Empirical Asset Pricing via Machine Learning

Final Presentation

IEOR E4576 Data Driven Methods in Finance

Presenters:
Ekow Andoh
Xuanyu Guo
Dong Xia
Peng Zhong

Dec 15, 2025

Project Overview

How can Machine Learning methods predict cross-sectional stock returns more accurately than linear models?

Goals:

- Improve out-of-sample prediction of asset returns.
- Examine implications: drawdown, Sharpe ratios, turnover.
- Calculate rolling covariances using empirical or PCA based methods
- Use an optimizer with constraints to generate portfolios
- Compare linear models with complex models/ Stratified portfolios and optimized portfolios

Methods

1. **Data Construction**:
    - Subset of 153 factors from the overall data for computational efficiency

2. **Methods Compared**:
    - Traditional: OLS
    - Regularized: Elastic Net
    - Non-linear: PLS/ PCR, Random Forest, GBRT, Neural Network (2 layer) , Neural Network (5 layer)

3. **Evaluation Metrics**:
    - Out-of-sample $\mathbf{R^2}$ (predictive power).
    - Portfolio return statistics (Sharpe ratio, turnover, alpha, drawdown).

Preprocessing

- Filter out penny stocks
- Only the independent variables with less than 25 per cent missing values are retained.
- Replace the missing values with the cross-sectional median at each month
- All monthly firm characteristics are winsorized at the 1% and 99% levels to ensure that the results are insensitive to outliers
- 214,016 observations left

20 Features List

Momentum

- **ret_1_0** (orig: mom1m) - Past 1 month return
- **ret_6_1** (orig: mom6m) - Past 6 months return (skip most recent 1 month)
- **ret_12_1** (orig: mom12m) - Past 12 months return (skip most recent 1 month)
- **ret_36_1** (orig: mom36m) - Past 36 months return
- **chmom** - Momentum change = ret_12_1 - ret_1_0

Reversal/Price Behavior

- **rmax1_21d** (orig: maxret) - Maximum return over past 21 days
- **rvol_21d** (orig: retvol) - Return volatility over past 21 days

Liquidity

- **turnover_126d** (orig: turn) - Average turnover over past 126 days
- **std_turn** - Turnover standard deviation
- **dolvol_126d** (orig: dolvol) - Dollar volume over past 126 days
- **bidaskhl_21d** (orig: baspread) - Bid-ask spread
- **zero_trades_252d** (orig: zerotrade) - Zero trade days ratio over past 252 days

Size/Valuation

- **me** (orig: mve) - Market capitalization
- **be_me** (orig: bm) - Book-to-market ratio
- **cashpr** - Cash price ratio

Profitability

- **qmj_prof** (orig: operprof) - Profitability
- **roeq** - Return on equity

Risk

- **beta_60m** (orig: beta) - 60-month Beta
- **betasq** - Beta squared
- **ivol_capm_252d** (orig: idiovol) - CAPM residual volatility

Industry

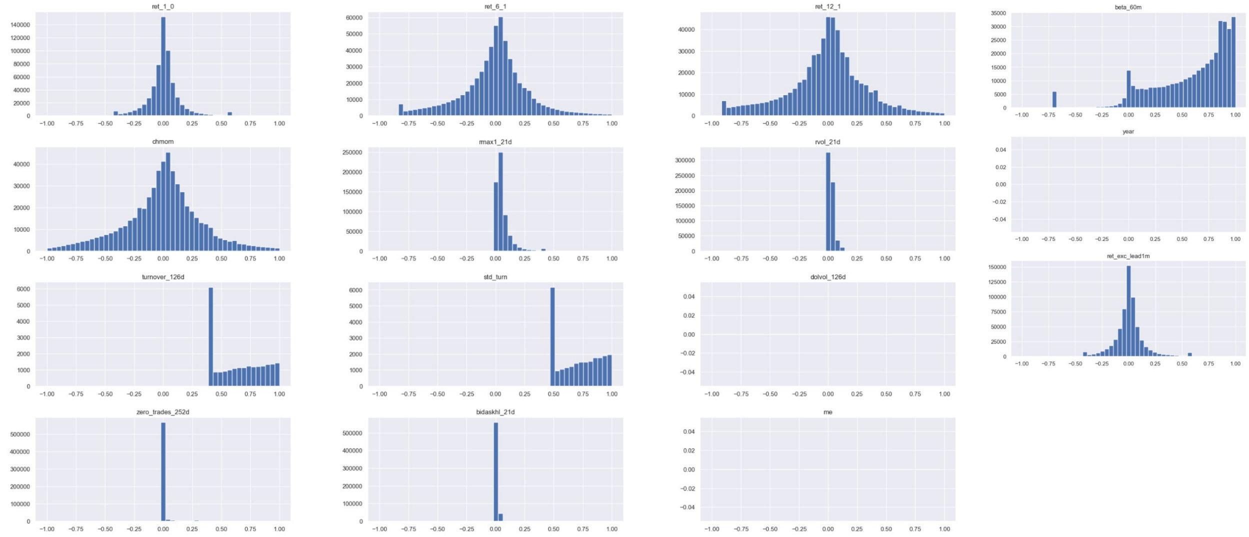
- **sic2** - 2-digit SIC industry code

Feature Name Mapping Table

Paper/Original Name	Project Variable Name	Description
mom1m	ret_1_0	Past 1 month return
mom6m	ret_6_1	Past 6 months return (skip most recent 1 month)
mom12m	ret_12_1	Past 12 months return (skip most recent 1 month)
mom36m	ret_36_1	Past 36 months return
maxret	rmax1_21d	Maximum return over past 21 days
retvol	rvol_21d	Return volatility over past 21 days
turn	turnover_126d	Average turnover over past 126 days
dolvol	dolvol_126d	Dollar volume over past 126 days
baspread	bidaskhl_21d	Bid-ask spread
zerotrade	zero_trades_252d	Zero trade days ratio over past 252 days
mve	me	Market capitalization
bm	be_me	Book-to-market ratio
beta	beta_60m	60-month Beta
idiovol	ivol_capm_252d	CAPM residual volatility
operprof	qmj_prof	Profitability

Exploratory Data Analysis

Distribution of each feature:



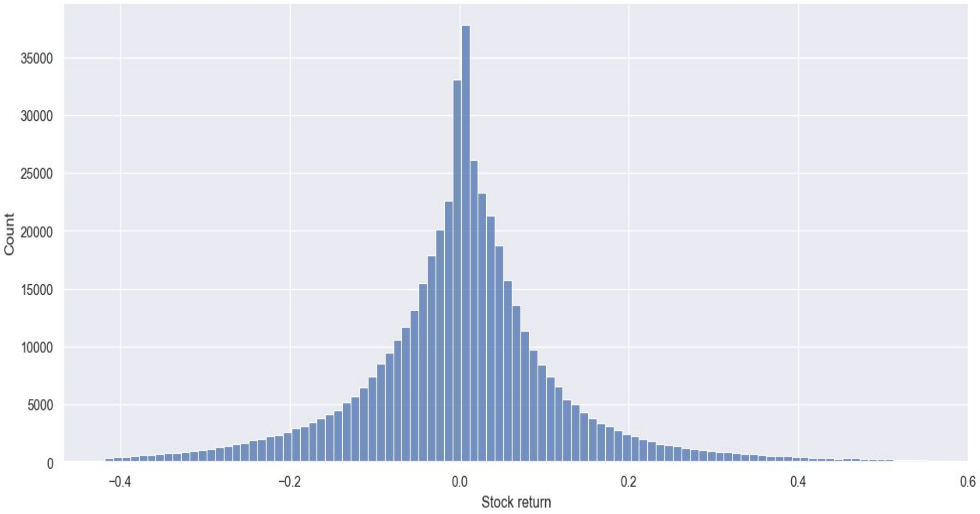
Basic Information of Data:

- Time Frame: 2019-01-01 - 2023-12-01
- Number of unique stocks: 12219

Exploratory Data Analysis

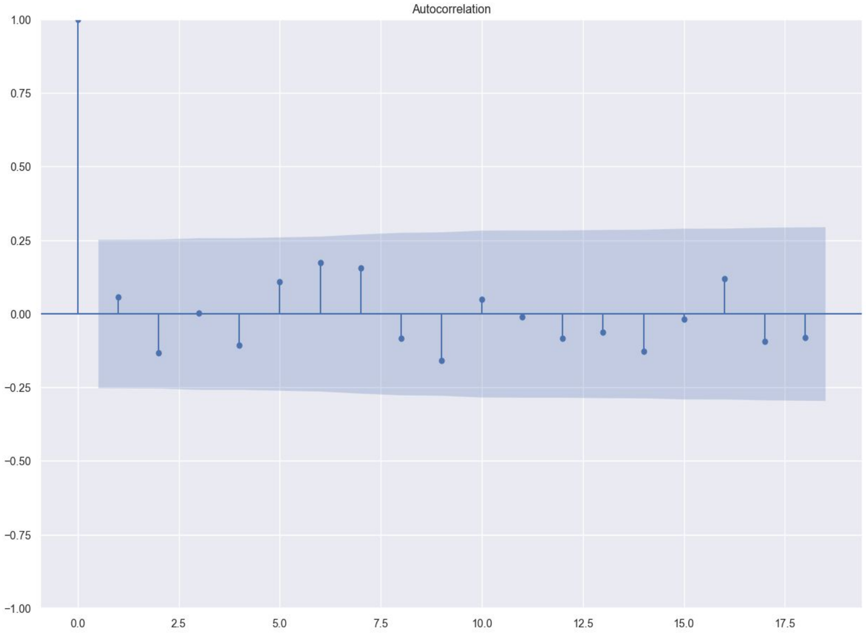
Equally weighted portfolio monthly returns over time:

Distribution of portfolio returns:

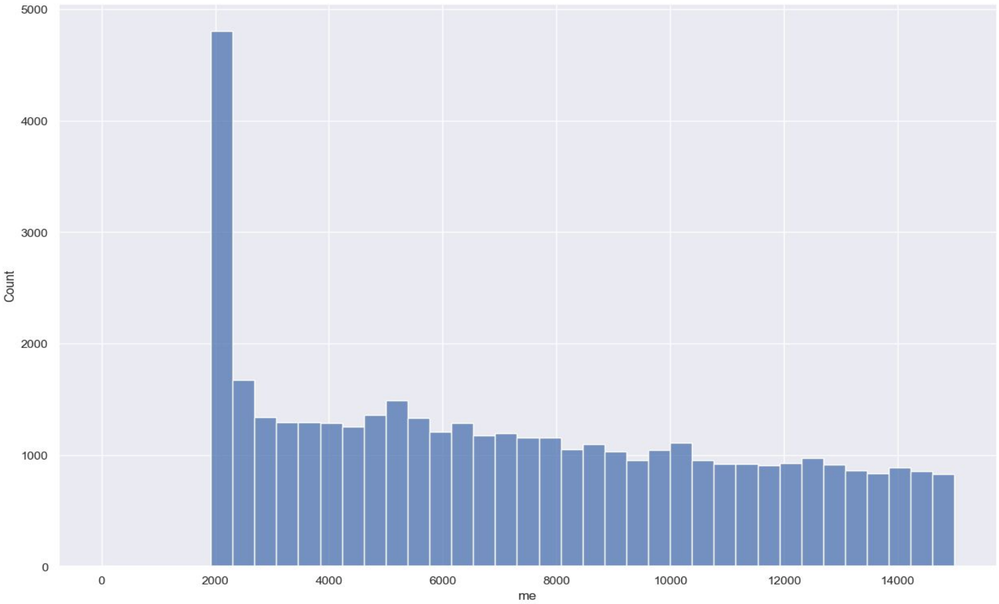


Exploratory Data Analysis

Portfolio autocorrelation graph:



Distribution of firm size:

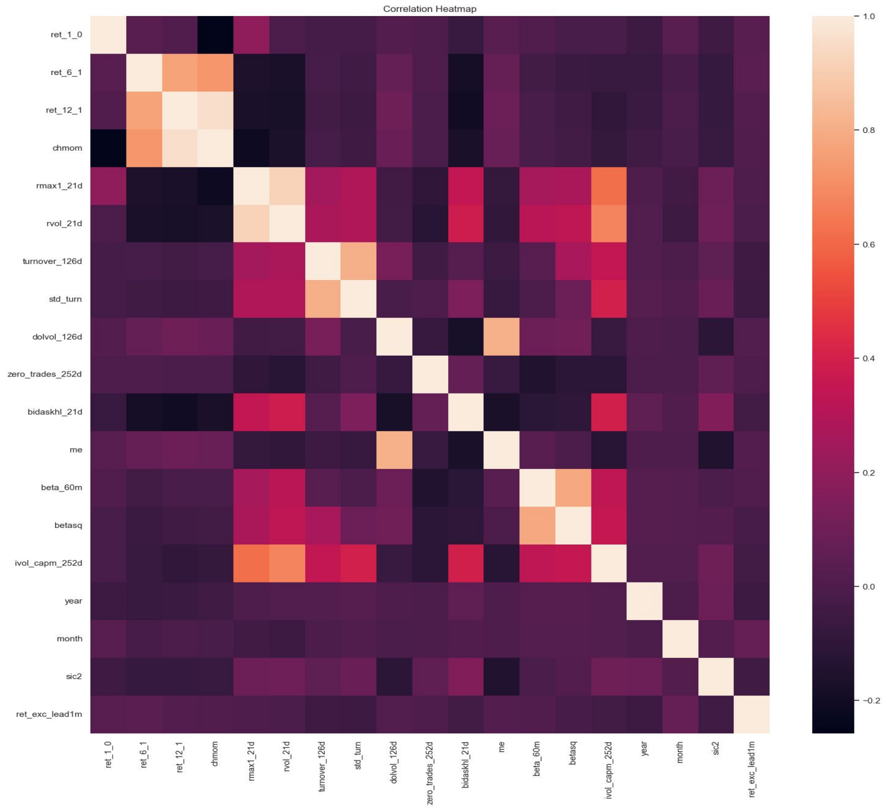


Exploratory Data Analysis

Top 10 correlated features:

	Variable 1	Variable 2	corr
0	ret_12_1	chmom	0.957494
1	rvol_21d	rmax1_21d	0.917116
2	me	dolvol_126d	0.813184
3	turnover_126d	std_turn	0.805406
4	betasq	beta_60m	0.785807
5	ret_6_1	ret_12_1	0.767594
6	ret_6_1	chmom	0.725515
7	rvol_21d	ivol_capm_252d	0.678469
8	rmax1_21d	ivol_capm_252d	0.617385
9	std_turn	ivol_capm_252d	0.398834
10	ivol_capm_252d	bidaskhl_21d	0.391477

Variable correlation heatmap:



OLS Model: Ordinary Least Squares Regression

The simplest baseline model for predicting stock returns: $r_{i,t+1} = z_{i,t}^T \beta + \epsilon_{i,t+1}$

No regularization is used, directly minimizing the sum of squared residuals

**The paper mentions a 920-dimensional feature space (94 stock features × 8 macro variables + constant), but this implementation uses 16 stock features only due to data availability.

OLS Model Prediction Performance Metrics:

```
======================================
r2_os                  :  -0.0234
mse                    :   0.0202
rmse                   :   0.1422
mae                    :   0.0922
correlation            :   0.0219
mean_pred              :   0.0026
mean_true              :  -0.0014
std_pred               :   0.0244
std_true               :   0.1406
n_observations         : 552772
======================================
```

3-Way Portfolio Construction:

```
======================================================================
Simple Stratification Portfolio Performance (Baseline)
======================================================================
annual_return            :     0.2586
annual_volatility        :     0.2241
sharpe_ratio             :     1.1538
cumulative_return        :     0.8877
n_months                 : 33
mean_monthly_return      :     0.0215
std_monthly_return       :     0.0647
======================================================================
Optimized Portfolio Performance (Empirical Covariance Method)
======================================================================
annual_return            :     0.4331
annual_volatility        :     0.3075
sharpe_ratio             :     1.4086
cumulative_return        :     1.8722
n_months                 : 33
mean_monthly_return      :     0.0361
std_monthly_return       :     0.0888
method                   : empirical
risk_aversion            :     1.0000
======================================================================
Optimized Portfolio Performance (PCA-based Covariance Method)
======================================================================
annual_return            :     0.4337
annual_volatility        :     0.3154
sharpe_ratio             :     1.3751
cumulative_return        :     1.8636
n_months                 : 33
mean_monthly_return      :     0.0361
std_monthly_return       :     0.0910
method                   : pca
risk_aversion            :     1.0000
======================================================================
```

OLS Model: Ordinary Least Squares Regression

Portfolio Performance Comparison:

```
================================================================================
PORTFOLIO PERFORMANCE COMPARISON
================================================================================
                 Method  Annual Return  Annual Volatility  Sharpe Ratio  Cumulative Return
Simple Stratification         0.258581           0.224110      1.153811           0.887744
Optimized (Empirical)         0.433141           0.307497      1.408600           1.872236
      Optimized (PCA)         0.433673           0.315371      1.375122           1.863605

================================================================================
```

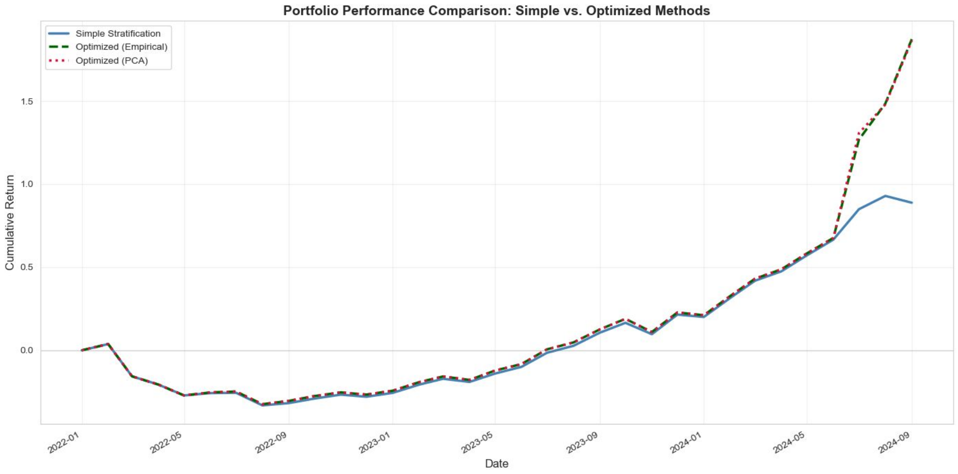
OLS Model Prediction Performance Metrics:

- Simple Stratification: Equal-weighted decile portfolios (baseline)
- Optimized (Empirical): Uses rolling empirical covariance + mean-variance optimization
- Optimized (PCA): Uses PCA-based covariance estimation + mean-variance optimization

The optimized methods consider:

- Stock correlations (covariance matrix)
- Risk-adjusted returns (mean-variance optimization)
- Constraints (max weight per stock, long-only)

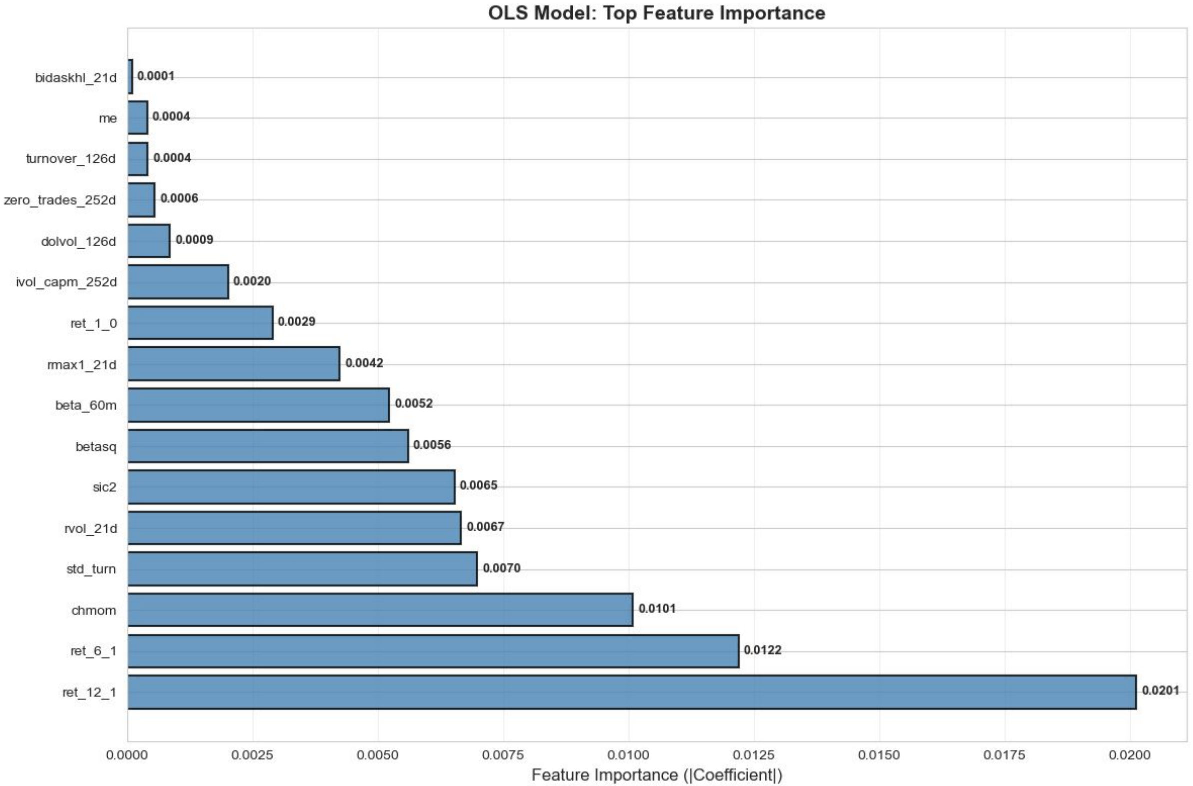
Portfolio Performance Comparison: Simple vs. Optimized Methods:



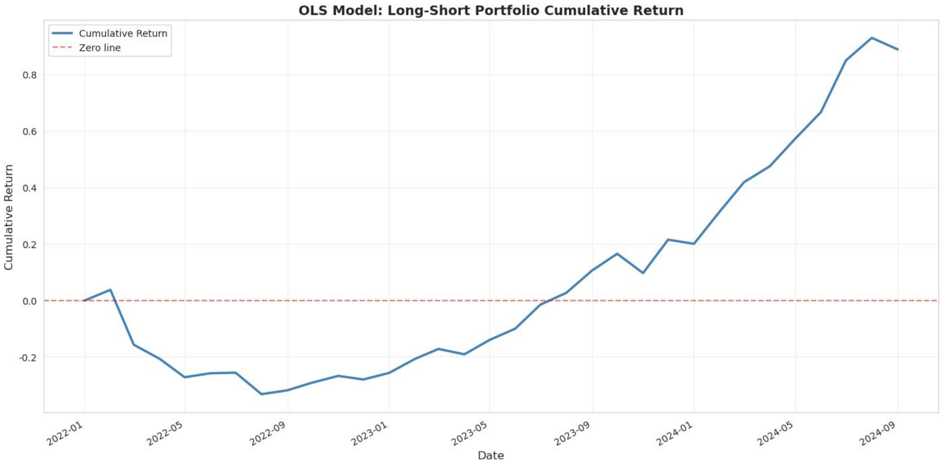

```
====================================================
Long-Short Portfolio Performance (OLS Model)
====================================================
annual_return            :     0.2586
annual_volatility        :     0.2241
sharpe_ratio             :     1.1538
cumulative_return        :     0.8877
n_months                 : 33
mean_monthly_return      :     0.0215
std_monthly_return       :     0.0647
====================================================
```

