Executive Summary

A systematic equity selection framework is developed to predict stock-level monthly excess returns and construct dynamically rebalanced long-short portfolios. The strategy is evaluated over the period J**anuary 2010 to December 2023** using out-of-sample predictions from **expanding-window models**.

Machine Learning Forecasting

The return prediction task is performed using multiple algorithms:

- **OLS, Ridge, Lasso, Elastic Net, XGBoost, AutoEncoder + Ridge**, and **IPCA** models are implemented.
- A Lasso-based feature selection procedure is applied to generate the OLS Reduced variant, enhancing robustness.
- An auxiliary **earnings surprise prediction** model is built using **Lasso** with **automatic feature selection and alpha tuning**, also under an expanding window setup.
- Model outputs are stored and used as forward-looking scoring signals.

Portfolio Construction

Two portfolio weighting schemes are tested:

- **Equally Weighted Strategy:** 50 long + 50 short stocks selected monthly;
- **Market Value Weighted Strategy:** Positions scaled by firm-level market equity.

A **dynamic threshold** mechanism is implemented:

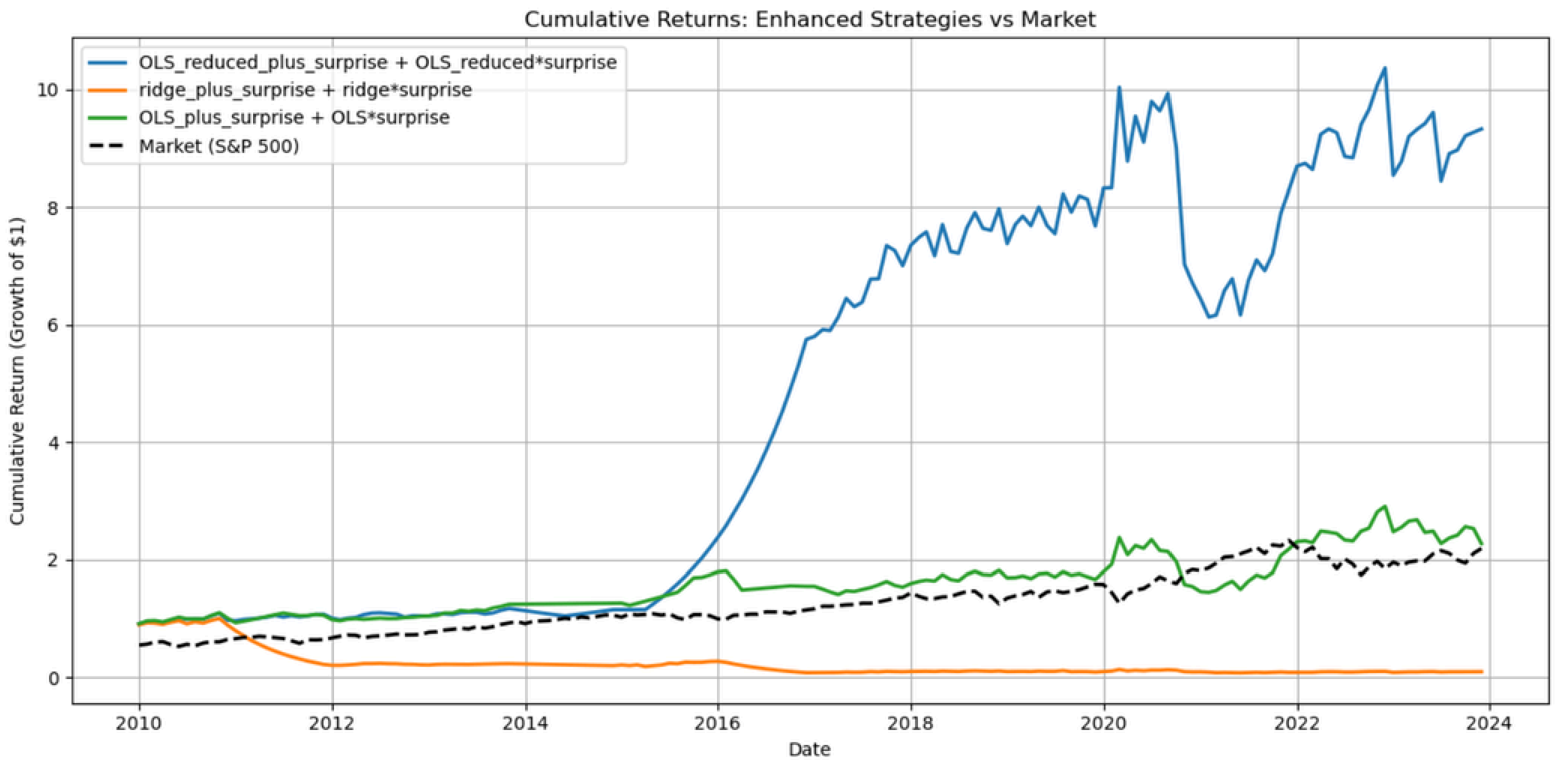
- Portfolio rebalancing occurs only when the signal spread (top 50 – bottom 50 average) **exceeds the 20th percentile of historical spreads**, reducing turnover and noise.

