Network-Tree Ensemble on Daily Path Features with Shrunk-Covariance and Residual Books

Deniel Nankov · JKP Common Task Framework submission · November 2026

Summary. Each month the model forecasts the risk-adjusted, factor-neutral return of every stock and holds a dollar-neutral portfolio that blends two books built from that forecast: a mean-variance book under a Ledoit-Wolf shrunk correlation matrix of daily returns, and a factor-neutral residual book. The forecast is a weighted average of three learners — a multi-task feed-forward network, gradient boosted trees and a ridge regression — fitted on the within-month ranks of the supplied characteristics and on features generated from the daily return file. Everything is re-estimated on an expanding window using only information available at the portfolio date. The submission runs on CPU only and takes about 80 to 110 minutes on 32 cores.

1. Data and timing

Only the three CTF tables are used. A row of `ctff_chars` dated month-end t carries the characteristics known at t and the excess return of month t+1. For the portfolio of month t the model uses characteristics at t, daily returns dated on or before t, and realised monthly returns of months up to t (rows dated t−1 or earlier). A model that scores month t is fitted on rows dated strictly before its refit month, so the most recent label it sees is the return of the month that has just ended. Features, scaling, binning, early stopping and every other fitted quantity obey the same rule. Refit dates are anchored on the first test month, so the weights of a given month do not change when later data are removed.

2. Risk models and targets

At each month-end a statistical factor model is estimated from the trailing 252 daily returns of the stocks then in the panel: the factors are the principal components of the sample correlation matrix whose eigenvalues exceed the Marchenko-Pastur edge, and each stock gets a loading vector and a specific-variance share. A second model uses 756 days with the systematic part shrunk by the Ledoit-Wolf intensity. Stock volatility is the square root of the average of the 21-, 63- and 252-day daily variances, scaled to a month.

The learning target is next month's excess return divided by this volatility and residualised by generalised least squares on a set of exposures, then standardised within the month and clipped at four standard deviations. Five targets are used, one per exposure set: (i) return factors and two-digit SIC industries, (ii) return factors and the principal components of the ranked characteristics, (iii) both, (iv) the 756-day factors and industries, and (v) industries only. Residualising removes the part of the return that no stock-level forecast can earn after hedging, which makes the target less noisy.

3. Features

Characteristics. All 402 supplied characteristics enter as within-month percentile ranks centred on zero (missing values at zero) for the network and ridge, and as raw values with missing values kept for

the trees. No characteristic is selected or dropped by hand.

Daily-return features (96). These are generated on a fixed grid rather than chosen one by one: cumulative returns over the last 1, 2, 3, 4, 5, 7, 10, 15 and 21 trading days, both raw and as standardised residuals from the 252-day factor model; older residual legs (days 22–63, 22–126, 22–252); extremes, volatility over several horizons and their ratios, downside share, fraction of up days, path efficiency and first-order autocorrelation; and, for six base signals, leave-one-out averages over five peer sets — the stock's two-, three- and four-digit SIC industry, its twenty nearest neighbours in characteristic-rank space, and its twenty nearest neighbours in residual-return correlation (estimated before the last month) — together with own-minus-peer spreads. All are ranked within the month.

Update recency (13). For each stock, the share of slow-moving characteristics whose value changed in each of the last twelve months, and the number of months since the last sizeable update. The set of slow-moving characteristics is determined by a rule from rows before the first scored month. These features tell the learners how fresh a stock's fundamental information is.

4. Learners

Learner	Inputs	Specification	Refit
Feed-forward network	ranks, daily features, update recency	128-64-32 SiLU units, dropout 0.1, AdamW (1e-3, weight decay 1e-4), batch 4096; one output per target (multi-task); five random initialisations averaged. Written in NumPy.	annual
Gradient boosted trees	raw characteristics, daily features, update recency	LightGBM, learning rate 0.02, 63 leaves, feature fraction 0.5, row fraction 0.8, L2 10, minimum leaf size equal to the median number of stocks per month.	annual
Ridge	ranks, daily features	penalty 1e-3 on normalised moments; an equal mix of expanding and exponentially weighted (120-month half-life) sufficient statistics.	monthly

All fits use an expanding window that starts with the first month of the data. The number of network epochs is the one with the best mean monthly information coefficient on the last 36 months of the window, and the number of boosting rounds is the one with the best book Sharpe ratio on the last 60 months; in both cases the held-out months are then returned to the training set and the model is refit. The network learns all five targets jointly; the trees and the ridge are fitted to target (iv), which is the one the portfolio uses. The three scores for that target are converted to within-month z-scores and averaged with weights 2 : 2 : 1 (network : trees : ridge).

