DATA 602 – Principles of Data Science

Fall 2025

Instructor: Dr. Fardina Alam

Factor-Based (FF5) Portfolio Strategy

(SPY vs Linear Fama–French vs XGBoost)

URL to Final Tutorial:
https://rayhanpatel.github.io/MSML-602-Final-Project-alphafoundry-ff5-sp500/

Group Members (UID):
Rayhan Basheer Patel (UID: 122087934)
Govind Singahl (UID: 117780413)
Chaithanya Sai Musalreddy (UID: 122257672)

Date of Submission: December 16, 2025

Contributions

Rayhan Basheer Patel *(A, B, C, D, E, F, G)* — Co-developed the project idea with the team, aiming for a finance/fintech project that demonstrates end-to-end modeling and is strong for a resume (A). Led the Fama–French 5 Factors data pull/cleaning and ensured factor definitions and dates aligned with the rest of the pipeline (B). Completed key exploratory checks (time-series trends, outliers/boxplots) and produced several figures used in the analysis and results discussion (C, F). Built and evaluated the XGBoost modeling workflow, including feature construction, training/testing logic, and performance reporting, and helped integrate findings into the final write-up (D, E, G).

Govind Singahl *(A, B, C, D, E, F, G)* — Co-developed the project idea with the team as a practical equity-factor/fintech-style study focused on portfolio-ready skills (A). Implemented the S&P 500 universe construction (Wikipedia) and supported merging/clean alignment of the stock universe with factor and returns panels (B). Performed EDA focused on correlation structure and factor–return relationships to motivate modeling choices (C). Implemented the rolling-window Fama–French regression baseline (beta estimation and walk-forward prediction) and contributed comparative result analysis and figures for the report (D, E, F). Co-wrote methodology/results sections and helped polish the final report (G).

Chaithanya Sai Musalreddy *(A, B, C, E, F, G, H)* — Co-developed the project idea with the team to build a finance/fintech project that highlights data engineering and ML validation skills (A). Constructed the monthly returns dataset using `yfinance` and ensured consistent resampling and date alignment across prices, returns, and factor series (B). Ran distribution diagnostics and pipeline sanity checks (missing data, ticker/month coverage, and transformation validation) and documented edge cases encountered during integration (C, H). Supported model evaluation by verifying split logic and reproducibility across reruns, and contributed figures/tables plus edits to the final conclusions and formatting (E, F, G).

Contents

1 Introduction

Factor-based investing plays a central role in modern asset pricing and portfolio construction. The Fama–French Five-Factor (FF5) framework explains equity returns through systematic sources of risk related to market exposure (Mkt–RF), firm size (SMB), valuation (HML), profitability (RMW), and investment behavior (CMA). While these factors are well established in academic literature, their real-world usefulness depends on how models are evaluated (to avoid look-ahead bias), how portfolios are formed, and whether performance survives transaction costs and changing market regimes.

This project evaluates whether factor-based models can generate a systematic long-only strategy that is competitive with a passive market benchmark. We compare three approaches:

1. **SPY Benchmark**: A passive buy-and-hold proxy for U.S. equity market performance.
2. **Linear FF5 Model**: A rolling-window linear regression that estimates time-varying factor loadings for each stock.
3. **XGBoost Model**: A nonlinear machine learning model trained in a walk-forward manner to capture interactions and regime-dependent relationships among factors and lagged signals.

Our primary research questions are:

- **Predictability:** Can FF5-based models predict or rank next-month stock returns using only information available up to the prior month?
- **Economic value:** Do model-driven portfolios improve cumulative and risk-adjusted performance relative to SPY, and are results robust to transaction costs?
- **Model complexity:** Does XGBoost provide meaningful improvements over an interpretable linear FF5 baseline, or does it primarily increase instability/overfitting risk?

To reflect a realistic deployment setting, all models are evaluated using a **walk-forward** protocol: models are trained on historical data up to time t and used to generate predictions for $t+1$, which are then compared against realized outcomes.

2 Data Curation

2.1 Data Sources

We use two publicly available data sources:

- **Fama–French Five Factors and Risk-Free Rate (RF)**: obtained from the Ken French Data Library and converted to a monthly time series.
- **S&P 500 constituent and price data**: the list of S&P 500 tickers is curated from Wikipedia, and monthly adjusted close prices are collected using the `yfinance` Python package.

