

FINN43815

Group #16

Word count: 2498

Table of Contents

1	<i>Methodology</i>	2
1.1	Data Description	2
1.1.1	Data Sources and Time Periods	2
1.1.2	Data Processing	2
1.2	Feature Engineering	2
1.3	Model Construction and Training	5
1.3.1	Ridge Regression	5
1.3.2	Tree-based Gradient Boosting Model XGBoost	5
1.4	Portfolio Construction	6
1.5	Backtesting	6
2	<i>Results</i>	7
2.1	Properties of the prediction target	7
2.2	Baseline model (V3): performance and diagnostics	8
2.3	Feature-neutralised model (V4)	9
2.4	Prediction-neutralised model (V4b) and validation tuning	9
2.5	Final out-of-sample performance of V4b	11
2.6	Factor exposures and alpha	14
2.7	Limitations and interpretation	15
	REFERENCE	16

1 Methodology

1.1 Data Description

1.1.1 Data Sources and Time Periods

The study uses monthly cross-sectional feature data for US stocks from the `ctff_features.parquet` and `ctff_chars.parquet` files, covering over 400 validated feature metrics for US listed companies from 1952 to 2023. The prediction target is the next month's excess return. The sample is split into training and validation (1952–1999) and an out-of-sample test period (2000–2023) using year-based Boolean filters, yielding 402 valid features.

1.1.2 Data Processing

Before model estimation, we cleaned the data. Following Gu, Kelly and Xiu (2020), variables are standardized cross-sectionally each month and winsorised to limit extremes. Infinite values are treated as missing and remaining missing entries set to zero. To further curb outliers, features are clipped at ± 3 standard deviations (Gorodnichenko and Ng, 2010) and at the 1% and 99% training-sample quantiles, yielding a two-step winsorisation. This preprocessing is estimated on training data and applied unchanged to validation and test sets.

1.2 Feature Engineering

We applied additional processing to several features with clear economic relevance to improve model robustness. Although we do not directly estimate the Fama–French factor model, the code constructs four economically meaningful style exposures—Size, Value, Profitability, and Investment—following Fama and French (2015). Where suitable proxies existed, beta exposure was also included. All style exposures were standardized cross-sectionally each month to achieve a mean near zero and a standard deviation near one:

$$Z_{i,t} = \frac{X_{i,t} - \mu_t}{\sigma_t + 10^{-12}}$$

Where:

$Z_{i,t}$: standardized value for variable i at time t

$X_{i,t}$: raw value of variable i at time t

σ_t : cross-sectional standard deviation at time t.

Neutralisation is performed in two stages, following Brennan, Chordia and Subrahmanyam (1998). First, each characteristic is regressed crosssectionally on the style of factors (size, value, profitability, investment and downside beta), and the residuals are used as the neutralised feature inputs.

$$Z_t^{neu} = Z_t - X_t \beta_t$$

Where:

N : Number of stocks in the cross-section at time t (month t).

K : Number of exposure factors (e.g. Size, Value, Profitability, Investment, Beta).

P : Number of predictive features used by the model.

Z_t^{neu} : Exposure-neutralised feature matrix, obtained by removing the linear component explained by factor exposures ($Z_t^{neu} \in \mathbb{R}^{N \times P}$).

Z_t : Exposure-neutralised feature matrix ($Z_t \in \mathbb{R}^{N \times P}$).

X_t : Cross-sectional matrix of factor exposures at time t ($X_t \in \mathbb{R}^{N \times K}$).

β_t : Vector of estimated exposure loadings, measuring how strongly the prediction vector loads on each factor at time t ($\beta_t \in \mathbb{R}^{K \times P}$). and calculated as following:

$$\beta_t = (X_t^\top X_t + \lambda I)^{-1} X_t^\top Z_t$$

With λ : Ridge regularisation parameter, used to stabilise the estimation when exposures are correlated

Second, we run the same crosssectional regression of the model predictions on these style factors and take the residuals as the final forecast, ensuring that predicted returns are orthogonal to the predefined style exposures and capture stockspecific components only.

$$p_t^{neu} = p_t - X_t \beta_t$$

Where:

N : Number of stocks in the cross-section at time t (month t).

