

Peer-Relative Boosted Forecasts in a 414-Factor Risk Model

Andrew Qi, Marcus Kwauk, Boyce Yang, Harley Chen

Yale SOM, MGT 924

October 2026

In one paragraph. Each month we group stocks into about 20 peer groups of look-alike firms, describe every stock by its 402 characteristics both in absolute terms and *relative to its peers*, forecast next-month returns with gradient-boosted trees (LightGBM), and hold the dollar-neutral mean-variance portfolio implied by those forecasts and a Barra-style risk model with 414 factors. Every design choice was made on 1970–1989 under rules written down before each run. **Test period 1990–2023: Sharpe ratio 4.03** (s.e. 0.22), 40% annual return at 10% volatility, maximum drawdown -6.2% , market beta ≈ 0 ; that would rank second of the 54 entries on the public leaderboard as of 2 October 2026. Performance has decayed by decade (8.8 in the 1990s, 2.0 in 2020–23); an investor should plan on the recent level.

1. Idea and economic rationale

Investors and analysts value a firm against comparable firms, not against the whole market: “cheap” means cheap relative to similar businesses (Bhojraj and Lee, 2002; Hoberg and Phillips, 2016). A book-to-market ratio of 0.8 says something different about a bank than about a software firm. We therefore give the forecasting model, next to each raw characteristic, a *peer-relative* version: the stock’s within-month rank minus the average rank of its peer group. Peers are found by the data each month, from similarity in all 402 characteristics, so no industry code or hand-made grouping is needed. Two further choices complete the design. Gradient-boosted trees capture the nonlinear interactions that a linear model misses (Gu, Kelly and Xiu, 2020). And a factor risk model turns forecasts into weights that hedge common industry and style risk (Jensen, Kelly, Malamud and Pedersen, 2024); in our tests it mattered as much as the forecast itself (Section 3).

2. The model

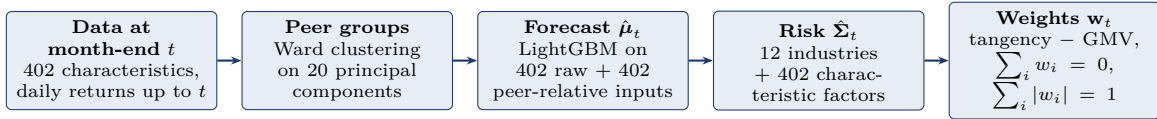


Figure 1: Monthly pipeline. Every input is observable at the formation date t .

Peer groups and features. Each month, every characteristic is ranked across stocks, $\tilde{x}_{ij,t} \in [-0.5, 0.5]$ (exact zeros at the bottom, missing at the median). Stocks are clustered by Ward’s (1963) method on the first 20 principal components of these ranks into $K_t = \text{round}(n_t/100)$ groups, clipped to $[10, 60]$ (median group size 93). The peer-relative characteristic is $\tilde{x}_{ij,t} - \bar{\tilde{x}}_{g(i)j,t}$, the rank minus the peer-group mean, set to missing where the raw value is missing. Group size and the number of components were fixed in advance and never tuned. Only month- t characteristics are used, no returns and no other months.

Forecast. LightGBM (Ke et al., 2017) on the 804 inputs (402 raw characteristics and 402 peer-relative ones). The target is $\arctan(z_{i,t+1})$, the next-month excess return standardised within the month. 1,000 trees, learning rate 0.01, 255 leaves, at least 2,000 stock-months per leaf, deterministic; refit every January on every return realised by the previous 31 December (expanding window from 1952). The tree settings are borrowed (Section 8).

Risk model (Barra-style; Menchero, Orr and Wang, 2011). Each trading day of month $t + 1$, a cross-sectional ridge regression (10^{-4}) of stock returns on 12 Fama–French industry dummies and the 402 standardised ranked characteristics known at t gives factor returns and residuals. At date t , using days up to t only, the factor covariance combines an exponentially weighted correlation (half-life 504 days over 2,520 days) with exponentially weighted volatilities (half-life 84); each stock’s specific variance is an exponentially weighted mean (half-life 84) of its last 252 squared residuals, with at least 200 observations: $\hat{\Sigma} = \mathbf{X}_t \hat{\Sigma}_f \mathbf{X}_t' + \text{diag}(\sigma_{e,i}^2)$.

