

Return Prediction and Portfolio Choice

An Ensemble Approach Using Ridge Regression, XGBoost, and LightGBM

1. Introduction

This report describes a portfolio construction method submitted to the Common Task Framework (CTF) competition proposed by Hellum et al. (2025). The goal is to use a high-dimensional set of 402 firm characteristics to predict monthly stock returns and build a long-short portfolio that maximizes the Sharpe ratio. The CTF dataset contains monthly observations of U.S. stocks from 1952 to 2023, with a pre-defined train period (1952-1989) and test period (1990-2023). I adopt an ensemble approach combining Ridge regression, XGBoost, and LightGBM, with rank-averaged predictions used to construct a dollar-neutral long-short portfolio.

2. Data Preprocessing

The raw characteristics have heavy tails and widely varying scales. I apply an Empirical CDF (ECDF) transformation cross-sectionally within each month, following the competition reference implementation:

- Observations are ranked within the cross-section and converted to percentiles (0 to 1).
- Exact zeros are preserved (they often carry special meaning, e.g. no R&D spending).
- Values are centered by subtracting 0.5, giving a bounded range of [-0.5, 0.5].
- Remaining NaN values are filled with 0 (the neutral/median value).

This transformation is robust to outliers and produces roughly uniform marginal distributions. Average missingness across the 402 features is about 10%. A Fama-MacBeth rank IC analysis on the training period identified the strongest univariate predictors (Figure 1), which include dividend yield ratios, value metrics, and volatility measures.

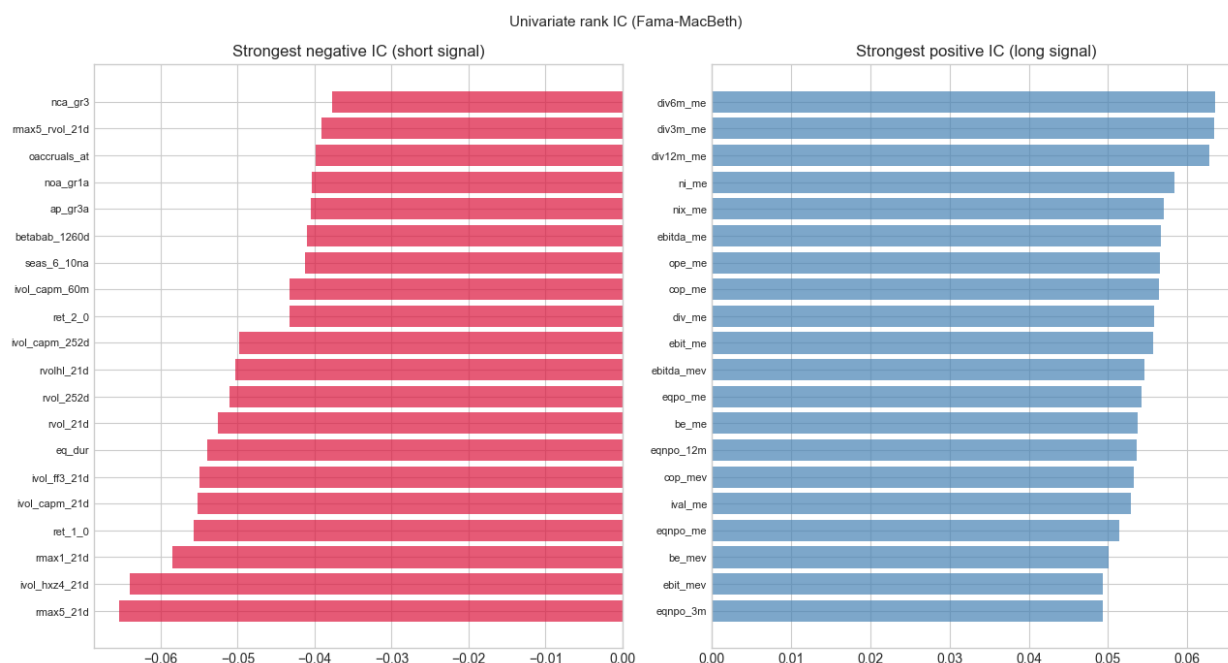


Figure 1: Univariate rank IC analysis. Left: features with strongest negative IC (short signal). Right: strongest positive IC (long signal).

3. Model Design

Ridge Regression. A regularized linear baseline. With 402 features and a very low signal-to-noise ratio (~12% cross-sectional std vs. <1% predictable), strong regularization is needed. Alpha tuned over {0.1, 1, 10, 100, 1000}; best: 1000.

XGBoost. Captures nonlinear interactions between characteristics. Searched 5 configurations varying depth (3-5), learning rate (0.05-0.1), and estimators (100-300). Best: depth=4, lr=0.05, 200 trees, subsample=0.8, colsample=0.3.

LightGBM. Leaf-wise growth, faster than XGBoost. Tuned leaves (15-63), learning rate, feature/bagging fractions. Best: 15 leaves, lr=0.05, 300 trees, feature_fraction=0.5, bagging=0.8.

Ensemble. Final predictions average the within-month rank-transformed outputs of all three models, following Gu, Kelly, and Xiu (2020). Rank averaging is robust to scale differences across models.