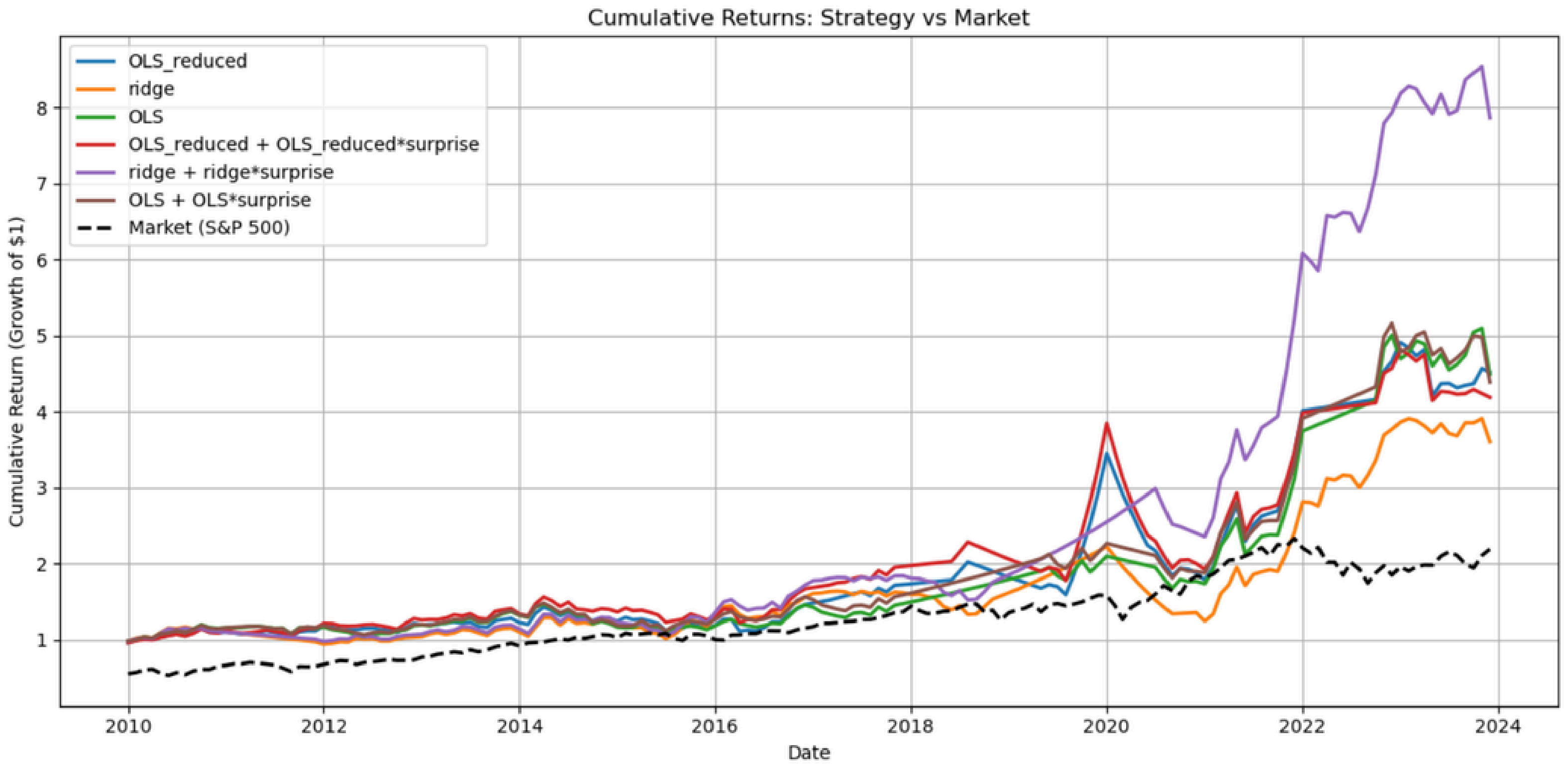
Signal Enhancement:

