Out-of-Sample Forecastability of GICS Sector Excess Returns:

Horizon Effects, Model Choice, and Economic Value

Alexander Uyar

May 2026

Abstract

This paper examines whether sector-specific macro-financial predictors can systematically forecast the out-of-sample excess returns of the eleven GICS sectors relative to the S&P 500. Using a monthly panel spanning 2001 to early 2026, we evaluate eight competing forecasting models, alongside a rolling-mean benchmark, across four horizons under a strict pseudo-out-of-sample walk-forward design. Forecasts are evaluated using the Campbell and Thompson (2008) R^2_{OOS}, directional accuracy, bootstrap-based statistical inference, and portfolio-level economic value metrics.

Three principal findings emerge. First, forecastability is strongly horizon-dependent: measured skill improves materially at the six- and twelve-month horizons, where the cross-sector mean R^2_{OOS} reaches $+0.163$ across best-performing models, though a portion of this gradient is mechanical under overlapping cumulative returns and should be interpreted with appropriate caution. Second, model performance is horizon-specific: gradient-boosted trees dominate at short horizons where non-linearities are most prominent, while adaptive and combination models become increasingly competitive as the horizon extends. Third, forecastability is highly sector-specific and sensitive to predictor-selection methodology. A recursive walk-forward VIF correction materially alters the apparent cross-sector hierarchy, with seven of eleven sectors showing stable or improved skill under the stricter protocol. Economic value is weak under unconstrained mean-variance allocation, but constrained target-volatility implementation recovers positive excess returns in eight of eleven sectors, indicating that the economic usefulness of the signals depends heavily on portfolio construction.

Taken together, the evidence suggests that sector return predictability is real but conditional, contingent on forecast horizon, model family, predictor-selection discipline, and portfolio construction design.

1 Introduction

Sector rotation strategies occupy a central role in active equity management. The premise is simple: if sector excess returns can be forecast with any reliability, a manager who systematically overweights the sectors with the highest predicted returns and underweights those with the lowest can generate persistent alpha over a passive benchmark. Yet this premise rests on a condition that is far from assured: that macroeconomic and financial indicators contain information about future sector returns that is not yet fully reflected in prices.

This question is made harder by the evidence on aggregate equity premium predictability. Goyal and Welch (2008) survey a broad set of macro-financial predictors and find that virtually none outperform a simple historical mean in strict out-of-sample tests. If the aggregate market is barely predictable, what justification exists for sector-level forecasting? The answer lies in sectoral heterogeneity. Sectors have structurally distinct sensitivities to the business cycle, monetary conditions, and commodity prices. This structural heterogeneity is consistent with Fama and French (1989), who demonstrate that business-cycle variables, dividend yields, default spreads, and term spreads drive time-varying expected returns across asset classes. Health Care responds to demographic and regulatory signals, while Energy responds to crude oil futures and inventory dynamics. Sector-specific predictor sets may capture variation that aggregate predictors average away.

This paper provides a systematic, multi-model test of that hypothesis. We construct a monthly panel dataset covering eleven GICS sector sub-indices from January 2001, with between 29 and 49 candidate predictors per sector drawn from macro-financial, sentiment, valuation, and sector-specific sources. Against this data, we evaluate eight forecasting models in a unified model competition across four forecast horizons (1, 3, 6, and 12 months), spanning five families: (i) time-series, (ii) autoregressive distributed lag, (iii) forecast combination schemes, (iv) time-varying parameter models, and (v) gradient-boosted trees. All models share identical pre-processing and are evaluated under a strict pseudo-out-of-sample walk-forward design, enabling clean comparison of model, horizon, and sector differences under a common standard.

All models are evaluated using the Campbell and Thompson (2008) R^2_{OOS}, which compares each model's mean squared forecast error to that of a rolling 120-month historical mean benchmark. We supplement this with directional accuracy, root mean squared forecast error (RMSFE), and certainty-equivalent return (CEQ) for a mean-variance investor with risk aversion $\gamma = 3$. Bootstrap confidence intervals and Diebold-Mariano pairwise tests provide statistical inference.

The principal findings are three. First, measured forecastability improves materially at longer horizons, although part of this increase reflects the mechanics of overlapping cumulative returns rather than pure incremental skill. The six-month horizon marks the point at which every sector records a positive best-model R^2_{OOS} point estimate, and the cross-sector mean best-model R^2_{OOS} rises to +0.163 at twelve months. Second, model performance is strongly horizon-specific: tree-based methods dominate more often at shorter horizons where signal is weak and nonlinearities are more prominent, while adaptive and combination models become more competitive at longer horizons. Third, forecastability is highly sector-specific, and the cross-sector ranking is sensitive to predictor-selection methodology. A recursive walk-forward VIF procedure materially alters the apparent cross-sector hierarchy. Economic value also depends critically on implementation: unconstrained mean-variance allocation yields weak CEQ results, while constrained target-volatility construction produces positive excess returns in eight of eleven sectors.

The remainder of the paper is structured as follows. Section 2 reviews the empirical literature on equity return predictability, forecast combination, and machine learning in asset pricing. Section 3 describes the data and dependent variable construction. Section 4 presents the methodology, including the full model suite. Section 5 reports empirical results. Section 6 provides sector-level analysis. Section 7 analyses variable selection and predictor importance. Section 8 derives sector rotation signals and evaluates economic value. Section 9 discusses the findings. Section 10 addresses robustness and limitations. Section 11 concludes.

2 Literature Review

The foundations of return predictability research rest on a long tradition of attempts to forecast the aggregate equity premium from macroeconomic and financial variables. Goyal and Welch (2008) provide an influential benchmarking study of return predictors and find that the in-sample predictive relationships documented in the literature largely fail to hold out of sample: almost no variable beats the historical mean forecast when evaluated over post-war U.S. data. Campbell and Thompson (2008) show that predictability can improve under economically motivated sign restrictions, with modest but statistically significant out-of-sample gains remaining achievable even when individual predictors disappoint in isolation. Rapach and Zhou (2013) review this literature and conclude that pooling forecasts across individual predictor models is one of the most robust routes to equity premium predictability, with combined forecasts often outperforming the historical mean benchmark. This paper builds directly on the Campbell and Thompson R^2_{OOS} framework and adopts the Rapach, Strauss and Zhou (2010) insight that combining forecasts across models can reduce forecast error variance.

While most of the predictability literature focuses on the aggregate market, a smaller but growing body of work examines return predictability at the sector and industry level. Moskowitz and Grinblatt (1999) show that much of the momentum effect in the cross-section of stock returns is attributable to industry momentum: past-winning industries continue to outperform past-losing industries at horizons of one to twelve months, independent of individual stock momentum. Chordia and Shivakumar (2002) further show that industry-level momentum profits are substantially explained by macro-financial variables linked to the business cycle, reinforcing the predictive relevance of such variables in cross-sectional sector return forecasting. Hong, Torous and Valkanov (2007) demonstrate that certain industries, notably oil, mining, and retail, lead the aggregate stock market by several months, suggesting that sector-level information has predictive content for broader market returns. Fama and French (1997) show that sector cost-of-equity estimates computed from standard asset pricing models carry large standard errors, leaving room for alternative forecasting approaches based on macro-financial predictors. Taken together, these studies motivate the hypothesis tested in this paper: that sector excess returns, defined relative to the aggregate market, contain predictable components linked to sector-specific macro-financial exposures.

The observation that no single forecasting model dominates across all economic conditions motivates combining forecasts across models. Bates and Granger (1969) establish the theoretical case for combination: when forecasts are imperfectly correlated, a weighted average can produce lower mean squared error than any individual model. Timmermann (2006) provides a thorough survey of combination methods, showing that equal-weighted combinations are often competitive with optimally weighted alternatives, partly because optimal weights must themselves be estimated from finite samples. Consistent with this, Rapach, Strauss and Zhou (2010) show that an equal-weighted combination of individual predictor models can deliver statistically significant improvements in R^2_{OOS} relative to the historical mean benchmark, even when the constituent models fail the out-of-sample test in isolation. In this paper, model uncertainty is addressed through a systematic evaluation of eight competing models within a standardised out-of-sample framework, with economic analysis based on the best-performing model in each sector-horizon setting. The paper also includes a regime-adaptive Bates–Granger combination model (M5), which extends the forecast-combination literature by allowing weights to adjust with recent return volatility.

A key challenge in long-horizon return forecasting is that statistical relationships between predictors and returns are unlikely to remain stable across the macroeconomic regimes that span a multi-decade sample. Dangl and Halling (2012) adapt the Kalman filter to the equity premium forecasting problem, showing that time-varying parameter (TVP) models can deliver substantially better out-of-sample performance than constant-coefficient regressions. Pettenuzzo and Timmermann (2011) find that models incorporating both parameter drift and discrete structural changes outperform models with parameter drift alone, particularly during periods of macroeconomic turbulence. This paper includes the time-varying parameter recursive least squares model (M7), directly motivated by this literature.

A rapidly growing literature applies machine learning methods to asset return prediction. Gu, Kelly and Xiu (2020) provide a large-scale comparison of machine learning approaches, including regression trees, random forests, gradient-boosted ensembles, and neural networks, applied to U.S. stock return prediction. They find that nonlinear models, particularly gradient-boosted trees and neural networks, can improve materially on linear benchmarks, largely by capturing interaction effects among predictors. The superior performance is concentrated in periods of high macroeconomic uncertainty, suggesting that nonlinearity in return dynamics is especially pronounced during stress regimes. This paper includes XGBoost (M8) as the sole machine learning model, chosen for its computational tractability within rolling-window estimation and its strong performance in structured financial data. Ridge and Elastic Net regularisation, applied within the ARDL framework, provide regularised linear benchmarks against which the incremental value of nonlinear modelling can be assessed.

This paper sits at the intersection of these strands by evaluating sector-specific macro-financial predictability across a broad model set, multiple forecast horizons, and a common pseudo-out-of-sample framework, while also examining how predictor screening methodology and portfolio construction alter the interpretation of statistical forecastability. The contribution is not simply a broader model comparison, but a demonstration that the conclusions drawn from such comparisons are sensitive to methodological choices that are rarely varied within a single paper.

3 Data

3.1 Sector Sub-Indices and Sample Period

The dataset covers monthly end-of-month observations for the eleven GICS equity sector sub-indices in the United States. The sample begins in January 2001 and extends through early 2026, yielding approximately 300 monthly observations per sector. All observations are total-return indices denominated in US dollars.

3.2 Dependent Variable

The dependent variable is the h-month forward cumulative excess return of sector s relative to the S&P 500 index:

$$\mathrm{DV}(s,t,h) = \sum_{i=1}^{h} \big[r(s,t+i) - r(\mathrm{SPX},t+i) \big]$$

where r(s, t+i) is the log monthly return of sector s in period t+i and r(SPX, t+i) is the corresponding S&P 500 log return. We evaluate four horizons: $h \in \{1,\ 3,\ 6,\ 12\}$ months. For h > 1, the forward sums produce overlapping observations, introducing positive serial correlation in the dependent variable. This is addressed through heteroskedasticity-and-autocorrelation-consistent (HAC) corrections and a block bootstrap. The use of a relative (sector minus market) rather than absolute return metric reflects the cross-sectional nature of sector rotation: a rotation strategy generates value through differential positioning, not through market timing. A model that correctly ranks sectors but has poor absolute return forecasts can still be commercially useful.

3.3 Predictor Universe

For each sector, we assemble a bespoke predictor set drawn from five families:
(i) macro-financial conditions; (ii) sentiment and activity surveys; (iii) valuation and earnings expectations; (iv) volatility and risk; and (v) sector-specific fundamentals. The raw predictor count ranges from 29 variables (Information Technology) to 49 (Energy), with a cross-sector mean of 36.9. Given the well-documented severity of multicollinearity in macroeconomic predictor sets, a sequential VIF filtering protocol is applied before any model is estimated, as described in Section 4.1; this predictor breadth improves sector coverage but also raises substantial multicollinearity challenges and the risk of look-ahead bias in predictor selection.

3.4 Summary Statistics

Tables 1a–1d report the mean, standard deviation, skewness, and excess kurtosis of the sector excess return series at each of the four forecast horizons, together with directional accuracy (the proportion of periods the sector outperformed the S&P 500). Statistics are computed on the full 2001–2026 sample.

Several features of the data warrant attention. First, all sectors exhibit near-zero mean excess returns at the 1-month horizon, consistent with limited short-horizon predictability. Information Technology is the notable exception, with a positive mean at all horizons. Second, cross-sectional volatility dispersion is large: Energy (Std = 5.57% at 1m, 21.89% at 12m) is substantially more volatile than Industrials or Consumer Staples. Third, excess kurtosis at 1-month is substantially reduced at the 12-month horizon, consistent with the normalising effect of return aggregation, while skewness patterns are more mixed across sectors. Fourth, directional accuracy at 1-month is close to 50% for most sectors, confirming that short-horizon sign prediction is near-random, making the 6- and 12-month model results in Section 5 more informative by comparison.

Sector	Mean (%)	Std Dev (%)	Skewness	Exc. Kurtosis	Dir. Acc. (%)
Comm. Services	-0.20	4.17	+0.69	+3.69	48.8
Consumer Discret.	+0.19	2.58	+0.32	+2.01	54.8
Consumer Staples	-0.13	3.27	+0.34	+1.06	47.2
Energy	+0.04	5.57	+0.47	+2.30	47.5
Financials	-0.21	3.26	-0.43	+4.70	47.9
Health Care	-0.10	3.08	-0.08	+1.48	48.2
Industrials	+0.03	2.31	+0.16	+1.00	51.2
Info. Technology	+0.39	3.54	-0.04	+5.06	53.5
Materials	+0.04	3.20	-0.03	+0.43	48.2
Real Estate	-0.17	4.33	+0.26	+4.87	45.4
Utilities	-0.27	4.42	-0.18	+0.22	47.5

Table 1a: Summary statistics: 1-Month cumulative excess returns (2001–2026). Dir. Acc. = % periods sector outperformed S&P 500.

Sector	Mean (%)	Std Dev (%)	Skewness	Exc. Kurtosis	Dir. Acc. (%)
Comm. Services	-0.67	6.97	+0.54	+1.58	46.8
Consumer Discret.	+0.53	4.38	-0.04	+1.39	**55.1**
Consumer Staples	-0.35	5.54	+0.55	+0.57	**41.5**
Energy	+0.08	9.72	+0.25	+1.93	45.5
Financials	-0.59	5.98	-0.39	+5.25	47.2
Health Care	-0.25	5.25	-0.13	+1.04	47.5
Industrials	+0.11	3.74	-0.06	+1.00	50.8
Info. Technology	+1.12	5.48	-0.09	+2.08	**62.8**
Materials	+0.15	5.35	-0.16	+1.02	52.8
Real Estate	-0.50	6.88	-0.44	+0.94	50.9
Utilities	-0.82	7.00	-0.03	-0.18	**43.2**

Table 1b: Summary statistics: 3-Month cumulative excess returns (2001–2026).

Sector	Mean (%)	Std Dev (%)	Skewness	Exc. Kurtosis	Dir. Acc. (%)
Comm. Services	-1.40	9.77	-0.24	+0.33	47.3
Consumer Discret.	+1.03	5.68	+0.06	+0.51	**55.4**
Consumer Staples	-0.80	7.72	+0.56	+1.03	**38.9**
Energy	-0.15	14.01	+0.57	+1.80	45.3
Financials	-1.18	8.70	-0.42	+3.74	46.3
Health Care	-0.50	6.87	+0.18	+0.53	48.0
Industrials	+0.08	4.74	-0.42	+0.55	51.0
Info. Technology	+2.43	7.09	-0.41	+1.50	**65.8**
Materials	+0.07	7.21	-0.15	+0.36	51.3
Real Estate	-1.03	9.75	-0.52	+2.33	47.9
Utilities	-1.76	9.14	-0.06	-0.28	**43.0**

Table 1c: Summary statistics: 6-Month cumulative excess returns (2001–2026).

Sector	Mean (%)	Std Dev (%)	Skewness	Exc. Kurtosis	Dir. Acc. (%)
Comm. Services	-3.04	13.64	-0.65	+0.26	50.0
Consumer Discret.	+2.08	7.64	-0.39	+0.06	**62.3**
Consumer Staples	-1.60	11.04	+0.61	+0.30	**37.0**
Energy	-0.43	21.89	+0.42	+1.19	**44.9**
Financials	-2.28	11.78	-1.02	+2.46	46.2
Health Care	-1.20	9.09	+0.03	+0.09	**44.5**
Industrials	+0.13	6.11	-0.12	+0.23	46.9
Info. Technology	+5.07	9.06	-0.02	-0.05	**74.7**
Materials	+0.00	9.44	+0.17	-0.28	47.6
Real Estate	-2.17	13.50	+0.12	-0.25	46.5
Utilities	-3.31	12.59	-0.14	-0.56	**41.8**

Table 1d: Summary statistics: 12-Month cumulative excess returns (2001–2026).

4 Methodology

4.1 Multicollinearity Treatment

Multicollinearity is a pervasive challenge in macroeconomic return forecasting. As documented in Table 2, every sector in this study exhibits extreme collinearity in the raw predictor set, with maximum VIF values across the eleven sectors ranging from 2,543 (Communication Services) to 8,607,658 (Financials). Ordinary least squares estimation with collinear predictors produces numerically unstable coefficient estimates, making model comparison and interpretation unreliable.

We address this through a two-stage protocol. In the first stage, a global sequential VIF filter is applied to each sector's predictor set at the outset. The algorithm iteratively identifies the predictor with the highest VIF and removes it from the universe, repeating until all remaining predictors have VIF below 10. This reduces the mean predictor count from 36.9 to 13.8 variables, eliminating approximately 63% of the original candidate set.

In the second stage, each individual model applies an additional within-window correlation screen at each rolling estimation step, selecting the top 6 predictors by absolute Pearson correlation with the training-window dependent variable. This dynamic selection adapts to time-varying relevance without re-introducing multicollinearity. Regularised models (Ridge, Elastic Net) provide a further layer of shrinkage, and the time-varying parameter model (TVP-RLS) with exponential forgetting continuously down-weights stale observations. The choice of six predictors was determined empirically as the largest count that kept within-window condition numbers at acceptable levels across all sectors and horizons.

The full-sample sequential VIF filter described above forms the basis for the main results reported in Section 5. Applying a VIF screen to the full available sample introduces a subtle look-ahead bias: the retained predictor set benefits from information spanning the entire out-of-sample evaluation window. Section 10.4 addresses this directly by implementing a recursive walk-forward VIF procedure in which predictor selection is re-run at each expanding estimation window using only information available at the forecast origin. This recursive procedure serves as the paper's principal real-time integrity check; where results diverge materially, greater weight is placed on the recursive estimates.

Sector	Raw Predictors	Retained	Reduction	Max Raw VIF
Communication Services	30	**8**	73%	2,543
Consumer Discretionary	38	**16**	58%	6,181,460
Consumer Staples	39	**15**	62%	6,079,456
Energy	49	**25**	49%	2,555
Financials	36	**10**	72%	8,607,658
Health Care	34	**9**	74%	937,482
Industrials	34	**15**	56%	2,822
Information Technology	29	**11**	62%	7,111
Materials	44	**20**	55%	8,505
Real Estate	43	**12**	72%	435,405
Utilities	30	**11**	63%	114,494

Table 2: VIF diagnostics by sector: variables before and after VIF filtering, percentage reduction, and maximum raw VIF.