2.2 Data Collation and Preparation

The final modeling dataset is constructed as a monthly panel by combining factor data with stock returns:

1. **Ticker universe:** A list of S&P 500 constituents is scraped from Wikipedia and cleaned to match `yfinance` ticker conventions.
2. **Monthly returns:** Adjusted close prices are downloaded for each ticker and resampled to month-end; simple returns are computed as
$$R_{i,t} = \frac{P_{i,t}}{P_{i,t-1}} - 1.$$
3. **Excess returns:** Stock and benchmark returns are converted to excess returns using the monthly risk-free rate:
$$R_{i,t}^{excess} = R_{i,t} - RF_t.$$
4. **Alignment/merge:** Factor observations and returns are merged on the month-end timestamp. Only months with available factor values and returns are retained.
5. **Missing data handling:** Tickers with insufficient history for rolling-window estimation are excluded from model training for those months. This avoids imputing returns and reduces look-ahead bias.

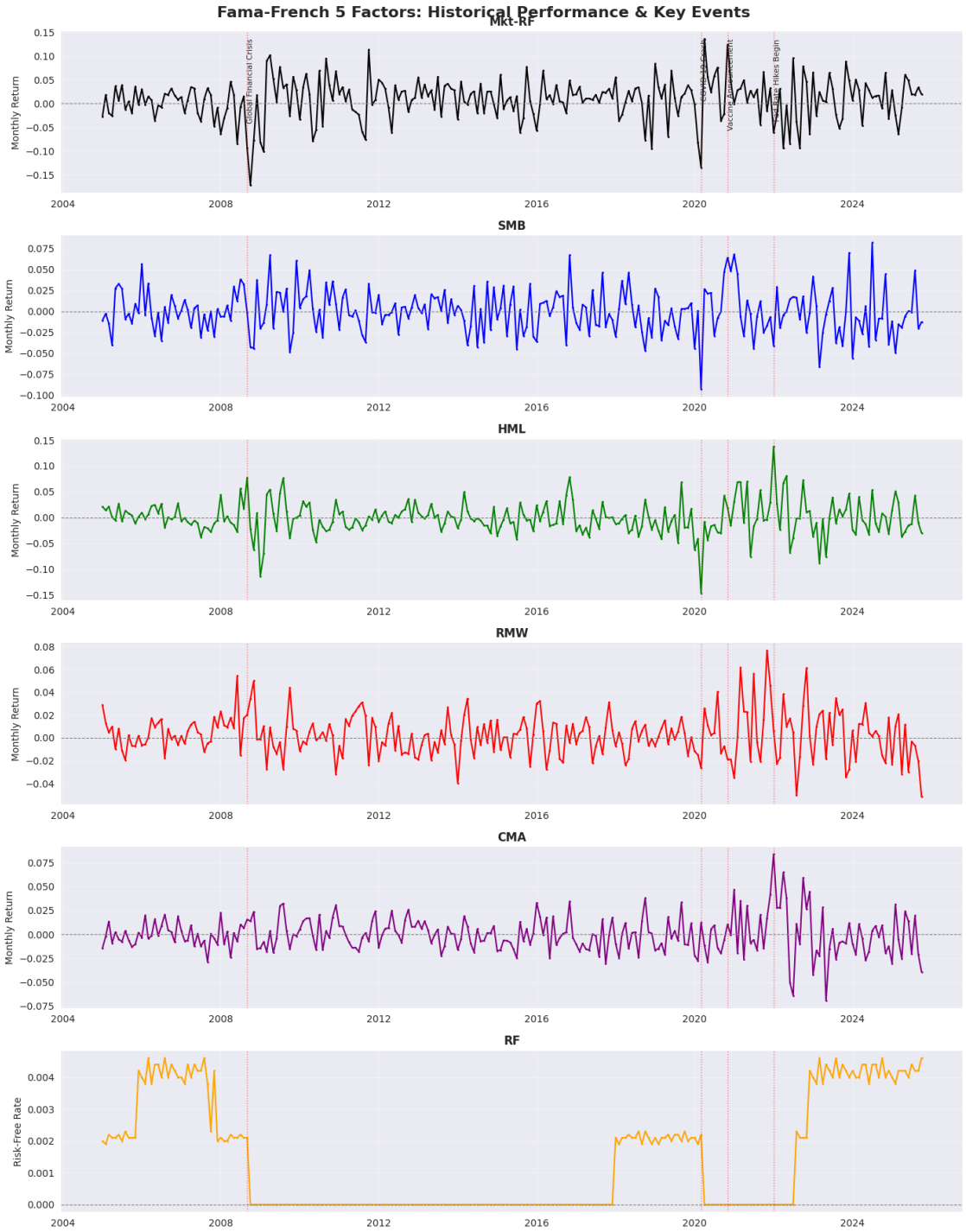

Figure 1: Time-series plots of the Fama–French Five Factors.