5. Portfolio

Two books are built each month from the ensemble score. **(a) Mean-variance book.** The score is first taken relative to its two-digit SIC industry mean. The sample correlation matrix of the last 504 daily returns of the stocks then in the panel (at least 250 observations each) is shrunk toward the identity with the Ledoit-Wolf intensity, and a ridge equal to the average variance is added. The book solves the mean-variance problem for the scores under dollar neutrality, in units of each stock's volatility forecast,

and is scaled to unit predicted risk. **(b) Residual book.** The score is projected, by generalised least squares, off the 252-day return factors, the dollar direction and the principal components of the ranked characteristics; positions are sized by inverse specific risk, again at unit predicted risk. **Position cap.** In both books each position, measured in units of the stock's volatility, is then limited to twice the book's median absolute position, and the book is not re-levered afterwards; in months when a few names would dominate, the book therefore carries less risk. **Blend.** The two books are blended every month with long-only, shrunk maximum-Sharpe coefficients computed from their own realised returns up to the previous month (exponential weights with a 60-month half-life, correlation shrunk halfway to zero, equal weights until 36 months of history exist). **Leverage.** The blended book is scaled by the average, to date, of the realised daily volatility of the books held in earlier months divided by its current value (exponentially weighted, 63-day half-life), limited to the range 0.75 to 1.5. So that the blend and the leverage rule have history at the first test month, the model starts producing books 120 months earlier; those months are not reported.

6. How the design was chosen

Every choice above — features, targets, learner settings, the composition of the ensemble and the portfolio rule — was made on the pre-1990 sample only, by walk-forward evaluation over 1965–1989 with the same code that produces the submission. Every full-sample run (to measure run time and to check determinism and the absence of look-ahead) was executed with performance reporting switched off. The test period was not used to select or tune anything. Table 1 reports the pre-1990 evidence behind the main choices.

Table 1. Pre-1990 walk-forward evidence (annualised Sharpe ratio)	**1965–89**	**1975–89**
A. What the inputs add (one residual book, one learner)		
Ridge on characteristic ranks	6.5	
Ridge, adding the daily-return features	8.9	
Network on characteristic ranks	7.3	
Network, adding the daily-return features	9.8	
Network, adding the update-recency features as well	10.1	
Network, all features and the three-horizon volatility	10.5	10.9
Trees on raw characteristics	6.6	
Trees, adding the daily-return features	9.7	
Trees, all features and the three-horizon volatility	10.2	10.8
B. Portfolio stage (all features, ensemble score)		
Residual book	10.6	11.1
Mean-variance book, 756-day window	10.1	11.7
Mean-variance book, 504-day window	10.0	12.2
Online blend of the residual and 504-day mean-variance books	11.3	12.2
Residual book with the position cap	10.9	11.3

Mean-variance book with the position cap and industry-relative score	10.3	12.6
Submitted model: online blend of the two capped books, with the leverage rule	**12.1**	**12.4**

Also tested before 1990 and not adopted because they did not help: calendar and market-state inputs, missing-value indicators, one-month rank changes, larger or smaller networks, more random initialisations, monthly refits, rolling 20-year training windows, hard industry or style constraints inside the mean-variance book, clipping or ranking the scores, covariance estimates from overlapping multi-day returns, neutralising market beta or size at the end, unit-gross scaling of the books, re-levering after the position cap (which removes its benefit), and leverage based on market volatility, on the back-cast volatility of the current weights or on monthly rather than daily realised volatility. The mean-variance book is the stronger of the two after 1975 and the weaker before, when the cross-section is thin; the blend is used because it holds up in both.

Sharpe ratios before 1990 are far above anything attainable later, because short-horizon return patterns were much stronger then; only the comparisons within the table are informative.

After the design was frozen, the model was run on the full sample and its weights were scored once with the leaderboard formula. The Sharpe ratio over the 408 test months is 5.12: 13.0 in 1990–1999, 7.2 in 2000–2009, 4.7 in 2010–2019 and 2.2 in 2020–2023. The same scoring code reproduces the published equal-weight benchmark of 0.4912.

Table 2. Risk profile of the submitted weights, 1990–2023

Sharpe ratio, 1990–2023 (95% interval)	5.12 (4.3–6.2)
Sharpe ratio by decade: 1990s / 2000s / 2010s / 2020–23	13.0 / 7.2 / 4.7 / 2.2
95% intervals	12.0–14.1 / 6.0–9.0 / 4.0–5.7 / 0.8–3.9
Correlation with the equal- / value-weighted market	0.14 / 0.14
Positions per month, average (effective number)	2,171 (1,267)
Largest position, share of gross exposure: average / maximum	0.35% / 1.47%
Net dollar exposure, largest share of gross	2.9%
One-way turnover per month, share of gross exposure	119%
Sharpe ratio net of proportional costs of 5 / 10 / 20 bps	4.63 / 4.14 / 3.16
Cost per traded dollar at which the average return is zero	55 bps
Largest drawdown / worst month, returns scaled to 10% volatility	-6.5% / -3.7%

Intervals are from a circular block bootstrap of monthly returns (12-month blocks). Turnover compares each stock's weight with its weight the month before, without adjusting for price drift. These statistics describe the weights after they were scored; nothing in the model was chosen from them. The leaderboard ranks returns before trading costs; the model holds about 2,000 names in an almost dollar- and market-neutral book and trades heavily, so its value after costs depends on execution.

