Stock Return Predictability With XGBoost and Random Forest

Group 3 06/05/2025
Hailey Cai, Linze Li, Shaoxiong Yuan

UCLA

Agenda

- Strategy Overview
- Model 1 - XGBoost
- Model 2 - Random Forest
- Conclusion

Strategy Overview

Question:

- Can combining momentum and fundamental factors produce a strategy that captures momentum gains while remaining effective over the long run?

Objective:

- Use XGBoost & Random Forest to forecast next month's stock return for individual stocks.

Our Approach:

- Utilize a comprehensive set of financial features from Compustat, then merge with CRSP stock data.
- Employ two powerful machine learning models: XGBoost and Random Forest.
- Benchmark performance against the Fama-French 3-Factor Model. (Machine Learning Non-Linear Model VS. Simple Linear Model)
- Evaluate models based on predictive accuracy (Mean Squared Error) and simulated portfolio performance (decile portfolios, cumulative returns, Sharpe Ratio).

Data Preparation

Data Sources

- CRSP (Center for Research in Security Prices): Stock prices, returns, shares outstanding, industry codes.
- CRSP/Compustat Merged Annual Fundament data: Fundamental accounting data. (We use annual data with *6-months lag* to prevent **look-ahead bias**)
- Fama-French Data: The 3-factor model with 'Excess market return', "SMB", "HML".

Key Feature Categories:

- **Stock Factors**: Lag Market Capital (lag_mkt_cap), 1-month rolling return (ret_1m), 12-month rolling return (rolling_ret_12m).
- **Fundamental Ratios**: Book-to-Market Ratio (bm), Earnings-to-Price Ratio (ep), Leverage Factor (lev), Cash Flow to Price Ratio (cfp), etc.
- **Growth/Investment Indicators**: Asset Growth Rate (agr), Investment (invest), Net Issuance (nincr), etc..
- **Total**: 18 fundamental factors.
- **Industry Dummies:** To capture industrial fixed effect. (Total: 74)

Time Period:

- 2012-2024 (Training: 2012-2019, Validation: 2020-2021, Testing: 2022-2024)
- **8 years training, 2 years validation, 3 years test.**

Performance Metrics & Portfolio Construction

Predictive Accuracy:

- **MSE:** The mean of the squared errors.

Portfolio Performance (Simulated Strategy):

- **Decile Portfolios:** Stocks are sorted into 10 deciles each month based on their predicted next returns.
- **Winner-Minus-Loser (WML) Portfolio:** We go long the top decile (highest predicted next returns) and short the bottom decile (lowest predicted next returns).
- **Evaluation Metrics for WML:**
  - **Cumulative Return:** Overall growth of the portfolio.
  - **Annualized Excess Mean Return %:** Average return above the risk-free rate.
  - **Annualized Standard Deviation %:** Volatility of returns.
  - **Sharpe Ratio:** Risk-adjusted return (higher is better).

Model Performance - Overview (2022 - 2024)

$$R_{i,t+1} = \alpha + \sum_{k=1}^{18} \beta_k F_{k,i,t} + \sum_{\ell=1}^{74} \gamma_\ell D_{\ell,i,t} + \varepsilon_{i,t+1},$$

	XGBoost	Random Forest	FF3(benchmark)
Test Set MSE	0.2514	0.495749	0.025160
WML VW Monthly Return	26.11%	14.54%	30.68%
Annualized Sharpe Ratio	0.96	0.65	1.31

Benchmark - Fama French 3 Factor:

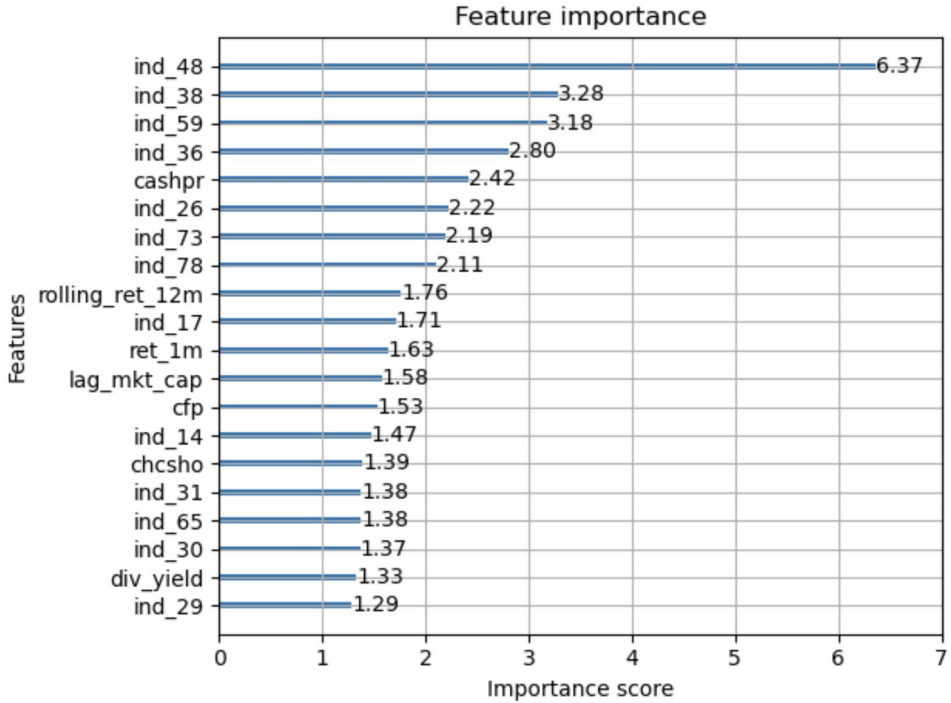
Decile	Excess Mean %	Std %	Sharpe
1	1.46	25.74	0.06
2	4.04	19.43	0.21
3	1.58	16.75	0.09
4	7.87	17.06	0.46
5	9.53	17.90	0.53
6	5.76	17.58	0.33
7	-3.81	16.70	-0.23
8	8.44	18.36	0.46
9	14.01	22.00	0.64
10	32.14	31.19	1.03
WML	30.68	23.37	1.31

Fama French is an ideal baseline for our model. The excess mean return is increasing from losers to winners.

The WML portfolio has the *excess mean return* for **30.68%** and the *Sharpe Ratio* is **1.31**!

Model 1 - XGBoost

XGBoost (Full 18 Factor)



TOP 3 Industry:

SIC 48: Communication

- Reason: In 2023, several major U.S. carriers (Verizon, AT&T, and T-Mobile) completed nationwide 5G coverage.