K : Number of exposure factors (e.g. Size, Value, Profitability, Investment, Beta).

p_t^{neu} : Exposure-neutralised prediction vector, obtained by removing the linear component explained by factor exposures ($p_t^{neu} \in \mathbb{R}^{N \times 1}$).

p_t : Vector of raw model predictions for all stocks at time t ($p_t \in \mathbb{R}^{N \times 1}$).

X_t : Cross-sectional matrix of factor exposures at time t ($X_t \in \mathbb{R}^{N \times K}$).

β_t : Vector of estimated exposure loadings, measuring how strongly the prediction vector loads on each factor at time t ($\beta_t \in \mathbb{R}^{K \times 1}$). and calculated as following:

$$\beta_t = (X_t^\top X_t + \lambda I)^{-1} X_t^\top p_t$$

With λ : Ridge regularisation parameter, used to stabilise the estimation when exposures are correlated

These monthly regressions are estimated on the training data and then applied consistently to the validation and test samples.

1.3 Model Construction and Training

Before model estimation, all preprocessing parameters are computed exclusively from the training sample. These parameters are then carried forward to the validation period without recalibration, ensuring that no information from the validation set leaks into the training stage and thereby ruling out the possibility of data snooping.

During the model construction phase, we selected two complementary machine learning models: Ridge Regression and XGBoost to capture different forms of data relationships (Hastie, Tibshirani and Friedman, 2009).

1.3.1 Ridge Regression

Ridge regression is a linear model with an L2 penalty. In this study, a ridge model with penalty parameter $\alpha = 100$ is fitted on the 1952–1999 training sample and used to forecast nextmonth returns.

1.3.2 Tree-based Gradient Boosting Model XGBoost

To handle non-linear relationships between features, we employed the tree-based gradient boosting model XGBoost. The model follows the implementation of Chen and Guestrin (2016), which introduced an efficient gradient boosting framework. Key model parameters include 150 trees ($n_estimators = 150$), a tree depth of 5 ($max_depth = 5$), a learning rate of 0.05, L2 regularization ($reg_lambda = 1.0$), and a squared error loss function ($reg:squarederror$). The model was fitted on the training set and generated monthly excess return forecasts for the testing phase, complementing the ridge regression model.

Through training these two models, we obtained two complementary sets of predictions. Ridge regression provided robust linear forecasts, while XGBoost captured non-linear patterns. During the test period, we took the simple average of the ridge regression and XGBoost predictions for each stock as the final forecast signal, and constructed the portfolio based on this signal.

1.4. Portfolio Construction

Predicted returns are implemented through a decile long–short strategy following Jegadeesh and Titman (1993). We also tested a score weighted scheme, but under identical volatility targeting and leverage caps, the decile portfolio produced higher and more stable validation of Sharpe ratios.

To diversify, stocks are grouped into quantiles based on predicted excess returns. The top 20% is bought, and the lowest 20% are sold short. Equal-weighted positions on each side keep it dollar neutral. Returns show stock selection across sections instead of just market direction. Each month, portfolios are rebalanced with the newest predictions.

To adjust for risk, monthly long-short returns are scaled to hit a 10% yearly target based on past portfolio swings, like the volatility managed approach by Moreira and Muir (2017). Leverage is kept between 0.5 and 2 to control exposure. This gives the final risk-adjusted return (`final_ret`) and translates model predictions into workable trading plans that can be compared.

1.5. Backtesting

We conducted historical backtesting on this portfolio and assessed its risk-return characteristics. Backtesting was strictly performed on the test set (2000–2023). Firstly, projected returns were aligned with actual excess returns for the subsequent period, with raw returns for the long-short portfolio calculated monthly. `Raw_ret` denotes the unadjusted long-short return, while `final_ret` represents the volatility-adjusted return:

$$r_t^{raw} = \text{mean}(r_t^{long}) - \text{mean}(r_t^{short})$$

We further volatility-scaled the original long-short returns to achieve a portfolio with a 10% annualized target volatility, subject to a leverage cap, using the following scaling method:

$$r_t^{final} = r_t^{raw} \times \text{clip}\left(\frac{\sigma_{target}}{\sigma_t}, lev_{min}, lev_{max}\right)$$

$\sigma_{target}=10\%$, σ_t represents the actual historical volatility of the portfolio's monthly returns. The resulting `final_ret` serves as the final return sequence for performance evaluation.