3 Exploratory Data Analysis

3.1 Summary Statistics and Trend Analysis

Table 1 reports summary statistics for the monthly Fama–French Five Factors and the risk-free rate (RF). The market excess return (Mkt–RF) exhibits the highest volatility, reflecting the dominant role of market-wide shocks. In contrast, RF is comparatively stable across time, consistent with its role as the baseline for excess-return construction.

From a trend-analysis perspective (see the factor time-series plot), factor realizations vary across market regimes: Mkt–RF shows sharp drawdowns and rebounds during crisis periods, while the style factors (SMB, HML, RMW, CMA) display smaller-amplitude but persistent swings. These dynamics motivate comparing an interpretable linear model to a nonlinear model that can adapt to regime-dependent behavior.

Table 1: Summary Statistics for Fama–French Factors (Monthly)

Factor	Mean	Std Dev	Median
Mkt–RF	0.008269	0.044389	0.013542
SMB	-0.000589	0.026621	-0.000866
HML	-0.000849	0.032316	-0.003298
RMW	0.002576	0.019001	0.002644
CMA	-0.000159	0.019188	-0.001038
RF	0.001418	0.001740	0.000000

3.2 Distributions, Outliers, and Correlations

Factor distributions are approximately centered around zero at the monthly frequency, with Mkt–RF showing the widest dispersion. Boxplots highlight the spread and extreme observations during volatile periods.