SIC 38: Measuring, Analyzing, and Controlling Instruments

SIC 59: Retail

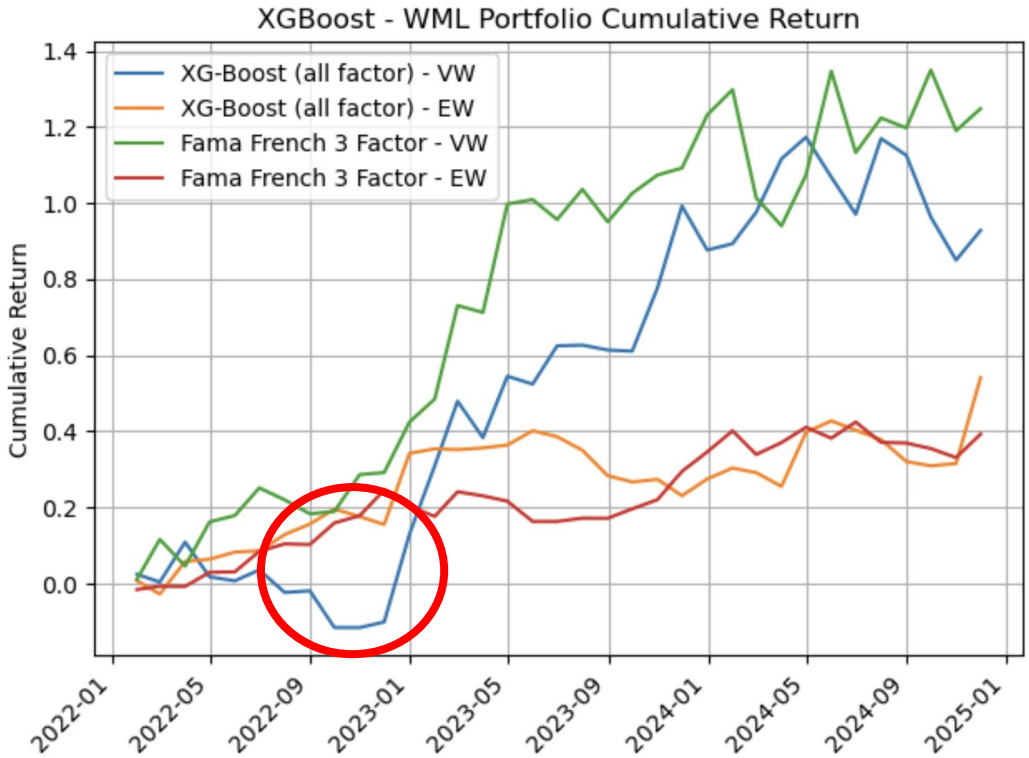
- Recovering from Post-Covid periods

XGBoost 18 factor - Decile Stats

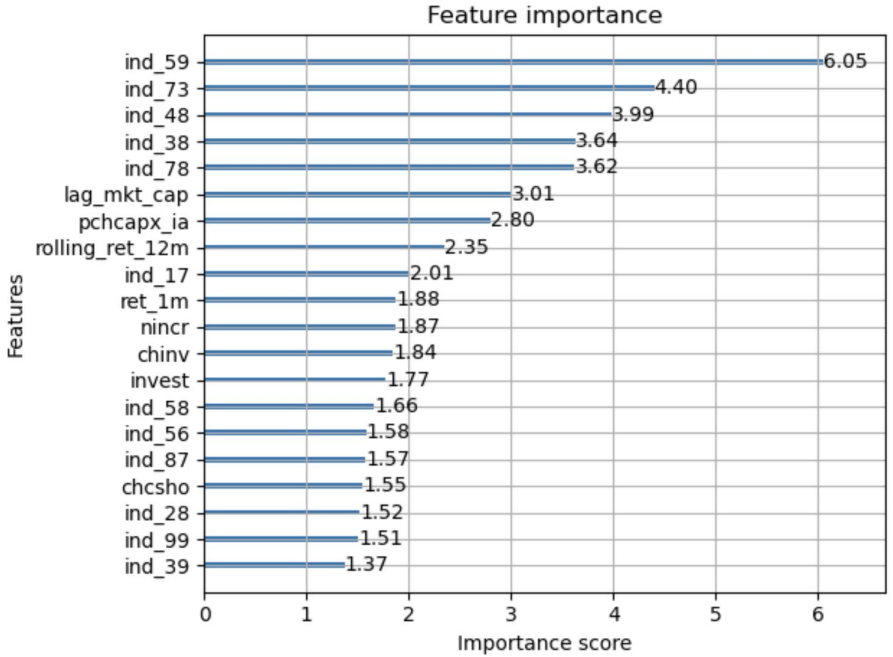
Decile	Excess Mean %	Std %	Sharpe
1	5.13	28.26	0.18
2	4.4	25.74	0.17
3	8.18	15.48	0.53
4	10.19	17.91	0.57
5	3.9	17.09	0.23
6	9.68	18.74	0.52
7	13.59	17.67	0.77
8	2.13	19.21	0.11
9	0.03	26.52	0
10	31.24	32.35	0.97
WML	**26.11**	**27.16**	**0.96**

XGBoost captured good performance industries, so the decile 10 shows an amazing result (31.24% and Sharpe Ratio is 0.97).