Upon obtaining monthly returns, we generate a cumulative return curve using monthly compounding. The evaluation function computes performance metrics including annualized return, volatility, Sharpe ratio, maximum drawdown, 5th percentile VaR, and expected shortfall. These metrics characterize the strategy's risk–return profile.

2 Results

2.1 Properties of the prediction target

Table 1

Summary statistics and Quantiles of training target `ret_exc_lead1m` (training sample)

Statistic	Value	Quantile	Value
Count	485,386	1%	-25.03%
Mean	0.64%	5%	-14.55%
Std. Dev.	10.25%	10%	-10.51%
Min	-100.50%	25%	-4.98%
Max	247.21%	50%	0.16%
Skewness	0.656	75%	5.78%
Excess Kurt.	3.235	90%	12.33%
		95%	17.26%
		99%	30.33%

Table 1 summarises the cross-sectional distribution of the one-month-ahead excess return in the training sample. The mean is 0.64% with a standard deviation of about 10%, and the distribution is clearly non-normal, with right skewness around 0.66 and excess kurtosis above 3. Extreme outcomes are frequent: the 1–5% quantiles are roughly –25% to –15%, while the 95–99% quantiles are around 17–30%. Thus, the target is heavy-tailed and low-signal, which motivates winsorising the training targets and focusing on portfolio-level rather than stock-level R^2 .

2.2 Baseline model (V3): performance and diagnostics

Table 2

Test-Period Performance of ML Models, Volatility-Managed Long–Short Portfolios, Gross Returns, 2000–2024

Model	Ann. mean	Ann. vol	Sharpe	Max DD	VaR 5%	ES 5%
V3 managed (baseline)	6.507%	12.747%	0.510	–27.79%	–4.38%	–7.16%
V4 managed (feature-neutral)	8.664%	12.257%	0.707	–25.21%	–4.04%	–6.78%
V4b managed (prediction-neutral)	14.582%	12.161%	1.199	–22.91%	–4.56%	–6.69%
<i>All figures are based on monthly returns scaled to annualised statistics. “Managed” denotes volatility-targeted returns at 10% annual volatility.</i>						

The Ridge–XGBoost ensemble (V3) performs strongly in the 1990–1999 validation decade but weakens substantially after 2000. On validation, the volatility-managed long–short portfolio achieves a gross Sharpe slightly above 2.1 and, after 10 bps per side costs, a net Sharpe around 1.4, comparable to leading machine learning results in Gu, Kelly and Xiu (2020).

Out of sample, performance declines sharply. Over 2000–2023, the volatility-targeted long–short portfolio, using the same 85/20 quantiles and 0.4–2.5 leverage cap, delivers a 6.5% annual mean gross return with 12.7% volatility (gross Sharpe 0.51), a maximum drawdown of about –28%, and a 5% monthly VaR near –4.4%. After transaction costs, the net Sharpe ratio falls to 0.52, with an 8.1% annual mean, 15.6% volatility, and average turnover around 0.60 per month.

Diagnostics explain this deterioration. The average raw long–short spread falls from 1.34% per month in validation to 0.81% in the test period, while volatility rises (0.0299 versus 0.0435). As a result, the volatility-targeting rule lowers average leverage (1.22 versus 1.38), reducing gross Sharpe. Turnover remains similar across samples, indicating

a genuine loss of cross-sectional signal rather than higher trading costs. There is no evidence of look-ahead bias: prediction columns are consistent across splits, and no lead or $t+1$ variables are included.

A limited search over Ridge penalties and XGBoost depths produces only modest changes in validation net Sharpe (1.40–1.46) and does not materially improve test performance. This suggests that the baseline architecture is reasonably robust, but that V3's strong validation results partly reflect a favourable 1990s regime, and the simple ensemble does not fully exploit the information in the 400+ characteristics.