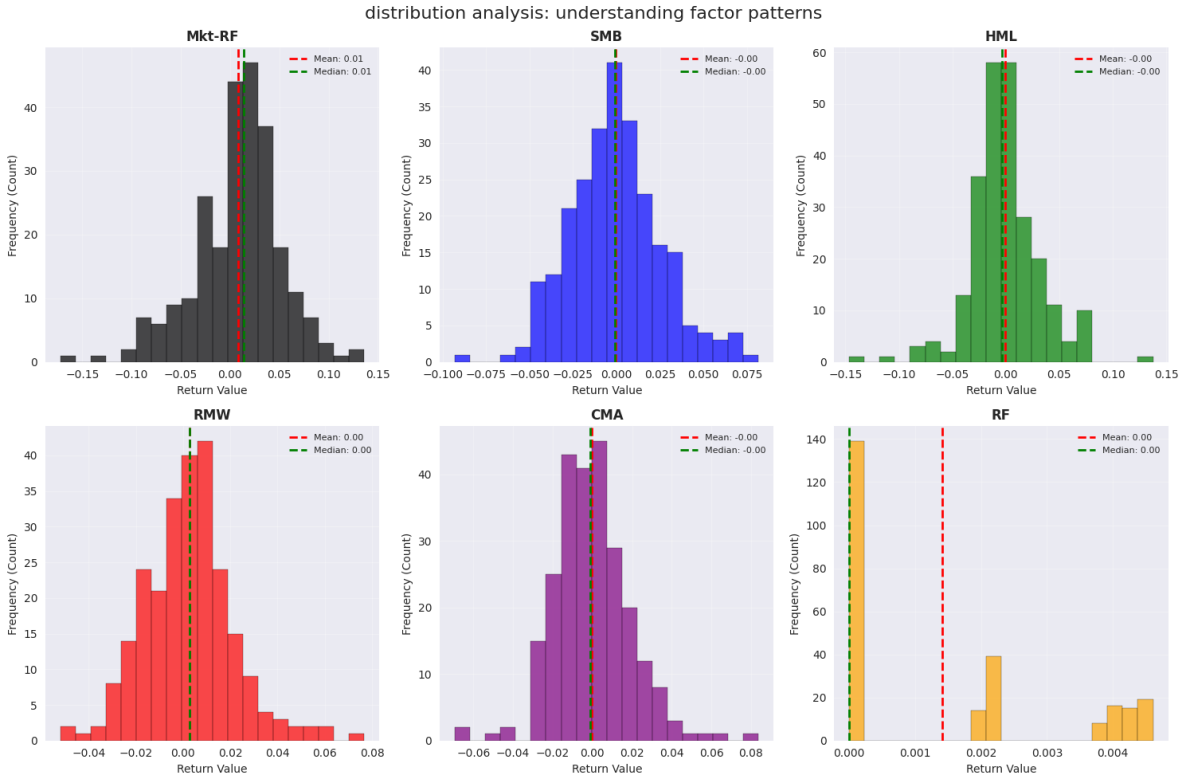

Figure 2: Distribution analysis of FF5 factors and RF (monthly).

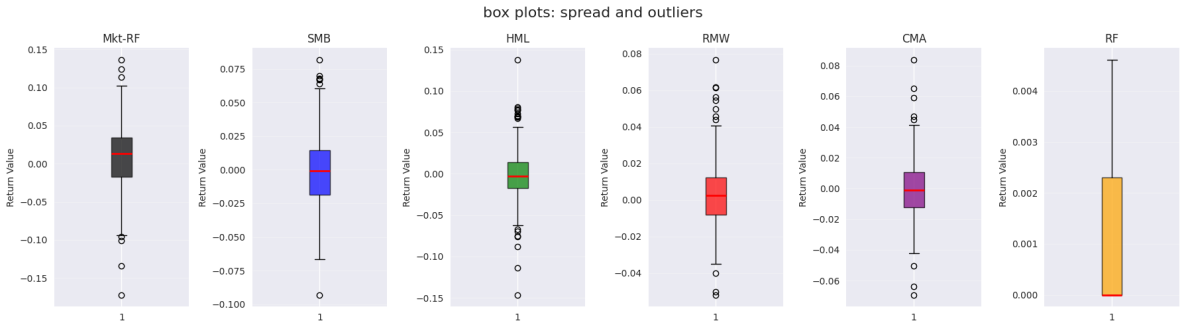

Figure 3: Boxplots showing spread and outliers for FF5 factors and RF.

Correlation analysis confirms that Mkt–RF is most strongly associated with SPY excess returns, while the remaining factors exhibit weaker linear relationships.

To quantify the strength of the market relationship, we conduct a Pearson correlation test between SPY excess returns and Mkt–RF:

$$H_0 : \rho = 0 \qquad \text{vs.} \qquad H_1 : \rho \neq 0.$$

The estimated correlation is $r \approx 0.990$ with a p-value of 1.172×10^{-214}, leading to rejection of H_0 at conventional significance levels. This result is expected because SPY closely tracks the market portfolio, but it also serves as a validation check that the merged dataset is correctly aligned.

Figure 4: Correlation heatmap between FF5 factors, RF, and SPY excess returns (monthly).

4 Models and Methodology

4.1 Walk-forward Evaluation Protocol

All strategies are evaluated using a walk-forward protocol to avoid look-ahead bias. At each month t, models are trained using only historical information available up to $t-1$, a score/signal is generated for month t, and portfolio weights are formed and applied to realized returns.

4.2 SPY Benchmark

SPY represents a passive, fully invested market portfolio and serves as the baseline for performance comparison.

4.3 Rolling Linear Fama–French Model (Time-varying Betas)

For each stock i, we estimate time-varying factor exposures using a rolling-window OLS regression:

$$R_{i,t}^{excess} = \alpha_i + \beta_{i,MKT}(Mkt-RF)_t + \beta_{i,SMB}SMB_t + \beta_{i,HML}HML_t + \beta_{i,RMW}RMW_t + \beta_{i,CMA}CMA_t + \epsilon_{i,t},$$

where $R_{i,t}^{excess} = R_{i,t} - RF_t$ is the monthly excess return. The regression is re-estimated over a trailing window (36 months) with a minimum number of observations to ensure stability.

