Risk Management Institute
Shaw Foundation Alumni House
11 Kent Ridge Dr #03-04
Singapore 119244

Can Traditional Factors Like Fama-French Power Modern Machine Learning Portfolio Design?

Michael Wynn[†]

Poreddy Saikiran Reddy[†]

Gaurav Agrawal[†]

Zheng Zhoudong[†]

Raditya[†]

Wang Yidong[†]

[†] National University of Singapore

Can Traditional Factors Like Fama-French Power Modern Machine Learning Portfolio Design?[*]

Michael Wynn[Ψ] Poreddy Saikiran Reddy[Ω] Gaurav Agrawal[ϕ]

Zheng Zhoudong[χ] Raditya[λ] Wang Yidong[ϖ]

April 2025

Abstract

This study is conducted to examine whether extending the Fama-French factor framework with machine learning (ML) techniques can improve automated equity portfolio construction. The analysis is based on a sample of nearly 1,000 U.S. firms across five industries between 2007 and 2024. Monthly stock returns are predicted using several ML models, including SVR, Lasso, and XGBoost, with OLS serving as the benchmark. While ML models offer greater flexibility, they do not outperform OLS in terms of forecast accuracy. Nonetheless, SVR forecasts, when combined with volatility estimates from EGARCH(1,1), are effective in capturing the relative risk and return characteristics of stocks for clustering. Stocks were grouped using KMeans, Agglomerative Clustering, and a median-based Grid method to identify forward-looking risk-adjusted portfolio groupings. These portfolios are subsequently further streamlined using top-N stock selection and constrained weight optimization with an aim to maximise portfolio Sharpe ratios. A robustness check is conducted across different parameter combinations. The empirical findings show that, despite sub-optimal return forecasts, the constructed portfolios often achieve Sharpe ratios in line with or exceeding the S&P 500 benchmark. These results suggest the approach has some potential in extracting forward-looking signals to be structured into viable equity investment strategies through clustering and optimization.

Keywords: Asset pricing model, stock markets, machine learning, portfolio formation
JEL classification: G12, G17, G11, C52

[*] The authors would like to thank Professor Chen Kan and Ding Li for delivering the course, and for regular consultations. The views and opinions expressed by the authors are their own and do not represent the views and opinions of the National University of Singapore or any of its subsidiaries or affiliates. Any mistakes or omissions are the authors' responsibility.

Authors' last names and emails:
[Ψ] Wynn, e1349581@u.nus.edu and Michael.wynn23@hotmail.com (Permanent)
[Ω] Poredddy, saikiran.poreddy@u.nus.edu and saikiranreddy.saikiran@gmail.com (Permanent)
[ϕ] Agrawal, e1348994@u.nus.edu and g.agrawal009@gmail.com (Permanent)
[χ] Zheng, e0843532@u.nus.edu and JayZheng00@gmail.com (Permanent)
[λ] Raditya, e1350738@u.nus.edu and raditya30597@gmail.com (Permanent)
[ϖ] Wang, e1349995@u.nus.edu and Wangyidong020321@gmail.com (Permanent)

I. Introduction[1]

Capital Asset Pricing Model: Market participants use asset pricing models to understand how risk factors drive average stock returns. Among the most prominent is the Capital Asset Pricing Model (CAPM), introduced by Sharpe (1964) and Lintner (1965), which leverages on the modern portfolio theory (MPT) by Markowitz (1952). MPT was the first framework to establish a strong mathematical and financial foundation for understanding the relationship between risk and return, benefiting both academia and practitioners. Even after several decades, CAPM is still being widely used to assess the performance of managed portfolios and estimate a firm's cost of capital (Fama and French 2003).

The CAPM model is represented by the following equation:

$$R_{i,t} - R_{f,t} = \beta_i \left(R_{m,t} - R_{f,t}\right) + \epsilon_{i,t} \tag{1}$$

Where $R_{i,t}$ is the return of the asset at time t, $R_{f,t}$ denotes the return of the risk-free asset at time t, β_i is the coefficient that proxies systematic risk and measures the i asset's sensitivity to expected excess market returns, $R_{m,t}$ corresponds to market rate of return at time t, and $\epsilon_{i,t}$ represents the error term of the linear regression model.

CAPM is a single-factor model that emphasizes the market factor, represented by the beta coefficient, as the sole driver of systematic risk influencing an asset's expected returns. Its appeal lies in its ability to provide clear and intuitive insights into risk measurement and the relationship between expected return and risk (Fama and French 2003). However, despite its theoretical elegance, CAPM has shown weak empirical support, primarily due to its incompleteness (Banz 1981; Basu 1983; Bhandari 1988; Chan and others 1991), raising concerns about its practical application.[2]

Three-Factor Fama-French: Since its introduction, the asset pricing literature has expanded considerably to address the limitations of the CAPM. Most notably, Fama and French (1992, 1993) challenged its empirical validity by showing that market beta alone does not fully explain stock returns. They introduced the three-factor model, which adds size (SMB) and value (HML) factors to the market risk premium, capturing cross-sectional variations in stock returns more effectively. Their findings demonstrated that smaller firms tend to generate higher returns than larger firms, and high book-to-market (value) stocks tend to outperform low book-to-market (growth) stocks. Specifically, they defined Small Minus Big (SMB) as the average return on the three smallest portfolios minus the average return on the three largest portfolios, capturing the size effect, while High Minus Low (HML) represents the average return on two high-value portfolios minus the average return on two growth portfolios, reflecting the value premium. The results of the three-factor model showed significant improvement in explaining the variation in stock returns.

The Fama-French three-factor model is represented in the following equation:

$$R_{i,t} - R_{f,t} = a_i + \beta_{1,i}\left(R_{m,t} - R_{f,t}\right) + \beta_{2,i} SMB_t + \beta_{3,i} HML_t + \epsilon_{i,t} \tag{2}$$

[1] As with Fama and French (2003), while all asset pricing models are capital asset pricing models, the finance profession reserves the acronym CAPM to the Sharpe–Lintner–Black model (Sharpe 1964; Lintner 1965; Black 1972). Thus, in this paper, we refer to Sharpe–Lintner–Black model as the CAPM.

[2] For more details, refer to Rossi (2016) for a critical literature review of the CAPM.

Where a_i is the intercept for asset i, SMB_t represents the excess returns of small over large firms at time t, HML_t represents excess returns of firms with high over low book-to-market value ratios, and β_2 and β_3 are sensitivities corresponding to the additional factors for asset i.[3]

The equations to calculate SMB_t and HML_t for the three-factor model are given by:

$$\begin{aligned} SMB_t \ &= \frac{1}{3} * (Small\ Value_t\ + Small\ Neutral_t\ + Small\ Growth_t) - \\ &\quad \frac{1}{3} * (Big\ Value_t\ + Big\ Neutral_t\ + Big\ Growth_t) \end{aligned} \tag{3}$$

$$\begin{aligned} HML_t \ &= \frac{1}{2} * (Small\ Value_t\ + Big\ Value_t) - \\ &\quad \frac{1}{2} * (Small\ Growth_t + Big\ Growth_t) \end{aligned} \tag{4}$$

Although the Fama-French three-factor model provided strong empirical support, some have questioned its predictive ability. For instance, Daniel and Titman (1996) were unable to validate the results in Fama and French (1993). Griffin and Lemmon (2002) pointed out that the three-factor model may be sensitive to country-specific effects, emphasizing the need to incorporate local factors to improve the accuracy of portfolio return predictions. In response, Fama and French (2012) incorporated both local and global risk factors to enhance the model's applicability to developing economies.

Five-Factor Fama-French: More recently, Fama and French (2015) expanded their three-factor model into a five-factor framework by introducing two additional factors—profitability and investment. These new variables included Robust Minus Weak (RMW), which captures the difference in returns between firms with strong and weak operating profitability, and Conservative Minus Aggressive (CMA), which reflects the investment patterns of firms.

The Fama-French five-factor model is represented in the following equation:

$$\begin{aligned} R_{i,t} - R_{f,t} \ = \ & a_i \ + \ \beta_{1,i}\left(R_{m,t} - R_{f,t}\right) + \ \beta_{2,i}\, SMB_t \ + \ \beta_{3,i}\, HML_t \ + \\ & \beta_{4,i}\, RMW_t \ + \ \beta_{5,i}\, CMA_t \ + \epsilon_{i,t} \end{aligned} \tag{5}$$

Where RMW_t represents the difference between the average returns of two portfolios with strong operating profitability and two portfolios with weak operating profitability at time t, CMA_t represents the difference between the average returns of two portfolios with conservative investment strategies and two portfolios with aggressive investment strategies at time t, and β_4 and β_5 are sensitivities corresponding to the additional factors for asset i.