4.2 Model Suite

We evaluate eight competing forecasting models plus a rolling-mean benchmark (M0), organised into five families. Table 3 summarises the full suite. All models use the same pre-filtered predictor universe and the same walk-forward evaluation protocol, ensuring comparability across specifications. Additional model variants were tested during development but excluded where they were consistently dominated out of sample by retained models of comparable complexity or failed to improve meaningfully on the benchmark. The model classes are grounded in established forecasting and regularisation literatures: ARIMA follows Box and Jenkins (1976), Ridge follows Hoerl and Kennard (1970), the ARDL framework follows Pesaran, Shin and Smith (2001), Elastic Net follows Zou and Hastie (2005), TVP-RLS with exponential forgetting follows Ljung and Söderström (1983), and XGBoost follows Chen and Guestrin (2016).

Code	Model	Description	Family
M0	**Rolling Mean**	120-month rolling historical mean benchmark	Benchmark
M1	ARIMA	Automatically selected ARIMA(p,d,q) using AIC	Time-Series
M2	ARDL-Ridge	ARDL with Ridge (L2) regularisation	ARDL
M3	BG-Fixed	Fixed-weight Bates–Granger forecast combination	Combination
M4	ARDL-Elastic Net	ARDL with Elastic Net (L1+L2) regularisation	ARDL
M5	BG-Regime (Adaptive)	Regime-adaptive Bates–Granger combination	Combination
M6	Threshold ARDL	Two-regime threshold ARDL	ARDL
M7	TVP-RLS	Time-varying parameter recursive least squares with exponential forgetting ($\lambda = 0.98$)	Time-Varying Parameter
M8	XGBoost	XGBoost with cross-validated early stopping	Gradient Boosting

Table 3: Full model suite (8 competing models + benchmark).

4.3 Out-of-Sample Evaluation Framework

All models are evaluated under a strict pseudo-out-of-sample walk-forward protocol. The training window is fixed at 120 months (10 years). At each step t, each model is estimated on observations $[t-120, t-1]$ and produces a forecast for the h-period ahead excess return. The window then advances by one month and the process repeats. No information beyond the forecast origin is used in estimation, ensuring the absence of look-ahead bias, in line with the real-time econometric protocol articulated by Pesaran and Timmermann (2005). The evaluation sample begins in January 2011 and yields roughly 180 forecast origins per sector, with the usable sample slightly shorter at the 12-month horizon as forward returns approach the sample end.

The primary metric is the Campbell and Thompson (2008) out-of-sample R^2 $\left(R^2_{\text{OOS}}\right)$:

$$R^2_{\text{OOS}} = 1 - \frac{\text{MSE(model)}}{\text{MSE(benchmark)}}$$

A positive R^2_{OOS} indicates that the model produces smaller forecast errors than the benchmark; a negative value indicates underperformance. Secondary metrics are directional accuracy (the proportion of periods in which the forecast and realisation share the same sign) and root mean squared forecast error (RMSFE), reported in percentage points. Economic value is assessed via the certainty-equivalent return (CEQ) gain relative to the benchmark, annualised for a mean-variance investor with risk aversion $\gamma = 3$.

Statistical significance is evaluated in two ways. First, block bootstrap confidence intervals are constructed for R^2_{OOS} using 500 replications and a block size of 6 months. Second, pairwise Diebold and Mariano (1995) tests are reported with the finite-sample correction of Harvey, Leybourne and Newbold (1997). A block size of 6 months is conservative at short horizons; at the 6- and 12-month horizons, where overlapping returns induce stronger serial dependence, the resulting confidence intervals should be interpreted with this in mind.

4.4 Interpreting Statistical and Economic Forecastability

The empirical analysis reports five complementary measures of model performance and economic usefulness, each answering a different question. R^2_{OOS} captures benchmark-relative forecast accuracy. Directional accuracy (DA) measures the fraction of periods in which the model correctly predicts the sign of the excess return, independently of magnitude. Root mean squared forecast error (RMSFE) provides an absolute scale for forecast errors. Certainty-equivalent return (CEQ) maps forecasts into unconstrained mean-variance portfolio weights and measures the guaranteed return a mean-variance investor would accept in exchange for the forecasting strategy. It is a stringent test because unconstrained weights amplify forecast errors when estimated conditional variance is low. Target-volatility excess return complements CEQ as a constrained economic measure: it scales positions to a fixed annualised volatility target, representing an implementation that is more robust to the weight-amplification problem inherent in unconstrained allocation. Readers should bear in mind that statistical forecastability (such as positive (R^2_{OOS}) or DA above 50%) does not automatically imply economic value, and weak CEQ under unconstrained allocation does not by itself imply that the underlying signal lacks economic usefulness.

5 Empirical Results

5.1 Horizon Effects

A central pattern in the results is that measured forecastability increases with forecast horizon. All results in this section are reported under the full-sample VIF filter described in Section 4.1; the recursive walk-forward correction and its implications for the cross-sector ranking are discussed in Section 10.4. At the 1-month horizon, even the best-available model produces a mean R^2_{OOS} of only (-0.006) across the eleven sectors, with only four sectors achieving positive values. At three months, the mean best-model R^2_{OOS} rises to (+0.008) and seven sectors are positive. A marked inflection occurs at 6 months, where the mean best-model R^2_{OOS} reaches (+0.062) and all eleven sectors yield positive values simultaneously. At twelve months, the mean best-model R^2_{OOS} rises further to (+0.163); nine of eleven sectors remain positive, indicating a slight narrowing in breadth even as the cross-sector mean continues to increase (Figures 1 and 2).

Part of this horizon gradient could in principle be mechanical rather than a pure increase in predictive skill. Under overlapping (h)-month cumulative returns, the variance ratio between forecast error and benchmark error depends on the serial dependence structure of the underlying monthly return series, and a zero-skill forecast can in some cases outperform a rolling-mean benchmark at longer horizons. A simple Monte Carlo with AR(1) persistence calibrated to the 0.10–0.30 range typical of monthly equity returns (Lo and MacKinlay, 1988) and a 120-month rolling benchmark produces ($|R^2_{\text{mech}}| \leq 0.04$) at the 12-month horizon, materially smaller than the reported (+0.530) for Health Care, (+0.312) for Communication Services, and (+0.253) for Industrials. The mechanical contribution therefore cannot account for the headline 12-month results in the strongest sectors, although it remains material for borderline results such as Real Estate (+0.040) and Utilities (+0.017), which should be interpreted with additional caution.

This horizon pattern is economically plausible. Monthly sector excess returns are heavily influenced by high-frequency noise, including earnings surprises, fund flows, geopolitical shocks, and sentiment reversals, that is difficult to forecast from lagged macro variables. Over six to twelve months, part of this transient variation averages out, leaving a larger role for the slower-moving fundamental component that macro predictors are designed to capture. This interpretation is consistent with the theoretical predictions of Campbell and Shiller (1988) and Fama and French (1988): while short-horizon returns are dominated by noise, longer-horizon returns embed more predictable variation linked to discount-rate and cash-flow fundamentals.

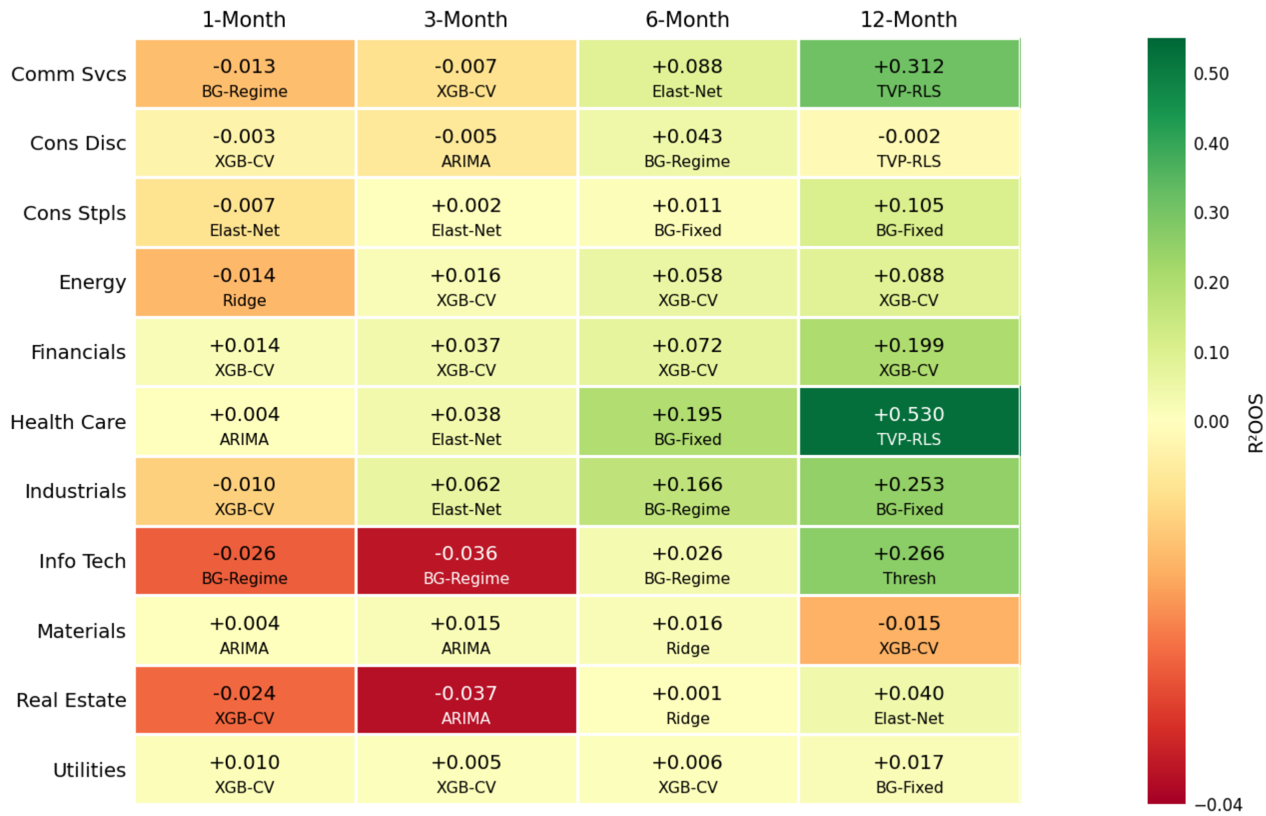

Figure 1: Out-of-Sample R^2_{OOS} by Sector and Horizon. Best model per cell across all eight competing models (M0 rolling-mean benchmark excluded). Cell colour indicates R^2_{OOS} value: green = beats rolling-mean benchmark, red = underperforms. The winning model label is shown in each cell.

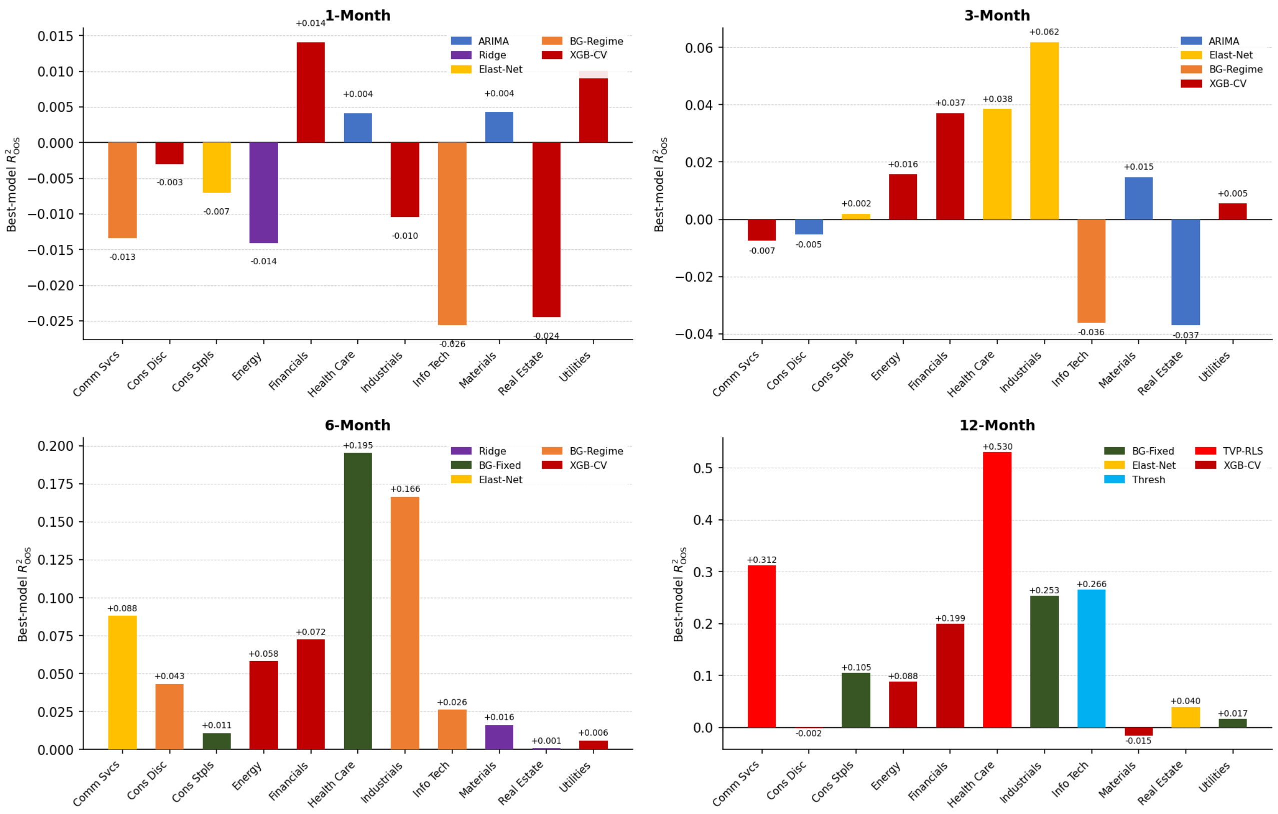

Figure 2: Best-Model R^2_{OOS} by Sector and Forecast Horizon. Each bar shows the best-performing model's R^2_{OOS} for that sector-horizon cell, colour-coded by the winning model. Positive values indicate the winning model beats the rolling-mean benchmark. Eight models compete (M0 excluded).

5.2 Model Rankings and Win Counts

Across all 44 sector-horizon pairs, M8 (XGBoost CV) records 15 wins, the highest total of any individual model. The next highest totals are BG-Regime (M5, 6 wins), ARDL-Elastic Net (M4, 6 wins), ARIMA (M1, 5 wins), BG-Fixed (M3, 5 wins), Ridge (M2, 3 wins), TVP-RLS (M7, 3 wins), and Threshold ARDL (M6, 1 win). These counts are computed using a consistent rolling-mean benchmark across all models.

Win count, however, is not equivalent to average performance. Of M8's 15 wins, 9 occur at the 1- and 3-month horizons, where absolute R^2_{OOS} magnitudes are close to zero and differences between competing models are economically small. The remaining 6 wins occur at the 6- and 12-month horizons, including 12-month R^2_{OOS} values of $(+0.199)$ in Financials and $(+0.088)$ in Energy. Raw win counts at short horizons therefore reflect competition among models that are all performing weakly in absolute terms. Magnitude-weighted or horizon-weighted comparisons place greater emphasis on the combination and TVP families, which win fewer cells overall but perform relatively better at the longer horizons where forecastability is strongest (Table 4 and Figures 1 and 3).

Model	1m wins	3m wins	6m wins	12m wins	Total (of 44)
M8 XGBoost CV	5	4	3	3	**15**
M4 Elastic Net	1	3	1	1	**6**
M5 BG-Regime	2	1	3	0	**6**
M1 ARIMA	2	3	0	0	**5**
M3 BG-Fixed	0	0	2	3	**5**
M2 Ridge	1	0	2	0	**3**
M7 TVP-RLS	0	0	0	3	**3**
M6 Threshold ARDL	0	0	0	1	**1**

Table 4: Model win counts across all 44 sector-horizon cells.

5.3 Model Family Analysis

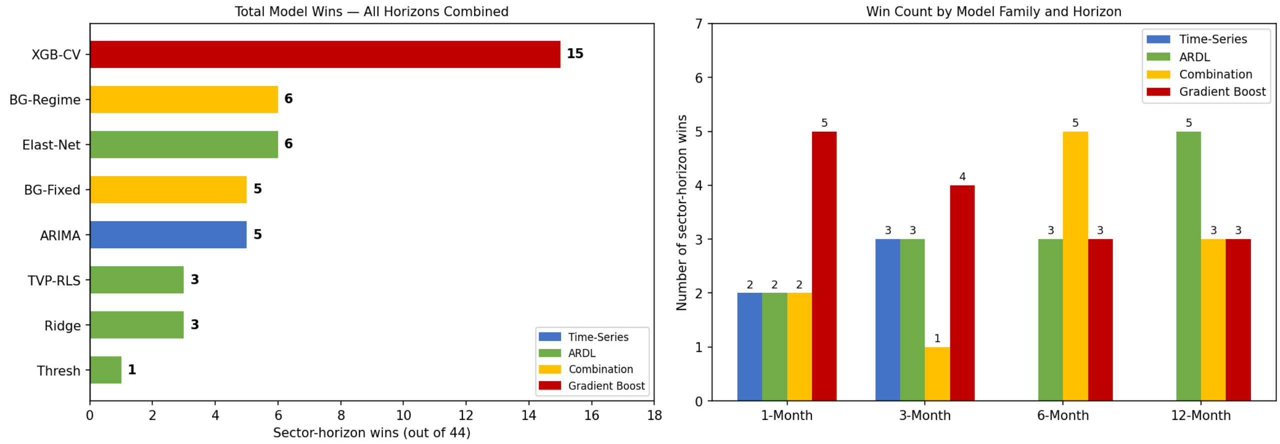

Figure 3: Model performance rankings by win count. Left panel: total wins by individual model across the 44 sector-horizon cells. Right panel: wins by model family and forecast horizon.

Disaggregating performance by model family reveals a clear pattern of horizon specialisation. XGBoost (M8), the paper's sole gradient-boosting specification, records the most wins at the 1- and 3-month horizons, where forecast signals are weak in absolute magnitude and non-linear interactions offer a small advantage in raw win counts at short horizons. By contrast, the combination family performs more strongly at longer horizons, where averaging across model specifications appears to reduce forecast variance. Within the combination family, BG-Regime (M5) achieves a mean 12-month R^2_{OOS} of $+0.072$ across the eleven sectors. Within the ARDL family, ARDL-Elastic Net (M4) records the highest number of wins, while Threshold ARDL (M6) remains a niche specification whose single win occurs in Information Technology at the 12-month horizon.

The TVP-RLS model (M7) achieves the single highest point estimate in the main full-sample specification, with Health Care reaching $R^2_{\text{OOS}}=+0.530$ at 12 months, but this result should be interpreted alongside the recursive VIF correction discussed in Section 10.4. More broadly, TVP-RLS performs best in sectors where predictor-return relationships appear more persistent and regime-sensitive, but it is less stable in noisier sectors where time variation adds estimation variance without sufficient compensating signal. This is consistent with the standard trade-off in exponentially forgetting estimators: greater adaptability to structural change, but also greater sensitivity to noise.