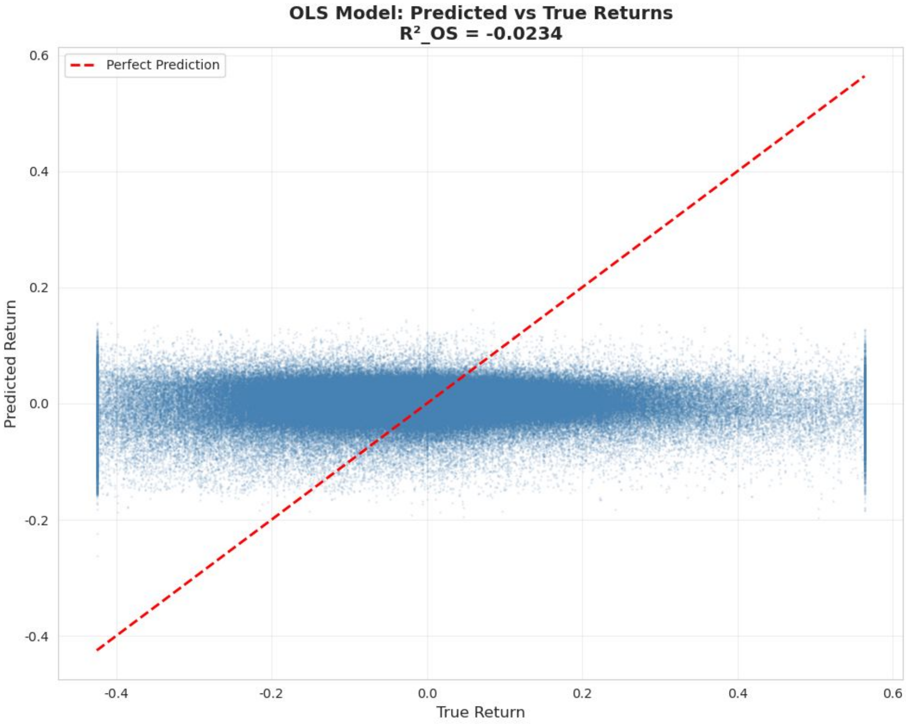
OLS Model: Ordinary Least Squares Regression

Feature Importance Analysis:



OLS Model: Ordinary Least Squares Regression





Cumulative Return Statistics:

- **Final Cumulative Return: 88.77%**
- **Max Drawdown: -36.93%**

OLS Model: Ordinary Least Squares Regression

Mean Monthly Return by Decile:

Elastic Net Model: Elastic Net Regression

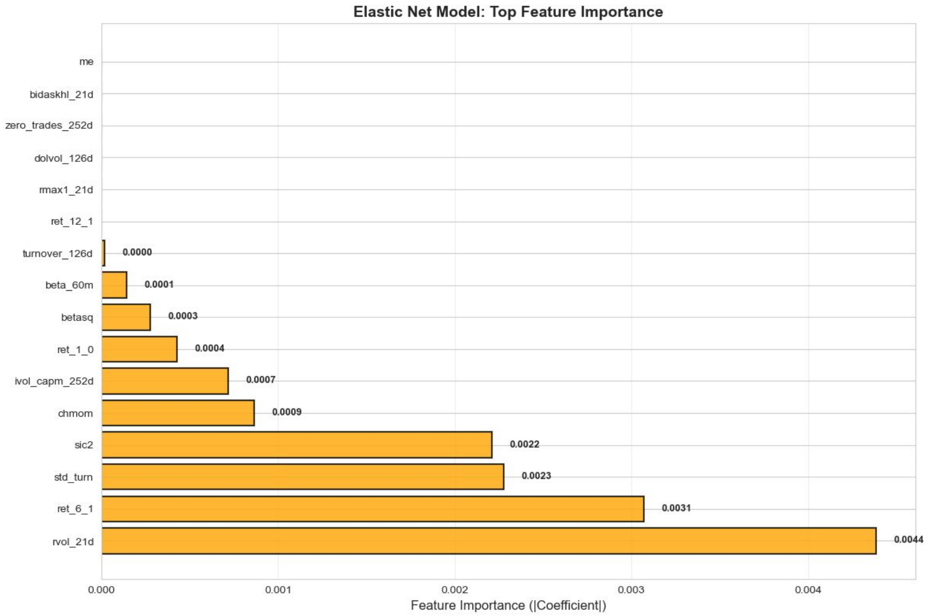
Elastic Net combines L1 (LASSO) and L2 (Ridge) regularization: $\min_{\beta} \sum (y - X\beta)^2 + \lambda[\alpha||\beta||_1 + (1-\alpha)||\beta||_2^2]$

- Can perform feature selection and prevent overfitting simultaneously
- Better performance on high-dimensional data

Elastic Net Model Prediction Performance Metrics:

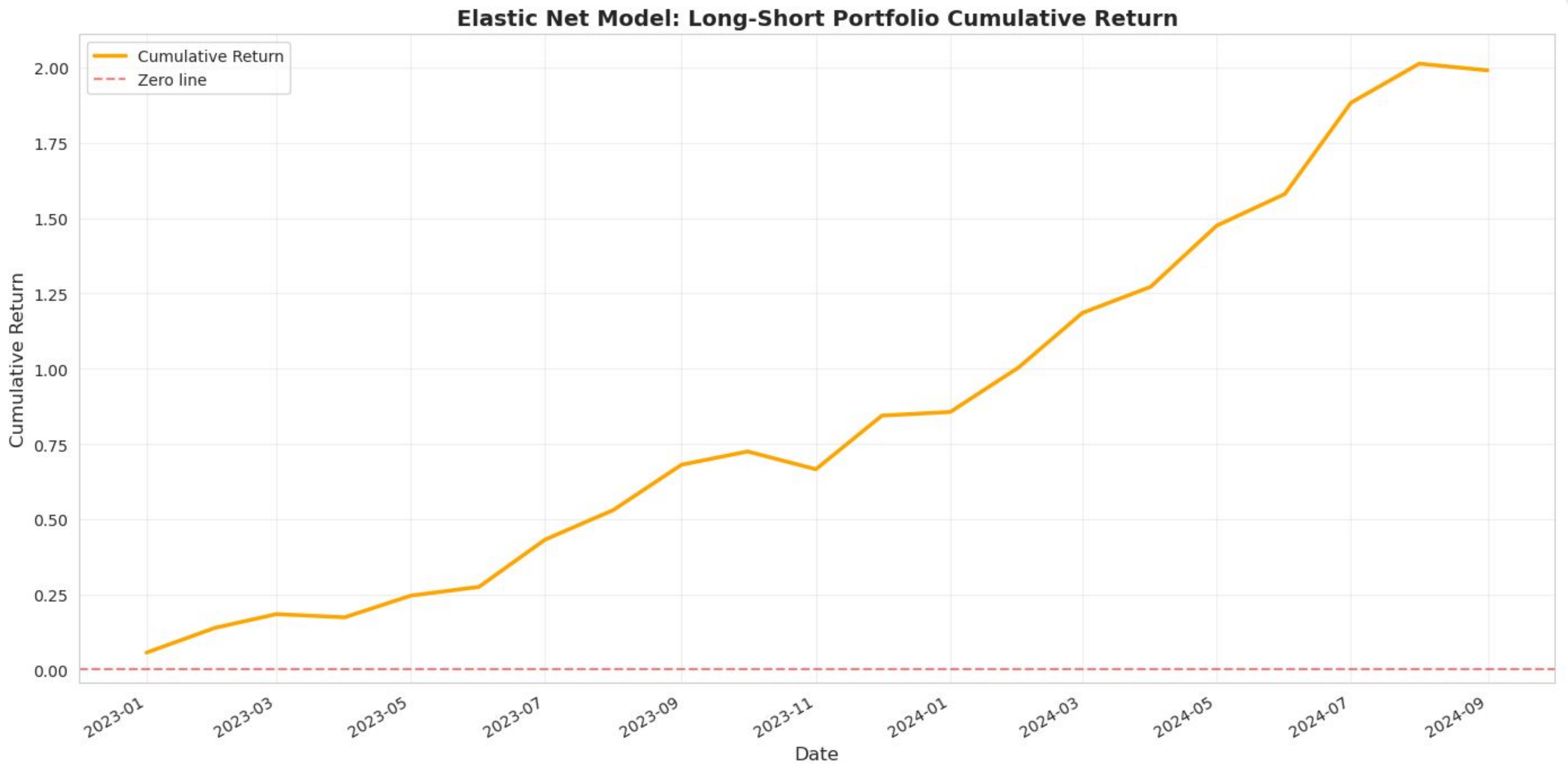
```
======================================================================
Elastic Net Model Prediction Performance Metrics
======================================================================
Metric                     Value
----------------------------------------------------------------------
R²_OS                      -0.0011
RMSE                        0.1406
MAE                         0.0910
Correlation                 0.0426
Mean Predicted             -0.0004
Mean True                  -0.0014
Std Predicted               0.0135
Std True                    0.1406
N Observations              552772
======================================================================

======================================================================
Best Hyperparameters Across Folds
======================================================================
           alpha   l1_ratio
count  10.000000  10.000000
mean    0.037000   0.380000
std     0.043474   0.193218
min     0.010000   0.100000
25%     0.010000   0.300000
50%     0.010000   0.300000
75%     0.077500   0.450000
max     0.100000   0.700000

Most frequently selected parameters:
  Alpha: 0.01 (appears 7 times)
  L1_Ratio: 0.3 (appears 6 times)
======================================================================
```

Feature Importance Analysis:



**Number of features regularized to zero: 6/16

Elastic Net Model: Result Visualization



Elastic Net Model: Elastic Net Regression

Portfolio Performance Comparison：

PORTFOLIO PERFORMANCE COMPARISON

Method	Annual Return	Annual Volatility	Sharpe Ratio	Cumulative Return
Simple Stratification	0.652955	0.150429	4.340621	1.990832
Optimized (Empirical)	0.919247	0.266145	3.453939	3.491434
Optimized (PCA)	0.911499	0.286823	3.177910	3.406145

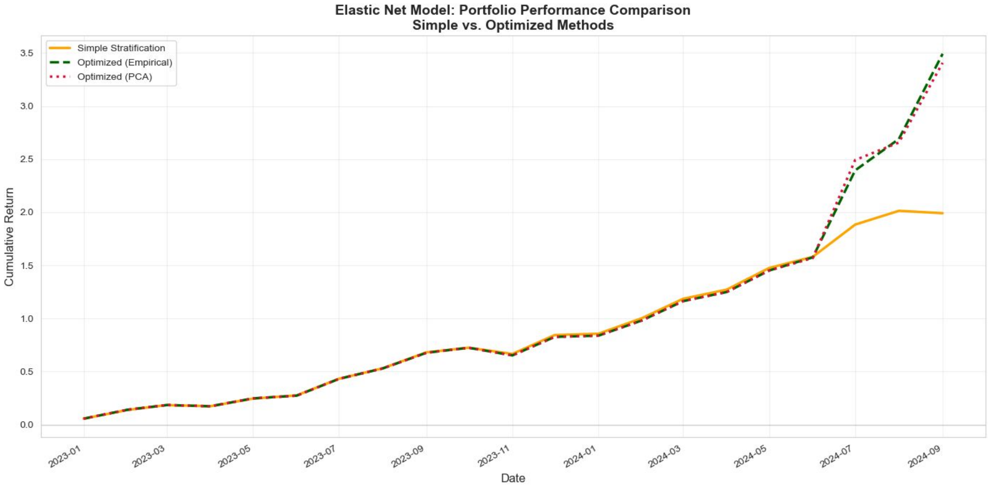
Elastic Net Model Prediction Performance Metrics:

1. Simple Stratification: Equal-weighted decile portfolios (baseline)
2. Optimized (Empirical): Uses rolling empirical covariance + mean-variance optimization
3. Optimized (PCA): Uses PCA-based covariance estimation + mean-variance optimization

The optimized methods consider:

- Stock correlations (covariance matrix)
- Risk-adjusted returns (mean-variance optimization)
- Constraints (max weight per stock, long-only)

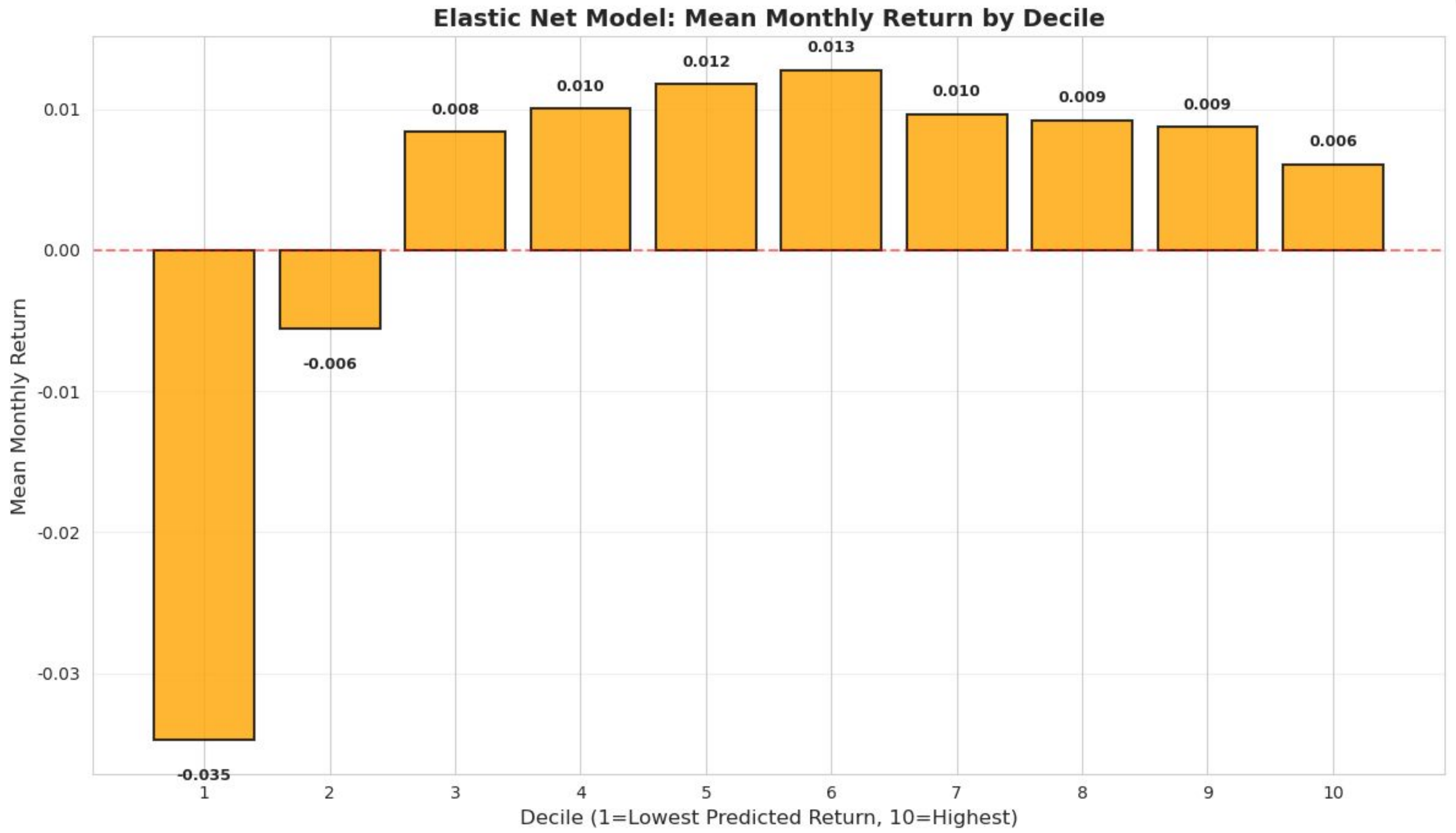
Portfolio Performance Comparison: Simple vs. Optimized Methods:



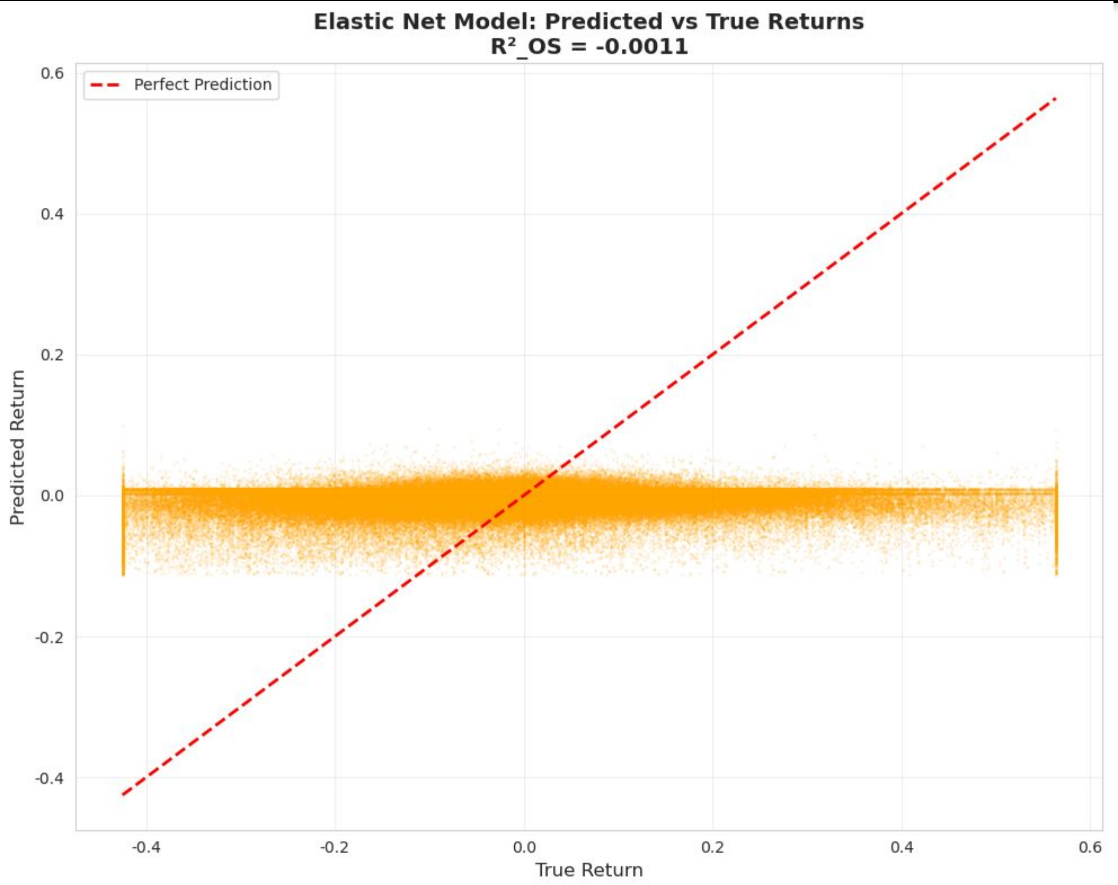
Final Cumulative Returns:

- Simple Stratification: 199.08%
- Optimized (Empirical): 349.14%
- Optimized (PCA): 340.61%

Elastic Net Model: Result Visualization



Elastic Net Model: Result Visualization



PLS/PCR Models: Partial Least Squares and Principal Component

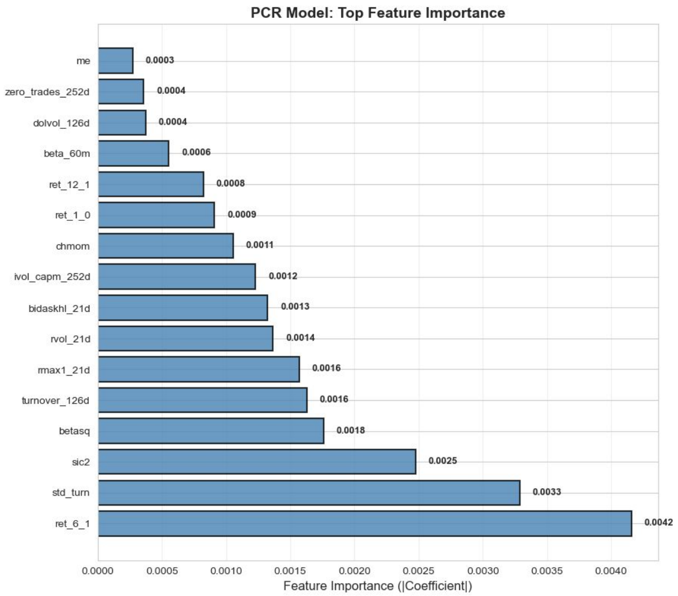
PCR (Principal Component Regression)

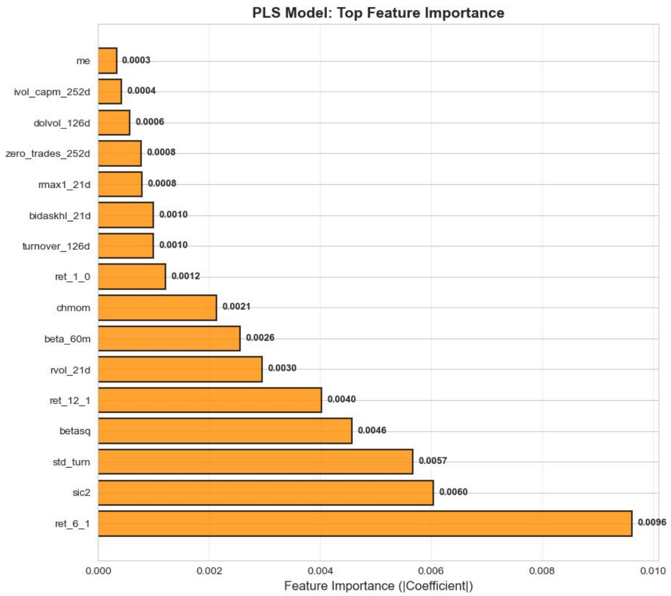
1. Apply PCA to feature matrix X, extract top k principal components
2. Perform OLS regression on principal components

PLS (Partial Least Squares)

1. Consider both X and y correlations for dimensionality reduction
2. Extract components that are more "prediction-oriented"

Feature Importance Analysis:





- Both methods require selecting the number of principal/PLS components (n_components)

PLS/PCR Models: Partial Least Squares and Principal Component

Portfolio Performance Comparison:

```
================================================================================
PCR MODEL: PORTFOLIO PERFORMANCE COMPARISON
================================================================================
               Method  Annual Return  Annual Volatility  Sharpe Ratio  Cumulative Return
Simple Stratification       0.319204           0.221086      1.443799           1.231409
Optimized (Empirical)       0.589963           0.326797      1.805289           3.308865
      Optimized (PCA)       0.595240           0.367653      1.619025           3.234022
================================================================================

================================================================================
PLS MODEL: PORTFOLIO PERFORMANCE COMPARISON
================================================================================
               Method  Annual Return  Annual Volatility  Sharpe Ratio  Cumulative Return
Simple Stratification       0.400855           0.197552      2.029115           1.811041
Optimized (Empirical)       0.581515           0.294370      1.975455           3.313956
      Optimized (PCA)       0.588023           0.301493      1.950367           3.370283
================================================================================
```