The equations to calculate RMW_t and CMA_t for the five-factor model are given by:

$$\begin{aligned} RMW_t \ &= \frac{1}{2} * (Small\ Robust_t\ + Big\ Robust_t) - \\ &\quad \frac{1}{2} * (Small\ Weak_t\ + Big\ Weak_t) \end{aligned} \tag{6}$$

[3] For a full description of the factor returns, refer to Fama and French (1993).

$$CMA_t = \frac{1}{2} * (Small\ Conservative_t + Big\ Conservative_t) - \frac{1}{2} * (Small\ Aggressive_t + Big\ Aggressive_t) \tag{7}$$

It is also worthwhile to note that in the Fama-French five-factor model, the calculation of the SMB_t is more refined compared to the three-factor model. Instead of a single SMB_t measure based on size, the five-factor model decomposes SMB into three components: one based on book-to-market $(SMB_{t,B/M})$, another on operating profitability $(SMB_{t,OP})$, and a third on investment $(SMB_{t,INV})$. Each component is computed as the difference between the average return of small stock portfolios and the average return of large stock portfolios within the respective sorting criteria. The final SMB_t factor is the average of these three components:

$$SMB_{t,B/M} = \frac{1}{3} * (Small\ Value_t + Small\ Neutral_t + Small\ Growth_t) - \frac{1}{3} * (Big\ Value_t + Big\ Neutral_t + Big\ Growth_t) \tag{8}$$

$$SMB_{t,OP} = \frac{1}{3} * (Small\ Robust_t + Small\ Neutral_t + Small\ Weak_t) - \frac{1}{3} * (Big\ Robust_t + Big\ Neutral_t + Big\ Weak_t) \tag{9}$$

$$SMB_{t,INV} = \frac{1}{3} * (Small\ Conservative_t + Small\ Neutral_t + Small\ Aggressive_t) - \frac{1}{3} * (Big\ Conservative_t + Big\ Neutral_t + Big\ Aggressive_t) \tag{10}$$

$$SMB_t = \frac{1}{3} * \left(SMB_{t,B/M} + SMB_{t,OP} + SMB_{t,INV}\right) \tag{11}$$

The above factor equations for both three- and five-factor Fama-French models are obtained from the Kenneth French Data Library.[4] The five-factor model was observed to perform better than the three-factor model in capturing average stock returns (Fama and French 2015).

Machine Learning (ML) in Asset Pricing Models: While traditional factor models like the Fama-French framework have been widely studied, empirical findings from the Fama-French model in the U.S. market indicate that the GRS test does not validate the five-factor model (Gibbons, Ross and Shaken 1989).[5] Cakici (2015) analysed the Fama-French five-factor model across 23 markets, finding that while it performs similarly to the U.S. model in North America, Europe, and globally, the profitability and investment factors are insignificant in Japan and Asia Pacific. Given these limitations, the asset pricing literature has expanded into exploring non-linear techniques. For instance, Dittmar (2002) applied the non-linear pricing kernel technique, while Gogas, Papadimitriou, and Karagkiozis (2018) used support vector regression (SVR) within the CAPM and Fama-French models, finding it more effective than Ordinary Least Squares (OLS). Other notable studies include Gu, Kelly and Xiu (2018) and Diallo, Bagudu and Zhang (2023), among others.

[4] https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html

[5] The GRS test is used to evaluate mean-variance efficiency and serves as a standard for assessing asset pricing models. It helps determine whether a given asset pricing model effectively explains the expected returns of an asset or portfolio.

When it comes to modelling volatilities, traditional methods are dominated by Generalized Autoregressive Conditional Heteroskedasticity (GARCH) family of models. Such studies include Brandt and Jones (2012), Ezzat (2012), Sinha (2012), and Aliyev, Ajayi, and Gasim Ajayi (2020). Nevertheless, ML approaches are emerging in this space as well, such as in Filipović and Khalilzadeh (2021) and Chatterjee, Bhowmick and Sen (2022), among others.

Automated Portfolio Construction: In addition to predicting returns and volatilities, researchers and practitioners have sought to use clustering algorithms to optimize stock market decisions in constructing optimal risk-adjusted portfolios from a large number of active stocks. Lemieux and others (2015) highlight that the choice of clustering technique can significantly influence perceived portfolio risk, cautioning against relying on a single method in visual analytics tools. León and others (2017) demonstrate that clustering-based portfolio construction yields more stable and lower-volatility portfolios than classical mean-variance optimization, with hierarchical clustering delivering the best trade-off between returns and risk-adjusted performance.

This paper adds to the literature by employing a range of ML and statistical methods to estimate a sample of U.S. stock returns across five industries, using Fama-French factors, with OLS serving as the benchmark. Volatility is modelled using advanced Autoregressive Conditional Heteroskedasticity (ARCH) models. The best-performing models for return and volatility are selected through a comparative evaluation and then applied to the sample of U.S. stocks. Based on these predictions, several clustering techniques were implemented to group stocks by risk-return profiles, followed by portfolio optimisation using predicted returns and volatilities. The resulting portfolios are evaluated in terms of Sharpe ratio performance relative to the Standard & Poor's 500 (S&P) Index. The paper is structured as follows: Section II describes the data, Section III outlines the methodology, Section IV presents the results, and Section V concludes.

II. Data[6]

As mentioned, the Fama-French three- and five-factor data are obtained from the Kenneth French Data Library. Stock price data for U.S. firms are retrieved from LSEG Datastream, with the initial universe comprising 9080 companies. In filtering the data, the sample is restricted to stocks of type equity; country of headquarters in the U.S; non-missing Standard Industrial Classification (SIC); and only stocks listed within the five major U.S. exchanges: NASDAQ/NGS (Global Select), NASDAQ/NMS (Global Market), NASDAQ Capital Market, NYSE, and NYSE MKT LLC (formerly AMEX). The Fama-French factors from the Kenneth French Data Library are appropriate for this sample as they are constructed to represent broad U.S. equity market dynamics, capturing systematic risk factors that apply across all major U.S. exchanges regardless of listing venue. Monthly frequency of the Fama-French factors is chosen given their limited short-term variability—particularly for SMB and HML—which makes them less effective at higher frequencies. Hence, monthly data better aligns with their typical application in risk decomposition and return explanation.

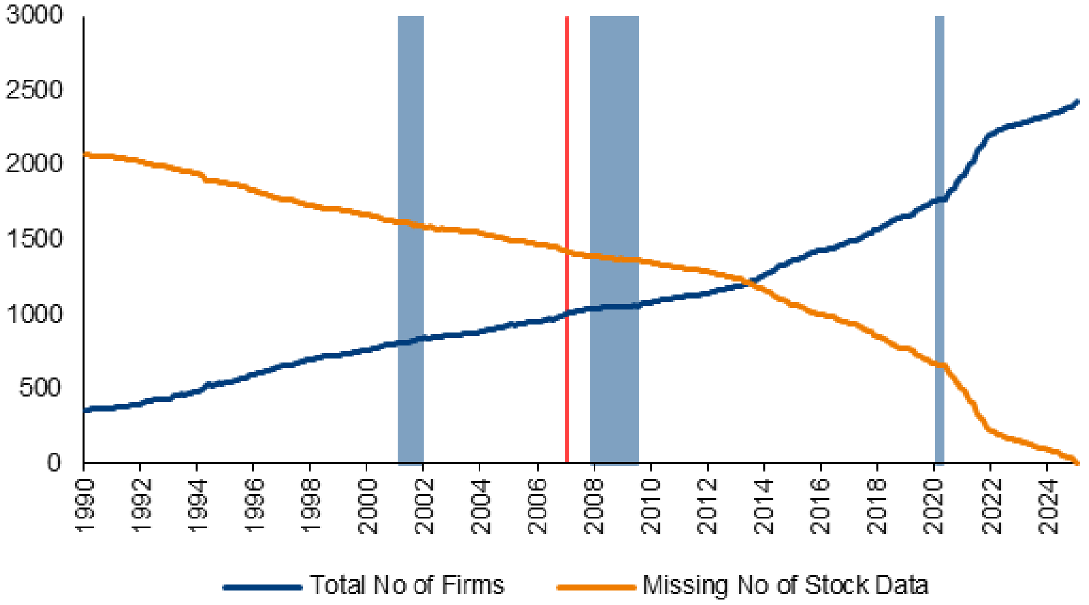
Figure 1 illustrates the evolution of the sample over time, including the total number of firms with stock price data and the number of firms with missing price data. To ensure a balanced panel with adequate time-series coverage, an arbitrary cut-off on 1st January 2007 is imposed