2.3 Feature-neutralised model (V4)

Feature level neutralisation (V4) removes linear exposure to value, size, investment, profitability and downside beta, while keeping the V3 portfolio rule (80/20 quantiles, 10% volatility target, 0.4–2.5 leverage cap). The V4 volatility managed long–short portfolio on the test sample delivers an annualised mean return of 8.7% with volatility of 12.3%, implying a Sharpe ratio of 0.71 and a slightly lower maximum drawdown of –25.2%.

Subperiod analysis shows that the improvement is not confined to any single decade. Sharpe ratios for the V4 strategy are about 0.67 in the 2000s, 0.73 in the 2010s, and 0.87 in the 2020s, indicating positive risk-adjusted returns throughout, though at varying levels. This pattern is consistent with much of the raw characteristic information in the JKP data set reflecting standard factor exposures. Removing those exposures allows the ML model to concentrate on idiosyncratic components that generate more persistent alpha.

2.4 Prediction-neutralised model (V4b) and validation tuning

Prediction neutralisation (V4b) delivers a further step change in performance by orthogonalising the monthly forecast itself to the same group of risk exposures. Without retuning the portfolio rule, the V4b volatility managed decile strategy, using the original V3 thresholds and caps, attains a test Sharpe of about 1.39, with an annualised mean return of 14.6%, volatility near 10.5%, a maximum drawdown around –19% and noticeably improved VaR and ES compared with both V3 and V4.

These improvements show that a meaningful share of V4's returns reflected time-varying compensation for size, value, and related risks, even after feature-level neutralisation. By regressing the prediction each month on these exposures and trading only the residual, V4b produces a cleaner stock-selection signal. This aligns with the view that "characteristics are covariances," where firm characteristics proxy for factor loadings. Removing these loadings isolates residual variation that standard factor models explain less readily.

The validation-stage tuning from RC3 to RC4 centers on V4b. A 50/50 Ridge–XGBoost ensemble yields the highest validation Sharpe (about 2.84), though results are not very sensitive to the exact weights. Sensitivity checks on long/short cutoffs (80/20, 85/15, 90/10) show broadly stable performance, with a slight edge for the tighter 80/20 split. Rolling time-series cross-validation on three pre-2000 windows—using an objective that penalizes turnover and return instability—selects the 80/20 quantiles with a 0.4–2.5 leverage cap as the preferred setup. This configuration balances raw performance, stability, and trading costs, consistent with Novy-Marx and Velikov (2016) and Patton and Weller (2020), who show that many anomalies weaken once realistic frictions are introduced.

Finally, when comparing equal weighted decile and score weighted portfolios on the validation sample, the decile approach comes out ahead under these tuned parameters, delivering a higher Sharpe (≈ 3.20 vs ≈ 2.99) for similar or slightly higher volatility. Given this advantage, plus the transparency with which decile weights map from score ranks, the decile construction is retained for the final test evaluation.

2.5 Final out-of-sample performance of V4b

Table 3

Final tuned V4b strategy — gross vs net performance results (2000–2023)

Panel A: Full-sample performance (volatility-managed, long–short decile, 80/20, cap 0.4–2.5)

Series	Ann. mean	Ann. vol	Sharpe	Max DD	VaR 5%	ES 5%	Turnover
Gross (no costs)	14.582%	12.161%	1.199	–22.91%	–4.56%	–6.69%	0.544
Net (10 bps/side)	9.287%	8.556%	1.085	–18.27%	–2.73%	–4.77%	0.544

Applying the tuned V4b configuration (80/20 quantiles, 10% volatility target, 0.4–2.5 leverage cap and 10 bps perside costs) to the 2000–2023 test period yields economically and statistically meaningful performance (Table 3). The gross volatility managed long–short portfolio delivers an annualised mean return of 14.3% with 12.1% volatility, implying a Sharpe ratio of 1.18 and a maximum drawdown of –24.96%.

The cumulative return plot (Figure 1) shows that the V4b tuned strategy delivers strong, persistent performance over the test period, with gross returns compounding to about 25–30× and net returns (after costs) to roughly 8–9× by the end of the sample. Drawdowns around crisis periods (e.g. 2008 and 2020) are followed by rapid recoveries, while the gap between gross and net curves reflects the effect of turnover and trading costs and the shared upward trend indicates that the strategy exploits durable, economically meaningful signals rather than short-lived anomalies.