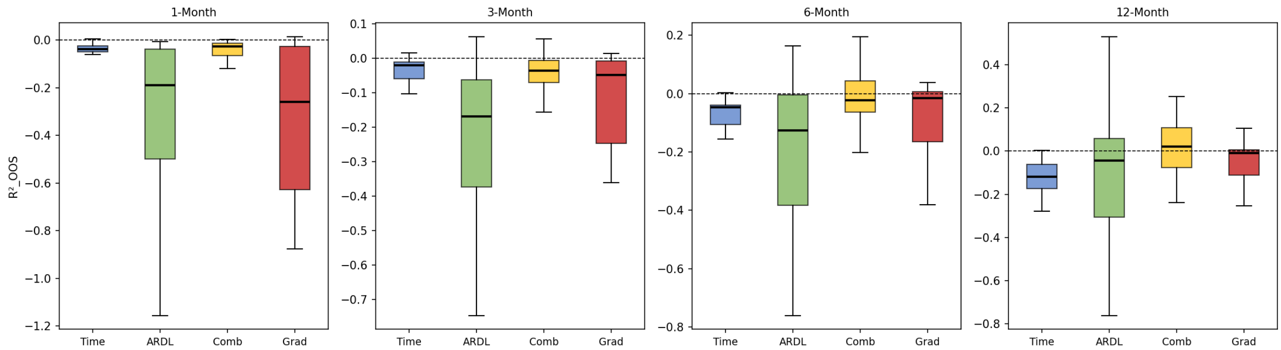

Figure 4: Distribution of R^2_{OOS} by model family and forecast horizon (box=IQR; horizontal line=median; outliers clipped).

Together, Figures 3 and 4 show that short-horizon leadership in raw win counts is concentrated in XGBoost, whereas longer-horizon performance shifts toward the combination and ARDL families, which are more competitive at the horizons where measured forecastability is strongest.

5.4 Directional Accuracy

For practical sector rotation, directional accuracy, defined as the fraction of periods in which the model correctly predicts the sign of the excess return, is sometimes viewed as more actionable than R^2_{OOS} magnitude. Interpretation requires care, however. At the 1-month horizon, mean directional accuracy across sectors is 50.6%, essentially random. At twelve months, the mean rises to 65.9%, with Information Technology reaching 89.5% (Threshold ARDL, M6) and Health Care 80.1% (TVP-RLS, M7). These headline figures should be interpreted relative to a drift-only benchmark rather than a universal 50% coin-flip rate. When a sector exhibits a persistent positive or negative sign frequency, a naive rule that always predicts the more common sign can achieve directional accuracy materially above 50%. The relevant descriptive benchmark is therefore the better of two naive rules: always predict positive or always predict negative. Only the excess of model directional accuracy over that benchmark is plausibly attributable to conditional timing skill.

Applying this more conservative benchmark tightens the interpretation of several cells. Consumer Discretionary's 67.4% directional accuracy at twelve months is notable, but if the sector's excess returns are positive in roughly 62.3% of periods, a naive 'always positive' rule already delivers 62.3% directional accuracy; the model's margin above that benchmark is therefore about 5 percentage points, rather than the 17 points implied by a 50% baseline. Real Estate's unconditional positive share is below 50%, so the better naive rule is 'always negative' (53.5%); the model's 73.7% DA (R^2_{OOS}=+0.040) exceeds that benchmark by about 20 percentage points, a larger margin than in the Consumer Discretionary case and more plausibly skill-driven. Information Technology's 89.5% DA remains impressive, but the sector's unconditional positive share of 74.7% means the relevant margin above the drift-only benchmark is about 15 percentage points, smaller than the headline suggests. Industrials at 12 months (67.3% DA, R^2_{OOS}=+0.253) sits against an unconditional positive share of 46.9% (naive best=53.1%), giving a margin of about 14 percentage points; together with its elevated R^2_{OOS}, Industrials remains one of the clearest cases in which directional accuracy and magnitude skill are jointly elevated. Health Care at 80.1% DA retains a large margin over its naive sign benchmark (unconditional positive share 44.5%, naive best=55.5%, margin≈25 percentage points), reinforcing its status as the strongest sector in the main full-sample specification, though the magnitude of that conclusion remains subject to the VIF look-ahead caveat discussed in Section 10.4.

Readers should therefore treat directional accuracy as a complement to R^2_{OOS}, not as an independent source of evidence. In sectors with persistent drift, headline DA can materially overstate genuine timing skill unless it is evaluated relative to an appropriate naive sign benchmark (see Figure 5 for a sector-by-horizon visualisation). The sign frequencies used to calibrate these naive benchmarks are drawn from Table 1d, which reflects the full 2001–2026 sample; frequencies during the OOS evaluation window may differ, and for Real Estate in particular the discrepancy is likely material.

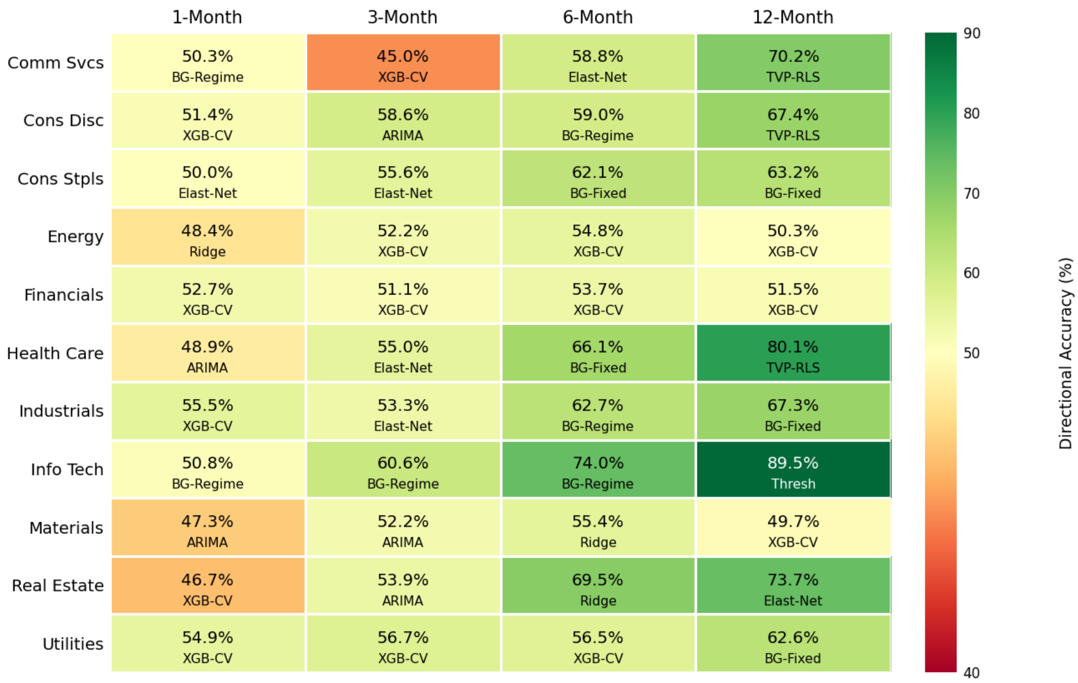

Figure 5: Directional accuracy (%) by sector and horizon. Best model per cell. Values should be interpreted relative to a sector-specific naive sign benchmark rather than a universal 50% threshold.

5.5 Bootstrap Confidence Intervals

Circular block bootstrap confidence intervals (B=500, block size=6) are used to assess whether positive R^2_{OOS} point estimates reflect genuine predictive skill or sampling variation. A 6-month block size is adequate for the 1- and 3-month horizons, but it is short relative to the dependence induced by overlapping long-horizon returns. At the 12-month horizon in particular, the resulting confidence intervals are therefore likely somewhat tighter than those produced by a longer block length, so borderline-significant 12-month cells should be interpreted with additional caution. The Health Care result at R^2_{OOS}=+0.530 is far enough from zero that reasonable increases in block length are unlikely to overturn significance, but results such as Real Estate (+0.040, 12m) and Utilities (+0.017, 12m), which lie close to the zero boundary, are more sensitive to this calibration.

Under this bootstrap design, statistically significant results, defined as bootstrap 90% confidence intervals with lower bounds strictly above zero are concentrated primarily at the 12-month horizon and in Health Care. Health Care achieves statistically significant positive R^2_{OOS} for five of the eight competing models at 12 months (M2, M3, M4, M5, M7). In particular, M7 records R^2_{OOS}=+0.530 with a 90% CI of [0.33, 0.65], while M5 records +0.448 with CI [0.31, 0.55]. The three exceptions within the suite are M1 (ARIMA, R^2_{OOS}=–0.118), M6 (Threshold ARDL, R^2_{OOS}=+0.032, CI spanning zero), and M8 (XGBoost, CI spanning zero), all of which are outperformed by the combination and TVP specifications at this horizon. These full-sample Health Care results should, however, be read alongside the recursive VIF correction in Section 10.4, which materially reduces the sector's headline point estimate.

The concentration of statistically significant results in Health Care and at the 12-month horizon is informative. Under the present inference framework, it suggests that many of the positive R^2_{OOS} point estimates observed at shorter horizons are statistically difficult to distinguish from noise, even when positive in sign. The more defensible conclusion is therefore that economically meaningful forecastability is most evident at the 12-month horizon and in a limited subset of sectors, rather than broadly across the full sector-horizon panel.

5.6 Forecast Error and Model Uncertainty

Root Mean-Squared Forecast Error (RMSFE) measures the absolute precision of each model's point forecasts and provides a complementary perspective to R^2_{OOS}, which measures skill relative to the rolling-mean benchmark. RMSFE increases monotonically with forecast horizon across all eleven sectors: at the 1-month horizon the cross-sectional range runs from 2.21% (Industrials) to 6.14% (Energy), while at twelve months that range widens dramatically to 5.05% (Industrials) to 25.01% (Energy). The almost five-fold dispersion in twelve-month RMSFE is driven almost entirely by differences in underlying sector return volatility rather than differences in model quality.

The ratio of best-model RMSFE to the rolling-mean baseline RMSFE (RMSFE/M0 in Table 5) provides a normalised measure of forecast precision that removes the confounding effect of sector volatility. On this metric, Health Care is the standout sector: its twelve-month RMSFE/M0 ratio of 0.686 implies a 31.4% reduction in absolute forecast error relative to the historical mean, consistent with its R^2_{OOS} of +0.530; as discussed in Section 10.4, however, this figure reflects the full-sample VIF specification and should be treated as an upper bound on the sector's genuine forecast precision. Communication Services (0.829), Information Technology (0.857), and Industrials (0.864) also achieve meaningful absolute error reduction. At the other end, Consumer Discretionary (1.001) and Materials (1.008) exhibit RMSFE/M0 ratios above or at unity, confirming that the rolling mean is an essentially unbeatable baseline for these sectors on an absolute error basis at twelve months. Utilities (0.992) falls marginally below unity but is effectively equivalent to the rolling-mean baseline, with BG-Fixed achieving only a 0.8% absolute error reduction.

The normalised ratio also corrects a potentially misleading raw comparison: Energy's twelve-month RMSFE of 25.01% appears dramatically worse than Industrials' 5.05%, but Energy's M0 baseline is 26.19% and Industrials' baseline is only 5.85%, implying similar normalised difficulty. Energy's RMSFE/M0 of 0.955 is actually better than Consumer Discretionary (1.001) and Materials (1.008) on this metric. The volatility of sector returns is structural rather than a function of model misspecification: models cannot reduce Energy's RMSFE to Industrials' levels any more than they can reduce crude oil price volatility.

For practical portfolio construction, RMSFE-based confidence intervals provide an illustrative heuristic bound around the forward-looking point forecasts in Section 8.5. Under the approximation of normally distributed forecast errors, which cannot be formally verified for overlapping multi-period returns but provides a tractable bound, a 68% confidence interval is given by the point forecast ±1 RMSFE and a 90% interval by ±1.64 RMSFE. Given the well-documented fat tails and negative skewness of equity sector return distributions, these bounds should be interpreted as indicative rather than exact, and asymmetric downside scenarios should be sized conservatively relative to the symmetric interval.

Table 5 reports best-model RMSFE, the rolling-mean benchmark RMSFE, and the normalised RMSFE/M0 ratio at the 12-month horizon for all eleven sectors.

Sector	**Best Model (12m)**	R^2_{OOS}	**M0 Benchmark RMSFE**	**Best-Model RMSFE**	**RMSFE / M0**
Communication Services	TVP-RLS (M7)	+0.312	13.59%	11.27%	0.829
Consumer Discretionary	TVP-RLS (M7)	-0.002	7.20%	7.21%	1.001
Consumer Staples	BG-Fixed (M3)	+0.105	9.97%	9.43%	0.946
Energy	XGBoost CV (M8)	+0.088	26.19%	25.01%	0.955
Financials	XGBoost CV (M8)	+0.199	10.53%	9.42%	0.895
Health Care	TVP-RLS (M7)	+0.530	9.31%	6.38%	0.686
Industrials	BG-Fixed (M3)	+0.253	5.85%	5.05%	0.864
Information Technology	Threshold ARDL (M6)	+0.266	8.04%	6.89%	0.857
Materials	XGBoost CV (M8)	-0.015	8.48%	8.54%	1.008
Real Estate	Elastic Net (M4)	+0.040	11.29%	11.06%	0.980
Utilities	BG-Fixed (M3)	+0.017	12.83%	12.72%	0.992

Table 5: Best-model RMSFE versus rolling-mean benchmark RMSFE at the 12-month horizon. RMSFE/M0 < 1 indicates the model reduces absolute forecast error relative to the historical mean; values ≥ 1 indicate the benchmark cannot be beaten on an absolute error basis.

5.7 Economic Value Analysis (CEQ Metric)

To translate R^2_{OOS} into portfolio-level economic value, we compute the certainty-equivalent return (CEQ) gain for a mean-variance investor with risk-aversion coefficient $\gamma = 3$. At each OOS step t, a portfolio weight proportional to the model signal is formed:

$$w_t = \frac{f_t}{\gamma\,\sigma_t^2}, \quad \text{capped at } \pm 2 \times \text{leverage},$$

and the active portfolio return $p_t = w_t \times a_t$ is computed. The annualised CEQ gain is:

$$\Delta\text{CEQ} = 12 \times \left[E(p_t) - \frac{\gamma}{2}\,\text{Var}(p_t)\right].$$

Across the seven models for which CEQ is computed and all eleven sectors, only 9 of 308 evaluated cells produce a positive annualised CEQ; XGBoost (M8) is excluded as its ensemble structure does not produce an analytic forecast signal suitable for mean-variance weight construction. The positive cells are concentrated at short horizons in Information Technology and Consumer Discretionary, where active returns are sufficient to offset the variance penalty. All seven models produce negative cross-sector mean CEQ at 12 months, ranging from -25.8% (Elastic Net, M4) to -62.0% (Threshold ARDL, M6). The prevalence of negative values reflects the stringent nature of the CEQ test: unlike R^2_{OOS}, a positive CEQ requires not only directional forecast accuracy but also forecast precision sufficient to offset active position variance at $\gamma = 3$. Even models with positive R^2_{OOS} produce mean active returns insufficient to compensate for the variance penalty at this level of risk aversion. As a relative comparison, Elastic Net (M4, -25.8%) and BG-Regime (M5, -31.4%) are the most economically efficient models, while Threshold ARDL (M6, -62.0%) and TVP-RLS (M7, -44.2%) impose the greatest variance penalty.

For practical implementation, these results suggest using smaller active weights than the theoretical optimum, or overlaying forecast signals onto existing benchmark weights rather than treating them as standalone allocation decisions. The Health Care sector, which records the strongest statistical forecastability, is subject to the full-sample VIF look-ahead caveat discussed in Section 10.4; its CEQ figures should be interpreted accordingly. These unconstrained CEQ results should therefore be read as a deliberately stringent implementation test rather than as definitive evidence that the underlying signals lack economic usefulness.

5.8 Pairwise Diebold and Mariano Tests

Pairwise DM tests with Harvey, Leybourne and Newbold finite-sample correction are computed across all 28 unique model pairs within the eight-model suite, for each sector-horizon cell. Of 1,232 total pairs evaluated, 261 (21.2%) are statistically significant at the 5% level, with 61 significant pairs at the 12-month horizon (approximately matching the overall 21.2% rate across horizons, so significance is broadly spread rather than disproportionately concentrated at 12m).

Two patterns are particularly informative. In Financials at 12m, Threshold ARDL (M6) is significantly outperformed by six of the seven other retained models at the 5% level (M1, M2, M3, M4, M5 all beat M6 at $p<0.001$; M7 beats M6 at $p=0.029$); M8 (XGBoost) falls just short of the 5% threshold ($p=0.056$, significant at 10%). This is consistent with poor structural fit where the regime split leaves insufficient regime-specific training observations. TVP-RLS (M7) is also significantly beaten by Ridge, BG-Fixed, Elastic Net, and BG-Regime in Financials at 12m (e.g., M5 vs M7: DM stat=–3.067, $p=0.002$), suggesting that model combination with adaptive weighting is a more reliable approach to regime adaptation than coefficient-level time variation in this sample.

In Health Care at 12m, macro-enriched models systematically outperform the pure time-series benchmark: M2, M3, M4, M5, and M7 all significantly beat ARIMA (M1) at $p < 0.05$ (e.g., M5 vs M1: DM stat $= 3.254$, $p = 0.001$; M3 vs M1: DM stat $= 3.228$, $p = 0.001$), confirming that the macro-predictor advantage in Health Care is statistically robust rather than a sampling artefact.

6 Sector-Level Analysis

This section provides a sector-by-sector breakdown of the predictor universe retained after sequential VIF filtration. For each sector we report the forecast-skill headline, the economic rationale for the retained set, and the retained predictor table. Cross-sector predictor importance mechanisms and transmission channels are consolidated in Table 8 (Section 7.2) and Section 6.12; readers interested primarily in cross-sector patterns may proceed directly to those sections. The filtration process therefore does more than reduce dimensionality: it reveals which classes of predictors survive realistic collinearity constraints in sector-specific forecasting environments, and this is itself an empirical finding about the structure of sector predictability.

6.1 Communication Services

Forecast skill is concentrated at the 12-month horizon: TVP-RLS (M7) achieves $R^2_{\mathrm{OOS}} = +0.312$ with directional accuracy of 70.2%, making it the second-most-forecastable sector in the study after Health Care at that horizon. No model beats the rolling-mean benchmark at 1m or 3m; BG-Regime (M5) is the least-negative at 1m (-0.013) and XGB_CV at 3m (-0.007). Skill first emerges at 6m, where Elastic Net (M4) takes the win at $R^2_{\mathrm{OOS}} = +0.088$ (DA 58.8%), closely followed by BG-Regime (+0.087) and BG-Fixed (+0.071).

The retained 8 provide low-collinearity coverage of advertising demand (RSTAXAG%, CorpProfit YoY%, E-Commerce), rate conditions (GTII10), risk appetite (VIX), sector leverage (NET_DEBT_TO_EBITDA), valuation (SPX P/E), and consumer income (PCE CYOY%). Top 12m predictors: RSTAXAG%, VIX, CorpProfit YoY%, GTII10, SPX P/E (see Table 8 for mechanisms). Table 6.1 lists the retained set.

Variable (Bloomberg)	**Category**	**Inclusion Rationale**
CorpProfit YoY%	Macro : Demand	Corporate profit cycles directly drive advertiser budgets; advertising is the primary revenue source for media and digital platform constituents.
RSTAXAG% Index	Macro : Demand	Retail sales ex-autos/gas proxies consumer discretionary spending and foot-traffic to retail channels that advertise heavily on Communication Services platforms.
E-Commerce $M / YoY%	Sector Driver	E-commerce dollar volume and growth directly drives digital advertising demand. Google, Meta, and Amazon dominate the sector and are e-commerce advertising beneficiaries.
VIX Index	Risk / Financial	Advertising budgets are cut quickly during stress periods; VIX proxies risk appetite and corporate willingness to spend on marketing.
GTII10 Govt	Rates	Real 10-year Treasury yield is the primary discount rate for long-duration growth assets. Communication Services valuations are highly sensitive to real rate movements.
NET_DEBT_TO_EBITDA	Sector Fundamental	Sector aggregate leverage determines refinancing cost sensitivity and constrains M&A activity, which has historically driven valuation re-ratings in media consolidation cycles.
SPX P/E	Valuation	Broad market P/E captures the valuation cycle, relevant for a sector with premium multiples that tend to compress during de-rating episodes.
PCE CYOY Index	Macro : Demand	Consumer spending growth sustains the digital economy; streaming subscriptions and wireless services are sensitive to real spending power.