[6] This section uses several Python libraries, including *pandas* and *numpy* for data handling, *matplotlib* and *seaborn* for visualization, *statsmodels* for stationarity testing.

to ensure enough firms and time periods, while also capturing at least one regime of high financial stress for robust modeling of stock return dynamics. In other words, firms included in the final sample, by design, must have been incorporated by 1st January 2007, with continuous and complete stock price data until 31st Dec 2024. Firms delisted, suspended, or had any missing stock price observations arising from other reasons during this period are excluded. Monthly returns are computed using logarithmic differences of adjusted closing prices to account for stock splits and dividend payouts, ensuring consistency across time. The resulting return series are tested for stationarity using the Augmented Dickey-Fuller (ADF) test; firms with non-stationary return series are also removed. Monthly transformation is chosen to minimize data loss from the transformation process. The restriction to firms with uninterrupted trading records from 2007 to 2024 may introduce survivorship bias, as it excludes firms that exited the market due to bankruptcy, mergers, or delisting.

The final sample consists of 975 U.S. firms across five industry groups (Table 1). The data spans from February 2007 (inclusive) to December 2024. The industry groups are mapped following the four-digit SIC of the five industry portfolios as defined in the Kenneth Data Library. While the sample set ensures sectoral diversity, it is notably skewed toward the Hi-Tech sector which comprises 24.9 percent of firms but accounts for 46.8 percent of total market capitalization. This concentration may influence portfolio clustering results in the later section.

Figure 1: Sample Coverage and Data Completeness Over Time



Source: LSEG Datastream (accessed on March 17, 2025); and authors' calculations.
Note: Shaded areas refer to high financial stress regimes namely, the Dot-com bubble, Global Financial Crisis, and COVID-19 pandemic. The green vertical line represents the cut off arbitrarily chosen.

Table 1: Data Representation by Industry

Industry	Number of Firms	Percent (By Count)	Total Market Cap (USD Billions)	Percent (By Market Cap)
Hi-Tech	243	24.92	13442.64	46.75
Other	210	21.54	3505.05	12.19
Manufacturing	188	19.28	3567.61	12.41
Healthcare	187	19.18	3159.38	10.99
Consumer	147	15.08	5076.65	17.66

Source: LSEG Datastream (accessed on March 17, 2025); and authors' calculations.

III. Methods

This section is divided into three parts. The first part discusses the ML approaches undertaken to predict monthly returns for each individual U.S stock, using both three- and five-factor Fama French equations. The ML models used in this study are estimated using several different ML algorithms, namely regularized regressions (Ridge, Lasso, and Elastic Net), tree-based models (Decision Tree Regression (DTR), and extreme Gradient Boosting (XGBoost)), and Support Vector Regression (SVR), with OLS regression serving as the benchmark model. The second part of this section presents techniques used to predict stock volatility. The third part leverages on clustering techniques to identify the optimal risk-adjusted portfolio based on both predicted returns and volatilities for stocks in the sample.

A. Machine Learning Approach for Stock Return Prediction[7]

Equations (2) and (5) represent the risk (explanation) models for the three- and five-factor Fama French models respectively, explaining expected returns based on systematic risk factor exposures. In predicting returns for any given stock, both equations are converted into alpha (prediction) models suitable for ML, re-interpreting the factor exposures as predictive features:

Three-Factor Fama French Forecasting Model:

$$\mathbb{E}[R_{i,t+1} \mid \mathcal{F}_t] = f(\, ExcessMktReturn_t\, ,\, SMB_t\, ,\, HML_t\, ,\, R_{f,t}) \tag{12}$$

Five-Factor Fama French Forecasting Model:

$$\mathbb{E}[R_{i,t+1} \mid \mathcal{F}_t] = f(\, ExcessMktReturn_t\, ,\, SMB_t\, ,\, HML_t\, ,\, RMW_t\, ,\, CMA_t\, ,\, R_{f,t}) \tag{13}$$

where $\mathcal{F}_t$ denotes information set available at time t and $ExcessMktReturn_t$ refers to $(R_{m,t} - R_{f,t})$.

Hyperparameter Tuning: For each machine learning model, we performed hyperparameter tuning using an expanding window cross-validation framework.[8] The historical data is split into 85.0 percent for training (182 months), 10.0 percent for validation (21 months), and the remaining for testing.[9] During tuning, the model is trained on all available data up to 182

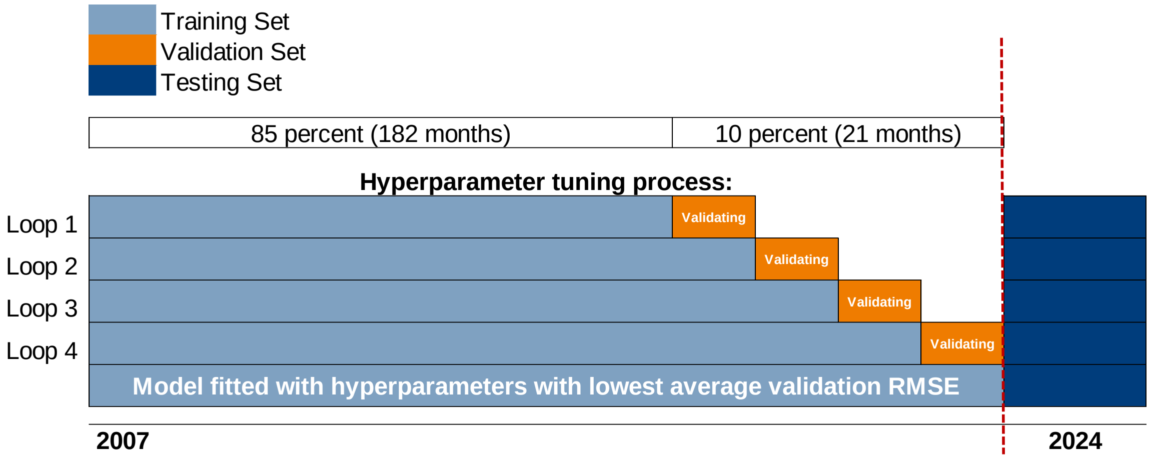
[7] Models in this section are implemented using *scikit-learn* and *xgboost* for extreme gradient boosting regression.

[8] Hyndman and Athanasopoulos (2018) describe the method as "evaluation on a rolling forecasting origin," while Cerqueira, Torgo, and Mozetič (2020) refer to it as the "prequential method." We adopt the more commonly used term "expanding window validation", which aligns with the sequential structure and cross-validation objective of this exercise.

[9] Stock returns are shifted backwards by one month in the Fama-French alpha models, and therefore the entire dataset spans a total of 214 months.

months and evaluated on a three-month validation window for the next 21 months. A three-month validation window is selected to balance generalization performance and computational efficiency, providing enough data to evaluate performance reliably while keeping the tuning process tractable. This process is repeated by expanding the training window forward (Figure 2). The average validation Root Mean Square Error (RMSE) across all validation windows are computed for each hyperparameter configuration, and the best-performing model is selected based on the lowest average RMSE. This procedure is applied consistently across all models, using the predefined ranges of hyperparameters listed in Table 2.

Figure 2: Expanding Window Cross-Validation Framework



Source: Authors.
Note: Figure is not drawn to scale and is only for illustrative purposes. For instance, there would be seven loops (21 months divided by three months rolling validation period).

