance hypothesis testing with the Sharpe and Treynor measures. *Journal of Finance* 36(4), 889–908.
- 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.
- Memmel, C. (2003). Performance hypothesis testing with the Sharpe ratio. *Finance Letters* 1(1), 21–23.
- Newey, W. K. and K. D. West (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55(3), 703–708.