XGB Full Factor and FF3

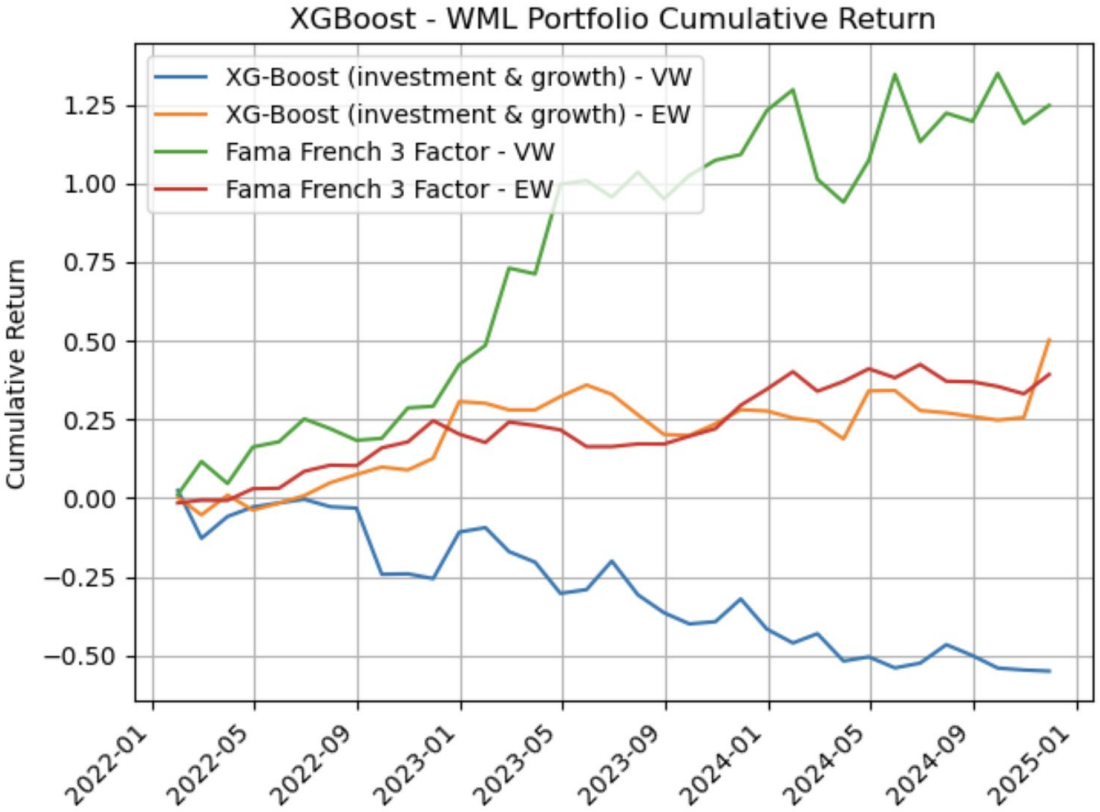


XGBoost (Investment and Growth Factor)



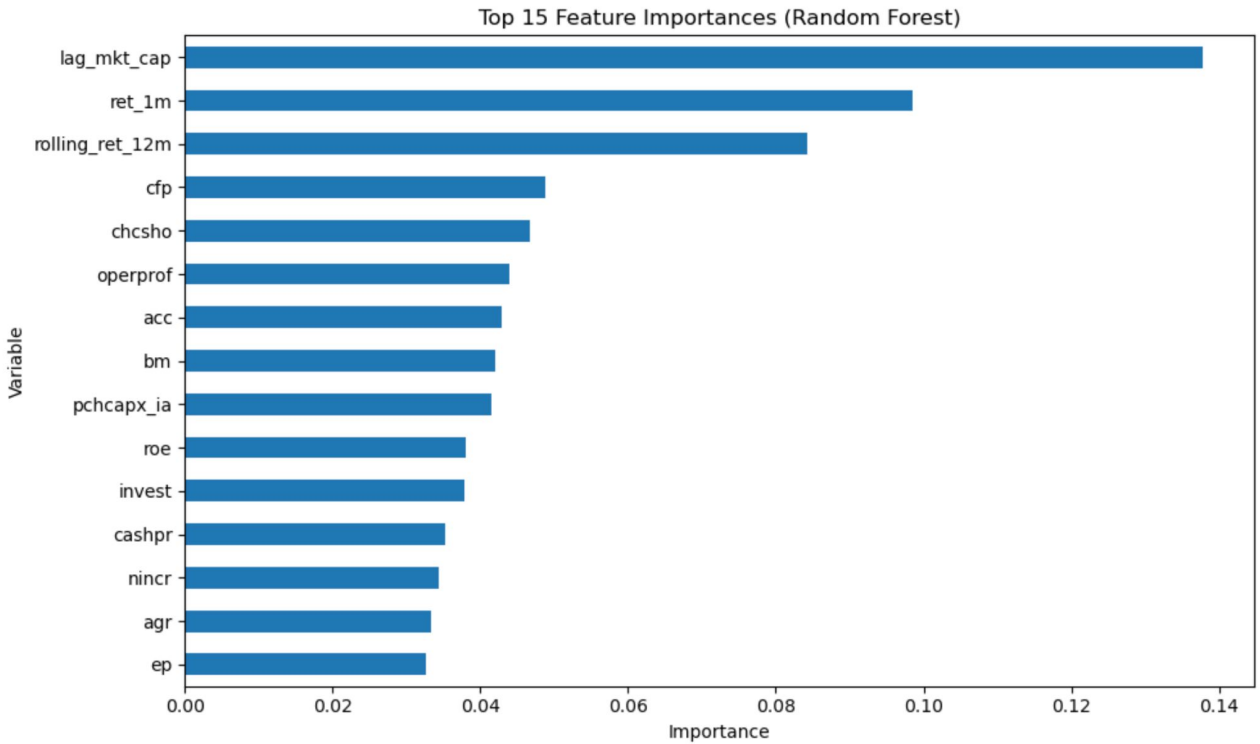
Decile	Excess Mean %	Std %	Sharpe
1	21.7	27.89	0.78
2	3.29	25.01	0.13
3	0.5	20.05	0.03
4	10.17	17.19	0.59
5	3.86	16.86	0.23
6	11.33	18.46	0.61
7	6.85	17.79	0.39
8	3.49	18.57	0.19
9	16.62	26.1	0.64
10	-0.56	38.01	-0.01
WML	**-22.27**	**31.12**	**-0.72**