Table 2: Fama French Models: Hyperparameter Tuning Options

Model	Parameters	Description	Options
Ridge	α (alpha)	L2 regularization strength	0.01, 0.1, 1.0, 10.0, 100.0
Lasso	α (alpha)	L1 regularization strength	0.01, 0.1, 1.0, 10.0, 100.0
ElasticNet	α (alpha)	Regularization strength	0.01, 0.1, 1.0, 10.0, 100.0
	l1_ratio	Mix between L1 and L2 penalty	0.1, 0.5, 0.9
SVR	C	Regularization parameter	0.1, 1, 10
	ε (epsilon)	Epsilon in epsilon-tube loss function	0.01, 0.1
DTR	max_depth	Maximum depth of the regression tree	2, 4, 6, 8, 10
XGBoost	n_estimators	Number of boosting stages	50, 100, 200
	learning_rate	Shrinkage step size	0.1, 0.05, 0.01

Source: Authors.
Note: SVR = Support Vector Regression; DTR = Decision Tree Regression; XGBoost = eXtreme Gradient Boosting.

Returns Prediction: Using the selected hyperparameters, we evaluate the predictive performance of all models over the testing set for each stock through a one-month rolling forecast. At each iteration, the model is trained on all data available up to time t, and returns at $t+1$ are predicted using the factor values at time t. This process continues until the final observation in the test set, and predictions are stored for each stock-model pair. Model performance is similarly assessed based on RMSE, which is computed over the full test window.

B. Volatility Prediction[10]

The paper employs several volatility models for each U.S stock in modelling stock return volatilities. A comparative modeling approach is undertaken, evaluating multiple models based on their information criterion. Specifically, four widely used volatility specifications were considered: the standard Generalized Autoregressive Conditional Heteroskedasticity model (GARCH), the Exponential GARCH model (EGARCH), which captures asymmetric effects in volatility, the Heterogeneous ARCH model (HARCH), which accounts for volatility clustering across different time horizons, and the Fractionally Integrated GARCH model (FIGARCH), which accommodates long memory in volatility dynamics. These models are estimated under the assumption of normally distributed residuals for each individual return series.

For each model, a set of lag orders $(p, q) \in \{(1,1), (1,2), (2,1), (2,2)\}$ are considered, a configuration that captures a broad range of volatility dynamics while remaining computationally tractable. Due to the additional complexity and convergence sensitivity of the FIGARCH model, its estimation is limited to the (1,1) specification. All models are fitted using maximum likelihood estimation, ensuring consistency in the estimation approach. Each asset's return series is modelled independently, with Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) values computed for each model specification. AIC and BIC are information-theoretic criteria that balance model fit and complexity: lower values indicate better trade-offs between capturing the data and avoiding overfitting. While AIC tends to favor more complex models, BIC imposes a stricter penalty on model complexity, often selecting more parsimonious specifications. The model yielding the lowest AIC and lowest BIC is selected as the best-fitting model for that asset, while the model with the second-lowest values is also recorded to facilitate robustness checks and comparative analysis.

In addition to tracking the best and second-best models on an individual basis, the frequencies with which each model specification is selected as the best or second-best across the entire sample set are also tallied. This frequency analysis provides insight into the relative performance and generalizability of each model type. Any time series exhibiting excessive missing values or insufficient variability are excluded from estimation.

C. Clustering for Automated Portfolio Construction[11]

To address the limitations set out in Wu, Wang and Wu (2022), this paper adopts a data-driven approach that identify attractive groups of stocks based on forward-looking predictions, rather than historical correlations alone. This section leverages predicted return and volatility estimates, combined with clustering techniques, to automate the classification of stocks into distinct risk-return profiles. Predicted returns and predicted volatilities are taken from the best-performing models that are selected based on their overall out-of-sample performance. One selected model for each return and volatility predictions are applied across the entire stock sample rather than on a stock-by-stock basis. This ensures consistency and comparability across firms and reduces the risk of overfitting that could arise from tailoring models to individual stocks.

[10] All models in this sub-section are implemented in Python using the *arch* package.

[11] Clustering methods in this sub-section are implemented in Python using the *scikit-learn* package. KMeans and Agglomerative Clustering are performed using *sklearn.cluster.KMeans* and *sklearn.cluster.AgglomerativeClustering.*

Clustering is performed monthly using these predicted returns and volatilities estimates from the best selected models. The objective is to group stocks into similar risk-return profiles and isolate those that fall into the high-return, low-volatility region of the return-volatility space. These stocks are identified as the optimal cluster. Three clustering techniques are applied: KMeans, Agglomerative Clustering, and a median-based grid method. KMeans forms clusters by minimizing intra-group variation, while Agglomerative Clustering builds groups incrementally based on pairwise similarity.[12] The grid method offers a simpler rule-based approach that partitions the return-volatility space into quadrants using medians.

Outliers may arise from estimation instability in certain volatility models, particularly during periods of low return variation or heightened market stress, which can lead to unrealistically high or low volatility predictions. Additionally, stocks with very limited trading activity or thin liquidity may exhibit near-zero predicted volatility, further skewing the clustering process Therefore, stocks with extreme or zero volatility are filtered out to prevent distortion in cluster formation. Each clustering method is applied monthly over the test set period.

D. Streamlining Optimal Clusters for Practical Portfolio Construction

Following the clustering process and removal of outlier stocks, the resulting optimal clusters can still contain large groups of firms with attractive risk-adjusted profiles. Therefore, practical constraints necessitate further refinement to ensure feasible portfolio implementation. This section operationalizes the stock selection process that converts these clusters into stocks of more manageable and investable sets. For each monthly optimal cluster, identified separately by each clustering technique, a range of top N stocks—where $N \in [5,10]$ based on market capitalisation—is selected based on their computed Sharpe ratios using predicted return and volatility estimates. The Sharpe ratio is used as a measure of risk-adjusted performance by identifying stocks with high predicted expected returns relative to their predicted risk, within the optimally identified cluster. This ensures that portfolio construction aligns not only with algorithmic clustering but also with core investment principles from MPT. The forward-looking Sharpe ratio is computed as the following:

$$SharpeRatio_{p,t} = \frac{\mu_{p,t}}{\sigma_{p,t}} = \frac{\sum_{i=1}^{n} w_{i,t}\mu_{i,t}}{\sqrt{w_t^T \Sigma_t w_t}} \tag{14}$$

where $w_{i,t}$ is the portfolio weight assigned to stock i at time t, $\mu_{i,t}$ is the predicted return of stock i at time t, w_t is a vector of portfolio weights, Σ_t is the covariance matrix at time t $\mu_{p,t}$ is the portfolio's predicted return at time t, and $\sigma_{p,t}$ is the predicted portfolio volatility at time t.

The covariance matrix Σ_t is constructed using the predicted volatilities and rolling historical return correlations:

$$\Sigma_t = \mathrm{D}_t \rho_t \mathrm{D}_t \tag{15}$$

where $\mathrm{D}_t = diag(\sigma_{1,t}, \sigma_{2,t}, \sigma_{3,t}, \sigma_{4,t}, \sigma_{5,t}, \ldots, \sigma_{n,t})$ is a diagonal matrix of monthly volatilities, and ρ_t is the historical correlation matrix of returns up to time t.

Portfolio weights are then obtained by solving the following constrained optimization problem:

[12] Feature standardization is performed using z-scores for the Agglomerative Clustering method, ensuring predicted returns and volatilities are treated on the same scale.

$$w_t^* = \underset{w_t}{\arg\max}\left(\frac{w_t^T \mu_t}{\sqrt{w_t^T \Sigma_t w_t}}\right) \quad s.t. \quad \textstyle\sum_{i=1}^{n} w_{i,t} = 1, w_{i,t} \geq \theta_i \; \forall i \in \{1, \ldots, n\} \tag{16}$$

Where w_t^* is the optimal portfolio weight vector at time t, μ_t is the vector of predicted monthly returns at time t, Σ_t is the monthly covariance matrix of predicted returns at time t, $\sum_{i=1}^{n} w_{i,t} = 1$ ensures full capital allocation with no cash left uninvested, and $\theta_i \in [0.05, 0.06, 0.07, 0.08, 0.09, 0.10]$ is the minimum allowable weight for stock i, to encourage diversification and avoid individual stock concentration. The range, in discrete 1.0 percent increments, ensures that every stock position in the portfolio is sufficiently sizable to be practically investable, while also limiting over-diversification. A 5.0 percent lower bound helps avoid excessive fragmentation into very small positions, which can be inefficient or costly to manage in real-world portfolios. The 10.0 percent upper range offers additional diversification discipline in smaller clusters or when stock selection is more concentrated. This balance reflects standard portfolio construction practices, particularly in institutional settings where minimum position sizes are required for liquidity, compliance, and/or trading cost considerations.