Figure 1.

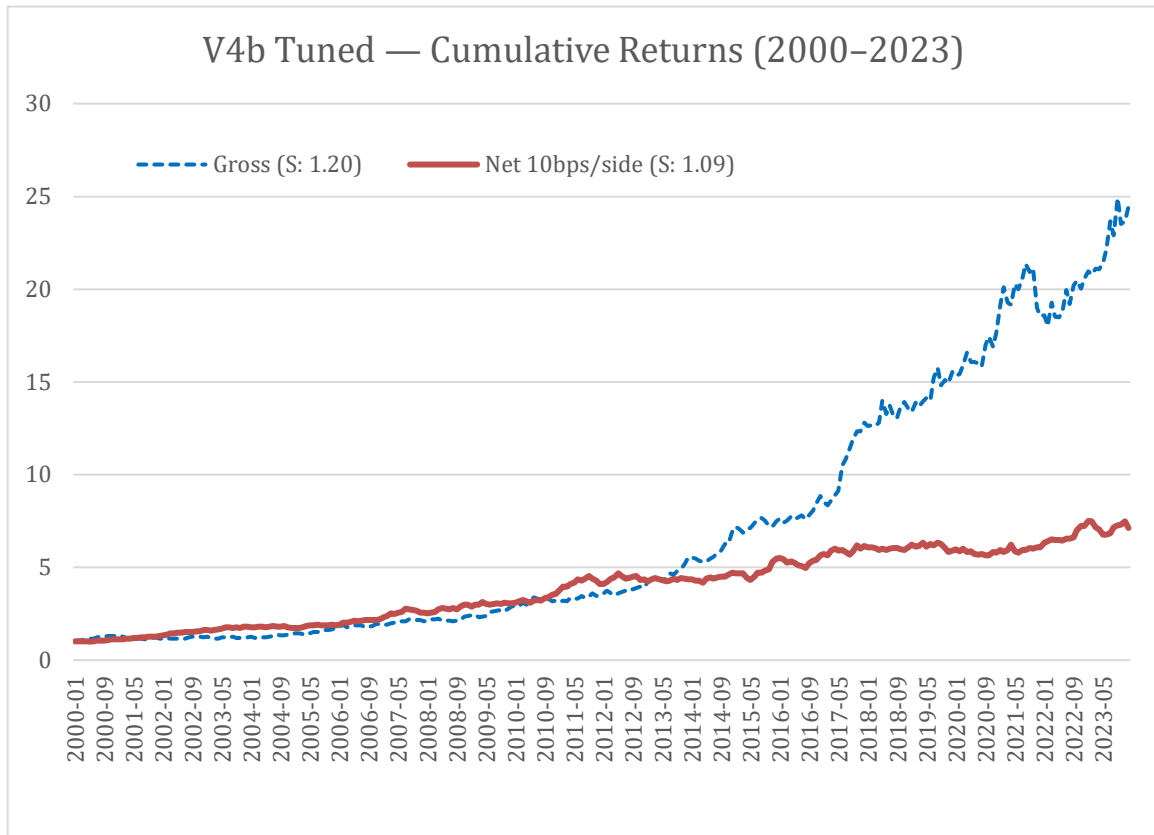
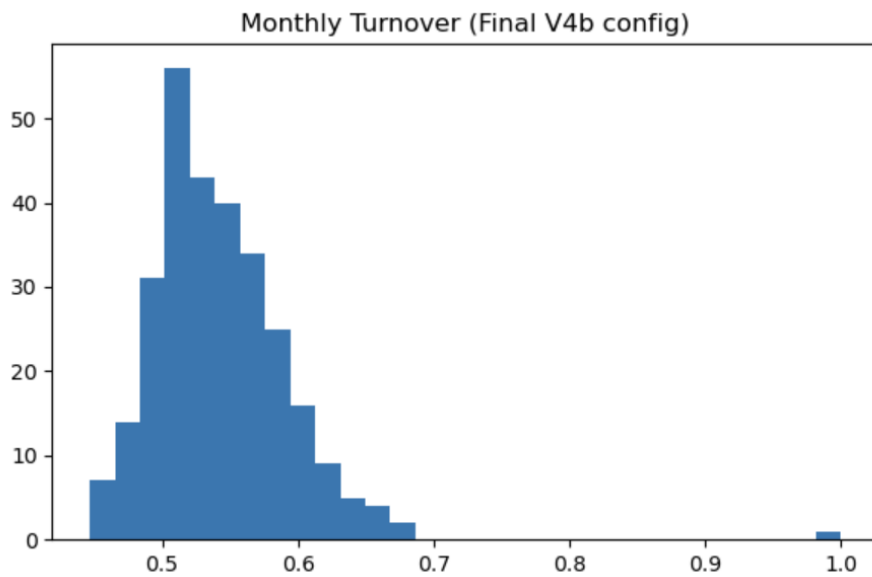