Performance is significantly improved by **combining return scores with earnings surprise** predictions through nonlinear interaction:

- **Best Equally Weighted Score:**
  - Score=0.7 · (OLS_Reduced Plus Surprise)+0.3 · (OLS_Reduced×Surprise)
- **Best Market Value Weighted Score:**
  - Score=1.5 Ridge + 0.3 Ridge×Surprise

The two graphs below compare cumulative returns of different strategies under equal-weighted (top) and market-value-weighted (bottom) portfolio construction. Both plots show that the two enhanced strategies — OLS_reduced_plus_surprise and OLS_plus_surprise — consistently outperform the benchmark (S&P 500) and other models over time.





Investment Strategy

Strategy Type

This is a **long-short strategy**. Every month, stocks are ranked by predictive signals. Positions are opened on the top 50 (long) and bottom 50 (short) stocks. Portfolio construction is dynamically controlled based on signal spread strength.

Predictive Signal Design

Two variations of signal construction are implemented:

Strategy	Signal Formula
Equal-Weighted Strategy	$0.7 \cdot (\text{Predicted Return}+\text{Surprise})+0.3 \cdot (\text{Predicted Return}\times\text{Surprise})$
Market-Weighted Strategy	$1.5 \cdot \text{Predicted Return}+0.3 \cdot (\text{Predicted Return}\times\text{Surprise})$

Key difference:

- Equal-Weighted version uses a hybrid additive + multiplicative signal
- Market-Weighted version uses a linear-enhanced signal with stronger weight on return

Dynamic Portfolio Construction

For both strategies, a model-specific dynamic threshold governs whether to update portfolio holdings:

For each month, compute the signal **spread**:

- Spread_t=Mean(Top 50)−Mean(Bottom 50)

Determine a **20th percentile threshold** from historical spreads.

- If spread ≥ threshold:
  - Rebuild long-short portfolio (top 50 & bottom 50)
- If spread < threshold:
  - Reuse previous month's portfolio (i.e., hold)

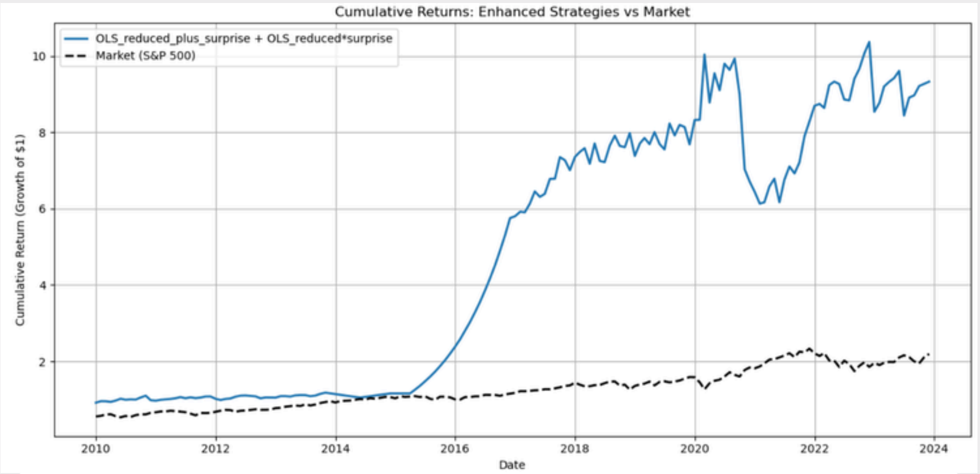
Top 10 Most Frequently Held Long and Short Positions- Equally -Weighted Method