Elastic Net Model Prediction Performance Metrics:

1. Simple Stratification: Equal-weighted decile portfolios (baseline)
2. Optimized (Empirical): Uses rolling empirical covariance + mean-variance optimization
3. Optimized (PCA): Uses PCA-based covariance estimation + mean-variance optimization

The optimized methods consider:

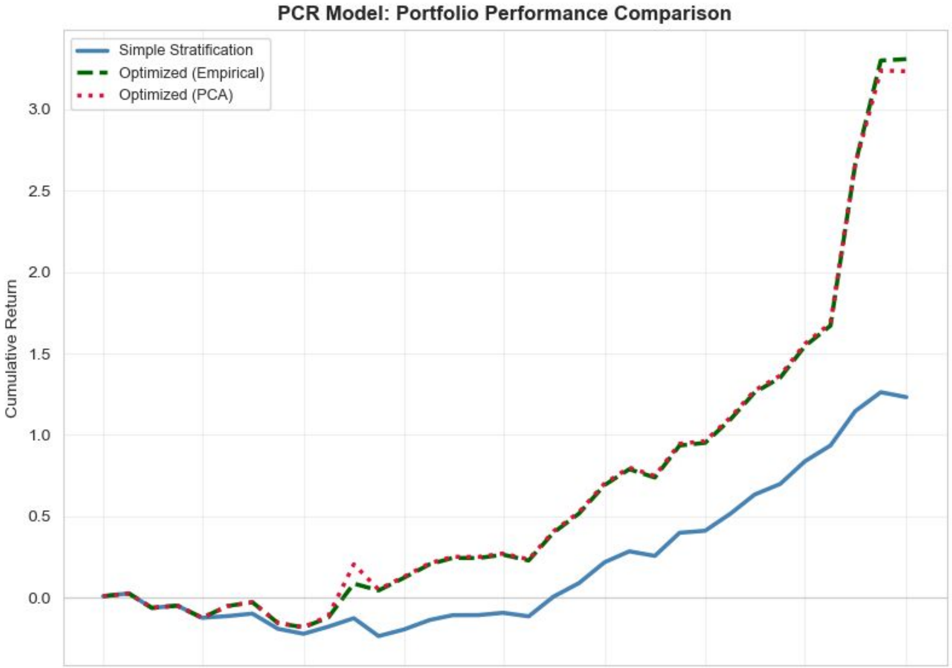
- Stock correlations (covariance matrix)
- Risk-adjusted returns (mean-variance optimization)
- Constraints (max weight per stock, long-only)

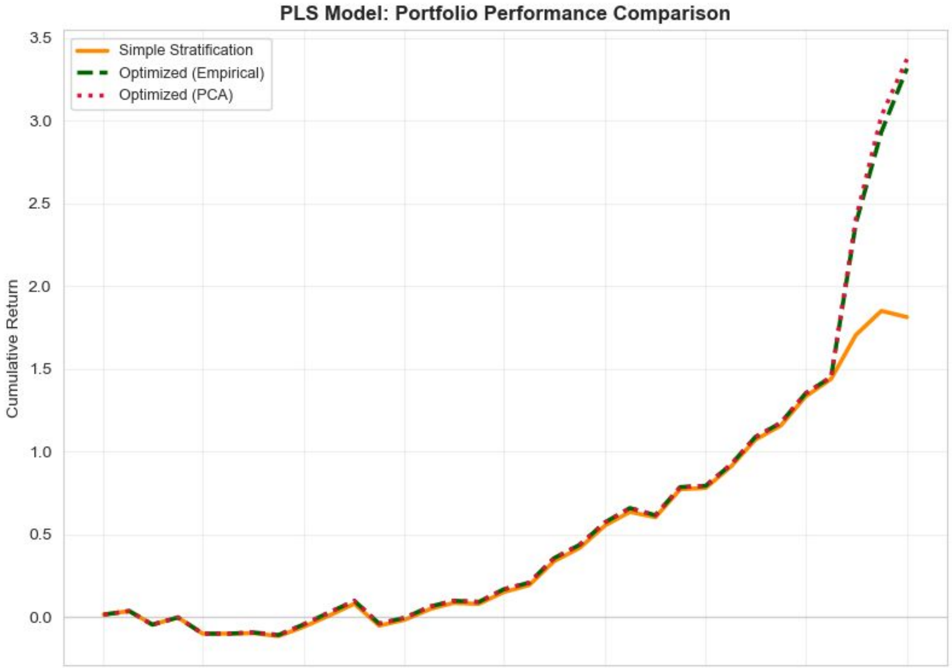
Long-Short Portfolio Returns:

Metric	PCR Model	PLS Model
Annual Return	31.92%	**40.09%**
Annual Volatility	22.11%	**19.76%**
Sharpe Ratio	1.44	**2.03**
Cumulative Return	123.14%	**181.10%**
Test Period	33 months	33 months

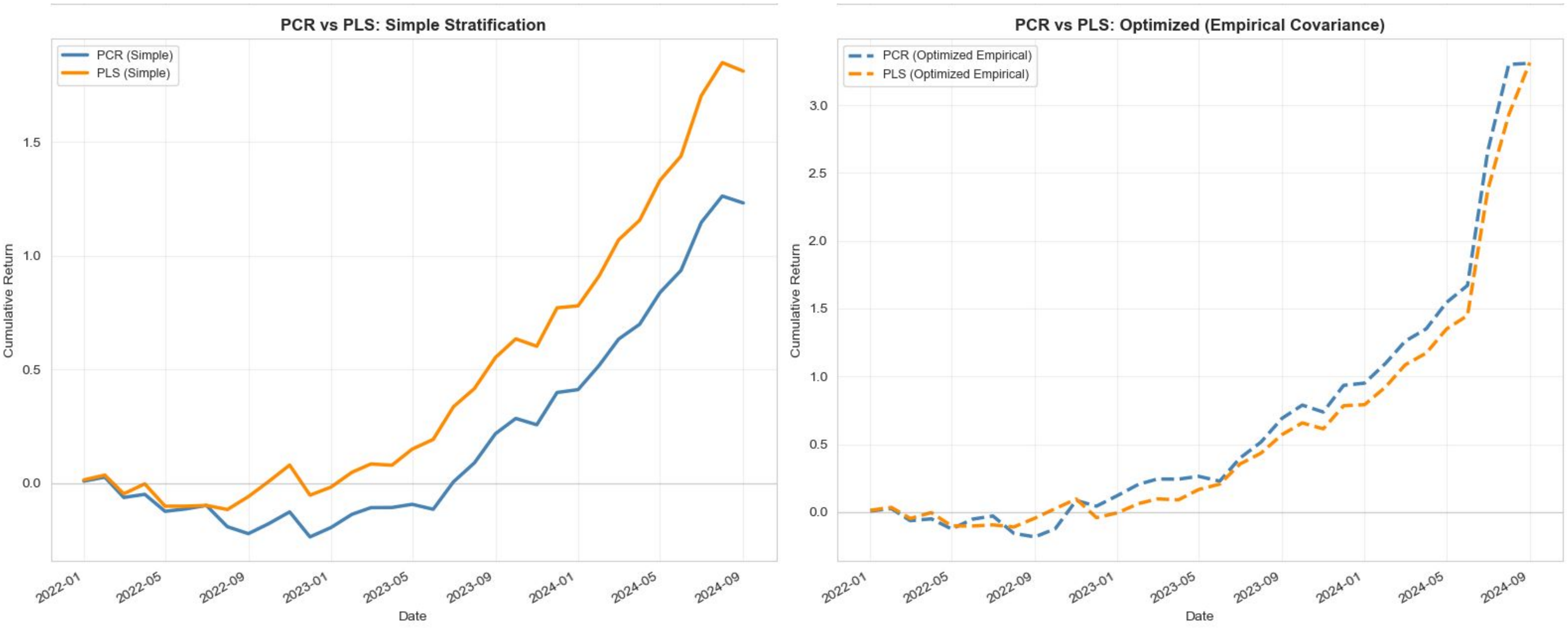
- **PLS significantly outperforms PCR** in portfolio construction
- PLS achieves 25% higher annual returns (40.09% vs 31.92%)
- PLS has lower volatility (19.76% vs 22.11%), resulting in superior risk-adjusted returns
- Sharpe ratio for PLS (2.03) is 41% higher than PCR (1.44), indicating outstanding risk-adjusted performance