The combination of both $N \in [5,10]$ and $\theta_i \in [0.05, 0.06, 0.07, 0.08, 0.09, 0.10]$ is applied as part of a robustness check rather than a fixed rule. The minimum weight constraint can be further tailored depending on portfolio size, investment strategy, or regulatory guidelines. It should also be noted that the current implementation does not incorporate transaction costs, rebalancing frictions, or capital gains taxes, which could materially affect the net performance and turnover characteristics of the strategy in live settings. The optimization is performed using the Sequential Least Squares Programming algorithm.[13] The process from Equations (14) to (15) is repeated for each optimal cluster across all months and clustering techniques. Portfolio Sharpe performance using derived weights on returns and volatilities are applied on realized returns and volatilities and are then benchmarked against the Sharpe ratio of that of Standard & Poor's 500 Index (S&P 500).

IV. Results

A. Stock Return Prediction Results

Hyperparameter Tuning: The results of the hyperparameter tuning process can be found in Table 3. Strong regularization is a consistent theme across Ridge, Lasso, and ElasticNet. Simpler models that favour generalization also tend to dominate across the board. For instance, shallow trees are most frequently selected in DTR, while SVR generally prefer low values of C and ε. Similarly, Ridge and Lasso models often favour higher values of α, and XGBoost perform best with low learning rates combined with a high number of estimators. Most importantly, the hyperparameter selections for the five-factor Fama-French model closely mirror those of the three-factor model. The close alignment in hyperparameter choices suggests that the addition of two factors in the five-factor Fama-French does not substantially alter the predictive structure of the ML models. At least in the context of the historical dataset that the models are trained on in this paper, it appears that these two additional factors provide limited incremental predictive power beyond the original three factors.

[13] Optimization is performed using the *scipy.optimize.minimize* function in Python.

Table 3: Hyperparameter Tuning Results

Model	Hyperparameters	Firm Count (3FF)	Firm Count (5FF)
Ridge	α=0.01	90	60
	α=0.1	1	8
	α=1.0	5	0
	α=10.0	40	42
	α=100.0	839	865
Lasso	α=0.01	121	80
	α=0.1	64	78
	α=1.0	269	307
	α=10.0	521	510
ElasticNet	α=0.01, l1=0.1	55	34
	α=0.01, l1=0.5	1	0
	α=0.01, l1=0.9	34	21
	α=0.1, l1=0.1	37	41
	α=0.1, l1=0.5	2	5
	α=0.1, l1=0.9	33	35
	α=1.0, l1=0.1	84	88
	α=1.0, l1=0.5	62	63
	α=1.0, l1=0.9	154	185
	α=10.0, l1=0.1	70	77
	α=10.0, l1=0.5	429	420
	α=10.0, l1=0.9	14	5
	α=100.0, l1=0.1	0	1
SVR	C=0.1, ε=0.01	293	236
	C=0.1, ε=0.1	258	246
	C=1, ε=0.01	133	174
	C=1, ε=0.1	122	145
	C=10, ε=0.01	87	94
	C=10, ε=0.1	82	80
DTR	depth=2	785	787
	depth=4	125	133
	depth=6	43	30
	depth=8	14	13
	depth=10	8	12
XGBoost	est=50, lr=0.1	46	52
	est=100, lr=0.05	43	71
	est=200, lr=0.01	886	852

Source: Authors' calculations.

Note: 3FF = Three-factor Fama French; 5FF = Five-factor Fama French; SVR = Support Vector Regression; DTR = Decision Tree Regression; XGBoost = eXtreme Gradient Boosting; est = n_estimators; lr = learning_rate.

Stock Returns: Across both the three- and five-factor Fama-French alpha models, ML techniques do not yield statistically significant improvements over the benchmark OLS model in forecasting one-month ahead stock returns. Results show that SVR and Lasso are more or almost as frequently selected at the individual stock level (Table 4). However, their RMSE improvements over OLS are marginal (Figures 3 and 4). OLS remains highly competitive under both alpha models, indicating that added model complexity does not translate into meaningful predictive gains. This suggests that the limited performance differentials reflect the fundamental limitations of factor exposures as return predictors, rather than shortcomings in the modelling techniques themselves.

RMSE levels are uniformly high across all models. This reflects the inherent difficulty of forecasting monthly stock returns, even with flexible learning algorithms. The high error levels underscore the noisy and idiosyncratic nature of monthly log returns data and highlight the limited explanatory power of factor-based features for short-horizon forecasting. Although differences are not large, the manufacturing sector exhibits the lowest average RMSEs under both alpha model specifications, suggesting a relatively more stable operating environment or stronger factor alignment. However, even in this “best-performing” sector, prediction errors remain considerable, reinforcing the broader conclusion that factor-based inputs lack sufficient alpha-generating signal. While healthcare firms make up nearly 20.0 percent of the stock universe by count, they represent only about 11.0 percent of total market capitalization, indicating a greater concentration of smaller, potentially more volatile firms. This size imbalance contributes to more erratic return behaviour and, in turn, higher forecast errors.

The comparison between the three- and five-factor Fama-French models reveals minimal differences in predictive performance. This suggests that the two additional factors in the five-factor Fama-French model; RMW and CMA, do not materially enhance short-term forecast accuracy. This likely reflects the slow-moving nature of these factors, which evolve gradually over time and are more suited to explaining long-term return differences rather than capturing short-horizon dynamics. As a result, while RMW and CMA may improve the explanatory power of factor models in a risk explanation context, they appear to offer little incremental predictive value for alpha generation over a monthly horizon.

Mid-cap and small-cap stocks tend to exhibit higher RMSEs than micro-cap and large-cap firms, likely resulting more from idiosyncratic volatility and thinner coverage in the mid/small-cap space, making returns harder to predict. In contrast, large-cap stocks may benefit from greater informational efficiency, while micro-caps may reflect simpler structures or less competitive pricing. Regardless of size category, high error levels persist across the board, further emphasizing the limited predictive value of the factors.

While these findings point to the weakness of Fama-French factors for short-term return prediction, they nonetheless provide a coherent framework when combined with volatility estimates for relative performance classification. Although ML methods do not statistically outperform OLS, we proceed with SVR predicted returns as the input for clustering. Given that the clustering objective is to identify relative positions in the risk-return space rather than generate precise forecasts, SVR’s outputs paired with volatility estimates remain suitable for mapping optimal portfolios, particularly those in the high-return, low-volatility quadrant or cluster. We aim to test that despite the relatively high RMSEs observed, SVR predicted returns can still serve as a coherent and interpretable input for identifying stock clusters with potentially attractive risk-adjusted trade-offs.

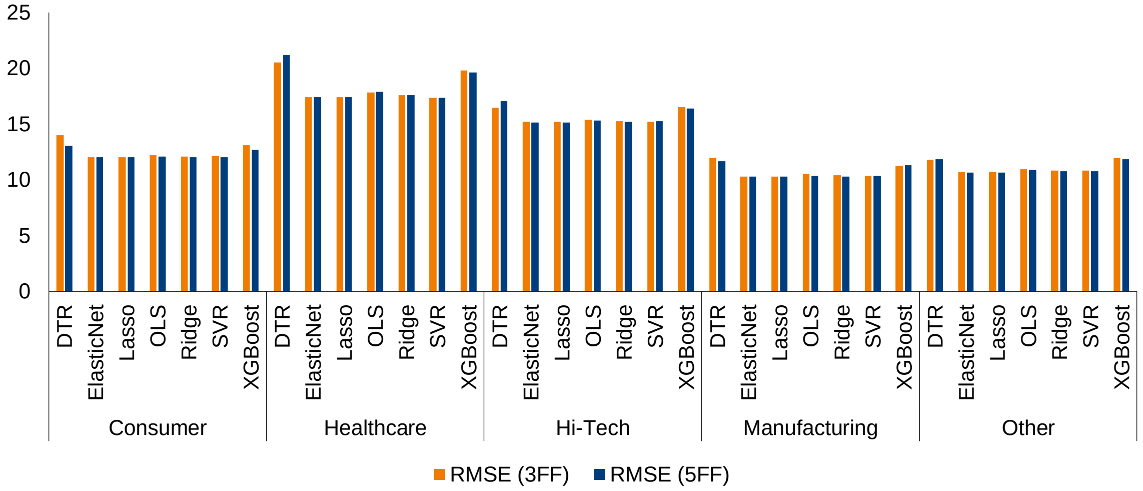
Table 4: Best Model by Firm Count

Model	Count (3FF)	Count (5FF)
DTR	146	133
ElasticNet	26	23
Lasso	205	178
OLS	162	209
Ridge	46	59
SVR	**225**	**213**
XGBoost	165	160

Source: Authors' calculations.