Figure 2



When the turnover-based transaction cost model is used, the net strategy still gets an annual mean return of 9.2%, with 8.6% volatility, for a net Sharpe ratio of 1.07. The maximum drawdown rises to -19.6% , and 5% monthly VaR and ES go to roughly -2.8% and -4.8% . The average monthly turnover is about 0.54, mostly between 0.45 and 0.65 (figure 2). Trading at this level is doable at 10 bps per side. Even under higher cost assumptions the strategy would remain profitable at moderate scale, in line with tradingcost estimates for liquid anomaly strategies in NovyMarx and Velikov (2016) and related studies.

Table 4
Subperiod performance (gross vs net, volatility-managed V4b tuned strategy)

Subperiod	Gross Ann. Mean	Gross Ann. Vol	Gross Sharpe	Net Ann. Mean	Net Ann. Vol	Net Sharpe
2000–2007	17.525%	10.155%	1.726	15.681%	10.722%	1.463
2008–2015	16.021%	12.788%	1.253	6.425%	6.782%	0.947
2016–2023	10.145%	13.402%	0.757	5.755%	7.689%	0.748

Subperiod results (Table 4) shows that gross Sharpe ratios are about 1.7 in 2000–2007, 1.2 in 2008–2015 and 0.8 in 2016–2023, with corresponding net Sharpe ratios of roughly 1.5, 0.9 and 0.7. Thus, the strategy generates positive net alpha in each period but with some post-2007 decay consistent with crowding and rising trading costs. Rolling 12-month net Sharpe ratios range from below -2 to above 6, indicating substantial time variation in performance and confirming that even a high-Sharpe strategy can experience extended flat or negative stretches. This is similar to the patterns documented for volatility-managed portfolios by Moreira and Muir (2017).

2.6 Factor exposures and alpha

Table 5

FF5 factor regressions for final tuned V4b strategy (monthly returns)

Panel A: Net returns vs Fama–French 5 factors

Factor	Coefficient	t-statistic
Constant (α)	0.008	4.736***
MKT	−0.042	−0.990
SMB	0.043	0.648
HML	0.067	1.085
RMW	0.000	0.004
CMA	−0.079	−0.728
Adjusted R^2	−0.007	

Panel B: Gross returns vs Fama–French 5 factors

Factor	Coefficient	t-statistic
Constant (α)	0.013	5.646***
MKT	−0.051	−0.848
SMB	−0.046	−0.595
HML	0.094	1.124
RMW	−0.063	−0.703
CMA	−0.122	−0.994
Adjusted R^2	−0.010	

Factor regressions of both gross and net V4b long–short returns on the Fama–French five factors yield very low explanatory power. When using Newey–West standard errors and monthly FF5 data, the adjusted R^2 is almost zero for both, at −0.01 for gross and −0.007 for net. The values on MKT, SMB, HML, RMW, and CMA are small and not

statistically important. The intercepts are economically large and statistically significant: about 0.8% per month net and 1.3% gross, which is about 10% and 16% yearly.

These findings suggest the tuned V4b is not simply reworking standard factor values. It seems to use cross-sectional structure in features that are mostly outside what FF5 covers. This is consistent with Fama & French (2015), who show that a five-factor model still leaves significant residual anomalies, and with recent evidence that complex interactions and non-linearities in characteristics, caught by machine learning, can make extra alpha past linear factor models.