After the score was known we tested further ideas against the same pre-1990 rehearsal. A change would have been adopted only if it raised the Sharpe ratio in both 1965–89 and 1975–89 by more than re-running the model with a different random seed does; Table 3 reports every idea tested. None passed, so the submitted model is unchanged. We also built a variant with a leverage rule that steadies the

expected return over time; because the idea came from the decline visible in the test-period score, and it made no difference before 1990, it was not submitted.

Table 3. Ideas tested after scoring: change in the pre-1990 Sharpe ratio of the submitted strategy	**1965–89**	**1975–89**
Re-run with a different random seed (two seeds)	−0.20, −0.15	−0.10, −0.14
Information-event features: largest residual shocks and the drift after them, seasonal timing of each stock's news, returns around fundamental updates (all three)	−0.15	−0.14
Recency-weighted copies of the learners with pay-off-based ensemble weights (60 / 120-month half-life)	+0.10 / +0.01	+0.02 / +0.02
Nonlinear shrinkage of the correlation matrix (ridge 1 / 0.5 / 0)	−0.80 / −0.75 / −0.63	−0.92 / −0.85 / −0.72
Correlations from volatility-standardised returns (21 / 63-day half-life)	−0.20 / −0.10	−0.22 / −0.06
Third, factor-exposed book from a raw-return target	−0.39	−0.25
Auxiliary network targets two and three months ahead	+0.03	+0.19
Ensemble weights learned inside the walk-forward	−0.09	−0.22
Leverage scaled by the model's own validation results	−1.94	−2.58

Each row changes one element of the submitted model and re-runs the full pre-1990 rehearsal. The first row is the noise floor: the seed alone moves the result by 0.1 to 0.2. Settings that were themselves chosen on this rehearsal, such as the 2 : 2 : 1 ensemble weights, have an in-sample advantage over rules that learn them as they go, which works against the adaptive rows.

7. Reproducibility and compute

- **Determinism.** The network, ridge and portfolio stages run single-threaded inside worker processes, and the trees use a fixed number of threads in LightGBM's deterministic mode, so the results do not depend on the number of cores. Seeds are fixed, the network is trained in double precision, and each month is scored in its own batch. Two full-sample runs on the same processor model produced bit-for-bit identical weights for all 885,698 test stock-months, and a third run on a different processor family (AMD instead of Intel) differed by at most 1.3×10^{-15}. On a 243-month extract, runs with 4 and with 32 cores were bit-for-bit identical, and a run on another architecture and operating system (ARM, macOS) agreed to within 10^{-12}.
- **No look-ahead.** The weights from the full sample were compared with those from samples cut at July 1996, December 2005 and March 2015, in which the forward return of the final month, unknown in real time, was blanked or replaced by noise. No weight up to the cut date changed: the runs on the same processor family were bit-for-bit identical, and the one on a different family agreed to within 2.2×10^{-16}. On the 243-month extract the same comparison at 14 cut dates, covering every phase of the annual refit cycle, also left every weight unchanged.
- **Resources.** CPU only; no GPU requested. 80 to 110 minutes wall-clock on 32 cores, peak memory about 65 GB. `requirements.txt` pins numpy, pandas, joblib and threadpoolctl at the versions used

and adds `lightgbm==4.7.0`, the one package not pre-installed; none of them, nor the packages they pull in, has a known vulnerability in the OSV database (checked 15 November 2026).

- **Small samples.** The code takes the feature list, the number of months and the test flags from its inputs and was run end to end on a 243-month extract with fewer than 150 stocks a month.

8. Sources

The residual target and the unit-risk residual book follow *Residual Books (LightGBM)* by Felix Gehling (CTF leaderboard, 2026). The Ledoit-Wolf mean-variance book with a ridge, the 756-day shrunk factor model and the leverage rule based on the realised daily volatility of held books follow *Residual Signal, Diversified Books* by Abhinav Keshri (CTF leaderboard, 2026). The use of short-horizon daily path features and characteristic-space peers, and the idea of bounding individual positions, follow the *Topology-Informed Decomposition Ensemble* by Hema Srikar Ankem (CTF leaderboard, 2026); see also *FU-HSA* by Vaidya et al. (CTF leaderboard, 2026). Each borrowed element was re-evaluated on pre-1990 data before being adopted; several others from the same entries were tested and left out (Section 6). We studied these entries because they ranked highest on the public leaderboard, so their test-period scores were known to us; for that reason no element was adopted on the strength of a leaderboard score, only on its pre-1990 evidence. Background: Gu, Kelly and Xiu (2020), “Empirical asset pricing via machine learning”; Jensen, Kelly and Pedersen (2023), “Is there a replication crisis in finance?”; Ledoit and Wolf (2004), “A well-conditioned estimator for large-dimensional covariance matrices”; Marchenko and Pastur (1967); Corsi (2009), “A simple approximate long-memory model of realized volatility”; Moreira and Muir (2017), “Volatility-managed portfolios”.