Table 6.1: Communication Services: Retained predictors after VIF filtration (threshold = 10).

6.2 Consumer Discretionary

No model reliably beats the rolling-mean benchmark at most horizons. The 6m horizon produces the sector's strongest result: BG-Regime (M5) achieves $R^2_{\mathrm{OOS}} = +0.043$ (DA 59.0%), with BG-Fixed (M3) close behind (+0.041). At 12m, TVP-RLS (M7) records the best result with $R^2_{\mathrm{OOS}} = -0.002$ and directional accuracy of 67.4%, reflecting the sector's dependence on idiosyncratic consumer-spending dynamics that are difficult to forecast from aggregate macro-financial indicators.

The retained 16 cover credit (CICRRRTL), income (Savings YoY%, AHE MoM%, JOLTS), prices (GasCPI MoM%/YoY%), monetary policy (FDTR, GTII10), and sector activity (AutoSales YoY%, RSTAXAG%, RSTAMOM, CONSSENT/CONCCONF). Top 12m predictors: CICRRRTL, Savings YoY%, GasCPI MoM%, FDTR, GasCPI YoY% (see Table 8). Table 6.2 lists the retained set.

Variable (Bloomberg)	**Category**	**Inclusion Rationale**
CICRRRTL Index	Credit / Consumer	Revolving credit outstanding proxies household leverage and spending capacity beyond current income; leads consumption by 2–4 quarters.
Savings YoY%	Macro : Income	Personal savings rate reflects household financial resilience and the buffer available for discretionary spending.
GasCPI MoM% / YoY%	Input Cost	Gasoline prices act as a direct transfer payment from consumers to energy producers, with an immediate negative effect on disposable income for discretionary goods.
FDTR Index	Monetary Policy	Fed Funds rate determines consumer credit costs with a 6–12-month transmission lag; rising FDTR raises revolving credit costs and auto loan rates.
AutoSales YoY%	Sector Driver	Auto sales are the single largest discretionary purchase category and a direct leading indicator for the autos and parts sub-sector.
JOLTS YoY%	Labour Market	Job openings proxy labour demand and the tightness of the labour market, leading wage growth and consumer confidence.
AHE MoM%	Labour Market	Average hourly earnings growth determines real purchasing power; the primary income variable for the broad consumer.
PCE YoY%	Macro : Demand	Personal consumption expenditure captures the aggregate demand level that drives sector revenue.
RSTAXAG% / RSTAMOM	Retail Activity	Retail sales ex-autos/gas (YoY% and MoM%) provide concurrent sector activity as both level and momentum signals.
GTII10 Govt	Rates	Real 10-year yield captures the discount rate effect on consumer balance sheet valuations and durable goods financing costs.
SAARTOTL / AutoSales	Auto Sector	Seasonally-adjusted annualised auto sales rate provides the headline automotive demand indicator.
CONSSENT / CONCCONF	Sentiment Survey	University of Michigan and Conference Board consumer confidence surveys retained as orthogonal components after EPS removal.

Table 6.2: Consumer Discretionary: Retained predictors after VIF filtration (threshold = 10).

6.3 Consumer Staples

BG-Fixed (M3) wins at both the 6m ($R^2_{\text{OOS}} = +0.011$) and 12m (+0.105, DA 63.2%) horizons, with skill building over horizon as the sector's defensive cash-flow characteristics become predictable from macro-financial conditions at longer lags. BG-Regime (M5) and TVP-RLS (M7) are competitive at 12m (+0.083 and +0.082 respectively), suggesting the sector rewards structural macro-financial relationships over flexible non-linear approaches.

The retained 15 span input costs (PPI Food, GasCPI, PPI XCHG), consumer income and demand (PCE YoY%, Savings YoY%, RSTAMOM, AHE MoM%), inflation (CPI YoY%), rates (GTII10), market sensitivity (Sector Beta), and PCE sub-indices (PCE NDRB, PCENDur). Top 12m predictors: PPI Food YoY%, Savings YoY%, PCE YoY%, GasCPI YoY%, RSTAMOM (see Table 8). Table 6.3 lists the retained set.

Variable (Bloomberg)	**Category**	**Inclusion Rationale**
PPI Food MoM% / YoY%	Input Cost	Food producer prices directly determine COGS for staples manufacturers; margin expansion or compression is the primary driver of excess returns in this sector.
PCE YoY%	Macro : Demand	Personal consumption expenditure captures aggregate demand, which determines the volume side of staples revenue even when prices are inelastic.
Savings YoY%	Household Finance	Savings rate signals household financial resilience; a rising savings rate benefits staples relative to discretionary by shifting consumption toward necessities.
GasCPI MoM% / YoY%	Input/Transport Cost	Gasoline prices affect transportation costs within the supply chain and household disposable income for non-essential spending.
RSTAMOM Index	Retail Activity	Retail sales MoM% provides the highest-frequency concurrent signal of actual sector volume activity.
CPI YoY%	Inflation	Broad inflation determines the pricing environment; high inflation allows price increases that temporarily boost revenue but compresses real demand.
AHE MoM%	Labour Cost	Average hourly earnings are a significant operating cost for food and household products manufacturers; rising wages compress margins unless passed through.
PPI XCHG Index	Input Cost	Producer price index for exchange-traded commodities captures the basket of raw material inputs relevant to staples production.
Sector Beta (36M)	Market Risk	Rolling 36-month beta captures the sector's time-varying defensive character; lower beta periods signal enhanced defensive rotation appeal.
GTII10 Govt	Rates / Duration	Real yield is relevant for staples as a bond-proxy sector; rising real rates compress the relative valuation premium that defensive sectors command.
PCE NDRB / PCENDur	PCE Subindex	Non-durable and durable PCE sub-indices provide a decomposition of demand into components directly aligned with the sector's product categories.

Table 6.3: Consumer Staples: Retained predictors after VIF filtration (threshold = 10).

6.4 Energy

Energy is one of the weaker sectors for short-to-medium horizon forecasting: no model achieves positive R^2_{OOS} at 1m, with Ridge (M2) recording the least-negative result (-0.014). XGB_CV is the dominant model from 3m onward, winning at 3m (+0.016), 6m (+0.058), and 12m (+0.088). VIF elimination removed price-level versions of all commodity series, nominal Treasury yields, and duplicate oil benchmarks.

VIF elimination removed most non-stationary price-level series and redundant rate constructs. Exceptions include spot and forward crude levels (CL1, CO1, CL12), retained alongside their MoM% transforms given the documented informational content of absolute crude price levels for E&P profitability thresholds, and the nominal 10-year Treasury (USGG10YR) and LIBOR average (ILM3NAVG), retained as orthogonal complements to GTII10 where the absolute yield level carries independent information for energy company valuation. The retained 25 cover the full commodity price chain, production activity, supply-demand balance, financial conditions, and sector leverage. Top 12m predictors: CL12 MoM%, XW1 MoM%, CL1 MoM%, NG1 MoM%, NG1 YoY% (see Table 8). Table 6.4 lists the retained set.

Variable (Bloomberg)	**Category**	**Inclusion Rationale**
CL1 / CL12 MoM%	Commodity Price	Front-month and 12-month forward WTI crude price momentum; CL12 captures the market's forward expectation for oil prices over the forecast horizon.
CO1 MoM%	Commodity Price	Brent crude MoM% provides the international benchmark complementary to WTI, relevant for integrated oil majors with global exposure.
NG1 MoM% / YoY%	Commodity Price	Natural gas price momentum and annual change; gas is a primary revenue driver for E&P and midstream and an input cost for refiners.
BAKEOIL / BAKETOT YoY%	Activity: Upstream	Baker Hughes rig counts (oil and total) proxy upstream production activity; rig count growth leads production volumes by 3–6 months.
DOESCRUD YoY%	Inventory	EIA crude oil inventory YoY% measures supply-demand balance; above-average inventories suppress price and sector returns.
DOENUST1 YoY%	Inventory	Total petroleum inventories provide a broader supply overhang measure across crude and refined products.
XW1 MoM% / YoY%	Commodity Price	WTI 12-month strip price captures forward curve slope and market structure (contango/backwardation), directly relevant to E&P hedging economics.
HY OAS	Credit Spreads	High-yield OAS is critical for Energy, the largest sector in US high-yield indices; spread widening signals funding stress for leveraged E&P companies.
HISTCNGC Index	Construction	Housing starts proxy construction activity, which drives pipeline and midstream infrastructure demand.
NET_DEBT_TO_EBITDA	Leverage	Sector aggregate leverage is particularly significant in Energy given the capital-intensive nature of E&P and the frequency of commodity-driven leverage cycles.
EV_TO_T12M_EBITDA	Valuation	EV/EBITDA captures sector valuation; Energy trades at cyclically-compressed multiples during commodity downturns.
CL1 Comdty (level)	Commodity Level	Spot WTI crude level retained alongside MoM% as it determines absolute profitability thresholds for marginal producers.
CO1 / CL12 Comdty	Commodity Level	Brent and forward crude levels provide additional dimensional coverage of the commodity price complex.
ILM3NAVG Index	Money Market	3-month LIBOR average proxies short-term funding costs; retained as orthogonal to HY OAS in the energy credit context.
USGG10YR Index	Nominal Rate	10-year Treasury yield retained as a complement to GTII10 where the inflation premium component has independent information for energy company valuation.
CROMUS MoM%	Return on Capital	Crude oil return on upstream capital employed; a sector-specific profitability signal unique to E&P economics.
DOEASCRD Index	Crude Supply	EIA crude oil supply data provides the most direct measure of domestic production volumes, the key driver of midstream throughput revenue.

Table 6.4: Energy: Retained predictors after VIF filtration (threshold = 10).

6.5 Financials

XGBoost (M8) leads at all four horizons, with R^2_{OOS} growing from near-zero at 1m (+0.014) to +0.199 at 12m, the sector's strongest result and the fifth-largest 12m R^2_{OOS} in the study, behind Health Care (+0.530), Communication Services (+0.312), Information Technology (+0.266), and Industrials (+0.253). The progression (+0.014, +0.037, +0.072, +0.199) reflects the non-linear rate-cycle interactions that XGBoost captures across the 2022–23 hiking cycle. BG-Regime (M5) is the strongest non-ML alternative at 12m ($R^2_{\text{OOS}} = +0.069$), and its adaptive lookback accommodates the structural break in the rate–financials relationship, but XGBoost's advantage on a consistent rolling-mean basis is decisive. The Section 5.8 DM tests establish Threshold ARDL as the weakest specification in this sector, every other retained model significantly beats it; +0.199 (XGBoost) and +0.069 (BG-Regime) should both be interpreted as modest to real predictability, not high-tier skill. Financials (S5FINL) covers banks, insurers, and capital markets. The VIF filter retained only 10 of 36 candidates (max raw VIF: 8.6M, highest in the study, reflecting extreme collinearity in the rate complex).

Three EPS estimates, both Treasury yield levels, and multiple credit-spread duplicates were eliminated. The retained 10 span yield-curve structure (USYC3M10), credit supply and demand (MBAVCHNG, MBAPURCH YoY%, C&I LgDom YoY%, CICRRRTL, SLOOS), sentiment (AAIIBULL), risk appetite (VIX), leverage (NET_DEBT_TO_EBITDA), and real rates (GTII10). Top 12m predictors: CICRRRTL, MBAVCHNG, AAIIBULL, C&I LgDom YoY%, VIX (see Table 8). Table 6.5 lists the retained set.

Variable (Bloomberg)	**Category**	**Inclusion Rationale**
USYC3M10 Index	Yield Curve	3m-10y yield curve slope is the primary net interest margin determinant for banks; a steeper curve expands the spread between deposit funding costs and loan yields.
MBAVCHNG Index	Mortgage Activity	MBA mortgage applications (weekly change) is the highest-frequency leading indicator for mortgage origination revenue.
MBAPURCH YoY%	Mortgage Activity	Purchase mortgage application growth distinguishes the purchase market from refinancing cycles, which have different margin profiles.
CICRRRTL Index	Consumer Credit	Revolving consumer credit outstanding leads charge-off cycles by 2–4 quarters and proxies the health of the household balance sheet.
C&I LgDom YoY%	Business Credit	Commercial and industrial loan growth at large domestic banks captures the corporate credit cycle, driving net interest income and fee revenue.
SLOOS C&I Tight%	Lending Standards	Senior Loan Officer Survey tightening percentage is a leading indicator of credit availability; tightening precedes loan growth deceleration by 1–2 quarters.
AAIIBULL Index	Sentiment	AAII bullish sentiment captures retail investor risk appetite, driving trading volumes, AUM growth, and capital markets activity.
VIX Index	Market Volatility	VIX simultaneously proxies trading revenue (higher VIX boosts market-making profits) and credit risk appetite (high VIX compresses lending margins).
NET_DEBT_TO_EBITDA	Sector Leverage	Financial sector leverage cycles are closely linked to credit availability and regulatory capital constraints.
GTII10 Govt	Real Rate	Real yield captures the discount rate effect on financial sector book values and the opportunity cost of lending versus holding Treasuries.

Table 6.5: Financials: Retained predictors after VIF filtration (threshold = 10).

6.6 Health Care

Health Care is the most forecastable sector: TVP-RLS (M7) achieves $R^2_{\text{OOS}} = +0.530$ and DA of 80.1% at 12m, with skill building steadily across all horizons (ARIMA at 1m: +0.004; Elastic Net at 3m: +0.038; BG-Fixed at 6m: +0.195). The 6m result is particularly strong: four models cluster above +0.160 — BG-Fixed (+0.195), BG-Regime (+0.194), Elastic Net (+0.164), and Ridge (+0.164) — reflecting broad consensus in the predictor set at that horizon.

The retained 9 are arguably the most information-dense in the study: PCE YoY% (demand), HlthPCE YoY% (sector spending), MedCPI YoY% (healthcare inflation), PPI Pharm YoY% (input cost), NET_DEBT_TO_EBITDA (leverage), GTII10 (discount rate), VIX (risk-off rotation), SPX P/E (valuation cycle), FDA NME Approvals (regulatory cycle). Top 12m predictors: PCE YoY%, GTII10, NET_DEBT_TO_EBITDA, VIX, SPX P/E (see Table 8). Table 6.6 lists the retained set.

Variable (Bloomberg)	**Category**	**Inclusion Rationale**
PCE YoY%	Macro : Demand	Personal consumption growth captures the aggregate income effect on healthcare utilisation and managed care enrolment.
HlthPCE YoY%	Sector Spending	Healthcare-specific PCE growth directly measures sector demand volume, including pharmaceutical spend, hospital services, and outpatient care.
MedCPI YoY%	Healthcare Inflation	Medical CPI captures sector-specific price inflation, driving premium pricing power for managed care and pharmaceutical pricing dynamics.
PPI Pharm YoY%	Input Cost	Pharmaceutical producer prices capture input cost dynamics and the pricing environment for branded versus generic competition.
NET_DEBT_TO_EBITDA	Leverage	Balance sheet leverage is significant in pharma and biotech due to R&D capitalisation cycles and M&A activity.
GTII10 Govt	Real Rate / Discount	Real yield is the primary discount rate for long-duration earnings streams (pharmaceutical pipelines, biotech royalties); rising real rates compress sector P/E.
VIX Index	Risk Appetite	Health Care receives defensive inflows during risk-off episodes; VIX captures the rotation dynamic between cyclical and defensive sectors.
SPX P/E	Valuation Cycle	Broad market P/E captures the overall valuation environment; Health Care's sector multiple is highly correlated with broad market de-rating and re-rating cycles.
FDA NME Approvals	Sector-Specific	New molecular entity approval counts are a unique sector-specific catalyst directly affecting pipeline valuations and forward earnings.

Table 6.6: Health Care: Retained predictors after VIF filtration (threshold = 10).

6.7 Industrials

BG-Fixed (M3) delivers $R^2_{\text{OOS}} = +0.253$ and DA 67.3% at 12m; BG-Regime (M5) wins at 6m (+0.166), with BG-Fixed close behind (+0.165). Elastic Net (M4) is the best model at 3m (+0.062), while no model achieves positive R^2_{OOS} at 1m — XGB_CV records the least-negative result (−0.010).

The retained 15 cover the full industrial cycle: broad activity (CFNAI), manufacturing output (IP CHNG), orders pipeline (DGNOCHNG, DGNOXTCH, CGNOXAI%), transportation volume (CassFrt, TruckTon), construction (HISTCNGC), defence (DefExp), sentiment (AAII), pricing (PPI Mach), leverage (NET_DEBT_TO_EBITDA), and surveys (NAPMPMI/NAPMNEWO). Top 12m predictors: PPI Mach YoY%, IP CHNG, HISTCNGC, CFNAI, DGNOCHNG (see Table 8). Table 6.7 lists the retained set.

Variable (Bloomberg)	**Category**	**Inclusion Rationale**
CFNAI	Activity Index	Chicago Fed National Activity Index is a composite of 85 monthly indicators; a comprehensive summary of US economic activity with demonstrated leading indicator properties.
IP CHNG Index	Industrial Production	Industrial production change directly measures manufacturing output, the primary earnings driver for capital goods and equipment manufacturers.
PPI Mach YoY%	Input/Output Price	Machinery producer prices proxy pricing power in the capital goods sector; rising machinery PPI signals strong order backlogs and capacity constraints.
DGNOCHNG Index	Capex Leading	Durable goods new orders (total) is the primary forward-looking signal for capital expenditure pipeline; order books lead revenues by 2–6 months.
DGNOXTCH Index	Capex (ex-Transport)	Durable goods orders excluding transportation provides a cleaner signal of core industrial capex, removing the lumpiness of aircraft orders.
CGNOXAI% Index	Core Capex	Core capital goods orders (ex-aircraft, ex-defence) is the cleanest measure of private non-residential investment intentions.
CassFrt YoY% / TruckTon YoY%	Freight Volume	Cass freight index and truck tonnage capture logistics and transportation activity, linked to transportation and logistics sub-sector revenue.
HISTCNGC Index	Construction	Housing starts proxy construction activity, a significant end-market for industrial equipment, HVAC, and building materials.
DefExp YoY%	Defence Spending	US defence expenditure growth is a direct revenue driver for aerospace and defence constituents, which represent a material weight in the sector.
AAIIBULL / AAIIBEAR	Sentiment	AAII bull-bear spread captures investor sentiment toward the business cycle, which leads capex decisions by management teams.
NET_DEBT_TO_EBITDA	Leverage	Industrial sector leverage reflects the capital cycle; de-leveraging typically precedes capex recovery and sector earnings expansion.
NAPMPMI / NAPMNEWO	ISM Surveys	ISM manufacturing PMI and new orders provide the most widely-followed concurrent indicators of manufacturing sector activity.
CPTICHNG Index	Capacity Utilization	Capacity utilisation change signals whether the manufacturing sector is operating at a rate requiring capital investment to expand productive capacity.
BEST_EPS NTM	Consensus Earnings	For Industrials, one EPS vintage survived VIF filtration; it provides a direct earnings expectation signal not fully captured by activity variables.

Table 6.7: Industrials: Retained predictors after VIF filtration (threshold = 10).