Top 10 Long-Leg Holdings:

	num_months_in_long	count
0	14593	63
1	91233	55
2	64194	52
3	48506	50
4	86580	50
5	85621	49
6	85663	48
7	48486	47
8	76795	46
9	11533	43

Top 10 Short-Leg Holdings:

	num_months_in_short	count
0	91968	77
1	79859	68
2	12073	60
3	50017	44
4	22517	41
5	12558	41
6	80072	41
7	60599	39
8	36397	38
9	52898	37



PERMNO 14593 and 91968 were the most frequently held stocks in the long and short legs, appearing in 63 and 77 months respectively. The equal-weighted long-short strategy achieved over 10× capital growth between 2010 and 2023, significantly outperforming the S&P 500 benchmark with consistent compounding and strong mid-horizon acceleration.

Top 10 Most Frequently Held Long and Short Positions - Market-value Weighed Method

Top 10 Long-Leg Holdings:

	num_months_in_long	count
0	27828	104
1	12490	65
2	82598	63
3	40125	63
4	85914	60
5	91461	59
6	27983	57
7	92293	56
8	49373	55
9	16678	55

Top 10 Short-Leg Holdings:

	num_months_in_short	count
0	93436	84
1	67847	72
2	75241	68
3	92239	63
4	86288	61
5	87056	57
6	13641	57
7	90178	55
8	89393	52
9	84723	49

In the market-weighted strategy, PERMNO 27828 and 93436 appeared in the long and short legs for 104 and 84 months respectively, indicating strong persistent signals in the model ranking. The market-value weighted strategy based on the Ridge + Surprise signal delivered steady and robust outperformance, growing to 8× capital by 2023 while preserving stability through size-adjusted weighting.

Data and Methodology

External Data Source:

Firm-level predictors are combined with **macroeconomic variables** to construct **interaction term** and model both earnings surprise and stock return. The macro predictors are retrieved from Amit Goyal's standardized dataset (macro_monthly.csv), as introduced in Goyal & Welch (2008).

There are **8 monthly macro variables** included in macro_monthly.csv:

dp, ep, bm, ntis, svar, dfy, tms, and tbl

Among them, dp and ep are log-difference constructs of dividends/earnings relative to price. The macro panel is cleaned and filtered to span from 2000.01 to 2024.01.

Firm-level data includes **147** standardized accounting-based characteristics. For stock return prediction, interaction terms between each firm variable and each macro predictor are constructed, yielding over 1,000 additional features per month. Given this high dimensionality, **chunked standardization** (variable blocks) is applied to reduce memory load, and rely on **Lasso** to perform **automated feature selection**.

Machine Learning Models and Motivation

Stock return prediction is performed using multiple ML and econometric models, including:

- OLS / OLS Reduced (low-variance baseline)
- Lasso / Ridge / Elastic Net (regularized linear models)
- XGBoost (nonlinear tree-based model)
- Autoencoder + Ridge (nonlinear compression followed by regression)
- IPCA (Instrumented PCA, dimension reduction on interaction space)

Earnings surprise is modeled separately using only firm-level features. Lasso is chosen for its built-in sparsity, which facilitates feature selection without extensive grid search. Ridge is also tested for stability.

Training Methodology

All return models follow an expanding-window training approach:

- 8 years of training
- 2 years of validation (for tuning hyperparameters like alpha or number of trees)
- 1 year of test set
- This rolling scheme is repeated over 14 rounds, producing OOS predictions from 2010 to 2023.

Each model is tuned on the validation set using either:

- Lasso/Ridge/ElasticNet: Grid search over alpha
- XGBoost: Grid + early stopping
- Autoencoder: epoch tuning + loss monitoring

To reduce feature space, only Lasso-selected variables are passed into Ridge, Elastic Net, and XGBoost. IPCA is also applied on the same reduced set to control runtime.