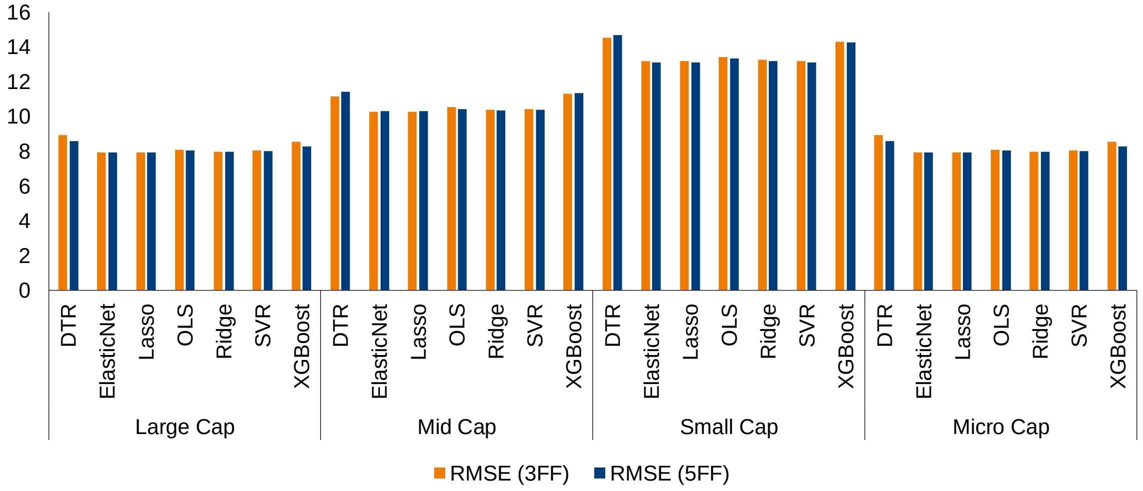
Note: Best model chosen by each U.S stock here is determined by the lowest average RMSE in the test set.

Figure 3: Average RMSEs Across All Models by Industry
(Percent)



Source: LSEG Datastream (accessed on March 17, 2025); and authors' calculations.
Note: 3FF = Three-factor Fama-French; 5FF = Five-factor Fama-French; OLS = Ordinary Least Squares; and DTR = Decision Tree Regression. The RMSEs calculated here are performed on monthly stock returns in test set using a one-month rolling prediction.

Figure 4: Average RMSEs Across All Models by Size
(Percent)



Source: LSEG Datastream (accessed on March 17, 2025); and authors' calculations.
Note: 3FF = Three-factor Fama-French; 5FF = Five-factor Fama-French; OLS = Ordinary Least Squares; and DTR = Decision Tree Regression. The RMSEs calculated here are performed on monthly stock returns in test set using a one-month rolling prediction.

B. Volatility Results

Overall, EGARCH and GARCH models are preferred in terms of model selection frequency, though HARCH models gain prominence under BIC criteria. EGARCH(1,1) emerged as the most frequently selected model, appearing either as the best or second-best model in 521 firms by AIC and 504 firms by BIC respectively. GARCH(1,1) follows closely, with 402 selections based on AIC and 396 based on BIC. Notably, while HARCH(1,1) performs modestly under AIC, it stands out under BIC, being the top-performing model in 330 cases, indicating its strength in penalizing model complexity more heavily. Hence with these results, EGARCH (1,1) was chosen as the default model for volatility prediction for the sample of 975 stocks.

Table 5: Best and Second-Best Model by Firm Count

Model	AIC		BIC	
	Best	Second-Best	Best	Second-Best
EGARCH(1,1)	252	235	269	0
EGARCH(1,1)	0	0	0	235
EGARCH(1,2)	75	0	0	77
EGARCH(1,2)	0	77	36	0
EGARCH(2,1)	113	106	39	106
EGARCH(2,2)	58	72	18	72
FIGARCH(1,1)	8	23	0	23
GARCH(1,1)	269	133	263	133
GARCH(1,2)	44	70	1	0
GARCH(1,2)	0	0	0	70
GARCH(2,1)	20	41	0	41
GARCH(2,1)	0	0	4	0
GARCH(2,2)	6	15	0	15
HARCH(1,1)	105	56	0	56
HARCH(1,1)	0	0	330	0
HARCH(1,2)	0	105	0	105
HARCH(2,1)	25	17	0	17
HARCH(2,1)	0	0	15	0
HARCH(2,2)	0	25	0	25

Source: Authors' calculations.

EGARCH improves over GARCH by adding asymmetry in how volatility responds to shocks, being a better predictor when residuals have fat tails and giving more flexibility with fewer constraints. The conditional variance equation in the EGARCH (1,1) model is represented in the following:

$$\ln(\sigma_t^2) = \omega + \beta \ln(\sigma_{t-1}^2) + \alpha \left|\frac{\sigma_{t-1}}{\epsilon_{t-1}}\right| + \frac{\gamma \cdot \sigma_{t-1}}{\epsilon_{t-1}} \qquad (17)$$

Where σ_t^2 is the conditional variance at time t, β is the persistence parameter and controls how persistent volatility is over time, α is magnitude effect (symmetric) which measures the impact of the magnitude of shocks on volatility., ϵ_{t-1} is residual at time $t-1$ and is assumed to follow student-t distribution, $\left|\frac{\sigma_{t-1}}{\epsilon_{t-1}}\right|$ is standardized residual, ω is the constant term, γ is asymmetry or leverage effect, if γ < *0*, negative shocks increase volatility more than positive shocks. Log-variance ensures $\sigma_t^2 > 0$ automatically,

The prediction period spans January to December 2024 (out-of-sample period), using monthly returns data for all 975 U.S. firms. Since monthly returns are used, the model produces monthly volatility estimates. Some extreme outliers were observed for specific stocks, likely due to illiquidity in micro-cap firms or potential model overfitting. To account for heavy tails in the return distribution and reduce the impact of such noise, a Student-t distribution is used. Additionally, upper-end outliers are mitigated by clipping values exceeding five times the median, while lower volatility values are left unadjusted. Model predictions are generated using a rolling window approach, where actual returns are incrementally added to forecast the next period's returns and volatilities.

Similarly, RMSE is used as the test metric to evaluate the predicted returns against the actual returns. Average volatilities and RMSEs were computed by industry and by market capitalization (Table 6 and 7). The results indicate generally healthy RMSE values. Higher RMSE values were observed for micro-cap and small-cap stocks compared to large-cap stocks, likely due to greater idiosyncratic volatility and lower liquidity, which result in more erratic returns. Among industries, the healthcare sector exhibited the highest volatility. Unsurprisingly, sectors with higher volatility also tend to show higher RMSE values.

Table 6: Average RMSE and Volatility by Industry

Industry	Average RMSE	Average Volatility
Hi-Tech	0.15	0.20
Consumer	0.12	0.15
Healthcare	0.19	0.28
Manufacturing	0.10	0.13
Other	0.11	0.15

Source: LSEG Datastream (accessed on March 17, 2025); and authors' calculations.

Table 7: Average RMSE and Volatility by Size

Market Cap	Average RMSE	Average Volatility
Large Cap	0.08	0.11
Mid Cap	0.10	0.15
Small Cap	0.13	0.17
Micro Cap	0.21	0.28

Source: LSEG Datastream (accessed on March 17, 2025); and authors' calculations.

C. Clustering Results[14]

The clustering results for January and November 2024 (Figures 5, 6, and 7), which represent the first and last months of the test set, show consistent dispersion patterns across all three clustering methods. After filtering out extreme volatility values, the predicted return-volatility space shows a widespread along the volatility axis and a tighter concentration along the return axis. This resulted in dense clusters of stocks around the median in both axes. Each method partitions the data into four clusters reliably, although the clarity and shape of these clusters vary depending on each clustering algorithm.