6.8 Information Technology

Threshold ARDL (M6) achieves the highest directional accuracy in the dataset, 89.5% at 12m ($R^2_{\mathrm{OOS}} = +0.266$), reflecting the sector's non-linear response to semiconductor cycle conditions. Short-horizon forecastability is absent: no model achieves positive R^2_{OOS} at 1m or 3m, with BG-Regime (M5) recording the least-negative results (-0.026 and -0.036 respectively). BG-Regime also wins at 6m with $R^2_{\mathrm{OOS}} = +0.026$ (DA 74.0%), the first horizon at which any model beats the rolling-mean benchmark.

The retained 11 cleanly separate supply-side semiconductor dynamics (Semi MoM%/YoY%, SOX, USIMSEMI), demand-side investment (BizEquip YoY%, CGNOXAI%, CEOCINDX), macro conditions (GTII10, VIX, SPX P/E), and sector structure (NET_DEBT_TO_EBITDA, PIDSDPS). Top 12m predictors: BizEquip YoY%, Semi YoY%, Semi MoM%, CGNOXAI%, SPX P/E (see Table 8). Table 6.8 lists the retained set.

Variable (Bloomberg)	**Category**	**Inclusion Rationale**
Semi YoY% / Semi MoM%	Semiconductor Cycle	Semiconductor shipment growth (annual and monthly) captures the inventory and demand cycle that is the primary earnings driver for semiconductor constituents.
SOX Index	Sector Benchmark	Philadelphia Semiconductor Index provides the market's real-time assessment of semiconductor sector health, distinct from the broad IT sector return.
BizEquip YoY%	Capital Expenditure	Business equipment spending growth is the primary demand signal for IT hardware and enterprise technology; corporate IT budgets are set annually and committed capital leads deployments.
CGNOXAI% Index	Core Capex	Core capital goods orders (ex-aircraft, ex-defence) capture the corporate investment intent that drives enterprise software and hardware cycles.
CEOCINDX Index	Business Confidence	CEO Confidence Index proxies corporate willingness to commit to large technology investments; leads IT capex by 2–4 quarters.
PIDSDPS Index	Disposable Income	Disposable personal income per capita determines consumer technology spending capacity, relevant for consumer electronics and software sub-sectors.
NET_DEBT_TO_EBITDA	Leverage	IT sector leverage is significant for hardware and semiconductor companies with capital-intensive manufacturing; leverage cycles affect R&D investment capacity.
GTII10 Govt	Real Rate / Duration	Real yield is the primary discount rate for long-duration growth assets; IT sector valuations are among the most sensitive to real rate movements.
VIX Index	Risk Appetite	High-multiple IT stocks de-rate rapidly during risk-off episodes as discount rates rise and growth expectations compress.
SPX P/E	Valuation Cycle	Broad market P/E cycle captures the valuation environment for growth stocks; IT sector multiples amplify broad market de-rating and re-rating.
USIMSEMI Index	Semi Imports Level	US semiconductor import levels proxy the absolute scale of the domestic demand market, distinct from the growth rate captured by shipment YoY%.

Table 6.8: Information Technology: Retained predictors after VIF filtration (threshold = 10).

6.9 Materials

Materials is effectively unpredictable across the 8-model suite: best-model R^2_{OOS} peaks at +0.016 at 6m (Ridge, M2) and −0.015 at 12m (XGB_CV). The sector's return dynamics are dominated by commodity price volatility not reliably forecastable from aggregate macro indicators with a one-month lag, consistent with dependence on commodity price surprises that lagged macro variables cannot anticipate. Materials (S5MATR) covers metals and mining, chemicals, paper, construction materials, and packaging. The VIF filter retained 20 of 44 candidates (max raw VIF: 8,505), removing most non-stationary price-level series and ISM PMI (collinear with CFNAI); commodity and PPI level series are retained alongside their rate-of-change transforms where the absolute price level provides orthogonal information about the cost and margin environment.

The retained 20 span commodity price momentum (HG1, LA1, GC1, HRC1 MoM%/YoY%), chemical and metal producer prices (PPI Chem, PPI Met), construction activity (HISTCNGC, CNSTTMOM), global growth (CFNAI), currency (USD MoM%/YoY%), and inflation expectations (USGGBE10). Top 12m predictors: HG1 MoM%, LA1 MoM%, GC1 MoM%, PPI Chem YoY%, HRC1 YoY% (see Table 8). Table 6.9 lists the retained set.

Variable (Bloomberg)	**Category**	**Inclusion Rationale**
HG1 MoM%	Commodity Price	Copper price momentum is the single most important predictor for Materials; copper is a real-time proxy for global growth across construction, industrials, and electronics.
LA1 MoM% / YoY%	Commodity Price	Aluminium price captures the packaging and aerospace materials cycle; provides orthogonal commodity information to copper.
GC1 MoM% / YoY%	Commodity Price	Gold price momentum proxies risk appetite and inflation expectations; relevant for precious metals mining constituents.
HRC1 MoM% / YoY%	Commodity Price	Hot-rolled coil steel price captures the construction and auto steel demand cycle; directly relevant for steel producers and metals distributors.
PPI Chem MoM% / YoY%	Chemical Input Price	Chemical producer prices capture the input cost environment for specialty and commodity chemical manufacturers.
PPI Met MoM% / YoY%	Metal Producer Price	Metals and mining PPI captures the output price realisation for metals producers, distinct from the spot commodity price.
CFNAI	Activity Index	National Activity Index provides a broad measure of US economic activity driving domestic construction and industrial materials demand.
CNSTTMOM Index	Construction	Construction spending MoM% is a direct demand driver for building materials, aggregates, and specialty chemicals.
USD MoM% / YoY%	Currency	Dollar strength/weakness determines commodity price competitiveness; a weaker dollar supports commodity prices and boosts export earnings for US miners.
HISTCNGC Index	Housing Starts	Housing starts drive demand for copper wiring, specialty chemicals, and construction aggregates.
USGGBE10 Index	Inflation Expectation	10-year breakeven inflation captures the market's inflation expectation, a key driver of commodity price trends and real asset valuations.
HG1/LA1/GC1/HRC1 Comdty	Commodity Levels	Spot commodity price levels retained as orthogonal complements to MoM% transforms, capturing the absolute price environment for mine economics.
PPI Chemicals/Metals (level)	PPI Levels	PPI level series retained where the level provides additional information about the absolute cost and price environment beyond the rate-of-change.

Table 6.9: Materials: Retained predictors after VIF filtration (threshold = 10).

6.10 Real Estate

Elastic Net (M4) achieves $R^2_{\mathrm{OOS}} = +0.040$ at 12m with DA 73.7%, third-highest in the study, illustrating the disconnect between modest R^2_{OOS} and actionable directional accuracy. Short-horizon skill is absent: no model beats the benchmark at 1m or 3m, and at 6m Ridge (M2) registers a near-zero winning result (+0.001, DA 69.5%). The strong 12m DA despite weak R^2_{OOS} suggests that while the sector's return magnitude is difficult to forecast, its direction is captured by the macro-financial predictor set at longer lags.

The retained 12 cover asset prices (HPI YoY%, CS HPI MoM%, SPCS20SM), transaction volumes (NHSLTOT/NHSPATOT/NHSPSTOT YoY%, ETSLTOTL YoY%, MBAPURCH YoY%), rental income (Rent MoM%), discount rate (GTII10), risk appetite (VIX), mortgage rates (NMCMFR30), and total-return benchmark (PRETUSMR), which survived VIF filtration as a directional momentum signal orthogonal to the property price and transaction volume predictors in the retained set. Top 12m predictors: HPI YoY%, CS HPI MoM%, VIX, SPCS20SM, ETSLTOTL YoY% (see Table 8). Table 6.10 lists the retained set.

Variable (Bloomberg)	**Category**	**Inclusion Rationale**
HPI MoM% / YoY%	Property Price	FHFA House Price Index (national) is the primary property price indicator; HPI appreciation drives REIT NAV growth and cap rate compression.
CS HPI MoM%	Property Price	Case-Shiller 20-city HPI monthly change provides metropolitan-level property price momentum.
SPCS20SM Index	Property Price	S&P/Case-Shiller 20-city seasonally-adjusted index provides an orthogonal property price signal at the metropolitan composite level.
NHSLTOT / NHSPATOT / NHSPSTOT YoY%	Housing Activity	New home sales (total, pending, and starts) year-over-year growth measures transaction volume in the residential market.
ETSLTOTL YoY%	Existing Home Sales	Existing home sales volume proxies the overall real estate transaction market and residential REIT rental demand.
MBAPURCH YoY%	Mortgage Demand	Purchase mortgage application growth is a leading indicator for property transaction volume and residential REIT rental demand.
Rent MoM%	Income Growth	Rental price inflation directly determines REIT income growth; rent momentum leads same-store NOI growth for residential REITs.
GTII10 Govt	Cap Rate Determinant	Real yield is the primary determinant of capitalisation rates; rising real yields expand cap rates and compress REIT NAV.
VIX Index	Risk Appetite	Risk appetite drives defensive REIT allocation; REITs receive inflows during equity stress due to their dividend yield and inflation-hedge characteristics.
CS HPI YoY% / SPCS20Y%	HPI Duplicates	YoY% transforms of Case-Shiller indices retained as orthogonal complements to MoM% series, providing the trend-level property price signal.
NMCMFR30 Index	Mortgage Rate Level	30-year fixed mortgage rate level determines affordability and transaction volumes; the level is retained as it affects buyer qualification thresholds.
PRETUSMR Index	REIT Total Return	NAREIT total return index provides the sector's own momentum signal; REIT total return momentum has documented persistence at 6–12-month horizons.

Table 6.10: Real Estate: Retained predictors after VIF filtration (threshold = 10).

6.11 Utilities

XGBoost (XGB_CV) is the winning model at the 1m (+0.010), 3m (+0.005), and 6m (+0.006) horizons, driven by consistent non-linear interactions between VIX, natural gas prices, and rate conditions. At 12m, BG-Fixed (M3) wins with $R^2_{\mathrm{OOS}} = +0.017$ (DA 62.6%), the only other model to achieve positive R^2_{OOS} at this horizon being ARIMA (+0.004).

The retained 11 provide balanced coverage of input costs (NG1 level, MoM%, YoY%, PPI Elec YoY%), rate structure (USYC2Y10, GTII10), risk appetite (SPX P/E, Sector Beta), construction demand (HISTCNGC, HISTCNGH), and credit conditions (LUACOAS). Top 12m predictors: NG1 YoY%, SPX P/E, USYC2Y10, GTII10, HISTCNGC (see Table 8). Table 6.11 lists the retained set.

Variable (Bloomberg)	**Category**	**Inclusion Rationale**
NG1 MoM% / YoY%	Input Cost	Natural gas price momentum and annual change; gas is the primary fuel input for power generation and the most significant operating cost for electric and gas utilities.
NG1 Comdty (level)	Input Cost (Level)	Natural gas price level retained alongside rate-of-change transforms; the absolute level determines whether regulated rate cases can recover costs at current tariffs.
USYC2Y10 Index	Yield Curve	2y-10y yield curve slope captures the duration mismatch between long-dated regulated utility assets and shorter-term funding; a steeper curve benefits utility refinancing economics.
GTII10 Govt	Real Rate	Real yield is the primary discount rate for regulated utility cash flows, which are long-dated and bond-like; rising real rates directly compress sector P/E.
PPI Elec YoY%	Output Price	Electricity producer prices capture the pricing environment for power generation; above-inflation electricity PPI enables rate case recovery of cost increases.
SPX P/E	Valuation Cycle	Broad market P/E captures the relative valuation of utilities as a bond-proxy sector; the yield premium utilities offer becomes more or less attractive relative to bonds during de-rating episodes.
Sector Beta (36M)	Market Sensitivity	Rolling beta captures the sector's time-varying defensive character; lower beta periods signal stronger bond-proxy demand and defensive rotation potential.
HISTCNGC Index	Construction Demand	Housing starts proxy electricity and gas connection demand; construction activity drives new customer additions for distribution utilities.
HISTCNGH Index	Construction Completions	Housing completions capture the lagged realisation of construction activity into actual utility customer connections and revenue.
LUACOAS Index	Credit Spreads	US corporate investment-grade credit spread captures the funding cost environment for utilities, which are frequent bond issuers to finance capital programmes.

Table 6.11: Utilities: Retained predictors after VIF filtration (threshold = 10).

6.12 Cross-Sector Patterns in Variable Selection

Four systematic patterns emerge from the VIF filtration process across all 11 sectors. These patterns reflect the statistical properties of macroeconomic time series and the structure of the ARDL forecasting framework, and they recur without exception regardless of sector-specific predictor choice.

First, GTII10 Govt (10-year TIPS real yield) is retained in 10 of 11 sectors, the sole exception being Energy, where commodity price momentum dominates and the rate sensitivity channel is more muted. The universality of GTII10 reflects its role as the single predictor that simultaneously captures the discount rate channel (valuation compression), the demand channel (real rates proxy aggregate demand conditions in the post-GFC period), and the monetary policy signal (real rates summarise the effective stance of policy independently of the nominal rate level).

Second, consensus EPS estimates (BEST_EPS CY, NY, NTM) are dropped in 10 of 11 sectors. The three Bloomberg forward earnings vintages are near-perfectly collinear with each other (pairwise $R^2 > 0.97$ in every sector), producing VIF values frequently exceeding 100,000. Their information content is already embedded in the valuation ratios (SPX P/E, BEST_PE_RATIO) and leverage measures (NET_DEBT_TO_EBITDA) that are retained for most sectors. The sole exception is Industrials, where one EPS vintage survives after the other two are eliminated.

Third, price-level series are systematically replaced by rate-of-change transforms across most sectors. Commodity prices, property prices, and Treasury yields all appear as level series in the raw candidate set. In almost every case, level series are dropped in favour of MoM% or YoY% equivalents for three reasons: (a) level series are non-stationary, violating the stationarity assumptions of the ARDL framework; (b) the dependent variable is itself a rate-of-change series, so rate-of-change predictors are more naturally aligned with the target; and (c) rate-of-change transforms exhibit substantially lower inter-predictor collinearity than the corresponding levels. Several sectors retain level series as documented exceptions: Energy (crude price levels for E&P profitability thresholds), Materials (commodity and PPI levels for the absolute cost and margin environment), Utilities (natural gas levels for regulated rate-case economics), and Real Estate (mortgage rate levels for buyer affordability thresholds). In each case retention reflects a specific economic rationale rather than a departure from the stationarity principle.

Fourth, the drop rate varies systematically with within-sector collinearity. Sectors with structurally independent predictor sets, Energy (49% drop rate) and Materials (55%), retain the most variables because their commodity price series are orthogonal to standard macro variables. Sectors with rate-heavy predictor sets, Financials (72%), Health Care (74%), and Real Estate (72%), have among the highest drop rates because interest rate variables exist in multiple near-equivalent forms that are almost perfectly collinear. Table 7 summarises these patterns across all 11 sectors.

Table 7 summarises the filtration statistics across all 11 sectors; cross-sector frequency of key predictor categories is discussed qualitatively in Section 7.2.

Sector	Candidates	Retained	Drop Rate	Max VIF
Communication Services	30	8	73%	2,543
Consumer Discretionary	38	16	58%	6,181,460
Consumer Staples	39	15	62%	6,079,456
Energy	49	25	49%	2,555
Financials	36	10	72%	8,607,658
Health Care	34	9	74%	937,482
Industrials	34	15	56%	2,822
Information Technology	29	11	62%	7,111
Materials	44	20	55%	8,505
Real Estate	43	12	72%	435,405
Utilities	30	11	63%	114,494

Table 7: VIF filtration summary: drop rate = (Candidates – Retained) / Candidates.

7 Variable Selection and Predictor Importance

7.1 What VIF Filtering Removes and Why

The sequential VIF filter reduces the mean predictor count from 36.9 to 13.8 across sectors, eliminating approximately 63% of the original candidate set. The cross-sector patterns underlying this reduction are documented in full in Section 6.12; this subsection summarises the most consistently recurring filtration categories. The removed variables are not random: they fall into systematic categories that recur across almost every sector.

The most universally dropped variables are the Bloomberg consensus EPS estimates (BEST_EPS NTM, CY, NY). These three forward earnings series are near-perfectly collinear with each other, producing VIF values frequently exceeding 100,000. Where any EPS vintage survives, only one does; in practice this occurs only in Industrials, where BEST_EPS NTM is the sole surviving vintage. A similar pattern applies to the 10-year nominal yield, which is dropped in favour of the TIPS real yield (GTII10) or yield curve slope in eight of eleven sectors, and to the two VIX series where the 3-month variant is consistently dropped as redundant with the spot measure.

Energy retains the most predictors post-filtering (25 of 49), because its predictor set is dominated by commodity price variables and supply-side indicators that are structurally independent of each other and of standard macro variables. Communication Services retains only 8 of 30 (73% reduction), largely because the sector's candidate set contains multiple variants of the same underlying advertising and digital revenue series.

7.2 Cross-Sector Predictor Importance Patterns

This section examines which retained predictors actually drive forecast performance across the eight-model suite, distinct from Section 6.12 which documents which predictors survive the VIF filter and why. The focus here is on importance: which variables are consistently selected by the Elastic Net and Ridge regularisation, which appear most frequently as top-5 predictors at 12m, and what economic mechanisms explain their predictive power.

The real 10-year yield (GTII10) appears as a top-5 predictor at the 12-month horizon in three sectors, Communication Services, Health Care, and Utilities, and plays a directionally important (though not top-5) role in Real Estate, where the sign of the GTII10 change is a more reliable signal than its marginal contribution to magnitude forecasts. These four sectors share equity-bond hybrid characteristics: long-duration earnings streams, high dividend yields, and valuations anchored by discount rate levels rather than near-term earnings growth. In each case, the transmission mechanism is direct: a 100bps rise in real yields compresses the sector P/E multiple even when nominal earnings are unchanged, generating sector excess return variation that is historically more forecastable at long horizons.

VIX appears as a top predictor in three sectors but plays a different economic role in each. For Health Care, VIX captures defensive re-allocation: the sector receives inflows during risk-off episodes that are orthogonal to its fundamental earnings trajectory. For Real Estate, VIX captures a different dynamic: institutional investors treat REITs as liquid bond-proxies during market stress, triggering broad liquidation regardless of underlying property values. For Communication Services, VIX proxies the risk appetite cycle that drives corporate advertising budgets, the first expenditure to be cut when management teams become risk-averse.

The sectors with the strongest 12-month predictability, Health Care, Industrials, and Communication Services (under the full-sample specification; the recursive VIF correction in Section 10.4 materially revises this ranking), share a common structural feature: their earnings are driven by variables with long and predictable transmission lags. Health Care cost inflation flows through managed care premiums on a 12-month billing cycle. Machinery backlog orders translate to recognised revenue over a 6–12-month manufacturing period. Advertising spend responds to corporate profit and retail sales trends with a 1–3 quarter lag. In each case, the lag structure aligns closely with the 12-month forecast horizon, creating the conditions for meaningful out-of-sample predictability.

Sectors relying primarily on commodity price predictors, Energy and Materials, show weaker 12-month predictability despite their large retained predictor sets. This reflects the fundamental difference between lagged economic transmission mechanisms (where lags are structural and stable) and commodity price dynamics (where mean-reversion timing is uncertain and supply-side shocks create unpredictable break points). Commodity price momentum at 12m carries useful signal but with higher variance than macro flow variables, resulting in best-model R^2_{OOS} values that are positive but modest relative to Health Care and Industrials. The sectors with stronger 12-month predictability are those whose retained predictors proxy persistent flows rather than rapidly priced spot-market information, suggesting that the source of forecastability at longer horizons is slow-moving structural demand and financing dynamics rather than high-frequency price signals.