PLS/PCR Models: Partial Least Squares and Principal Component

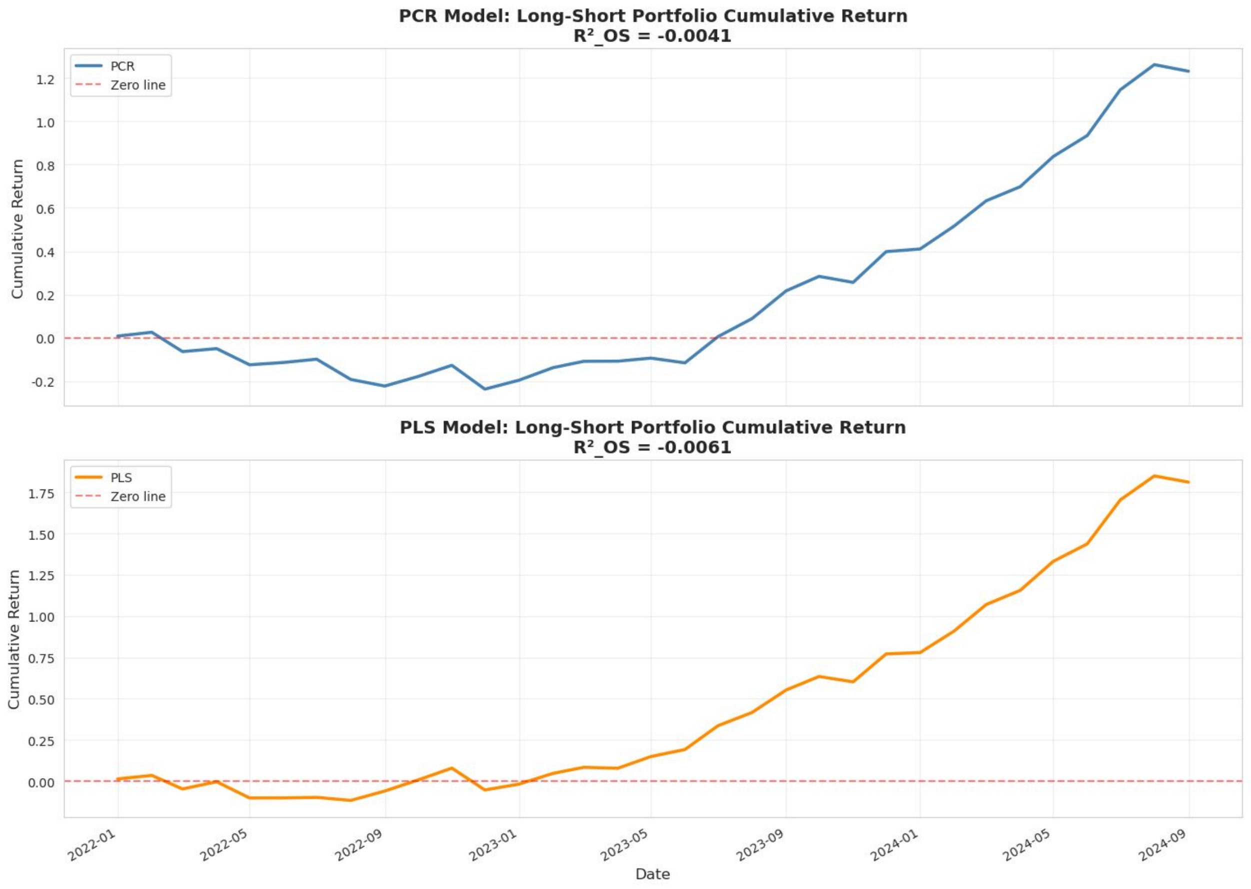




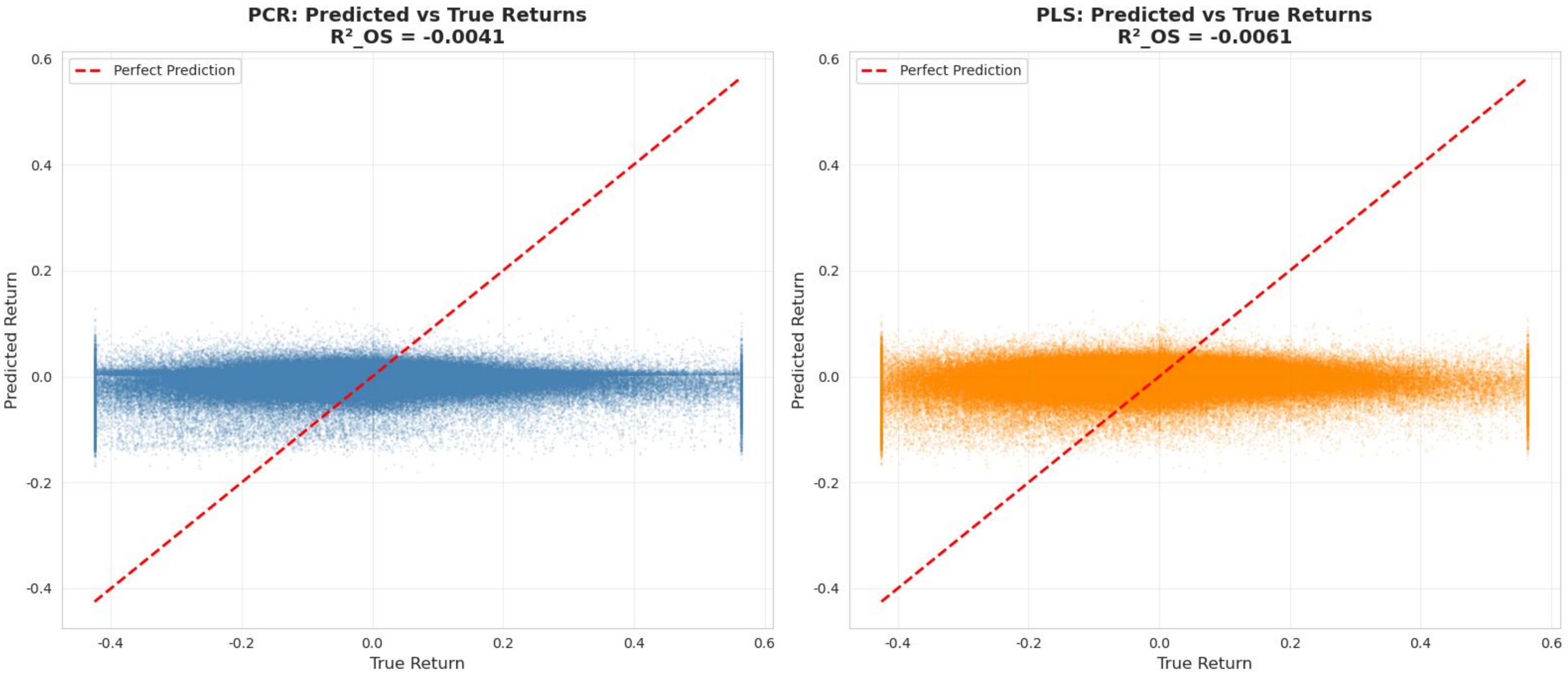
PLS/PCR Models: Partial Least Squares and Principal Component



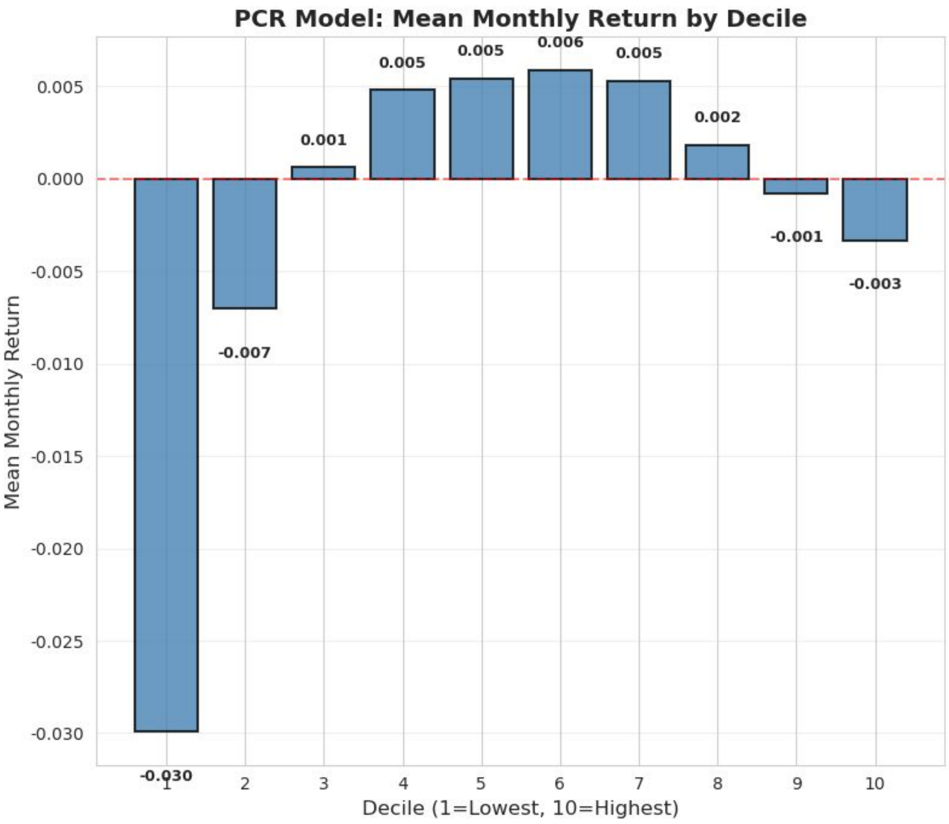
PLS/PCR Models: Result Visualization

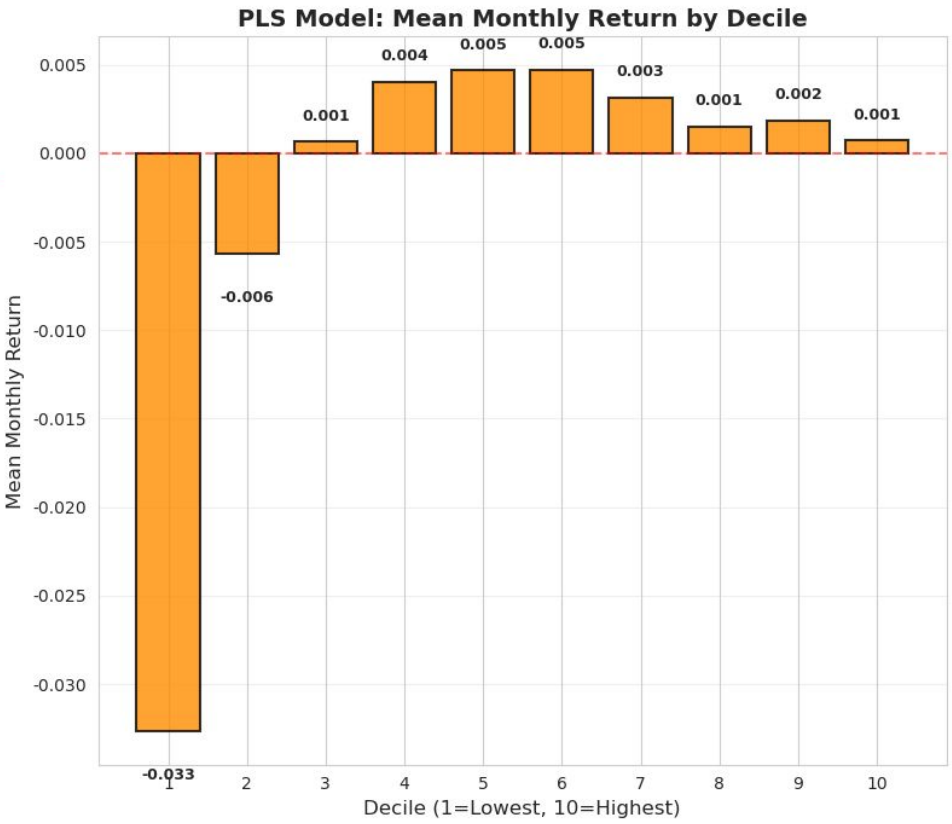


PLS/PCR Models: Result Visualization



PLS/PCR Models: Result Visualization





PLS/PCR Models: Key Conclusions

Model Comparison

1. **PLS outperforms PCR** in both prediction quality (correlation) and portfolio performance (Sharpe ratio)
2. PLS's advantage comes from considering X-y correlations during dimensionality reduction, making it more "prediction-oriented"
3. Both models show similar feature importance patterns, but PLS assigns more decisive weights

Practical Implications

1. **Dimensionality reduction is beneficial**: Both PCR and PLS improve upon simple OLS by reducing overfitting
2. **Fewer components are often better**: Models frequently select 1-5 components, suggesting that most predictive information is captured in a low-dimensional space
3. **Portfolio construction is viable**: Despite low R^2_OS values, both models generate economically significant long-short portfolio returns

Feature Insights

1. **Momentum effects dominate**: Past returns (ret_6_1, ret_12_1) are the strongest predictors
2. **Liquidity matters**: Turnover and bid-ask spread measures contribute significantly
3. **Risk factors are important**: Beta and volatility measures help distinguish stocks
4. **Industry effects exist**: Industry classification (sic2) plays a role in return prediction

Overall Assessment:

1. Both models demonstrate predictive ability despite challenging prediction task
2. PLS model is recommended for practical applications due to superior portfolio performance
3. Dimensionality reduction techniques (PCR/PLS) are effective for high-dimensional asset pricing problems
4. The models identify economically meaningful features and generate actionable investment signals

Random Forest

Model Intuition & Implementation:

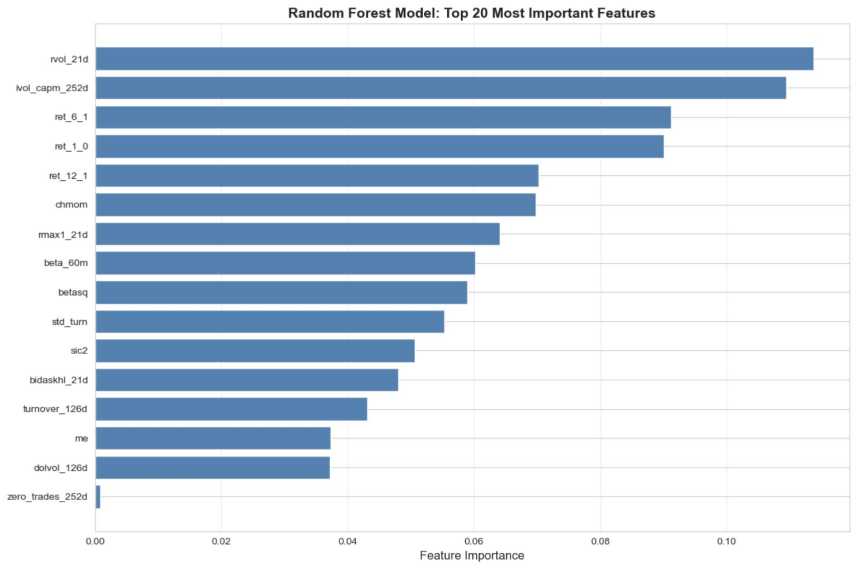
Ensemble of multiple decision trees

- **Bagging**: Each tree is trained on bootstrap samples
- **Feature Randomness**: Each tree only considers a random subset features
- **Non-linear Modeling**: Can capture complex interactions between features
- Limited hyperparameter tuning