4.4 Linear FF5 Signal Construction

We forecast each factor at month t using a rolling mean of the prior 36 months (shifted so it uses only information up to $t-1$). The score for stock i at month t is:

$$\widehat{r}_{i,t} = \sum_{j \in \{MKT,SMB,HML,RMW,CMA\}} \beta_{i,j,t}\, \widehat{F}_{j,t}.$$

4.5 XGBoost Learning-to-Rank Model

We use XGBoost (`XGBRanker`) to learn a nonlinear ranking function that maps factor-exposure features to a return-based relevance label. The feature vector is:

$$X_{i,t} = [\beta_{i,MKT,t}, \beta_{i,SMB,t}, \beta_{i,HML,t}, \beta_{i,RMW,t}, \beta_{i,CMA,t}],$$

and labels are derived from realized next-month excess returns and transformed into ordinal relevance bins for ranking. Training is performed in a walk-forward manner to prevent leakage.

4.6 Portfolio Construction

For both the Linear FF5 and XGBoost strategies, we form a monthly rebalanced long-only portfolio by selecting the top $K = 50$ stocks according to the model score. The portfolio is equal-weighted:

$$w_{i,t} = \begin{cases} \frac{1}{K} & \text{if } i \in \text{Top-}K, \\ 0 & \text{otherwise.} \end{cases}$$

5 Results

5.1 Aligned Evaluation Window

To ensure a fair comparison across strategies, we evaluate SPY, the Linear FF5 strategy, and the XGBoost strategy on the common aligned window used by the XGBoost pipeline:

2011-02-01 to 2025-10-01 (**177 monthly observations**).

All reported performance metrics and cumulative return curves in this section refer to this aligned evaluation period.

5.2 Predictions vs Actual (Signal Validation)

We compare model outputs to realized outcomes. For the Linear FF5 strategy, the predicted score $\widehat{r}_{i,t}$ is compared against the realized excess return $R_{i,t}^{excess}$. A positive relationship indicates that the score is informative for ranking.

Figure 5: Predicted score vs realized excess return (stock-month observations).

5.3 Overall Performance (Aligned Window)

Table 2 summarizes performance on the aligned window. XGBoost achieves the strongest risk-adjusted performance (Sharpe on excess returns) and the highest growth rate (CAGR), while the Linear FF5 baseline remains competitive and interpretable. We also report volatility, maximum drawdown, and Calmar ratio to capture risk beyond variance.

Table 2: Performance Summary (Aligned Window: 2011-02-01 to 2025-10-01; 177 months)

Model	Sharpe (Excess)	CAGR	Vol (ann.)	Max DD	Calmar
SPY	0.884	14.00%	14.44%	-23.97%	0.584
Linear	0.969	18.71%	18.03%	-32.46%	0.576
XGB	1.085	25.06%	21.59%	-27.06%	0.926

5.4 Cumulative Returns

Figure 6 presents aligned cumulative returns for all strategies. The XGBoost strategy achieves the highest terminal wealth relative to SPY and the Linear FF5 strategy, with the largest separation emerging after 2020.

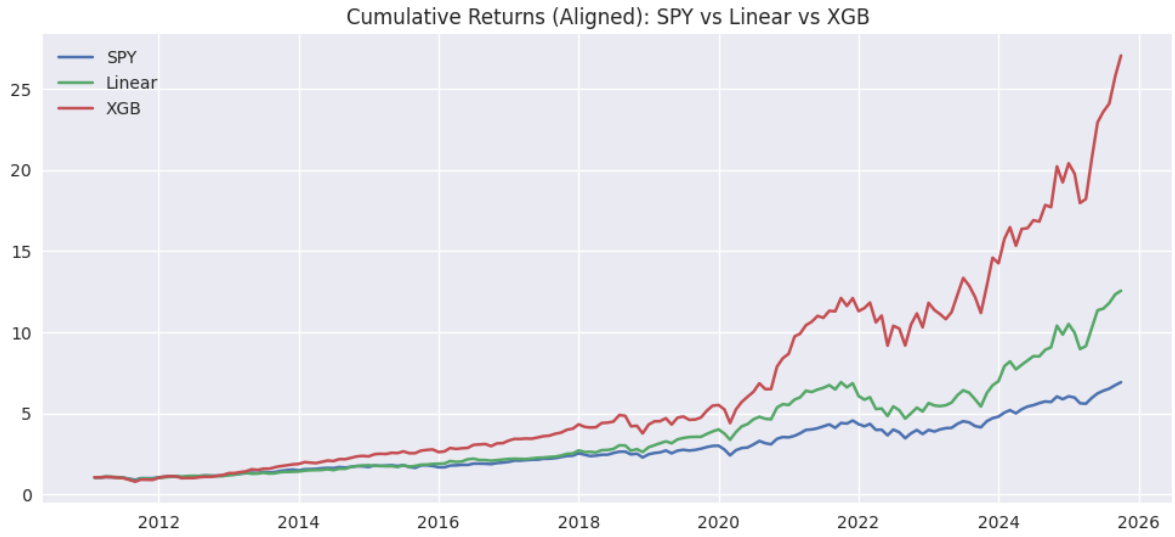

Figure 6: Aligned cumulative returns for SPY, Linear FF5, and XGBoost strategies (2011-02 to 2025-10).