In each method, the optimal cluster is visually highlighted in orange. The grid-based method offers the clearest segmentation of this group, identifying the top-left quadrant where predicted returns are high, and volatilities are low. This precision is due to its rule-based design, which assigns stocks into quadrants using monthly median splits. In contrast, the KMeans algorithm uses a centroid-based approach that favours predicted returns more strongly than volatility, and as a result includes stocks with higher volatilities. Agglomerative Clustering also tends to isolate the top-left region as the optimal group, but it does so less precisely, assigning a relatively large number of stocks to the cluster. Pseudo-centroids are computed as the mean of each cluster and show that the optimal cluster still reflects relatively high predicted returns, but its boundaries are more diffused.

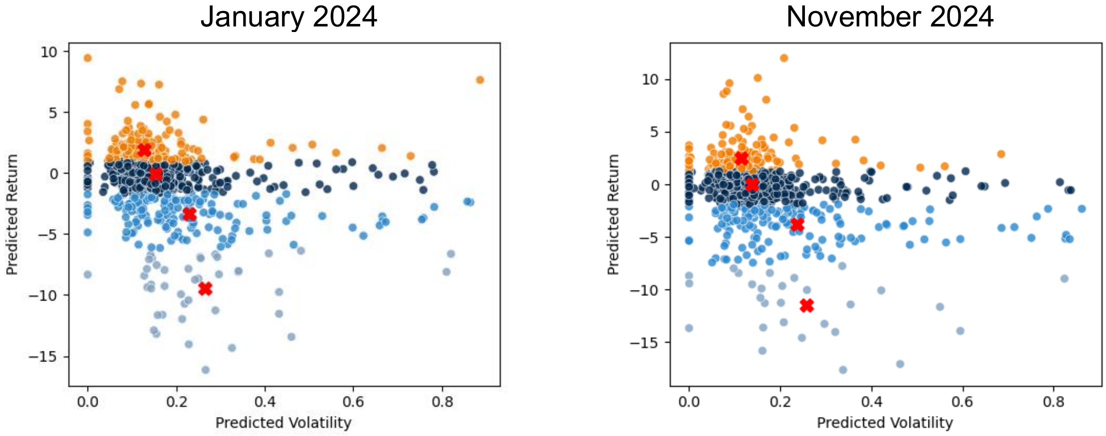
A fixed number of four clusters was used for each method. The elbow method, which is commonly used to determine the optimal number of clusters by evaluating the within-cluster sum of squares, was not applied in this analysis.[15] This was a deliberate choice as the objective was not to uncover the "true" number of underlying groups, but to reliably extract a subset of stocks with attractive forward-looking risk-return characteristics. Four clusters were sufficient to isolate one interpretable group in each method that served as a reasonable candidate for an optimal cluster.

The clustering results help to filter the sample of stocks, with each clustering method providing a distinct lens for identifying potentially attractive stocks. Although the grid-based cluster is clean and well-defined, it does not necessarily guarantee that the sub-sample of stocks are ideal. The relatively high RMSEs in return predictions, along with extreme or unstable volatility forecasts observed in certain months, underscore the importance of accounting for the forecast uncertainty. Allowing for this margin of error, the broader groupings produced by KMeans and Agglomerative Clustering help retain a wider pool of stocks, offering greater flexibility when computing portfolio weights based on predicted inputs and applying those weights to realised outcomes. This broader approach supports a more diversified and potentially more stable foundation for maximising the out-of-sample Sharpe ratio. Given this trade-off between precision and robustness, it is necessary to test the performance of each clustering method and evaluate which provides the most effective portfolio weights, based on predicted estimates, when applied to actual returns and volatilities.

[14] Only SVR return estimates under the three-factor Fama-French alpha model are analysed in this paper, as the SVR estimates from the five-factor alpha specification were found to be broadly similar and are available upon request from the authors.

[15] Within-cluster sum of squares refers to the sum of squared Euclidean distances between each data point and the centroid of its assigned cluster.

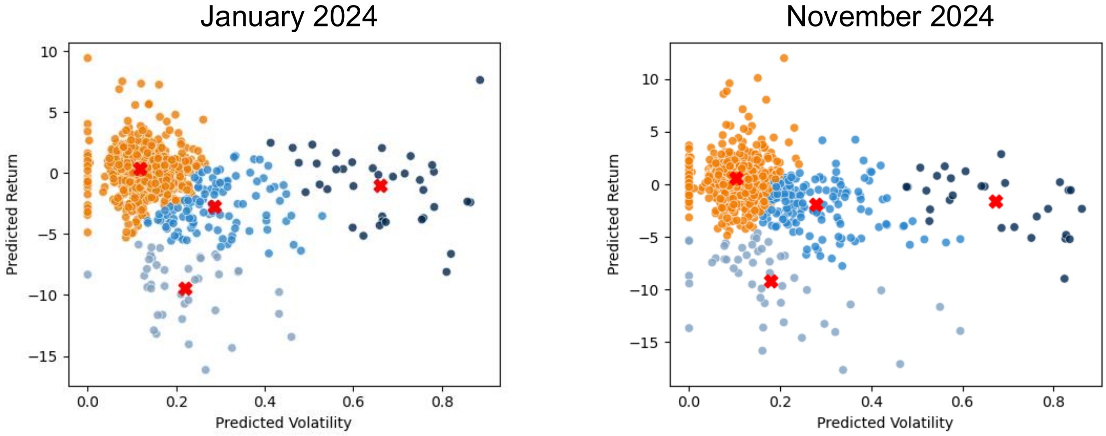
Figure 5: KMeans Clustering


January 2024

November 2024

Source: LSEG Datastream (accessed on March 17, 2025); and authors' calculations.
Note: Clusters above are formed using SVR predicted returns under the three-factor Fama-French alpha model. KMeans groups stocks into four clusters based on descending predicted returns. The orange cluster lies in the upper quadrant and is considered optimal, representing higher return. Cluster centroids, representing the average predicted return and volatility within each cluster, are shown as red 'X' markers.

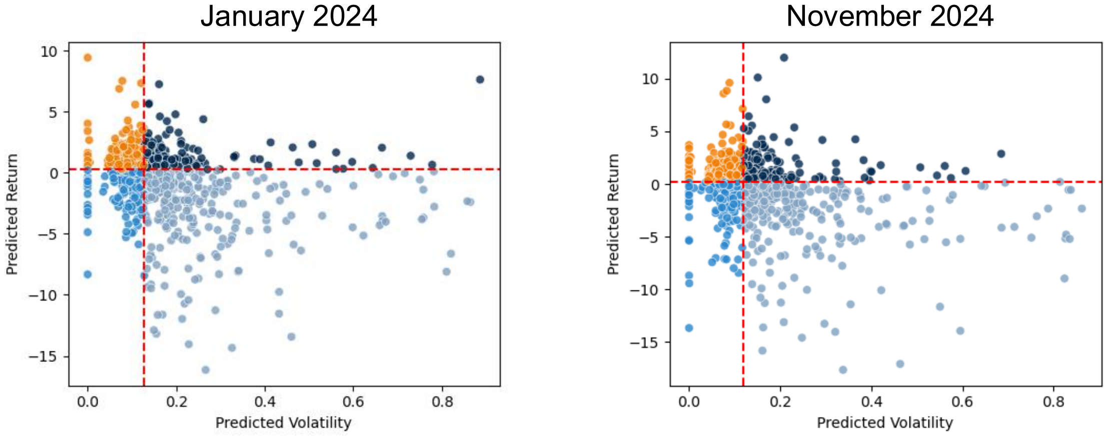
Figure 6: Agglomerative Clustering


January 2024

November 2024

Source: LSEG Datastream (accessed on March 17, 2025); and authors' calculations.
Note: Clusters above are formed using SVR predicted returns under the three-factor Fama-French alpha model. As Agglomerative Clustering is not a centroid-based algorithm, pseudo-centroids are calculated by taking the mean position of all points in each cluster, represented by the red 'X' markers. The orange cluster is considered optimal, representing comparatively higher returns.