Prediction performance Metrics:

```
==================================================
Random Forest Model Prediction Performance Metrics
==================================================
r2_os                :  -0.0198
mse                  :   0.0201
rmse                 :   0.1419
mae                  :   0.0925
correlation          :   0.0384
mean_pred            :  -0.0033
mean_true            :  -0.0014
std_pred             :   0.0258
std_true             :   0.1406
n_observations       : 552772
==================================================
```

Feature Importance Analysis:



Random Forest

Build Portfolio:

- **Baseline: Simple Stratification**

- **Rolling Covariance + Mean-Variance Optimization**:
  - **Rolling Covariance**: Calculate covariance matrix using a rolling window (12 months default)
  - **Methods**:
    - Empirical: Direct calculation from historical returns
    - PCA-based: Use Principal Component Analysis to reduce dimensionality before estimating covariance
  - **Optimization consider**:
    - Stock Correlations (covariance matrix)
    - Risk-adjusted returns (Mean-variance optimization)
    - Constraints (max weight per stock, long-only)

```
==================================================
Long-Short Portfolio Performance (Random Forest Model)
==================================================
annual_return            :     0.3963
annual_volatility        :     0.1656
sharpe_ratio             :     2.3930
cumulative_return        :     1.8211
n_months                 : 33
mean_monthly_return      :     0.0330
std_monthly_return       :     0.0478
==================================================
```

```
======================================================================
Optimized Portfolio Performance (Empirical Covariance Method)
======================================================================
annual_return            :     0.5846
annual_volatility        :     0.2496
sharpe_ratio             :     2.3418
cumulative_return        :     3.4673
n_months                 : 33
mean_monthly_return      :     0.0487
std_monthly_return       :     0.0721
method                   : empirical
risk_aversion            :     1.0000
======================================================================
```

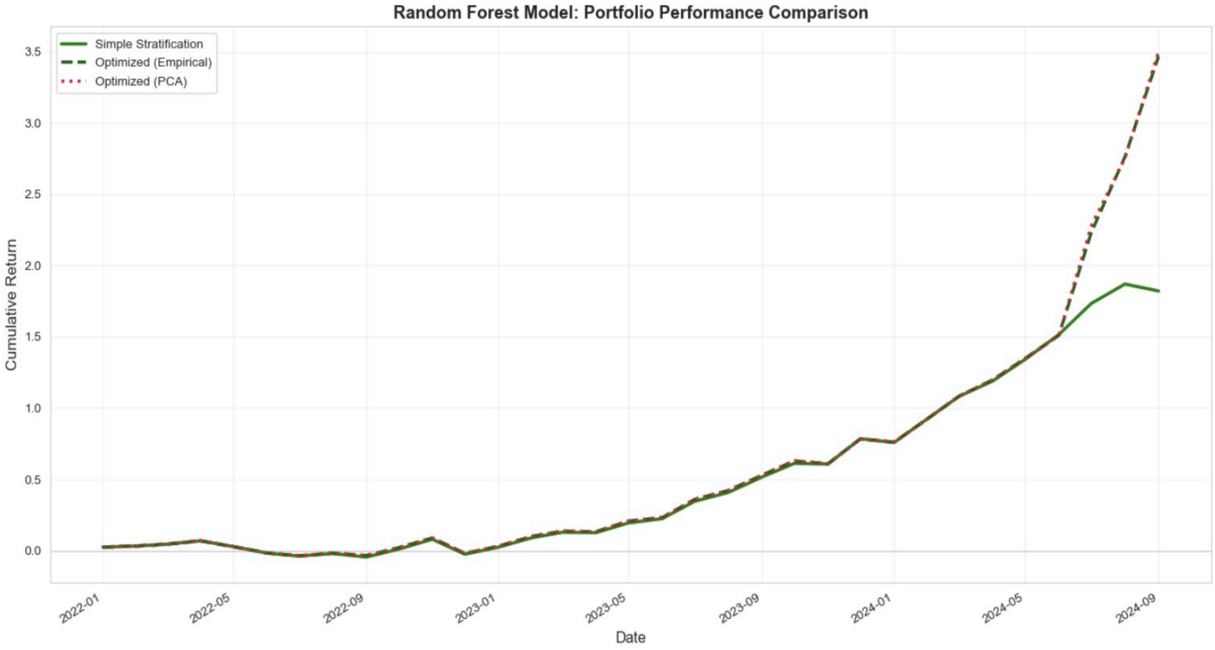
```
======================================================================
Optimized Portfolio Performance (PCA-based Covariance Method)
======================================================================
annual_return            :     0.5884
annual_volatility        :     0.2558
sharpe_ratio             :     2.3003
cumulative_return        :     3.4989
n_months                 : 33
mean_monthly_return      :     0.0490
std_monthly_return       :     0.0738
method                   : pca
risk_aversion            :     1.0000
======================================================================
```

Random Forest

Comparison:

optimized portfolio vs. simple stratification method

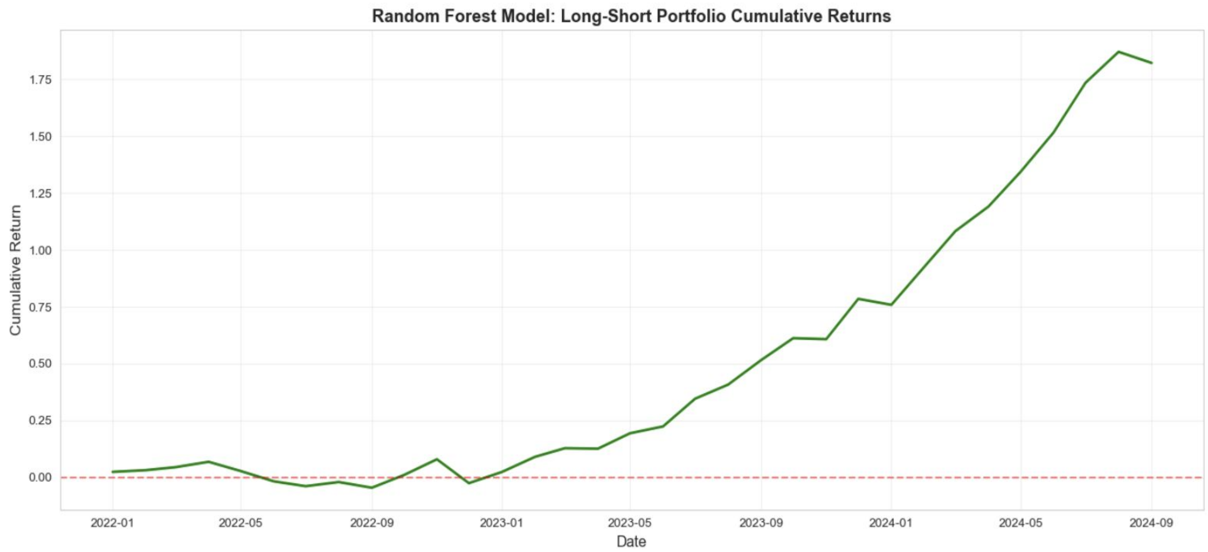
```
==========================================================================================
RANDOM FOREST MODEL: PORTFOLIO PERFORMANCE COMPARISON
==========================================================================================
                 Method  Annual Return  Annual Volatility  Sharpe Ratio  Cumulative Return
 Simple Stratification        0.396274           0.165599      2.392971           1.821106
 Optimized (Empirical)        0.584552           0.249622      2.341752           3.467252
       Optimized (PCA)        0.588377           0.255780      2.300320           3.498874
==========================================================================================
```

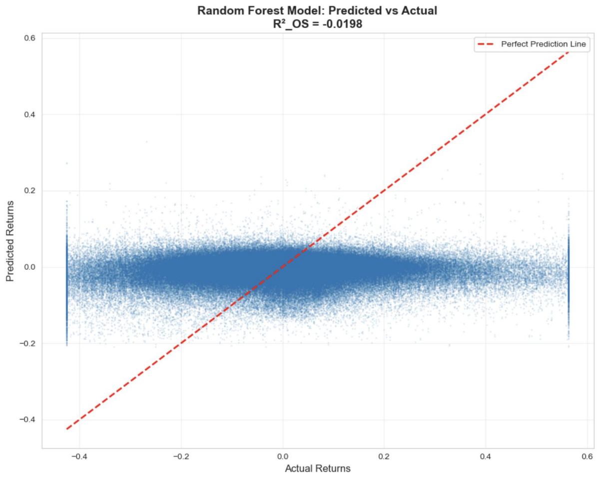



```
📊 Final Cumulative Returns Summary:
========================================
   Simple Stratification: 182.11%
   Optimized (Empirical): 346.73%
   Optimized (PCA):       349.89%
========================================
```

Random Forest: Result

Results:





Random Forest: Result

Key Findings:

1. Non-linear modeling has limited benefit
2. Ranking matters more than accuracy
3. Volatility is the key predictor
4. Good balance between model performance and computational efficiency

GBRT: Gradient Boosting Regression Tree

Model Intuition & Implementation:

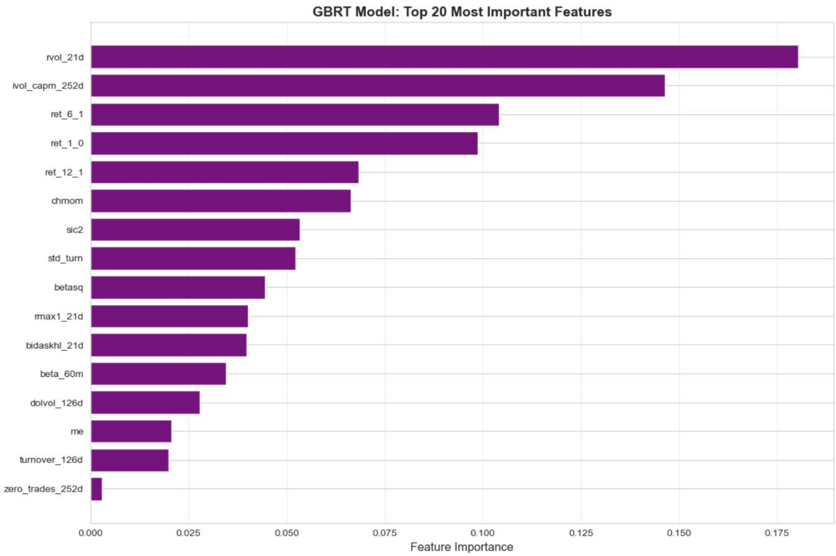
Fits residuals by sequentially adding weak learners (decision trees)

- **Boosting**: Each tree fits the residuals of the previous tree
- **Gradient Descent**: Uses gradient descent to optimize the loss function
- **Strong Non-linear Modeling**: Usually performs better than Random Forest

Prediction performance Metrics:

```
==================================================
GBRT Model Prediction Performance Metrics
==================================================
r2_os                : -0.0176
mse                  :  0.0201
rmse                 :  0.1418
mae                  :  0.0923
correlation          :  0.0512
mean_pred            : -0.0032
mean_true            : -0.0014
std_pred             :  0.0271
std_true             :  0.1406
n_observations       : 552772
==================================================
```

Feature Importance Analysis:

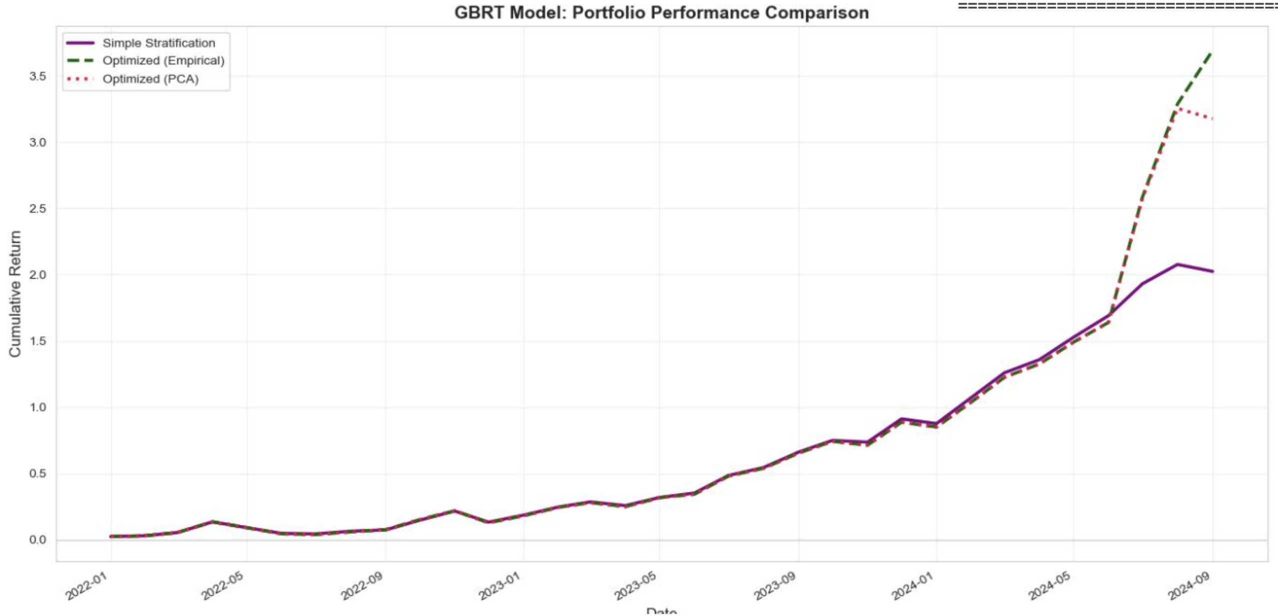


GBRT: Gradient Boosting Regression Tree

Comparison:

optimized portfolio vs. simple stratification method

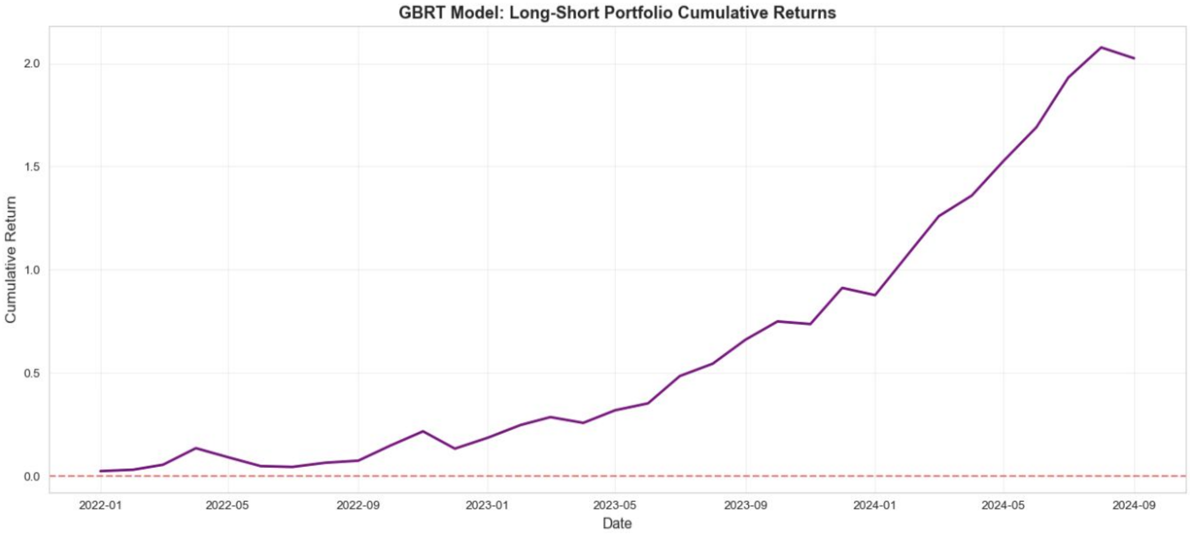
```
================================================================================
GBRT MODEL: PORTFOLIO PERFORMANCE COMPARISON
================================================================================
                 Method  Annual Return  Annual Volatility  Sharpe Ratio  Cumulative Ret
urn
Simple Stratification       0.420587           0.154481      2.722584           2.024
368
Optimized (Empirical)       0.605265           0.264628      2.287233           3.690
712
      Optimized (PCA)       0.560896           0.263596      2.127858           3.176
422
================================================================================
```

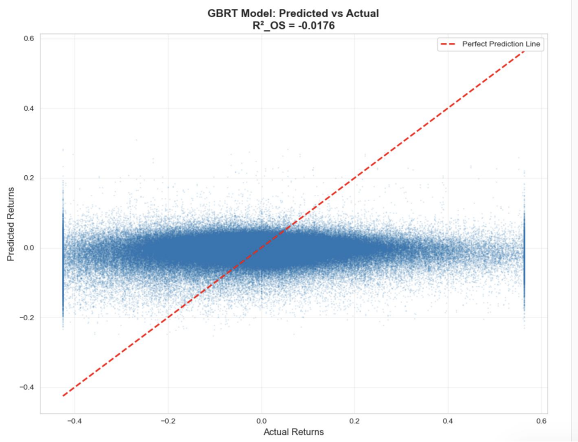



```
📊 Final Cumulative Returns Summary:
==================================================
   Simple Stratification: 202.44%
   Optimized (Empirical): 369.07%
   Optimized (PCA):       317.64%
==================================================
```

GBRT: Gradient Boosting Regression Tree

Results:





Neural Network

Model Architecture

2-Layer MLP: Input → Hidden (128 units, ReLU) → Output

5-Layer Deep MLP: Input → 4 Hidden layers (256 units, BatchNorm, Dropout) → Output

Training: PyTorch, Adam optimizer, early stopping, time-CV

Portfolio Performance (Simple Stratification)

Model	Annual Return	Sharpe Ratio	Cumulative Return
2-Layer	35.40%	2.35	153.57%
5-Layer	40.05%	2.23	183.32%

Prediction Performance

2-Layer: R^2OS = -0.0328, Correlation = 0.0395

5-Layer: R^2OS = -0.0354, Correlation = 0.0347

Insight: Negative R^2 common in asset pricing; ranking ability (correlation) matters more than absolute accuracy

Neural Network

Optimized Portfolio Results

2-Layer + Empirical/PCA: 48.20% annual return, Sharpe 2.27, 248% cumulative

5-Layer + Empirical/PCA: ~57.5% annual return, Sharpe ~2.25, ~280% cumulative

2-LAYER NEURAL NETWORK MODEL: PORTFOLIO PERFORMANCE COMPARISON

Method	Annual Return	Annual Volatility	Sharpe Ratio	Cumulative Return
Simple Stratification	0.353990	0.150707	2.348869	1.535671
Optimized (Empirical)	0.497860	0.237577	2.095571	2.582038
Optimized (PCA)	0.481951	0.211959	2.273792	2.477098

5-LAYER NEURAL NETWORK MODEL: PORTFOLIO PERFORMANCE COMPARISON

Method	Annual Return	Annual Volatility	Sharpe Ratio	Cumulative Return
Simple Stratification	0.400512	0.179712	2.228633	1.833171
Optimized (Empirical)	0.575006	0.255116	2.253897	3.335174
Optimized (PCA)	0.577042	0.256263	2.251758	3.355931

Key Findings

Deeper networks (5-layer) show higher returns but similar Sharpe ratios

Optimized portfolios outperform simple stratification by 60-80% in cumulative returns

PCA-based covariance provides better risk-adjusted returns (higher Sharpe) than empirical method

Non-linear models capture complex patterns but require careful regularization (dropout, batch norm)

Simple Stratification Portfolio Performance

Model	Annual Return	Sharpe Ratio	Cumulative Return
OLS	25.86%	1.15	88.77%
ElasticNet	65.30%	**4.34**	199.08%
PCR	31.92%	1.44	123.14%
PLS	40.09%	2.03	181.10%
RandomForest	39.63%	2.39	182.11%
GBRT	42.06%	**2.72**	202.44%
2-Layer NN	35.40%	2.35	153.57%
5-Layer NN	40.05%	2.23	183.32%

Key Insights:

ElasticNet has the highest Sharpe ratio (4.34) with simple stratification.

GBRT achieves the highest cumulative return (202%) among tree-based methods.

Regularized linear models (ElasticNet) outperform unregularized OLS.

Non-linear models (RandomForest, GBRT, Neural Networks) show similar performance levels.

Optimized Portfolio Performance (Empirical Covariance)

Model	Annual Return	Sharpe Ratio	Cumulative Return	Improvement
OLS	43.31%	1.41	187.22%	+111%
ElasticNet	91.92%	3.45	349.14%	+75%
PCR	59.00%	1.81	330.89%	+169%
PLS	58.15%	1.98	331.40%	+83%
RandomForest	58.46%	2.34	346.73%	+90%
GBRT	**60.53%**	2.29	**369.07%**	+82%
2-Layer NN	49.79%	2.10	258.20%	+68%
5-Layer NN	57.50%	2.25	333.52%	+82%

Key Insights:

Optimization improves cumulative returns by 68–169% across models.

GBRT achieves the highest cumulative return (369%) with optimized portfolios.

ElasticNet maintains strong performance (349% cumulative, Sharpe 3.45).

All models benefit from mean-variance optimization.

Optimized Portfolio Performance (PCA Covariance)

Model	Annual Return	Sharpe Ratio	Cumulative Return
OLS	43.37%	1.38	186.36%
ElasticNet	91.15%	3.18	340.61%
PCR	59.52%	1.62	323.40%
PLS	58.80%	1.95	**337.03%**
RandomForest	58.84%	2.30	**349.89%**
GBRT	56.09%	2.13	317.64%
2-Layer NN	48.20%	**2.27**	247.71%
5-Layer NN	**57.70%**	2.25	335.59%

Key Insights:

PCA-based covariance often provides similar or slightly better Sharpe ratios than empirical methods.

RandomForest with PCA achieves the highest cumulative return (349.89%).

2-Layer NN shows the best Sharpe ratio (2.27) among neural networks with PCA.

PCA is particularly effective for high-dimensional covariance estimation.

Main Takeaways

Prediction vs. Portfolio Performance:

Negative R^2 doesn't prevent strong portfolio performance

Ranking ability (correlation) is more important than absolute accuracy

All models generate economically significant returns despite poor R^2

Optimization Impact:

Mean-variance optimization is essential: improves returns by 60-170%

Rolling covariance estimation captures dynamic risk relationships

Both empirical and PCA methods are effective

Final Recommendations:

Use ElasticNet for best risk-adjusted returns (Sharpe 3.45)

Apply mean-variance optimization with rolling covariance for all models

Consider PCA-based covariance for high-dimensional settings

Focus on ranking ability rather than absolute prediction accuracy

Bottom Line:

Combining predictive models with modern portfolio optimization generates substantial economic value. The "ad-hoc" stratification method significantly underestimates potential returns. Stock correlations and risk-adjusted optimization are critical for portfolio construction.

Thank you!