XGBoost Cumulative Return (Investment and Growth Factor)



Model 2 - Random Forest

Random Forest (Full Factor)



From the feature importances plot, our Random Forest model weights Momentums more than Fundamentals.

However, the fundamentals are important if we consider fundamentals as groups.

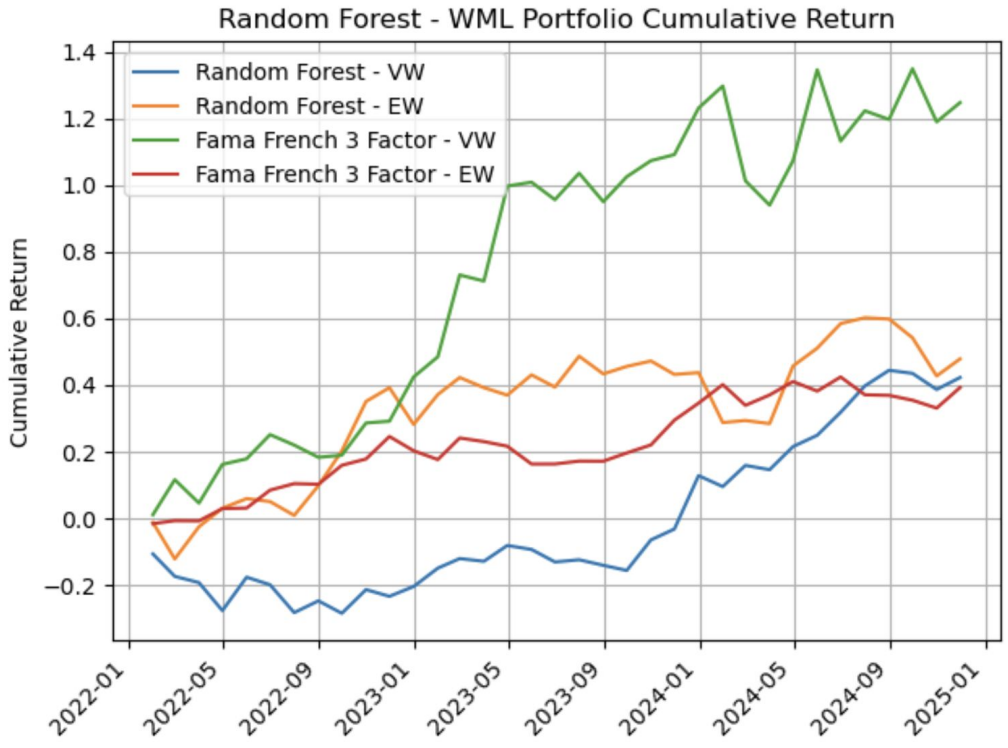
Random Forest (Full Factor) - Decile Stats

Decile	Excess Mean %	Std %	Sharpe
1	-4.73	27.93	-0.17
2	8.7	34.13	0.25
3	22.36	26.6	0.84
4	6.29	21	0.3
5	5.61	18.24	0.31
6	12.36	17.23	0.72
7	9.37	19.51	0.48
8	2.63	17.12	0.15
9	6.69	17.65	0.38
10	9.82	19.49	0.5
WML	**14.54**	**22.35**	**0.65**

Our Random Forest model excels at identifying Decile 1 (the lowest predicted-return group).

However, Deciles 3 and 6 deliver excess mean returns of 22.36% and 12.36%.