4. Hyperparameter Tuning

All hyperparameters were selected using expanding-window temporal cross-validation on the training period (1952-1989) with 3 folds and a 1-month gap to prevent look-ahead bias. Each monthly cross-section stays intact within one fold, as recommended for financial data. The evaluation metric is the annualized Sharpe ratio of the resulting long-short portfolio.

5. Portfolio Construction

Each month, stocks are ranked by predicted return, centered (subtract 0.5), and normalized so that $|w|$ sums to 2 (\$1 long, \$1 short). This is a dollar-neutral long-short portfolio rebalanced monthly. I chose rank-based weighting over mean-variance optimization because covariance estimation with 2,000+ stocks is noisy and degrades out-of-sample performance.

6. Results

On the out-of-sample test period (1990-2023):

Model	Sharpe	Ann. Ret	Ann. Vol	Max DD
Equal weight	0.49	9.3%	19.0%	-55.7%
Ridge	1.22	13.3%	10.9%	-34.2%
XGBoost	1.23	13.7%	11.1%	-28.6%
LightGBM	1.31	13.7%	10.5%	-28.6%
Ensemble	1.27	14.7%	11.6%	-32.6%

All three ML models roughly doubled the baseline Sharpe. The ensemble hit rate is 67.9% (positive months). Figure 2 shows cumulative returns on a log scale; the ensemble (purple) grows to over 100x the initial investment across 34 years.

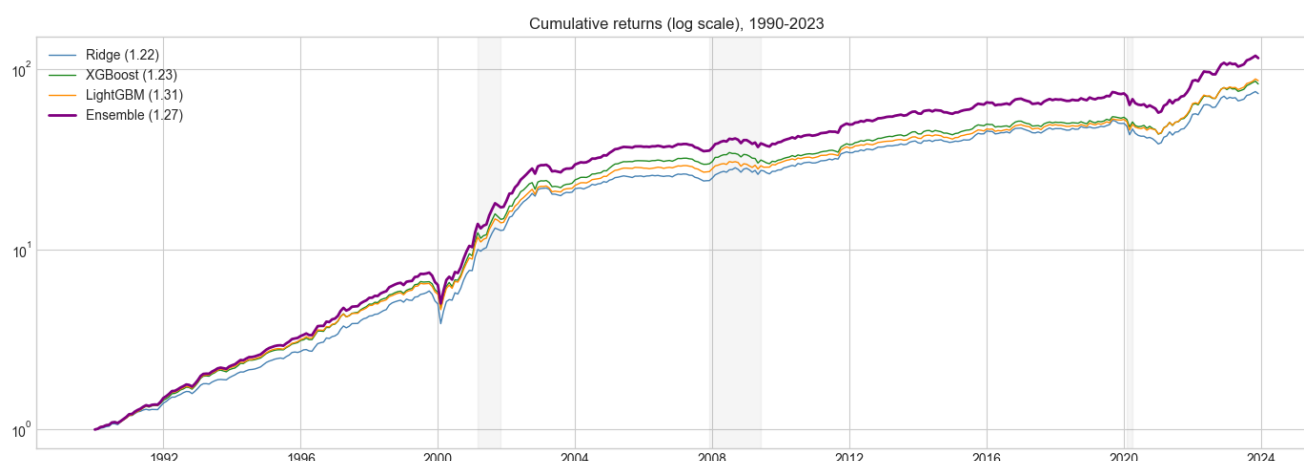


Figure 2: Cumulative portfolio returns (log scale) over the test period 1990-2023. Gray bands = recessions.

A decile analysis (Figure 3) confirms strong return monotonicity: D10-D1 spread is 24.1% annualized. The drawdown profile (Figure 4) shows the worst period around the dot-com crash (~-33%).



Figure 3: Average annualized excess return by predicted-return decile. Monotonically increasing pattern validates signal quality.

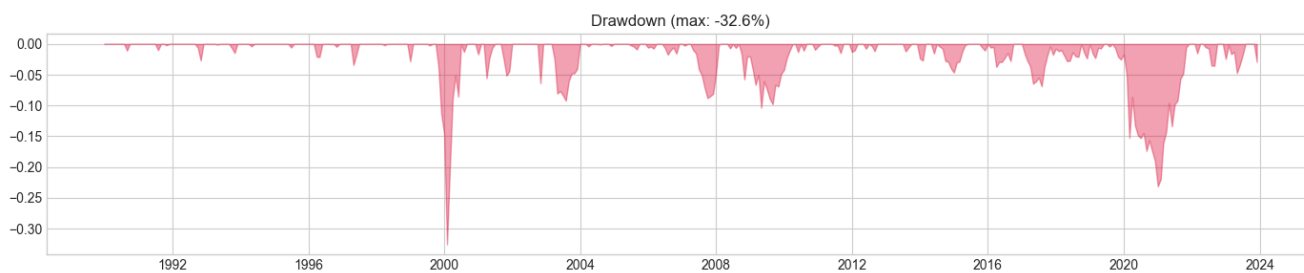


Figure 4: Ensemble portfolio drawdown over the test period. Maximum drawdown of -32.6%.

7. References

- [1] Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4), 745-766.
- [2] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5), 2223-2273.
- [3] Hellum, O., Jensen, T. I., Kelly, B. T., & Pedersen, L. H. (2025). The power of the common task framework. Available at SSRN 5242901.
- [4] Jensen, T. I., Kelly, B., & Pedersen, L. H. (2023). Is there a replication crisis in finance? *The Journal of Finance*, 78(5), 2465-2518.