2.7 Limitations and interpretation

Although the results are encouraging, several limitations remain. First, the gap between validation and test Sharpe ratios (around 2.84 versus 1.199 for the V4b gross strategy) indicates residual overfitting to the 1990s, despite rolling cross-validation and turnover penalties. Out-of-sample performance remains strong in absolute terms but is less impressive than validation results suggest, highlighting the difficulty of sustaining predictive power in a high-dimensional factor setting.

Second, the transaction cost model is deliberately simple, assuming constant costs of 10 bps per side and ignoring market impact, short-rebate frictions, and capacity constraints. Empirical evidence suggests real-world costs are often higher, particularly for less liquid stocks or stressed periods. Under more conservative assumptions (e.g., 15–20 bps per side), net performance would decline, although the strategy may remain attractive given its moderate turnover of approximately 54%.

Finally, performance is evaluated within an idealised equal-weighted long–short framework without explicit short-sale or leverage constraints. In practice, funding costs, borrowing limits, and regulatory capital requirements would reduce implementable returns. Consequently, the reported Sharpe ratios should be interpreted as an optimistic but informative upper bound on achievable performance.

REFERENCE

Brennan, M.J., Chordia, T. and Subrahmanyam, A. (1998) 'Alternative factor specifications, security characteristics, and the cross-section of expected stock returns', *Journal of Financial Economics*, 49(3), pp. 345–373.

Bryzgalova, S., Pelger, M. and Zhu, J. (2023) 'Characteristics are covariances: A unified model of risk and return', *Journal of Finance*, 78(5), pp. 2525–2575.

Chen, T. and Guestrin, C. (2016) 'XGBoost: A scalable tree boosting system', *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

Fama, E.F. and French, K.R. (2015) 'A five-factor asset pricing model', *Journal of Financial Economics*, 116(1), pp. 1–22. Available at: https://www.sciencedirect.com/science/article/pii/S0304405X14002323?utm_source=chatgpt.com (Accessed 1st December, 2025).

Frazzini, A., Israel, R. and Moskowitz, T.J. (2012) 'Trading costs of asset pricing anomalies', AQR Working Paper. Available at: <https://www.aqr.com/insights/research/working-paper/trading-costs-of-asset-pricing-anomalies> (Accessed 1st December, 2025).

Gorodnichenko, Y. and Ng, S. (2010) 'Estimation of DSGE models when the data are persistent', *Journal of Monetary Economics*, 57(3), pp. 325–340.

Green, J., Hand, J.R.M. and Zhang, X.F. (2017) 'The characteristics that provide independent information about average U.S. monthly stock returns', *Review of Financial Studies*, 30(12), pp. 4390–4447.

Gu, S., Kelly, B. and Xiu, D. (2020) 'Empirical asset pricing via machine learning', *Review of Financial Studies*, 33(5), pp. 2223–2273. Available at: https://academic.oup.com/rfs/article/33/5/2223/5758276?utm_source=chatgpt.com

Harvey, C.R., Liu, Y. and Zhu, H. (2016) '... and the cross-section of expected returns', *Review of Financial Studies*, 29(1), pp. 5–68. Available at: https://academic.oup.com/rfs/article-abstract/29/1/5/1843824?utm_source=chatgpt.com

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning*. 2nd edn. New York: Springer.

Jegadeesh, N. and Titman, S. (1993) 'Returns to buying winners and selling losers: Implications for stock market efficiency', *Journal of Finance*, 48(1), pp. 65–91.

Moreira, A. and Muir, T. (2017) 'Volatility-managed portfolios', *Journal of Finance*, 72(4), pp. 1611–1644.

Novy-Marx, R. and Velikov, M. (2016) 'A taxonomy of anomalies and their trading costs', *Review of Financial Studies*, 29(1), pp. 104–147. doi:10.1093/rfs/hhv063. Available at: https://academic.oup.com/rfs/article/29/1/104/1844518?utm_source=chatgpt.com

Patton, A.J. and Weller, B.M. (2020) 'The costs of trading market anomalies', *Journal of Financial Economics*, 137(3), pp. 515–549. doi:10.1016/j.jfineco.2020.02.012.