Model Performance and R^2 Results

r2_ols	r2_ols_reduced	r2_lasso	r2_ridge	r2_enet_simple	r2_xgb	r2_ae_ridge	r2_ipca_all	r2_ipca_tuned
-0.065726021	-0.032101644	-0.022427054	-0.021435332	-0.022072618	-0.075477934	-0.029014984	-0.233866838	-0.170122277

Despite the uniformly negative R^2values, relatively stronger performance is observed for Lasso, Ridge, and OLS Reduced, which are ultimately selected for trading strategy construction due to their relative stability and interpretability. While IPCA showed comparatively less negative R^2, it was excluded from downstream use because its predicted returns were flat across all stocks in a given month, rendering it unsuitable for any ranking-based portfolio strategy.

Portfolio Strategy Performance

Metrics	Equally Weighted (Best) 0.7 ols_reduced_plus_surprise + 0.3 ols_reduced * surprise	Market-Value Weighted (Best) 1.5 ridge_return + 0.3 ridge * surprise	S&P500 Benchmark
Annualized Return	17.80%	16.17%	11.52%
Annualized Volatility	18.53%	16.76%	14.82%
Sharpe Ratio	0.9126	0.9647	0.72
Annualized Alpha	20.80%	17.50%	0.00%
Information Ratio	1.1661	1.0472	N/A
Max Drawdown	-38.90%	-23.79%	-24.77%
Max 1-Month Loss	-21.86%	-10.37%	-12.51%
Turnover	34.57%	37.60%	N/A

Both strategies significantly outperformed the S&P 500 in terms of return and risk-adjusted metrics:

- **Annualized Return:**
  - Equal-Weighted strategy achieved 17.80%, and Market-Value Weighted returned 16.17%, both well above the S&P 500's 11.52%.
- **Sharpe Ratio & Alpha:**
  - Sharpe Ratios of 0.91 (equal) and 0.96 (value-weighted) indicate strong risk-adjusted returns.
  - The annualized CAPM Alpha is notably high—20.8% and 17.5%, suggesting substantial outperformance beyond market exposure.
- **Drawdown & Stability:**
  - The Market-Value Weighted strategy shows better downside control with lower drawdown (–23.79%) and smaller max loss (–10.37%) compared to both the Equal-Weighted strategy and S&P 500.
- **Turnover:**
  - Both strategies maintain moderate turnover levels (~35–37%), indicating reasonable trading intensity.

The Equal-Weighted strategy delivers higher returns and alpha, while the Market-Value Weighted approach offers a better risk-return balance with lower drawdowns. Both strategies consistently outperform the market across all key dimensions.

Expectation and Improvement

The strategy largely performed in line with expectations. The equal-weighted and market-value weighted portfolios both outperformed the S&P 500 over the 2010–2023 OOS period, delivering strong annualized returns and high CAPM alphas. The use of expanding-window training and dynamic thresholding ensured robustness over multiple market regimes. The risk-adjusted performance—measured by Sharpe and Information Ratios—also exceeded expectations, particularly for the market-weighted strategy.
However, performance volatility was higher than anticipated for the equal-weighted version, especially during turbulent periods like early 2020. This suggests sensitivity to idiosyncratic risk when position sizes are not adjusted by market capitalization.

Key signals contributing to portfolio performance

The hybrid predictive score combining firm-level return predictors and forecasted earnings surprise was the most significant contributor to performance.
Specifically:

- The OLS Reduced + Surprise and Ridge + Surprise models outperformed their standalone counterparts, suggesting that earnings surprises added orthogonal, forward-looking information to traditional cross-sectional return signals.
- The multiplicative term (e.g., ridge * surprise) played a key role in amplifying differences between top and bottom-ranked stocks, improving cross-sectional spread and portfolio alpha.

Most profitable stocks

Top holdings such as **PERMNO 27828, 91461, and 85914** were frequently selected in the long leg of the market-value weighted strategy and contributed meaningfully to total returns. These stocks exhibited:

- Consistent top rankings in the hybrid signal distribution
- Positive realized returns across many holding periods
- Strong alignment with surprise forecasts

Conversely, short-leg positions such as **PERMNO 93436 and 75241** repeatedly underperformed, validating the bottom-ranking effectiveness of the model.