Portfolio. With $\mathbf{A} = \hat{\Sigma} + \kappa \frac{\text{tr} \hat{\Sigma}}{n} \mathbf{I}$ ($\kappa = 1$), $\mathbf{a} = \mathbf{A}^{-1} \hat{\boldsymbol{\mu}}$, $\mathbf{b} = \mathbf{A}^{-1} \mathbf{1}$: $\mathbf{w}_t \propto \mathbf{a} - \frac{\mathbf{1}' \mathbf{a}}{\mathbf{1}' \mathbf{b}} \mathbf{b}$, scaled to $\sum_i |w_{i,t}| = 1$, the maximum-Sharpe portfolio among zero-cost portfolios; κ guards against estimation error (DeMiguel, Garlappi and Uppal, 2009). Stocks without a specific variance get zero weight; weights are rebalanced monthly.

3. How we arrived at the model

Protocol. All choices were made on 1970–1989 (240 months). Before each comparison we wrote down the decision rule (in the final round, in a version-controlled log, committed before each run): *a change is adopted only if its 1970–1989 Sharpe ratio rises by more than one standard error*. The 1990–2023 scoring period was used for reporting.

Step 1: choosing the base model. We first searched for the best combination of a return forecast and a portfolio

construction, holding everything else fixed (Table 1). Two lessons stand out. How forecasts become weights matters as much as the forecasts: with the same linear (ridge) forecast, the Sharpe ratio rises from 2.86 in a decile sort to 4.75 with a statistical covariance and 5.51 with the 414-factor Barra model, while a coarser 20-factor Barra model and the MAXSER optimiser of the leaderboard’s top entry do worse. And boosted trees forecast better than linear models: LightGBM with the Barra mean-variance portfolio reaches 7.42. Averaging ridge and LightGBM forecasts adds nothing (7.38), so the rule picks the simpler LightGBM + Barra combination as the base model.

Table 1: Step 1, development window 1970–1989 (240 months): forecast and portfolio combinations. Sharpe ratios annualised, s.e. from Lo (2002); maximum drawdown at 10% volatility. MV = mean-variance (tangency – GMV).

Forecast	Portfolio construction	Sharpe	s.e.	Max DD
—	Equal-weight market (long only)	0.35	0.22	–33%
Ridge	Decile long-short sort	2.86	0.26	–9%
Complex Factors (random features)	Own maximum-Sharpe weights	4.13	0.29	–11%
Ridge	MAXSER optimiser	3.96	0.29	–6%
Ridge	MV, Barra with 20 clustered factors	4.06	0.29	–8%
Ridge	MV, Ledoit–Wolf covariance	4.75	0.31	–8%
Lasso / elastic net	MV, Ledoit–Wolf covariance	4.90–4.94	0.32	–6%
Ridge	MV, Barra with 414 factors	5.51	0.34	–7%
$\frac{1}{2}$ Ridge + $\frac{1}{2}$ LightGBM	MV, Barra with 414 factors	7.38	0.40	–3%
LightGBM	MV, Barra with 414 factors (base model)	7.42	0.41	–1%

Step 2: improving the forecast with peers. Keeping the base model’s risk model and portfolio unchanged, we gave LightGBM information about each stock’s peers. The 402 peer-relative features raise the Sharpe ratio from 7.42 to 9.04 (+1.61, paired $t = 4.3$; Figure 2).