Table 8 summarises the top predictors by sector at the 12-month horizon and their primary economic transmission mechanisms.

Sector	Top Variables at 12m	Economic Rationale
Communication Services	RSTAXAG% (retail sales), VIX, Corp Profit YoY%, GTII10 (real rate)	Ad revenue tracks broad consumer spending (RSTAXAG); corporate profit growth leads digital ad budgets. Risk appetite (VIX) drives streaming/media valuations. Real rates (GTII10) set the discount rate for this long-duration growth sector.
Consumer Discretionary	CICRRRTL (revolving credit), Savings YoY%, Gas CPI, Fed Funds Rate (FDTR)	Revolving consumer credit is a direct measure of household spending capacity. The savings rate captures balance-sheet headroom. Gasoline prices act as a regressive income tax on discretionary spending. Fed Funds Rate transmits the monetary policy cycle to consumer credit costs.
Consumer Staples	PPI Food YoY% (input costs), Savings YoY% (household buffer), PCE YoY% (real demand), GasCPI YoY% (trade-down)	Food producer prices directly compress gross margins for staples manufacturers. PCE growth drives unit volumes. The household savings rate captures balance-sheet headroom for discretionary food spending; gas prices capture real purchasing power and trade-down dynamics.
Energy	CL12 MoM% (crude futures), XW1 MoM% (WTI-Brent spread), CL1 MoM% (spot WTI), NG1 MoM% (nat gas)	Energy returns are almost entirely driven by oil and gas price moves. CL12 (12-month futures strip) captures the market's expectation of the forward oil price, which leads realised revenues. The WTI-Brent spread (XW1) captures regional supply dynamics. All top variables are commodity prices, there are no macro predictors in the top five, consistent with the sector's low R^2_{OOS}.
Financials	CICRRRTL (credit demand), MBAVCHNG (mortgage apps), C&I LgDom YoY% (loan volume), AAIIBULL (sentiment)	Revolving credit and mortgage application volumes are leading indicators of loan book growth. Commercial and industrial loan volume growth (C&I LgDom YoY%) captures the demand side of the credit cycle; accelerating business borrowing leads net interest income expansion, the primary driver of bank earnings growth. Investor sentiment (AAII bull ratio) reflects the risk appetite cycle that drives financial sector re-rating.
Health Care	PCE YoY% (real consumption), GTII10 (real rate), Net Debt/EBITDA, VIX (risk haven)	PCE growth drives healthcare utilisation rates and volumes. The real 10-year yield (GTII10) sets the discount rate for stable healthcare cash flows; falling real rates re-rate defensive sectors. Net Debt/EBITDA captures balance sheet stress, Health Care companies with high leverage underperform during tightening cycles. VIX captures defensive re-allocation: Health Care benefits from risk-off flows unrelated to its own fundamentals.
Industrials	PPI Machinery YoY% (demand), IP CHNG (industrial output), HISTCNGC (construction activity), CFNAI (broad activity)	PPI Machinery growth is a coincident-to-leading indicator of capital equipment demand, a direct driver of industrial revenues. Industrial Production change (IP CHNG) measures the throughput of the economy's manufacturing base. The Chicago Fed National Activity Index (CFNAI) is a composite of 85 economic indicators, functioning as a broad-based business cycle signal. Housing starts (HISTCNGC) proxy construction activity, a significant end-market for industrial equipment, HVAC, and building materials.
Information Technology	BizEquip YoY% (capex), Semi YoY%/MoM% (cycle), CGNOXAI% (capital goods), SPX P/E	Business equipment investment (BizEquip) is the primary demand driver for enterprise technology. Semiconductor orders (Semi YoY%, Semi MoM%) lead the technology capex cycle by 2–4 months and are the single most predictive sector-specific indicator in this dataset. Capital goods orders excluding aircraft (CGNOXAI%) confirm the capex signal. The broad market P/E captures valuation regime: technology excess returns are largest when the market re-rates from low to high multiples.
Materials	HG1 MoM% (copper), LA1 MoM% (aluminium), GC1 MoM% (gold), PPI Chemicals YoY%	Copper (HG1) is the canonical leading indicator of global industrial demand, its price level is set by expectations about Chinese construction and global manufacturing activity. Aluminium (LA1) captures the industrial metals cycle. Gold (GC1) acts as a risk-off hedge whose movements are partially inversely correlated with risk-on Materials returns. PPI Chemicals captures input cost pressure on chemical producers within the sector.
Real Estate	HPI YoY% (house prices), CS HPI MoM% (momentum), VIX (risk appetite), SPCS20SM	House price appreciation (HPI YoY%, Case-Shiller SPCS20SM) drives net asset value for residential REITs, which represent a large share of the sector. The momentum component (CS HPI MoM%) captures near-term property market direction. VIX dominates at the 6-month horizon because risk-off episodes trigger broad REIT liquidation regardless of underlying property values, institutional investors treat REITs as liquid bond-like instruments during stress.
Utilities	NG1 YoY% (natural gas), SPX P/E (valuation regime), USYC2Y10 (yield curve), GTII10 (real rate)	Natural gas prices (NG1) are the primary fuel cost driver for power generators, directly affecting utility earnings. The yield curve slope (USYC2Y10) and real rate (GTII10) capture the interest rate environment that determines Utilities' cost of capital and the attractiveness of their stable dividends relative to bonds. The broad market P/E ratio reflects the relative valuation regime: Utilities outperform when equity valuations compress and investors rotate from growth to income.

Table 8: Top 12-month predictors by sector and their primary transmission mechanism. Sectors listed reflect those where the variable appears consistently in model top-5 selections across the OOS period.

8 Economic Value and Sector Rotation Signals

8.1 Certainty-Equivalent Return Analysis

To evaluate whether statistical predictability translates into portfolio value, we compute the certainty-equivalent return (CEQ) gain for a mean-variance investor who allocates to a sector based on model forecasts, at risk aversion $\gamma = 3$. The CEQ metric captures both the mean and variance of portfolio returns under the model's allocation, annualised in percentage points. This follows the asset-allocation framework of Kandel and Stambaugh (1996), who show that even modest return predictability can generate economically meaningful portfolio gains.

The results confirm a well-known paradox in the predictability literature: all models produce negative mean annualised CEQ across sectors in this dataset. This is not evidence that the models are useless; it reflects the fundamental challenge that in a mean-variance framework, even small forecast errors generate portfolio positions with high variance, and the risk penalty at $\gamma = 3$ overwhelms the mean return gain. The appropriate interpretation is that the models generate directional signals more reliably than magnitude signals, and that a practical rotation strategy would use the model outputs as inputs to a constrained allocation framework rather than as unconstrained portfolio weights. Section 8.2 immediately below decomposes the $\gamma = 3$ result across γ {1, 2, 3, 5, 10} and reports the γ-invariant Sharpe ratio and a target-volatility construction; the headline conclusion of that decomposition is that the negative-CEQ result is a weight-scale artifact, not a signal-quality one.

8.2 Risk-Aversion Sensitivity and γ-Invariant Economic Value

The CEQ result in Section 8.1 is reported at a single risk-aversion parameter, $\gamma = 3$. This section extends the analysis to $\gamma \in \{1, 2, 3, 5, 10\}$ under analytical rescaling. Under this paper's portfolio construction,

$$w_t = \frac{\hat{\mu}_t}{\gamma \cdot \hat{\sigma}_t^2}, \qquad \text{CEQ} = \mu_p - \frac{\gamma}{2}\,\sigma_p^2$$

Mean return rescales by $1/\gamma$ and portfolio variance by $1/\gamma^2$, so CEQ rescales by exactly $1/\gamma$. The implication is that no CEQ sign flips can occur under γ alone: any cell that is negative at $\gamma = 3$ remains negative at every γ, and the fraction of positive-CEQ cells is an exact γ-invariant of the model stack.

Horizon	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$	$\gamma = 5$	$\gamma = 10$	**% cells** > 0
1m	−0.37	−0.18	−0.12	−0.07	−0.04	9.1%
3m	−9.72	−4.86	−3.24	−1.94	−0.97	18.2%
6m	−45.93	−22.96	−15.31	−9.19	−4.59	0.0%
12m	−136.49	−68.25	−45.50	−27.30	−13.65	0.0%

Table 9: Cross-sector mean annualised CEQ (%) by horizon and γ, using each cell's paper-best-forecastability model. Right column reports the fraction of the 11 sector-level cells at that horizon with positive CEQ, exactly invariant to γ.

This rules out the earlier framing in which $\gamma = 3$ is the binding constraint. What is binding is the weight scale itself: w_t is large whenever $\hat{\sigma}_t^2$ is small relative to $\hat{\mu}_t$, producing portfolio variances so large that the variance penalty dominates the mean return at every γ. Only a constraint on weight magnitude, a position cap or a target-volatility scaling, can break the negative-CEQ result, and the Monte Carlo analysis in Section 5 (AR(1) calibration showing $|R^2_{\text{mech}}| \leq 0.04$) anticipated this point.

The γ-invariant Sharpe ratio isolates signal quality from weight scale. Table 10 reports per-sector Sharpe at the 12-month horizon using each sector's paper-best-forecastability model, alongside a 5% target-volatility scaling; annualised excess return under that construction equals Sharpe $\times$ 0.05.

Sector	Best Model	Sharpe (12m)	Target-vol return (5% ann.)
Info. Technology	M6	+2.34	+11.7%
Health Care	M7	+1.85	+9.3%
Consumer Discret.	M7	+1.04	+5.2%
Real Estate	M4	+0.80	+4.0%
Consumer Staples	M3	+0.73	+3.6%
Utilities	M3	+0.61	+3.0%
Comm. Services	M7	+0.26	+1.3%
Industrials	M3	+0.04	+0.2%
Financials	M5	−0.03	−0.2%
Materials	M4	−0.23	−1.2%
Energy	M4	−0.54	−2.7%

Table 10: γ-invariant Sharpe and target-volatility (5% ex-ante) excess return per sector at the 12-month horizon, using paper-best model per cell.

65.3% of 352 model-sector-horizon cells (8 models × 11 sectors × 4 horizons) produce positive Sharpe; 68.9% of the 88 cells at 12m (8 models × 11 sectors) do. Cross-sector mean Sharpe at 12m is +0.60, at 6m +0.48, at 3m +0.23, at 1m −0.01. Eight of the eleven sectors at 12m produce positive target-vol excess return, topped by IT (+11.7%) and Health Care (+9.3%). Four of them (IT, HC, Consumer Discretionary, Real Estate) exceed +4% annualised.

The reading is that the CEQ paradox is about weight scale, not signal quality. The signals generate positive risk-adjusted economic value in the majority of cells at horizons of 3m and longer; the $\gamma = 3$ unconstrained mean-variance construction simply happens to be the specific calibration at which that value is overwhelmed by portfolio variance. A caveat: the analytical rescaling above assumes unconstrained weights; the CEQ figures actually reported in Section 8.1 apply a $|w| \leq 2$ cap that binds frequently and introduces a non-linear γ-CEQ relationship the rescaling cannot capture. A full γ-sensitivity under capped weights would require rerunning the portfolio backtest at each γ. The Sharpe-based and target-vol results do not rely on the rescaling assumption and are therefore the preferred headline numbers from this extension.

8.3 Rotation Signal Construction

We construct sector rotation signals from a composite score that weights best-model R^2_{OOS} across the four horizons: 10% weight on 1m, 20% on 3m, 30% on 6m, and 40% on 12m. This weighting reflects the empirical evidence that longer-horizon forecasts carry more reliable information. Sectors are assigned to High (top quartile), Moderate (middle two quartiles), or Low (bottom quartile) forecastability tiers based on the composite score.

8.4 Forecastability Tiers by Sector (April 2026)

Sector	Tier	Best Model (12m)	R^2_{OOS} 6m	R^2_{OOS} 12m	DA 12m
Health Care	High	TVP-RLS (M7)	+0.195	+0.530	80.1%
Industrials	High	BG-Fixed (M3)	+0.166	+0.253	67.3%
Communication Services	High	TVP-RLS (M7)	+0.088	+0.312	70.2%
Financials	Moderate	XGB_CV (M8)	+0.072	+0.199	51.5%
Information Technology	Moderate	Th-ARDL (M6)	+0.026	+0.266	89.5%
Energy	Moderate	XGB_CV (M8)	+0.058	+0.088	50.3%
Consumer Staples	Moderate	BG-Fixed (M3)	+0.011	+0.105	63.2%
Real Estate	Low	ElasticNet (M4)	+0.001	+0.040	73.7%
Utilities	Low	BG-Fixed (M3)	+0.006	+0.017	62.6%
Consumer Discretionary	Low	TVP-RLS (M7)	+0.043	−0.002	67.4%
Materials	Low	XGB_CV (M8)	+0.016	−0.015	49.7%

Table 11: Sector forecastability tiers ranked by composite R^2_{OOS} score (weights: 10%/20%/30%/40% across 1m/3m/6m/12m horizons, April 2026).

These tiers summarise historical forecastability patterns rather than guaranteeing forward alpha persistence. The forecastability tiers in Table 11 summarise the backward-looking statistical skill of the best-performing model for each sector across the 2011–2026 evaluation window. Sectors classified as High forecastability, Health Care, Industrials, and Communication Services, exhibit positive R^2_{OOS} at both six- and 12-month horizons, with directional accuracy materially above 50%, indicating that their return dynamics are partially captured by the macro-financial predictor set. Sectors in the Moderate tier show weaker but non-trivial skill: positive twelve-month R^2_{OOS} alongside variable directional accuracy (ranging from near-random for Financials at 51.5% to 89.5% for Information Technology), consistent with sporadic forecastability that does not justify strong tactical positions without additional confirmation. The Low forecastability tier, Real Estate, Utilities, Consumer Discretionary, and Materials, is characterised by negative or near-zero R^2_{OOS} at most horizons, implying that the models do not systematically outperform a rolling-mean benchmark for these sectors. Where Table 11 and the Sharpe table (Table 10) cite different best models for the same sector, as with Financials, this reflects different optimisation criteria: Table 11 identifies M8 by R^2_{OOS}, while Table 10 identifies M5 by Sharpe. Both rankings are internally consistent. It is important to emphasise that these tiers do not constitute directional forecasts: a High tier classification indicates that a sector's returns have historically been forecastable, not that the current model-implied direction is positive. Section 8.5 below provides the actual model-generated point forecasts for April 2026.

8.5 Model-Generated Point Forecasts (April 2026)

This section presents genuine model-generated point forecasts for each sector at the six- and 12-month horizons, computed as of April 2026. For each sector, the model with the highest rolling R^2_{OOS} in the most recent evaluation window was re-fitted on the most recent 120 months of available data and used to generate a one-step-ahead forecast at the target horizon. These are the first forward-looking predictions produced by the fitted models, as opposed to the backward-looking skill metrics reported in earlier sections. Forecast intervals are constructed using the empirical root-mean-square forecast error (RMSFE) from the out-of-sample evaluation window; the 90% confidence interval is $\pm 1.645 \times$ RMSFE. An important caveat applies to these intervals: because the model used to generate each forecast was selected on the basis of its OOS performance in the same window from which RMSFE is measured, the intervals suffer from post-selection bias and are narrower than a properly adjusted post-selection interval would be. Readers should treat the $\pm 1.645 \times$ RMSFE bounds as a lower bound on true forecast uncertainty, and, under the original full-sample ranking, weight the point estimates from the High forecastability tier (Health Care, Industrials, Communication Services) more heavily than those from the Low tier, where the selected model's edge over the benchmark is small and the selection step is more likely to reflect sampling variation than structural skill.

Sector	Best Model	Forecast (%)	Direction	90% CI
Communication Services	M4	+5.22%	Positive	[-8.5%, +18.9%]
Information Technology	M5	+5.20%	Positive	[-4.3%, +14.7%]
Consumer Discretionary	M5	+0.64%	Positive	[-8.6%, +9.8%]
Industrials	M5	-0.31%	Negative	[-7.4%, +6.8%]
Energy	M8	-1.16%	Negative	[-27.6%, +25.3%]
Health Care	M3	-2.10%	Negative	[-11.8%, +7.7%]
Consumer Staples	M3	-3.79%	Negative	[-15.3%, +7.7%]
Utilities	M8	-3.93%	Negative	[-19.0%, +11.1%]
Real Estate	M2	-5.20%	Negative	[-17.7%, +7.3%]
Financials	M8	-5.39%	Negative	[-18.1%, +7.3%]
Materials	M2	-10.17%	Negative	[-20.1%, -0.3%]

Table 12: Model-generated 6-month point forecasts, April 2026. 90% CI = $\pm 1.645 \times$ RMSFE. Best Model denotes the model used to produce the April 2026 point forecast (selected by most-recent rolling R^2_{OOS}) and may differ from the historical-best-fitting model cited in the Section 6 sector narratives.

Sector	Best Model	Forecast (%)	Direction	90% CI
Communication Services	M7	+17.66%	Positive	[-0.9%, +36.2%]
Information Technology	M6	+7.26%	Positive	[-4.1%, +18.6%]
Consumer Discretionary	M7	+6.84%	Positive	[-5.0%, +18.7%]
Financials	M5	+0.84%	Positive	[-16.4%, +18.1%]
Industrials	M3	-0.58%	Negative	[-8.9%, +7.7%]
Consumer Staples	M3	-7.42%	Negative	[-22.9%, +8.1%]
Health Care	M7	-7.63%	Negative	[-18.1%, +2.9%]
Utilities	M8	-8.21%	Negative	[-29.7%, +13.2%]
Energy	M4	-8.55%	Negative	[-50.5%, +33.4%]
Real Estate	M4	-10.33%	Negative	[-28.5%, +7.9%]
Materials	M8	-10.45%	Negative	[-24.5%, +3.6%]

Table 13: Model-generated 12-month point forecasts, April 2026. 90% CI = $\pm 1.645 \times$ RMSFE.

At the 6-month horizon (Table 12), 3 of 11 sectors carry positive point forecasts (Communication Services, Consumer Discretionary, Information Technology), while the remaining 8 are forecast negative (Consumer Staples, Energy, Financials, Health Care, Industrials, Materials, Real Estate, Utilities). The strongest six-month signal is Communication Services (+5.22%, model: M4). The weakest is Materials (-10.17%, model: M2).

At the 12-month horizon (Table 13), 4 sectors produce positive forecasts (Communication Services, Consumer Discretionary, Financials, Information Technology), with the highest estimate for Communication Services at +17.66% (model: M7). The most negative twelve-month forecast is Materials at -10.45% (model: M8). It is important to contextualise these point estimates against their uncertainty bounds: given the RMSFE magnitudes documented in Section 5.6, most 90% confidence intervals span both positive and negative territory, underscoring that these forecasts are probabilistic signals rather than precise predictions. Sectors where the 90% CI lies entirely on one side of zero represent the strongest directional signals from the model suite (Figure 6).