Random Forest - All Factors

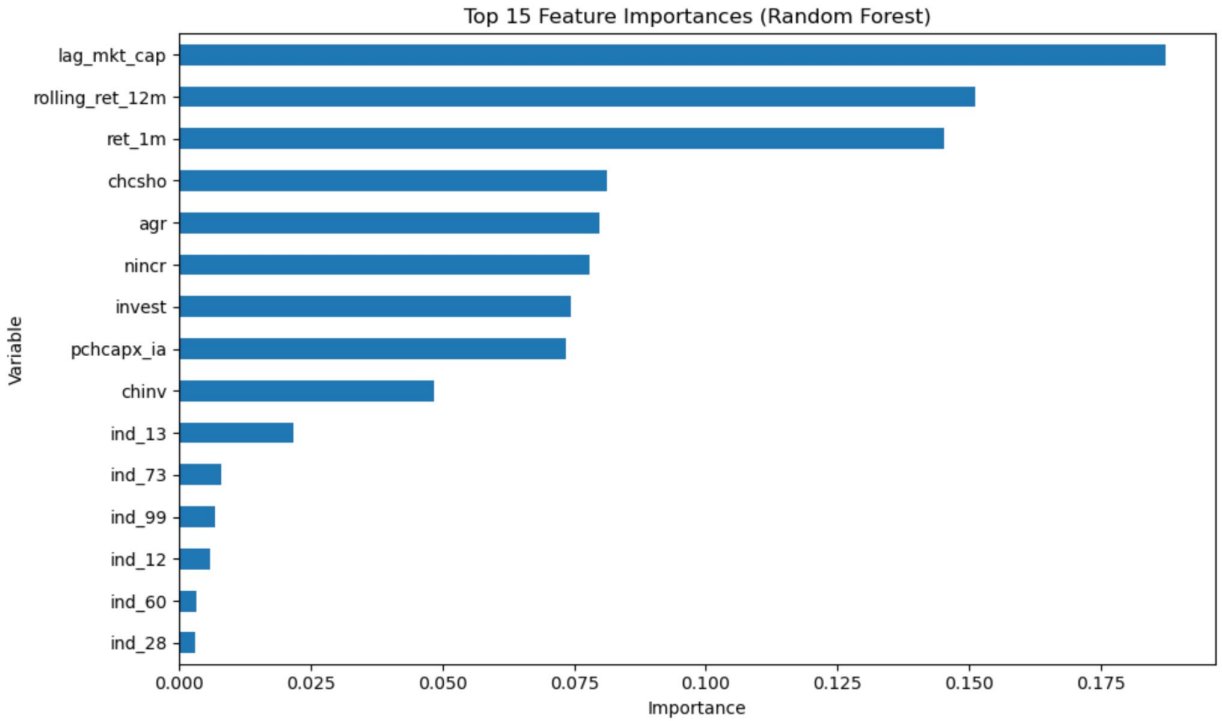


Although our model doesn't outperform the Fama–French three-factor model, we uncovered an interesting insight.

By including fundamental factors, our WML portfolio excludes "momentum-only" and "fundamental-only" stocks. (These stocks dominate most profit in stock market.)

As a result, Decile 10 contains stocks that are strong in both momentum and fundamentals, while Decile 1 contains those weak in both.

Random Forest (Invest & Growth Factor Only)



Same as all 18 factors model, the random forest weight momentum much more than other factors.

But still, combining all 6 investment and growth factor provide around 0.45 importance level.

RF - Investment & Growth Factor only

Decile	Excess Mean %	Std %	Sharpe
1	21.16	25.37	0.83
2	7.59	30.29	0.25
3	13.34	27.72	0.48
4	10.02	18.48	0.54
5	6.23	15.57	0.4
6	6.46	19.37	0.33
7	2.05	17.41	0.12
8	1.2	17.13	0.07
9	13.84	18.65	0.74
10	10.48	20.83	0.5
WML	**-10.68**	**19.66**	**-0.54**