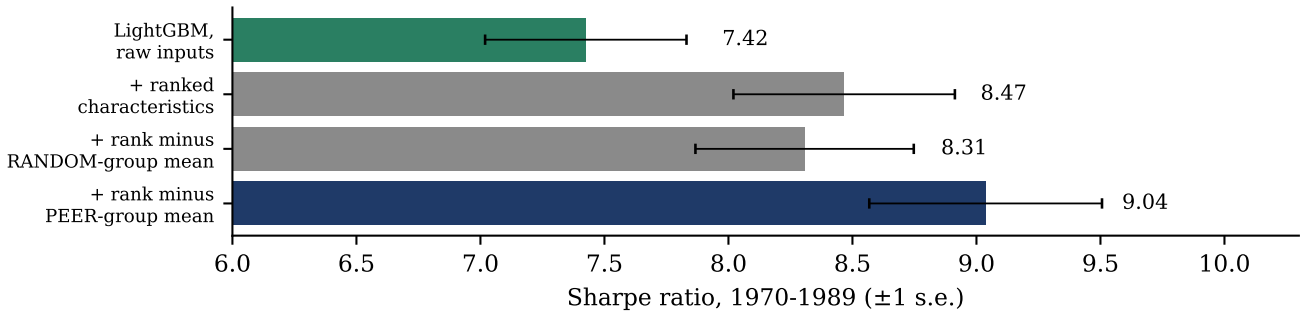


Figure 2: Where the gain from peer-relative features comes from (1970–1989, same LightGBM, risk model and portfolio).

Is it really the peers? A peer-relative feature does two things at once: it ranks the characteristic, and it subtracts the peer mean. We separated the two with a placebo, run under a rule fixed in advance: the same features but with each month’s group labels shuffled across stocks (same group sizes, random members). The placebo earned 8.31, also well above the raw-input model, so, before looking at any test results, we designed and pre-registered one further comparison: the ranked characteristics alone (8.47). Ranking adds about +1.0; subtracting a *random* group’s mean adds nothing (–0.16); subtracting the *real* peer group’s mean adds a further +0.57 over ranking alone and +0.73 over random groups (paired $t = 2.2$). Both steps clear the one-standard-error rule, so the peer-relative model was adopted. Another feature set, scores for 20 clusters of characteristics, earned 7.77 and was not adopted.

Checks. Shuffling the LightGBM forecasts across stocks within each month leaves no performance (Sharpe ratio 0.18), so nothing but the forecasts drives returns. The submitted code runs end to end on a 123-month sample with valid output, the uploaded weights reproduce the research results exactly (Sharpe ratios 9.04 and 4.03), and every portfolio is built only from data available at its formation date. The code is deterministic: fixed seeds, LightGBM’s deterministic mode with a fixed thread count, and deterministic Ward clustering. Under the deflated Sharpe ratio (Bailey and López de Prado, 2014), the best Sharpe expected from the 57 configurations we evaluated on 1970–1989, if all were worthless, is 3.1; the probability that the final model’s true Sharpe exceeds it is above 0.999 in both periods.

4. Test-period results, 1990–2023

Table 2: Test period (408 months). Sharpe ratios annualised; return and drawdown at 10% volatility; rank IC = mean monthly rank correlation of forecast and next-month return.

Strategy	Sharpe	s.e.	Return	Max DD	Turnover	Rank IC
Equal-weight market (long only)	0.49	0.17	4.9%	−34%	—	—
LightGBM on raw characteristics + Barra	3.75	0.22	37.5%	−7.8%	1.05	0.069
Peer-relative LightGBM + Barra (submitted)	4.03	0.22	40.3%	−6.2%	1.12	0.075