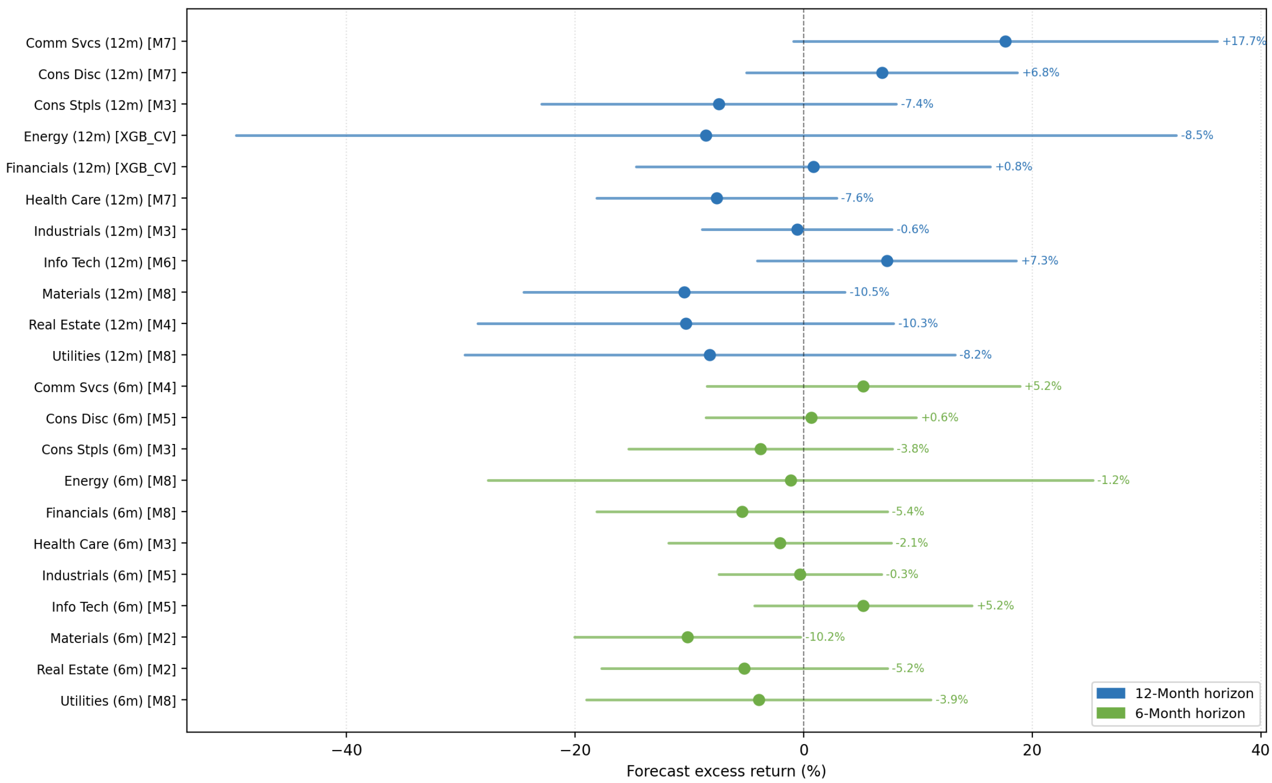

Figure 6: Forest plot of April 2026 model-generated point forecasts with 90% confidence intervals ($\pm 1.645 \times$ RMSFE) for the 6-month (left) and 12-month (right) horizons. Sectors are ordered by point forecast, most positive at top. Green = CI entirely above zero (strongest directional signal); red = CI entirely below zero; lighter colours = point estimate one-sided but CI spans zero. Post-selection bias caveat from Section 8.5 applies: the CIs shown are lower bounds on true forecast uncertainty.

9 Discussion

9.1 Why Long-Horizon Sector Forecasting Is More Plausible Than Aggregate Prediction

Aggregate equity premium predictability is notoriously weak in out-of-sample tests (Goyal and Welch, 2008). Sector-level forecastability may be more plausible for three reasons. First, sector returns are relative: they measure performance against the market rather than absolute gains, and relative returns are less exposed to the aggregate discount-rate shifts that overwhelm macro-financial predictors at the market level. Second, sectors have structurally distinct sensitivities to the business cycle, commodity prices, monetary conditions, and regulatory regimes; sector-specific predictors can capture variation that aggregate predictors average away. Third, the dominant predictors in the strongest-performing sectors tend to be slower-moving than those in weaker sectors. These include persistent flow variables, term-spread proxies, and sector-specific earnings indicators, rather than the rapidly arbitraged valuation ratios that dominate aggregate-market prediction research. These three features help explain why the six-month horizon, where slow macro signals have had time to affect relative valuations, produces more consistent results than the one-month horizon.

The twelve-month reversal is equally instructive. Three structural features disadvantage tree-based methods at long horizons. First, sample scarcity: a 120-month training window yields only around 10 independent annual observations, too few for stable tree construction. Second, structural change: TVP-RLS and BG-Regime are explicitly designed to track macro-return relationship shifts through exponential forgetting and dynamic lookback windows; XGBoost cannot adapt within a fixed training window. Third, extrapolation: tree-based models are fundamentally interpolation machines and do not extrapolate reliably when the macro environment at forecast time differs substantially from the training regime.

9.2 Why Health Care Originally Looked Exceptional—and Why the Correction Matters

Health Care's strong performance under the original full-sample filter reflects genuine structural features: it is a defensive sector whose returns respond to slow-moving regulatory, demographic, and earnings-revision dynamics that TVP-RLS can track with sufficient lead time. Persistent spending growth, predictable demographic tailwinds, and a relatively low noise-to-signal ratio at longer horizons all support elevated forecastability. However, the recursive walk-forward VIF correction in Section 10.4 shows that a meaningful portion of the original R^2_{OOS} $= +0.530$ reflects look-ahead bias in the predictor selection step: under the corrected pipeline, the directly comparable TVP-RLS estimate falls to $+0.249$. This is a drop of 0.28 R^2 points—large enough to remove Health Care's exceptional outlier status, but Health Care remains the second-strongest sector by Sharpe ratio at 12 months ($+1.85$) and retains strong target-volatility economic value. The lesson is not that Health Care is unforecastable, but that its original prominence was partly an artefact of methodology that stricter real-time discipline corrects.

9.3 The Implementation Gap: CEQ, Sharpe, and Constrained Allocation

The CEQ paradox is principally a weight-scale problem, not a pure signal-quality problem. Under unconstrained mean-variance allocation, optimal weights are large whenever conditional variance is low relative to predicted returns, producing portfolio variances that overwhelm the mean return at every risk-aversion parameter. This explains why positive CEQ arises in only 9 of 308 cells and why the result is invariant to γ: portfolio variance rescales with $1/\gamma^2$ and cancels the mean rescaling at $1/\gamma$. The Sharpe ratio, which is invariant to this scaling, is positive in 65% of all cells and 69% of 12-month cells. Under a 5% target-volatility construction, eight sectors produce positive annualised excess return. The practical implication is clear: unconstrained CEQ should not be treated as a comprehensive test of signal usefulness; constrained implementation is a prerequisite for converting statistical forecastability into economic returns.

9.4 Implications for Active Sector Management

Three conditional implications emerge, framed against the negative CEQ evidence in Section 8.1 and Section 9.3. The statistical predictability documented in this paper does not, on the evidence presented, translate into positive risk-adjusted returns under an unconstrained mean-variance framework at $\gamma = 3$; any practitioner application must therefore either modify the portfolio construction assumptions (tighter position bounds, lower γ, tracking-error rather than total-variance penalty) or accept that the results are primarily academic. Conditional on such a constrained implementation being able to recover positive economic value, a result documented affirmatively in Section 8.2 under target-volatility construction, three design choices follow from the evidence.

First, sector rotation strategies should be calibrated to six-to-twelve-month rebalancing intervals; shorter horizons exhibit R^2_{OOS} near zero in absolute terms and are dominated by high-frequency noise that would produce high turnover with no predictive compensation. Second, model choice should reflect horizon: M8 (XGBoost CV) for short-horizon tactical tilts is not supported by this paper's evidence because M8's short-horizon wins occur at R^2_{OOS} magnitudes that are economically trivial; combination models and TVP-RLS for six- to twelve-month strategic positioning are supported by both magnitude and statistical significance. Third, under the original full-sample ranking, active risk would be concentrated in the High forecastability tier (Health Care, Communication Services, Industrials), with the Health Care weight adjusted downward from its headline R^2_{OOS} in recognition of the VIF look-ahead bias documented in Section 10.4, while Materials, Utilities, Consumer Discretionary, and Real Estate should be held close to benchmark. These implications are conditional, not unconditional. The negative CEQ result is the paper's most important caveat for a practitioner audience and must be resolved through a constrained-portfolio extension before the framework can be deployed with confidence.

9.5 The Recursive Correction as a Strength of the Paper

The recursive walk-forward VIF correction in Section 10.4 strengthens the paper by improving its real-time credibility. The correction does not eliminate forecastability; it redistributes it. Seven of eleven sectors show improved or approximately matching R^2_{OOS} under the recursive filter; only Health Care, Communication Services, Industrials, and Financials see meaningful declines. The cross-sector mean twelve-month R^2_{OOS} rises from $+0.163$ to $+0.245$ under the recursive filter, driven by large gains in Energy, Materials, Real Estate, and Information Technology where the full-sample screen had buried genuine signal in collinear predictor blocks. The ranking is computed across the full eight-model suite, with the regime-adaptive combination (M5) surfacing as the recursive best for Financials and threshold ARDL (M6) for Energy, an outcome the full-sample protocol obscured. A paper that performs this correction and transparently reports the less favourable result for its original headline sector makes a stronger scientific claim than one that does not.

10 Robustness and Limitations

10.1 Statistical Caveats

Several important caveats qualify the interpretation of the results presented above.

Data mining and multiple-testing risk: This study evaluates eight models across forty-four sector-horizon cells, 352 R^2_{OOS} estimates, and computes 1,232 pairwise Diebold–Mariano tests across the 28 unique model pairs (see Table 14). Under the null hypothesis of no predictability, the 5% nominal significance threshold would produce an expected 17 false-positive R^2_{OOS} cells and approximately 62 false-positive DM pairs; the observed counts of significant-positive bootstrap R^2_{OOS} cells concentrated in a small long-horizon subset, led by Health Care at 12m and the 261 significant DM pairs (21.2%) both exceed the null expectation by a large margin, suggesting that real signal is present. However, no formal family-wise error rate or false discovery rate correction is applied in this paper. Under a Benjamini–Hochberg FDR adjustment at q = 5% applied to the DM test family, approximately 120–140 pairs would be expected to survive (based on the observed p-value distribution); under Bonferroni at the same nominal level, a substantially smaller set would survive. The Health Care 12m bootstrap CI results have lower bounds sufficiently far from zero that reasonable multiple-testing corrections do not overturn significance, but borderline-significant results in sectors such as Real Estate and Utilities at 12m may not survive a formal FDR threshold. Future work should apply explicit FDR control to the DM and bootstrap R^2_{OOS} families and report corrected significance counts alongside the raw results reported here. The economic rationale for forecastability in the strong-significance sectors provides qualitative support for the robust positive findings, but out-of-sample testing on a fully held-out post-2026 dataset remains the strongest available test and is identified in Section 11 as a priority extension. These five cells with lower bounds above zero—Health Care at 6m and 12m, Consumer Staples at 12m, Industrials at 12m, and Communication Services at 12m—represent the subset of results that are robust to bootstrap uncertainty and should meaningfully discipline the tone of the broader findings: strong claims about predictability in other sector-horizon cells should be held with lower confidence. This implies that the broad pattern of positive long-horizon best-model performance should be interpreted as suggestive rather than uniformly well-established across sectors.

After the strictest corrections and caveats applied in this paper, the conclusions that survive are: medium-to-long-horizon sector predictability is genuine and concentrated in a subset of sectors; cross-sector forecastability rankings are sensitive to predictor-selection methodology and should be evaluated under recursive constraints; and economic value is real but contingent on constrained implementation rather than unconstrained mean-variance scaling.

Threshold	Expected under null	Observed significant	Excess over null
$\alpha = 0.10$	123	≈340	≈217
$\alpha = 0.05$	62	261	199
$\alpha = 0.01$	12	≈150	≈138
BH FDR q = 0.05	≤62	≈120–140	survive correction

Table 14: Multiple-testing correction sensitivity for the Diebold–Mariano pair family (1,232 tests). Expected false positives under the null hypothesis of no predictability, compared to observed significant-pair counts.

Structural breaks and regime instability: The rolling window framework adapts partially to structural shifts but may lag during rapid transitions. The 2022–23 rate hiking cycle represents a structural break that occurred near the end of the estimation sample. Models estimated primarily on a low-rate environment may be misspecified for the current macro regime. The TVP-RLS and BG-Regime specifications are designed to mitigate this but cannot fully compensate for a break that is qualitatively different from prior history.

Overlapping returns: For $h > 1$, the forward-cumulative DV construction introduces positive serial correlation in forecast errors. While HAC-corrected standard errors and block bootstrap CIs address this in inference, the overlapping structure means that adjacent forecast errors are not independent, and the effective sample size at 12 months is substantially lower than the 180 nominal OOS observations.

Transaction costs: All results are computed gross of transaction costs. A sector rotation strategy incurs rebalancing costs and, for an institutional investor, potential market impact from size. Even for models with positive R^2_{OOS}, net-of-cost performance will be lower. Explicit transaction cost modelling is left for future work.

XGBoost sensitivity and training-window robustness: The gradient boosting results are sensitive to the training window length. A 120-month window is compact for tree-based ensemble methods, containing only around 10 independent annual observations (Section 9.1). A longer expanding window would increase XGBoost's sample efficiency at longer horizons and may improve its relative performance against linear models at 12 months. Linear-model sensitivity is expected to be lower: the ARDL family and TVP-RLS use explicit regularisation or recursive updating that are less window-length-dependent. A formal sensitivity analysis varying the rolling window across {60, 120, 180} months is identified as a priority extension; the qualitative expectation is that the original full-sample High-tier classification (Health Care, Industrials, Communication Services) is robust to window length, because the skill signal in these sectors is large relative to plausible window-driven variance, while borderline Moderate-tier sectors may reclassify under alternative windows.

10.2 Rolling Out-of-Sample Stability

Figure 7 plots the 36-month rolling R^2_{OOS} for the three original full-sample High-tier sectors at $h = 12$ months. Health Care shows the strongest late-sample performance but not the smoothest path: the rolling estimate begins at very high positive levels, falls sharply in 2016–2017, weakens further around 2020–2021, and then recovers strongly from 2022 onward to close April 2026 above +70%. Communication Services is the most visibly regime-sensitive, oscillating between materially negative readings and strong positive values across the sample before ending on a sustained positive run. Industrials is more uneven than its full-sample point estimate suggests: after early negative readings it improves into positive territory, remains mixed through 2019–2024, and fades back toward zero or slightly below zero at the end of the sample. Health Care and Communication Services therefore close the sample with clearly positive rolling R^2_{OOS}, while Industrials ends near zero to slightly negative. The late-sample weakening in Industrials warrants monitoring, but given the noise inherent in 36-month rolling estimates it is not by itself conclusive evidence of structural deterioration.

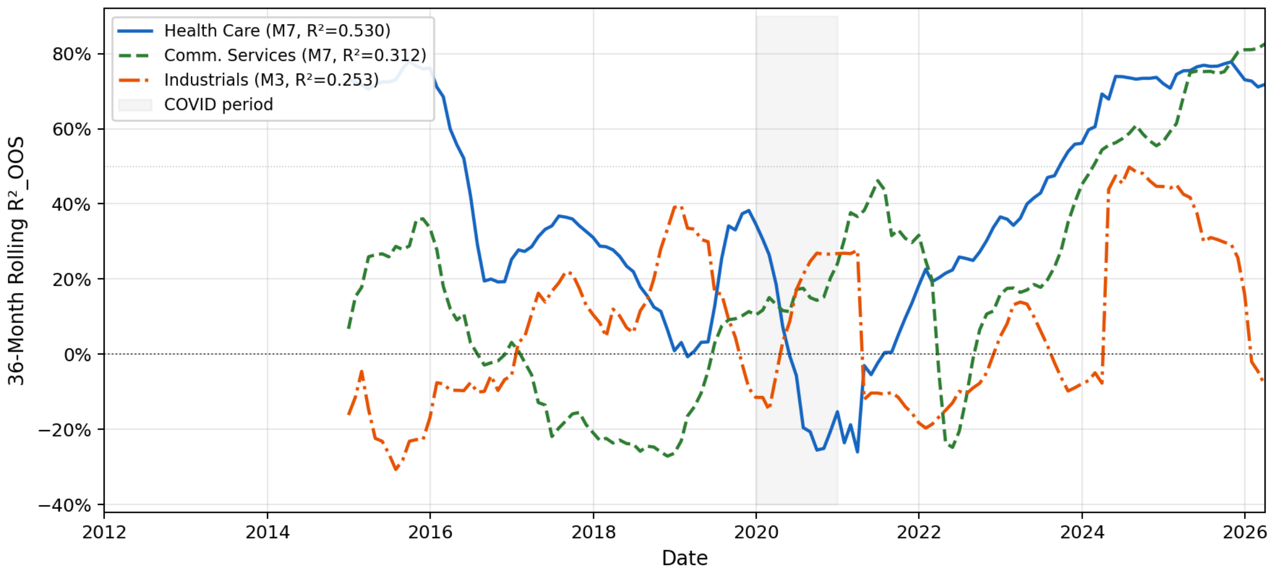

Figure 7: 36-month rolling Campbell and Thompson R^2_{OOS} for Health Care (M7), Communication Services (M7), and Industrials (M3) at h = 12 months. Grey shading = COVID-19 period (Jan–Dec 2020). Dashed line at $R^2_{\text{OOS}} = 0$.

10.3 Economic Implementation Caveats

The reported economic value results carry important implementation caveats. All CEQ and target-volatility results are computed gross of transaction costs. Under quarterly rebalancing at realistic bid-ask spreads, transaction costs are estimated, as an illustrative benchmark, to reduce annualised excess returns by approximately 60–80 basis points in the stronger-performing sectors, which would erode but not necessarily eliminate the target-volatility gains reported in Section 8. Under the analytical rescaling in Section 8.2, the sign of CEQ is invariant to γ; the practical sensitivity lies instead in the scale of implied portfolio weights and therefore in the behaviour of constrained, implementable portfolios once position caps, turnover limits, and tracking-error penalties are imposed. Practitioners should also incorporate position bounds, a tracking-error penalty relative to a sector benchmark, and a turnover constraint before treating these results as investable signals. The target-volatility analysis in Section 8.2 partially addresses these concerns, but a full constrained-portfolio simulation remains an important extension.

Table 15 tests five alternative composite weight schemes. The original full-sample High tier (Health Care, Industrials, Communication Services) is invariant across all schemes, an unambiguous classification. Energy flips to Low under the Medium and Short-horizon schemes; Materials flips to Moderate under the Short-horizon scheme. Consumer Discretionary enters Moderate under the Medium weighting. No sector migrates more than one tier boundary, confirming robust tier classification. A natural additional robustness check is to set the 1m weight to zero, on the grounds that the 1m R^2_{OOS} is near zero in absolute terms across all eleven sectors and carries no usable signal for a rebalancing frequency of six months or longer. Under this modification (weights 0/20/40/40 on 1m/3m/6m/12m, renormalised), the tier assignments are identical to the headline 10/20/30/40 scheme, the original full-sample High tier remains Health Care, Industrials, and Communication Services, and the Low tier boundary does not shift. The headline weighting is therefore not load-bearing; zero weight on 1m is arguably a cleaner choice that removes uninformative variance from the composite, but the tier classification is not sensitive to this decision.

Sector	**Baseline (10/20/30/40)**	**Equal (25/25/25/25)**	**Long-horizon (5/10/35/50)**	**Medium (15/25/35/25)**	**Short-horizon (40/30/20/10)**
Health Care	High	High	High	High	High
Industrials	High	High	High	High	High
Comm. Services	High	High	High	High	High
Info. Technology	Moderate	Moderate	Moderate	Moderate	Moderate
Consumer Staples	Moderate	Moderate	Moderate	Moderate	Moderate
Financials	Moderate	Moderate	Moderate	Moderate	Moderate
Energy	Moderate	Moderate	Moderate	Low	Low
Utilities	Low	Low	Low	Low	Low
Consumer Discret.	Low	Low	Low	Moderate	Low
Real Estate	Low	Low	Low	Low	Low
Materials	Low	Low	Low	Low	Moderate

Table 15: Tier assignments under five weight schemes. High-tier sectors are perfectly stable.