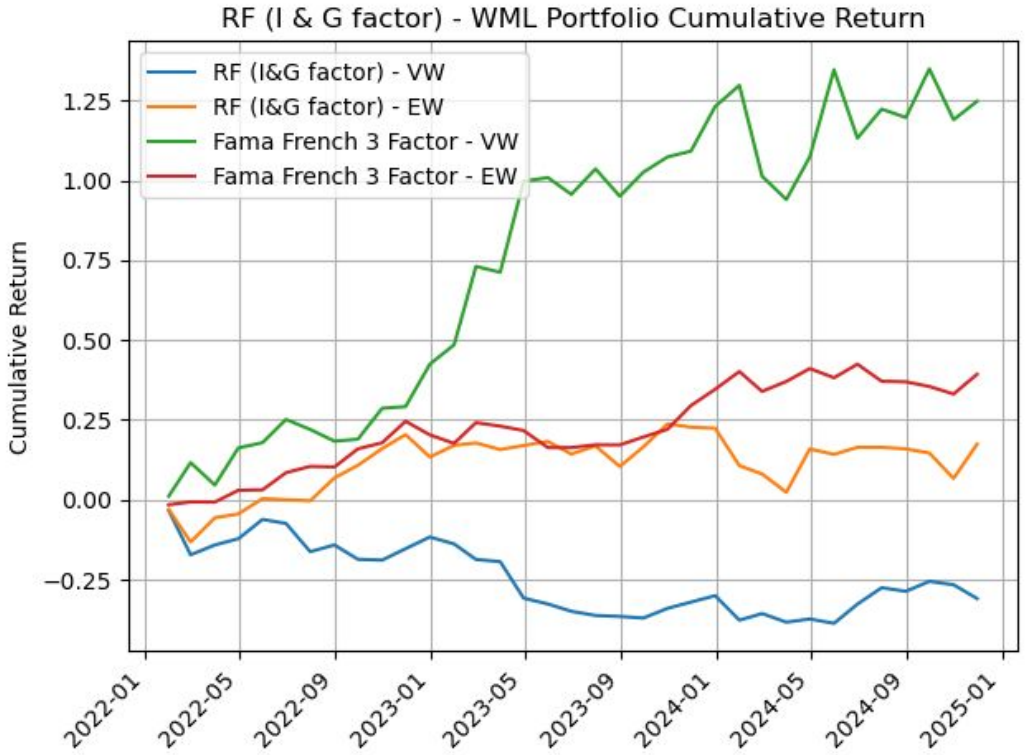
A Disaster Result:

This disappointing outcome may stem from the difficulty of accurately measuring ‘investment’ and ‘growth.’

We cannot simply assume that higher investment automatically leads to higher profits—some companies pursue overly aggressive investment strategies and even end up bankrupt.

The same measurement challenges apply to ‘growth’.

Random Forest - Invest & Growth



Conclusion

To Whom may Interested in this Study:

Random Forest:

1. Good for finding out new **factors**.
2. **Robust to high-dimensional factor combinations**: Cross-sectional analysis often requires inputting dozens or even hundreds of factors (especially when performing factor mining). RF inherently selects a random subset of features at each split, naturally providing “dimensionality reduction + de-correlation.”
3. Easier to explain compared with XGBoost.

XGBoost:

1. Good for finding out **industry fixed effect**.
2. It can capture extreme consequences, but need to care about overfitting problem more than Random Forest.
3. Harder to find the best parameters.

What You can Do with Our Model:

Including Any Factors You want:

Our original paper use about 94 factors with different types of measure (Fundamentals, Economic Effect and etc.) + 74 industry dummies.

You can not only include more factors, but also just looking at some specific factors.

Clean Data Once and Enjoy with Different Models

From our result, it shows Random Forest is good at finding out factors and XGBoost is good at finding out industry fixed effect. Then, you may ask why not we combining these two methods? We get this idea too late, so we did not have time to imply.

We also tried Neural Network (from layer 1 to 5), but none of them provide a strong result.

We provide a framework of Machine Learning and a simple example of it.

Areas of Improvement:

New Factors may Need to Chase Alpha

Since the original paper came out in 2020 and financial institution may imply these factors already, it may cause our models perform really weak.

Include Longer Periods

The original paper worked on total 60 years data (from 1957 - 2016). We only include 13 years data (from 2012 - 2024). Also, there are many ‘Black Swans’ happened in the time periods we chose (eg: Covid-19, Russia & Ukraine War, Bankruptcy of Silicon Valley Bank, etc).

More Detailed Data Cleaning

Our fundamental data uses a uniform six-month lag from the fiscal year. Although this already avoids “look-ahead” bias, each company’s actual release date is different. If we instead lag each company’s annual report release date by one week, the end result might be even better.

Thank You

Reference: Gu, Shihao; Kelly, Bryan; and Xiu, Dacheng. 2020. “Empirical Asset Pricing via Machine Learning.” The Review of Financial Studies 33(5): 2223–2273. doi:10.1093/rfs/hhaa009