Macro events supporting performance

The strategy benefited from periods of macro-driven dispersion, especially:

- Post-2008 recovery and QE cycles (2010–2014): reward for beta and reversion signals
- COVID rebound (2020–2021): earnings surprise forecasts aligned well with actual recoveries
- Rising rate regime (2022–2023): return predictors favored defensive/quality stocks

The use of Goyal-Welch macro predictors (e.g., term spread, dividend-price ratio) in return modeling helped contextualize firm-level signals, especially during macro turning points.

Potential improvements

- **Forecasting horizon refinement:** Instead of monthly signals, incorporating quarterly forward returns may improve signal alignment with earnings cycles.
- **Model ensembling:** A weighted blend of Lasso, Ridge, and OLS predictions may provide a more stable signal across market regimes.
- **Style exposure control:** Risk factor exposures (e.g., size, value, momentum) could be monitored and constrained to avoid unintended tilts.
- **Turnover reduction:** Imposing signal smoothing or ranking inertia may reduce trading costs and improve net performance.
- **Alternative surprise models:** Exploring nonlinear models (e.g., gradient boosting or transformers) for surprise prediction, while managing interpretability and overfitting risk.

Appendix

R^2 for All Models:

round	start_dat	end_date	r2_ols	r2_ols_reducec	r2_lasso	r2_ridge	r2_enet_simple	r2_xgb	r2_ae_ridge	r2_ipca_all	r2_ipca_tuned
1	2000/1/1	2011/1/1	-0.052933595	-0.047165574	-0.031638338	-0.03637388	-0.032720495	-0.062569687	-0.035080874	-0.427288709	-0.312875029
2	2000/1/1	2012/1/1	-0.027217525	-0.014948388	-0.000234442	-0.00036039	-0.000229903	-0.045492536	0.00185249935	-0.201459574	-0.190040121
3	2000/1/1	2013/1/1	-0.061047656	-0.033852255	-0.019804914	-0.019223181	-0.019811248	-0.119405833	-0.024783951	-0.074732974	-0.108495978
4	2000/1/1	2014/1/1	-0.153327103	-0.130262661	-0.113649753	-0.112870831	-0.113638933	-0.145832772	-0.128198362	-0.417647201	-0.401601347
5	2000/1/1	2015/1/1	-0.038436495	-0.025707083	-0.005219174	-0.00394911	-0.005222553	-0.090754345	-0.015829303	-0.296930159	-0.177519277
6	2000/1/1	2016/1/1	-0.177859129	-0.016955647	-0.005723156	-0.005573173	-0.005737832	-0.072004239	-0.012568917	-0.367386315	-0.176690149
7	2000/1/1	2017/1/1	-0.113858533	-0.00680499	-0.005435032	-0.004328571	-0.005435032	-0.002262573	-0.001058155	-0.22326382	-0.150459037
8	2000/1/1	2018/1/1	-0.037163027	-0.01958932	-0.017074057	-0.016468573	-0.016441274	-0.140458061	-0.026505903	-0.215025926	-0.143865917
9	2000/1/1	2019/1/1	-0.036403997	-0.028571272	-0.027647911	-0.025787321	-0.027652674	-0.065650805	-0.029214639	-0.436048838	-0.343943297
10	2000/1/1	2020/1/1	-0.087955009	-0.034872105	-0.029287526	-0.029310322	-0.029287526	-0.112739988	-0.03167891	-0.07307463	-0.057233344
11	2000/1/1	2021/1/1	-0.024596358	-0.016673121	-0.010354052	-0.010650281	-0.010354052	-0.020843151	-0.018006728	-0.217687167	-0.080209785
12	2000/1/1	2022/1/1	-0.001117349	0.00265188716	-0.00758719	-0.006831668	-0.00758719	-0.051861801	0.00079114051	-0.073960353	-0.03841545
13	2000/1/1	2023/1/1	-0.027410397	-0.026184834	-0.02988772	-0.024573104	-0.025648075	-0.045416079	-0.023715743	-0.190689818	-0.178392909
14	2000/1/1	2024/1/1	-0.080838122	-0.015734128	-0.010435486	-0.003794244	-0.009249865	-0.081399202	-0.001538322	-0.058940241	-0.02197024
			r2_ols	r2_ols_reducec	r2_lasso	r2_ridge	r2_enet_simple	r2_xgb	r2_ae_ridge	r2_ipca_all	r2_ipca_tuned
Overall			-0.065726021	-0.032101644	-0.022427054	-0.021435332	-0.022072618	-0.075477934	-0.029014984	-0.233866838	-0.170122277