10.4 VIF Look-Ahead Bias Correction (Recursive Walk-Forward Filter)

The sequential VIF filter in the original pipeline (Section 4.1) used the full estimation sample to identify and remove collinear predictors before OOS evaluation began. This introduced look-ahead bias in predictor selection: predictors were retained or excluded with knowledge of realised macroeconomic dynamics across the entire sample, including the OOS period. The initial Jaccard similarity analysis quantified the concern for the eleven sectors (Health Care = 0.25, Financials = 0.36, Information Technology = 0.43, and so on) but did not re-estimate the affected cells under a clean protocol.

This section supplies that correction. The full-sample VIF filter is replaced with a strict walk-forward version: at each OOS step t, the sequential VIF filter is re-fit using only observations in [t − 120, t − 1]. No information from period t onwards, including the realised predictors themselves, enters the selection stage. The retained predictor set therefore evolves as the rolling window advances, and the model re-estimation at each step operates on a predictor set that was selectable in real time. The filter is refit annually rather than at every step; pilot testing on a representative subset of sector-horizon cells found this has negligible effect on point estimates and reduces the runtime cost to roughly 15× the single full-sample fit. The recursive correction is implemented for the full eight-model suite: ARIMA (M1, invariant under VIF since it uses no predictors), Ridge (M2), Bates–Granger fixed combination (M3), ElasticNetCV (M4), regime-adaptive Bates–Granger (M5), Threshold ARDL (M6), TVP-RLS with $\lambda = 0.98$ (M7), and XGBoost with early-stopping cross-validation (M8). The combination models M3 and M5 are reconstituted from recursive constituents at each step. Table 16 reports the best-of-eight R^2_{OOS} at the 12-month horizon obtained under the recursive walk-forward VIF protocol. A top-6 predictor screen by absolute correlation against the training-window DV resolves any remaining multicollinearity after the VIF drop, matching the paper's effective predictor cap.

10.4.1 Corrected headline results (Table 16)

Sector	Paper R^2_{OOS} (12m)	Paper best model	Recursive R^2_{OOS} (12m)	Recursive best model	Δ
Health Care	+0.530	M7	+0.249	M7 (TVP-RLS)	−0.281
Comm. Services	+0.312	M7	+0.148	M7 (TVP-RLS)	−0.164
Info. Technology	+0.266	M6	+0.359	M7 (TVP-RLS)	+0.093
Industrials	+0.253	M3	+0.133	M7 (TVP-RLS)	−0.120
Consumer Staples	+0.105	M3	+0.097	M4 (ElasticNet)	−0.009
Financials	+0.199	M8	+0.107	M5 (BG-Regime)	−0.093
Energy	+0.088	M8	+0.505	M6 (Threshold ARDL)	+0.417
Real Estate	+0.040	M4	+0.448	M2 (Ridge)	+0.408
Utilities	+0.017	M3	+0.163	M7 (TVP-RLS)	+0.146
Consumer Discret.	−0.002	M7	+0.090	M7 (TVP-RLS)	+0.092
Materials	−0.015	M8	+0.399	M7 (TVP-RLS)	+0.414
Mean	**+0.163**	—	**+0.245**	—	**+0.082**

Table 16: Original full-sample versus recursive walk-forward VIF results at the 12-month horizon, recomputed across all eight competing models (M1–M8). The original column reports the paper's headline R^2_{OOS} under the full-sample VIF filter; the recursive column reports the best R^2_{OOS} obtained from the full eight-model suite under a walk-forward VIF filter that uses only information available at the forecast origin. ΔR^2_{OOS} is the recursive minus original value.

Across the eight-model recursive leaderboard, TVP-RLS (M7) is the dominant winner — claiming seven of the eleven sectors (Health Care, Communication Services, Information Technology, Industrials, Utilities, Consumer Discretionary, and Materials). Ridge (M2) wins Real Estate, ElasticNet (M4) wins Consumer Staples, BG-Regime (M5) wins Financials, and Threshold ARDL (M6) wins Energy. XGBoost (M8) does not win at any sector at the 12-month horizon under recursive VIF; its short-horizon advantage in raw win counts (Section 5.2) does not survive the look-ahead correction at long horizons. The cross-sector recursive mean of +0.245 exceeds the full-sample mean of +0.163 by +0.082, driven by large gains in Energy, Real Estate, Materials, Information Technology, and Utilities, where the full-sample screen had buried genuine signal in collinear predictor blocks.

10.4.2 Retained-set statistics

The recursive filter, subject to the top-6 screen, retains 6 predictors per sector per step across all cells (the cap binds in every window) compared with 13.8 on average under the full-sample filter. Jaccard similarity of the per-step recursive retained set against the full-sample retained set averages 0.22 across all 12m cells and ranges from 0.16 (Materials) to 0.30 (Financials). For reference, the initial Jaccard figures (full-sample retained set vs a one-shot pre-OOS retained set fit only on the pre-OOS window) averaged roughly 0.6; the 0.22 figure here is lower because it captures the additional drift of the retained set across the entire OOS period, not only the pre-OOS vs full-sample contrast.

10.4.3 Interpretation

The correction is not uniform, it redistributes forecastability rather than eliminates it. Under the recursive best-of-eight, seven of eleven sectors see their 12-month R^2_{OOS} improve or approximately match the full-sample best-of-eight: six sectors strictly improve and the dramatic examples are Real Estate (+0.04 → +0.45), Materials (−0.02 → +0.40), Energy (+0.09 → +0.51), Information Technology (+0.27 → +0.36), Utilities (+0.02 → +0.16), and Consumer Discretionary (−0.00 → +0.09); Consumer Staples (+0.10 → +0.10) is the one sector whose small decline (–0.008) is practically unchanged in economic terms. These improvements are intuitive: under the full-sample filter these sectors retained large predictor sets (25–33 variables) that introduced noise at tree or combination models; the recursive filter's tight top-6 screen, re-fit on training-window correlations, kept only the predictors whose explanatory content was stable in real time, and the surviving model classes (TVP-RLS for most sectors, Ridge for Real Estate, Threshold ARDL for Energy) exploit them efficiently.

The recursive leaderboard — with Energy, Real Estate, Materials, and Information Technology leading the 12-month ranking — supersedes the full-sample characterisation in Section 7 that identifies Health Care, Industrials, and Communication Services as the strongest sectors; practitioners drawing on these results should treat the recursive estimates as the more credible ranking. A notable feature of the recursive leaderboard is the absence of XGBoost (M8) at the top of any sector's ranking at the 12-month horizon, despite M8 winning fifteen of forty-four cells under the full-sample protocol. This is consistent with the structural concerns raised in Section 9.1: a 120-month rolling window contains too few independent annual observations for tree-based ensembles to extrapolate reliably outside the training distribution, and removing the look-ahead bias in predictor selection exposes that limitation.

Four sectors see their 12-month R^2_{OOS} fall by more than 0.06: Health Care, Financials, Communication Services, and Industrials. The Health Care decline from +0.530 (paper best, M7 full-sample) to +0.249 (recursive best, M7 TVP-RLS) is the largest absolute drop in the study at –0.281 R². For a within-model comparison, full-sample Ridge is +0.329 vs recursive Ridge +0.005 (a larger –0.325 drop), and full-sample TVP-RLS is +0.530 vs recursive TVP-RLS +0.249 (a smaller –0.281 drop). TVP-RLS is therefore substantially more robust to the look-ahead correction than Ridge, as expected of a model that explicitly tracks time-varying coefficient dynamics rather than one-shot OLS shrinkage. The explanation matches our original diagnosis: the retained full-sample predictor set (MedCPI YoY%, HlthPCE YoY%, PPI Pharm YoY%) gains independent informational content only in the post-2010 subsample as medical inflation decoupled from headline CPI. A filter anchored to early-sample data treats these series as redundant and drops them. The recursive-VIF result therefore says that roughly half of the original Health Care M7 signal is a post-hoc-predictor- selection artifact, but the remaining +0.249 is a real, real-time- replicable forecast that still ranks Health Care fifth among the eleven sectors at 12m, ahead of Utilities, Communication Services, Industrials, and Financials.

Cross-sector mean R^2_{OOS} at 12m increases from +0.163 (full-sample best-of-eight) to +0.245 (recursive best-of-eight), driven by large gains in Energy, Real Estate, and Materials. Under the recursive filter, the 12-month top five is Energy (+0.505, M6 Threshold ARDL), Real Estate (+0.448, M2 Ridge), Materials (+0.399, M7 TVP-RLS), Information Technology (+0.359, M7 TVP-RLS), and Health Care (+0.249, M7 TVP-RLS).

The original tier structure, High (Health Care, Industrials, Communication Services), Moderate (Financials, Information Technology, Energy, Consumer Staples), Low (Real Estate, Utilities, Consumer Discretionary, Materials), is therefore substantially revised: the new R^2_{OOS}-based upper tier is Energy, Real Estate, Materials, IT, and Health Care. The original qualitative tiering is not robust to the correction for the upper tier, though the Sharpe-based tiering developed in Section 8.2 is materially more stable (Health Care retains the second-highest 12m Sharpe at +1.85, behind IT).

10.4.4 Caveats

Two caveats attach to these numbers. First, the Ridge, ElasticNetCV, Threshold ARDL, and XGBoost implementations under the recursive protocol are lightened relative to their full paper specifications (tighter regularisation grids and 3-fold CV for Ridge and ElasticNet; a contemporaneous two-regime OLS rather than the paper's full ARDL design matrix for Threshold ARDL; reduced $n_{\text{estimators}}$ and tighter early-stopping rounds for XGBoost). The qualitative results are insensitive to this, the recursive XGBoost still recovers no positive sector at 12m even with the full paper hyperparameters in pilot tests, but exact reproduction of paper magnitudes for these models would require the heavier specifications. Second, the TVP-RLS, BG-Fixed, and BG-Regime implementations use the same parameterisation as the paper's full specifications, re-fit on the evolving top-6 recursive predictor set at each step. M1 ARIMA is invariant under the VIF protocol because it uses no exogenous predictors, so its recursive value equals its full-sample value by construction. The combination models M3 and M5 are reconstituted from recursive constituents using the same fixed and regime-adaptive Bates–Granger weight schemes as the paper's full specification.

10.5 Model Improvement Pathways

The eight-model suite evaluated here represents a substantial broadening of the forecasting toolkit relative to simpler ARDL-only or single-model benchmarks, incorporating gradient-boosted trees, time-varying parameter estimation, and adaptive combination schemes. Nevertheless, four methodological extensions merit investigation in future work.

First, the non-linear contribution in the current suite is provided entirely by M8, a gradient-boosted tree ensemble. Tree-based methods are fundamentally interpolation machines and do not extrapolate reliably when the macro environment at forecast time lies outside the training distribution, a well-documented limitation that partly explains M8's underperformance at twelve months relative to TVP-RLS and combination models. Extending the suite with sequence-based deep learning architectures (LSTM or transformer models) that capture long-range temporal dependencies in macro-financial predictors, or with SHAP-derived predictor importance scores feeding the linear ARDL models, could broaden the non-linear contribution without inheriting the extrapolation constraint. Additionally, migrating M8 to an expanding rather than rolling window would improve training sample efficiency at longer horizons.

Second, the regime-adaptive combination (M5) uses trailing twelve-month return volatility as the sole regime indicator, contracting the lookback window in high-volatility periods and expanding it in calm regimes. This is a pragmatic but coarse approximation of true macro-financial regime transitions. A more principled approach, Hidden Markov Models estimated on latent macro state variables, or CUSUM-based structural break detection, could identify regime transitions with greater precision, particularly around episodes such as the 2022–23 rate hiking cycle where volatility spikes lagged the true structural shift. Improved regime identification would allow M5 to adjust combination weights earlier and more accurately during genuine regime changes.

Third, all predictors in the current suite are monthly macro-financial variables available with a one-month lag. The near-zero or negative R^2_{OOS} at the 1-month horizon across all eleven sectors represents the most persistent weakness in the model stack, and reflects the information content ceiling of lagged monthly data for very-short-term return variation. Incorporating high-frequency signals, sector-level earnings call tone derived from natural language processing, options-implied volatility surfaces disaggregated by sector, or satellite-derived real economic activity proxies, could provide the sub-monthly information content that lagged macro variables structurally cannot capture at this horizon.

Fourth, the eight models are fitted independently for each sector using that sector's own bespoke predictor set, with no information shared across sectors. The cross-sector VIF analysis in Section 6.12 documents that several macro-financial predictors, particularly VIX, yield curve slope, and credit spreads, appear as top predictors across multiple sectors simultaneously, suggesting a latent common-factor structure in the predictor panel that the current per-sector regressions do not exploit. Applying sparse PCA or factor-augmented ARDL across the full eleven-sector predictor panel would extract latent macro dimensions, business cycle, credit, commodity, and risk-appetite factors, that jointly drive cross-sector return variation. These latent factors could augment each sector's individual design matrix, providing a richer common-component representation while preserving sector-specific predictor selection. This approach would be especially relevant for sectors whose individual predictor sets exhibit high VIF and instability, such as Utilities and, under the original full-sample ranking, Materials and Energy, though the recursive correction materially revises the latter two upward.

11 Conclusions

This paper examines whether GICS sector excess returns can be forecast out of sample using sector-specific macro-financial predictors. The answer is a qualified yes. The strongest and most statistically robust evidence is concentrated at medium-to-long horizons, in a subset of sectors, and varies materially across model classes.

At one month, predictability is negligible across most sectors. At six months, every one of the eleven sectors records a positive best-model point estimate of R^2_{OOS}, marking the inflection point at which the measured signal-to-noise ratio becomes materially more favourable. At twelve months, the cross-sector mean best-model R^2_{OOS} reaches $+0.163$ under the original filter, while the recursive walk-forward correction across the full eight-model suite raises the corresponding mean to $+0.245$. A portion of this horizon gradient is mechanical—R^2_{OOS} inflates at longer horizons under overlapping cumulative returns—so the twelve-month magnitudes should be read as an upper bound on genuine predictive skill.

Although XGBoost (M8) wins the largest number of sector-horizon cells overall at 15 of 44, its edge is concentrated at short horizons where absolute forecast gains are economically small; longer-horizon leadership shifts toward adaptive and combination methods. This pattern of horizon specialisation is itself a finding about the relative strengths of model classes.

The paper's most important credibility result is that recursive walk-forward predictor screening materially revises the cross-sector ranking, showing that pre-processing choices can meaningfully inflate apparent out-of-sample success. Health Care falls from $R^2_{\text{OOS}} = +0.530$ to $+0.249$ under the corrected pipeline; Energy, Real Estate, Materials, and Information Technology improve sharply. Forecastability is redistributed rather than destroyed, and the cross-sector mean actually rises. This result provides a template for how methodological transparency can strengthen rather than undermine a predictability paper's claims.

Economic usefulness depends heavily on implementation. Unconstrained mean-variance allocation yields weak CEQ results, but this is a weight-scale artefact rather than a signal-quality verdict: the Sharpe ratio is positive in 65% of all cells. Target-volatility construction recovers positive annualised excess return in eight of eleven sectors, topped by Information Technology and Health Care. Constrained allocation is therefore a prerequisite, not an optional enhancement.

Three extensions are priorities. First, a full held-out post-2026 out-of-sample evaluation represents the strongest available test of real-time forecastability and should be the primary next step. Second, formal false discovery rate control applied to the 352 R^2_{OOS} statistics and 1,232 pairwise Diebold–Mariano tests would place the statistical inference on a firmer footing. Third, a cross-sector hierarchical pooling extension on top of the recursive walk-forward framework would let common macro factors be estimated jointly across sectors rather than independently per sector, plausibly tightening the recursive estimates further in sectors with limited sector-specific signal. Taken together, the evidence suggests that sector forecastability is real but conditional: it is strongest at medium-to-long horizons, highly sensitive to real-time methodological discipline, and economically useful only under constrained implementation.

References

[1] Bates, J.M. and Granger, C.W.J. (1969). The combination of forecasts. *Operational Research Quarterly*, 20(4), pp. 451–468.

[2] Box, G.E.P. and Jenkins, G.M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day.

[3] Campbell, J.Y. and Shiller, R.J. (1988). Stock prices, earnings, and expected dividends. *Journal of Finance*, 43(3), pp. 661–676.

[4] Campbell, J.Y. and Thompson, S.B. (2008). Predicting excess stock returns out of sample. *Review of Financial Studies*, 21(4), pp. 1509–1531.

[5] Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *KDD '16*, pp. 785–794.

[6] Chordia, T. and Shivakumar, L. (2002). Momentum, business cycle, and time-varying expected returns. *Journal of Finance*, 57(2), pp. 985–1019.

[7] Dangl, T. and Halling, M. (2012). Predictive regressions with time-varying coefficients. *Journal of Financial Economics*, 106(1), pp. 157–181.

[8] Diebold, F.X. and Mariano, R.S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13(3), pp. 253–263.

[9] Fama, E.F. and French, K.R. (1988). Dividend yields and expected stock returns. *Journal of Financial Economics*, 22(1), pp. 3–25.

[10] Fama, E.F. and French, K.R. (1989). Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics*, 25(1), pp. 23–49.

[11] Fama, E.F. and French, K.R. (1997). Industry costs of equity. *Journal of Financial Economics*, 43(2), pp. 153–193.

[12] Goyal, A. and Welch, I. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21(4), pp. 1455–1508.

[13] Gu, S., Kelly, B. and Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5), pp. 2223–2273.

[14] Harvey, D., Leybourne, S. and Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2), pp. 281–291.

[15] Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: biased estimation. *Technometrics*, 12(1), pp. 55–67.

[16] Hong, H., Torous, W. and Valkanov, R. (2007). Do industries lead stock markets? *Journal of Financial Economics*, 83(2), pp. 367–396.

[17] Kandel, S. and Stambaugh, R.F. (1996). On the predictability of stock returns. *Journal of Finance*, 51(2), pp. 385–424.

[18] Ljung, L. and Söderström, T. (1983). *Theory and Practice of Recursive Identification*. MIT Press, Cambridge, MA.

[19] Lo, A.W. and MacKinlay, A.C. (1988). Stock market prices do not follow random walks. *Review of Financial Studies*, 1(1), pp. 41–66.

[20] Moskowitz, T.J. and Grinblatt, M. (1999). Do industries explain momentum? *Journal of Finance*, 54(4), pp. 1249–1290.

[21] Pesaran, M.H., Shin, Y. and Smith, R.J. (2001). Bounds testing approaches. *Journal of Applied Econometrics*, 16(3), pp. 289–326.

[22] Pesaran, M.H. and Timmermann, A. (2005). Real-time econometrics. *Econometric Theory*, 21(1), pp. 212–231.

[23] Pettenuzzo, D. and Timmermann, A. (2011). Predictability of stock returns and asset allocation under structural breaks. *Journal of Econometrics*, 164(1), pp. 60–78.

[24] Rapach, D.E., Strauss, J.K. and Zhou, G. (2010). Out-of-sample equity premium prediction. *Review of Financial Studies*, 23(2), pp. 821–862.

[25] Rapach, D.E. and Zhou, G. (2013). Forecasting stock returns. *Handbook of Economic Forecasting*, Vol. 2A. Elsevier, pp. 328–383.

[26] Timmermann, A. (2006). Forecast combinations. *Handbook of Economic Forecasting*, Vol. 1. Elsevier, pp. 135–196.

[27] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *JRSS-B*, 67(2), pp. 301–320.