5.5 Calendar-Year Returns

To provide additional interpretability and regime context, Table 3 reports calendar-year returns for each strategy over the aligned-window years (2011–2025). This highlights periods where the strategies diverge (e.g., strong post-2020 outperformance and the 2022 drawdown year).

Table 3: Calendar-Year Returns (Aligned Window Years)

Year	SPY	Linear	XGB
2011	-1.05%	-4.40%	-11.07%
2012	15.90%	17.47%	33.25%
2013	32.53%	24.37%	55.60%
2014	13.45%	27.13%	28.38%
2015	1.19%	5.23%	16.73%
2016	12.00%	14.37%	14.09%
2017	21.80%	17.57%	28.27%
2018	-4.64%	3.69%	-7.15%
2019	31.34%	48.76%	45.48%
2020	18.41%	43.40%	53.55%
2021	28.82%	23.22%	44.18%
2022	-18.26%	-25.41%	-14.82%
2023	26.24%	31.61%	41.60%
2024	24.97%	46.55%	31.92%
2025	17.79%	27.41%	40.61%

5.6 Transaction Cost Sensitivity

We apply proportional transaction costs in basis points (bps) to monthly turnover. Turnover is computed as

$$\text{Turnover}_t = \tfrac{1}{2} \sum_i |w_{i,t} - w_{i,t-1}| \,,$$

and net returns are defined as $R_t^{net} = R_t - c \cdot \text{Turnover}_t$, where c is the cost rate.

Note: Transaction-cost sensitivity depends on realized turnover and should be computed on the same aligned window (2011-02 to 2025-10). The final tutorial reports the transaction-cost tables produced from the latest rerun for full reproducibility.

5.7 Subperiod Analysis

To understand stability across market regimes, we also evaluate performance across subperiods (e.g., pre-2020 vs post-2020). The final tutorial reports the subperiod breakdown produced from the latest rerun, alongside implementation details and reproducibility notes.

6 Insights and Conclusions

This project evaluated whether factor-based models can produce a systematic long-only strategy that is competitive with a passive SPY benchmark under a realistic walk-forward setting. On the aligned evaluation window (2011-02 to 2025-10), model-driven ranking strategies add economic value: XGBoost achieves the strongest cumulative performance and the highest risk-adjusted performance (Sharpe on excess returns), while the Linear FF5 approach remains a stable, interpretable baseline that helps connect factor exposures to expected returns.

These findings highlight the trade-off between flexibility (nonlinear models capturing interactions and regime effects) and robustness (simpler models with more predictable behavior). Future work could incorporate broader baselines, additional trading frictions (slippage/liquidity), and a universe definition that reduces survivorship bias to further evaluate deployability.

7 Data Science Ethics

All datasets used in this project are publicly available and were accessed through legitimate sources (Ken French Data Library, Wikipedia constituents, and `yfinance` market data). No personally identifiable information (PII) is collected or used.

Key ethical and methodological risks include survivorship bias, regime dependency, overfitting, and data leakage. We mitigate these risks by using a walk-forward evaluation protocol to reduce look-ahead bias, making conservative modeling choices, and testing sensitivity to transaction costs. Results are reported transparently with clear assumptions and limitations.

References

[1] Ken French Data Library. *Fama/French 5 Factors (2x3) [Daily].* Accessed and resampled to monthly frequency for this project.
`https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`

[2] Wikipedia. *List of S&P 500 companies.*
`https://en.wikipedia.org/wiki/List_of_S%26P_500_companies`

[3] Ran Aroussi. *yfinance: Yahoo! Finance market data downloader (Python package).*
`https://pypi.org/project/yfinance/`

[4] XGBoost Documentation. *XGBoost Python API (including ranking objectives / ranker usage).*
`https://xgboost.readthedocs.io/`