R^2 for Earning Surprise Prediction:

year	alpha	val_r2	test_r2
2011	0.0162377673918	-0.023589009	0.004066592423785265
2012	0.0037926901907	-0.002829854	-0.000194245
2013	0.0037926901907	0.00477301205914	-0.006895801
2014	0.0162377673918	-0.002258274	0.00042076255359435 6
2015	0.0078475997035	-0.001270534	0.0008507328973348383
2016	0.0078475997035	0.00153342835015	0.00022656931798381574
2017	0.0037926901907	0.00179464181488	-0.003052968
2018	0.0026366508987	0.00067600743514	-0.009357391
2019	0.0026366508987	-0.006595588	0.00772182510188590 4
2020	0.0026366508987	-0.000206653	-0.053607767
2021	0.0001	-0.021127897	-0.0880957
2022	0.0001	-0.075953423	0.011570588884355137
2023	0.0001	-0.031126173	0.006104158317107755

Appendix

Optimal Equally Weighted Strategy Output

```
=== Evaluating mixed-enhanced model: ols_reduced_enhanced_mix ===
Model ols_reduced_enhanced_mix: Dynamic threshold = 0.0405

=== Evaluating mixed-enhanced model: ridge_enhanced_mix ===
Model ridge_enhanced_mix: Dynamic threshold = 0.0322

=== Evaluating mixed-enhanced model: ols_enhanced_mix ===
Model ols_enhanced_mix: Dynamic threshold = 0.0486
                          Sharpe   Alpha  Alpha (Annualized)  Alpha t-stat  \
Model
ols_reduced_enhanced_mix  0.9126  0.0173              0.2080        3.6406
ridge_enhanced_mix       -0.6319 -0.0076             -0.0915       -1.1083
ols_enhanced_mix          0.3772  0.0089              0.1068        2.2911

                              IR  Annualized Return  Annualized Vol  Turnover  \
Model
ols_reduced_enhanced_mix  1.1661             0.1780          0.1853    0.3457
ridge_enhanced_mix       -0.4236            -0.1351          0.2277    0.3095
ols_enhanced_mix          0.6512             0.0739          0.1732    0.4092

                          Max Drawdown  Max Monthly Loss
Model
ols_reduced_enhanced_mix       -0.3890           -0.2186
ridge_enhanced_mix             -0.9172           -0.1989
ols_enhanced_mix               -0.3898           -0.1966
```

Optimal Market-Value Weighted Strategy Output

```
=== Evaluating model with dynamic threshold: ols_enhanced ===
Model ols_enhanced: Dynamic threshold = 0.0682

=== Strategy Performance Summary ===
                  Model  Sharpe   Alpha  Alpha (Annualized)  Alpha t-stat  \
0           ols_reduced  0.6566  0.0089              0.1065        1.7178
1                 ridge  0.6132  0.0100              0.1197        2.1562
2                   ols  0.7301  0.0100              0.1197        2.5046
3  ols_reduced_enhanced  0.6274  0.0084              0.1014        1.6635
4        ridge_enhanced  0.9647  0.0146              0.1750        3.3671
5          ols_enhanced  0.7329  0.0099              0.1190        2.6002

       IR  Annualized Return  Annualized Vol  Turnover  Max Drawdown  \
0  0.5588             0.1261          0.1920    0.4266        0.6580
1  0.6908             0.1065          0.1737    0.3713        0.5797
2  0.7189             0.1212          0.1660    0.4549        0.3051
3  0.5302             0.1208          0.1925    0.4250        0.6918
4  1.0472             0.1617          0.1676    0.3760        0.2379
5  0.7302             0.1190          0.1624    0.4543        0.2820

   Max Monthly Loss
0           -0.1740
1           -0.1212
2           -0.1774
3           -0.1790
4           -0.1037
5           -0.1731
```