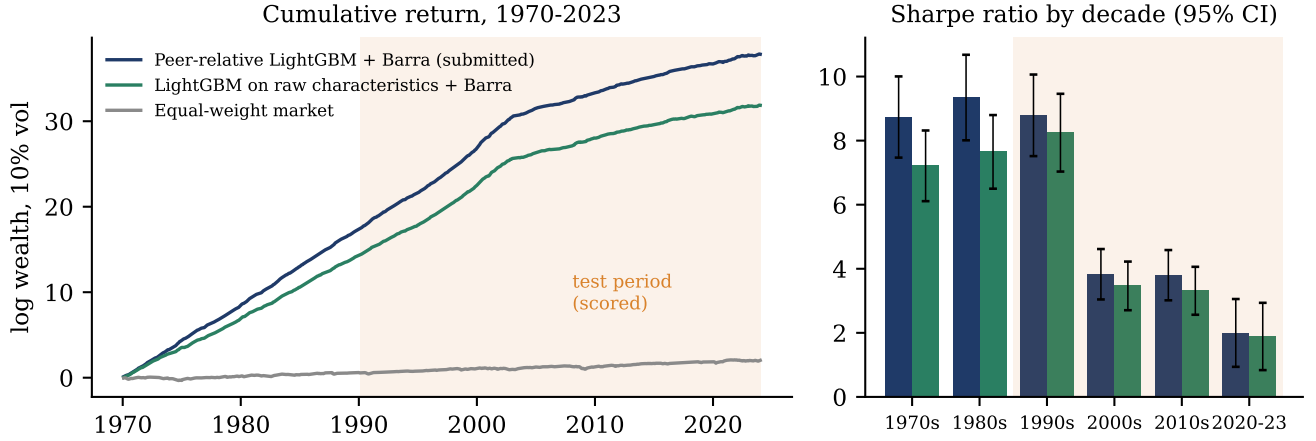


Figure 3: Left: log wealth at 10% volatility (scaled with 1970–89 volatility). Right: Sharpe ratio by decade with 95% confidence intervals; blue = submitted model, green = same model without peer-relative inputs; shaded = test period.

- **The peer gain holds out of sample, but smaller.** It is +0.28 (paired $t = 2.5$) on 1990–2023, positive in every decade, against +1.6 on 1970–1989. Plausible reasons: ranking matters most when raw values are least comparable over time, as in the early sample, and peer-relative signals became more widely traded later.
- **Market neutral, and a hedge in crashes.** Beta on the equal-weight market is -0.008 ($t = -1.4$). At 10% volatility the strategy gained during the 2000–02 and 2007–09 crashes (the market fell 37% and 55%) and in the first quarter of 2020 (market -31%). Its worst month (October 2020, -4.7%) ended its deepest drawdown (-6.2% , June to October 2020), during the sharp rotation after the COVID crash.
- **Decay.** The Sharpe ratio falls from 8.8 in the 1990s to 3.8, 3.8 and 2.0, as published anomalies weaken over time.

5. What an investor should expect

- **Expected Sharpe.** Our planning number is 2 before costs, not 4.03: the 2010s delivered 3.8 and 2020–23 delivered 2.0, and the trend is down.
- **Costs and capacity.** The CTF score ignores trading costs. Turnover is about 112% of gross exposure per month across about 1,900 positions; net returns will be materially lower, especially in small stocks.
- **Risks.** Sharp style rotations, as in mid-2020, are the main risk; peer groups are re-estimated monthly and can shift when the cross-section changes quickly.

6. Assumptions and disclosures

- **Data.** Only the CTF dataset. The accounting characteristics already embed a reporting lag, so we add none; we cannot check for Compustat restatements.
- **Borrowed.** The LightGBM settings, the tangency-minus-GMV mapping and κ come from Geertsema’s CTF entry; the Barra settings and the Fama–French 12-industry map from the CTF Markowitz-ML and Minimum-Variance benchmarks. We re-implemented them in our own code. The MAXSER solver, used only in the Step 1 comparison, was run unchanged from Huang’s public entry. We knew Geertsema’s leaderboard result when we chose LightGBM for the base model.

- **Looks at the test period.** Earlier designs were also evaluated on 1990–2023 for reporting; every such run is logged. The final selection steps used 1970–1989 only, with rules committed to version control before each run, and in the last round only the selected model’s test results were opened.
- **AI use.** The team used an AI coding assistant (Claude) for code, experiments and drafting; all modelling decisions are ours.
- **Reproducibility.** One Python file (numpy, pandas, scipy, scikit-learn, LightGBM; versions pinned), deterministic, with a fixed LightGBM thread count. Runtime about 14 hours on 16 cores (estimated from the per-year timings of our cluster run), mostly the 34 yearly refits.

References. Bailey and López de Prado (2014), The deflated Sharpe ratio, *JPM*. Bhojraj and Lee (2002), Who is my peer? A valuation-based approach to the selection of comparable firms, *JAR*. DeMiguel, Garlappi and Uppal (2009), Optimal versus naive diversification, *RFS*. Gu, Kelly and Xiu (2020), Empirical asset pricing via machine learning, *RFS*. Hoberg and Phillips (2016), Text-based network industries and endogenous product differentiation, *JPE*. Jensen, Kelly, Malamud and Pedersen (2024), Machine learning and the implementable efficient frontier, working paper. Ke et al. (2017), LightGBM, *NeurIPS*. Lo (2002), The statistics of Sharpe ratios, *FAJ*. Menchero, Orr and Wang (2011), The Barra US equity model (USE4), MSCI. Ward (1963), Hierarchical grouping to optimize an objective function, *JASA*.