Figure 7: Simple Grid Clustering


January 2024

November 2024

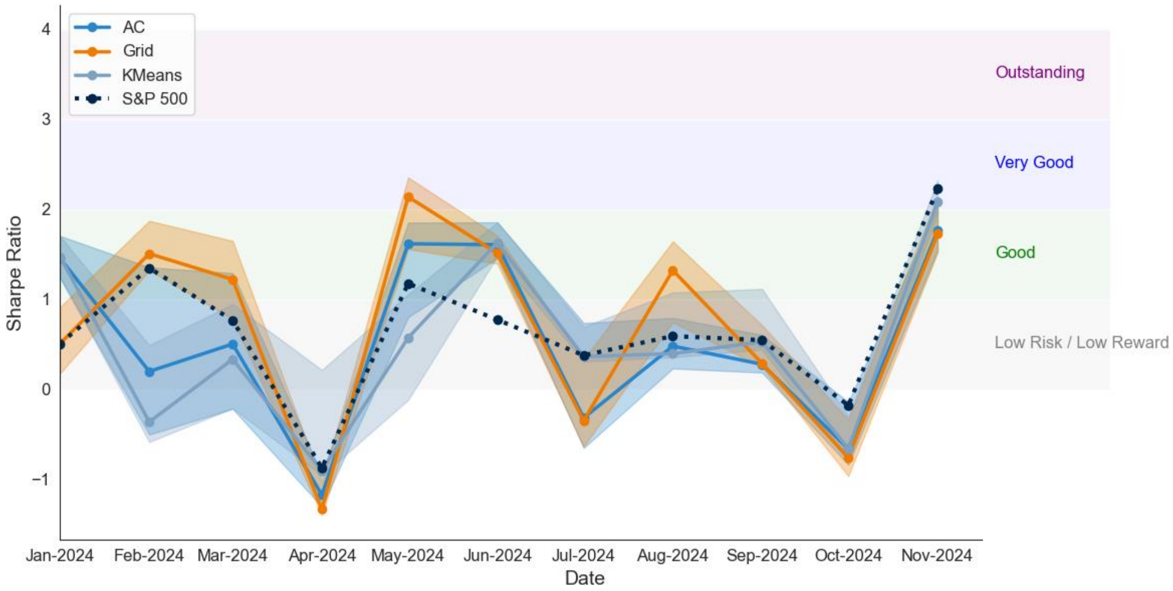
Source: LSEG Datastream (accessed on March 17, 2025); and authors' calculations.
Note: Clusters above are formed using SVR predicted returns under the three-factor Fama-French alpha model. The vertical and horizontal red dashed lines represent the median predicted volatility and return, respectively. These lines divide the plot into four quadrants. The orange cluster lies in the upper-left quadrant and is considered optimal, representing high return and low volatility U.S stocks. Other clusters are assigned based on their relative positions across the median splits.

D. Streamlining Optimal Clusters

Despite the limitations of forecast accuracy, the clustering-based framework shows potential for enhancing portfolio performance. Although return predictions exhibited relatively high RMSEs and some volatility estimates are distorted by outliers, particularly for illiquid stocks, the clustering process still succeeds in isolating stock groups with favourable predicted risk-return characteristics. The Grid and Agglomerative Clustering methods frequently produce portfolios with median Sharpe ratios that meet or exceed that of the S&P 500 benchmark for several periods over the test set (Figure 8).[16]

Comparative performance across different clustering methods shows the practical trade-offs between concentration and diversification. While the Grid method produces more concentrated clusters with sharper Sharpe ratio peaks, the broader definitions from KMeans and Agglomerative Clustering allow for greater flexibility under forecast uncertainty (arising from a wider selection of stocks within the optimal cluster). Moreover, the performance bands across all clustering techniques are not consistent across the different months, even though the number of top-N and minimum weight constraint combinations imposed as robustness check was limited. Despite this, the clustering and weighting framework consistently delivers risk-adjusted returns that outperform the benchmark in several periods. This suggests that, even with imperfect forecasts, the models appear to capture the general direction of expected returns well enough. Hence, there is potential for usefully structuring forward-looking signals through clustering and optimization to further improve portfolio outcomes.

Figure 8: Out-of-Sample Median Sharpe Ratio Performance by Clustering Method Against the S&P 500 Benchmark



Source: LSEG Datastream (accessed on March 17, 2025); and authors' calculations.
Note: Sharpe ratios are calculated by dividing the most recent monthly return within a 12-month rolling window by the monthly volatility, which is computed as the standard deviation of returns over that same window. Solid lines represent the median Sharpe ratio for each clustering method, while the shaded areas reflect the range between minimum and maximum Sharpe ratios. These values are computed across portfolios generated from all iterated combinations of top-N stock selections and minimum weight allocation constraints. The S&P 500 line serves as a benchmark reference and is plotted as a dotted line. Annotations are made accordingly to indicate performance levels based on Sharpe ratio ranges (Schwab, accessed 6 April, 2025). KMeans = KMeans Clustering; AC = Agglomerative Clustering; and Grid = Median-Based Grid Clustering.

[16] Note that S&P 500 index returns include dividend reinvestment, whereas the constructed portfolios exclude dividend payouts, which may understate their relative performance.

V. Conclusion

Building effective, automated portfolio construction tools requires integrating predictive modelling with practical strategies for filtering and allocating assets under uncertainty. This study contributes to that goal by extending the Fama-French factor framework through ML and clustering techniques. By predicting returns and volatilities across nearly 1,000 U.S. firms from five industries, and grouping stocks based on forward-looking risk-return profiles, the framework offers a structured approach to constructing risk-adjusted portfolios that yield positive portfolio outcomes.

The empirical findings show that while ML models such as SVR, Lasso, and XGBoost offer modelling flexibility, they do not significantly outperform OLS at short horizons. Fama-French factors, while effective in explaining returns, exhibit limited predictive power at monthly frequencies. Even though SVR return predictions were not highly accurate, they were reasonably effective in capturing the direction of returns across stocks. When combined with volatility estimates, these estimates help identify stocks into desirable clusters for portfolio selection. Results also indicate that volatility behaviour varies systematically across sectors and firm sizes, with EGARCH(1,1) emerging as the most robust model due to its ability to capture asymmetric effects and fat tails.

The clustering step plays a central role in portfolio formation by grouping stocks based on their predicted return and volatility profiles. Among the three techniques implemented, the Grid method consistently provided clear segmentation of high-return, low-volatility stocks. In contrast, KMeans and Agglomerative Clustering generated broader clusters that better accommodated forecast uncertainty and allowed for greater variety. Despite differences in precision, all three methods managed to isolate stocks with suitable risk-return profiles for a useful starting basis to begin the subsequent streamlining for investable portfolio construction.

The streamlining process applied top-N stock selection and constrained weight optimization to form investable portfolios. Results showed that the Grid-based portfolios achieved consistently strong and stable Sharpe ratios, frequently outperforming the S&P 500 benchmark. While KMeans occasionally delivered higher maximum Sharpe ratios, its performance was more volatile across months. Even with sub-optimal forecasts, the framework can translate model outputs into practical and competitive equity portfolios yielding positive outcomes.

These results highlight the value of integrating statistical models, machine learning, and clustering to improve portfolio design. However, challenges remain. Survivorship bias from excluding firms listed after 2007 limits market representativeness, while sectoral imbalances, especially the overweight in technology, may distort factor exposures. Applying a single return and volatility model across all firms oversimplifies stock-level differences, which could be addressed through firm-specific model selection. Real-world applicability would also benefit from incorporating seasonality, transaction costs, and rebalancing frictions.

In summary, this study presents a structured framework for integrating predictive signals into portfolio construction. The findings point to two key areas for advancement: improving prediction accuracy through firm-specific modelling and longer forecast horizons, and enhancing portfolio robustness through adaptive weighting schemes and more representative market samples. Together, these extensions can support developing more resilient data-driven strategies for asset allocation in practice.

References

Aliyev, Fuzuli, Richard Ajayi, and Nijat Gasim. 2020 “Modelling Asymmetric Market Volatility With Univariate GARCH Models: Evidence from Nasdaq-100.” *The Journal of Economic Asymmetries* 22: e00167.

Banz, Rolf W. 1981. “The Relationship Between Return and Market Value of Common Stocks.” *Journal of Financial Economics* 9(1): 3–18.

Basu, Sanjay. 1983. “The Relationship between Earnings Yield, Market Value, and Return for NYSE Common Stocks: Further Evidence.” *Journal of Financial Economics* 12(1): 129–156.

Bhandari, Laxmi Chand. 1988. “Debt/Equity Ratio and Expected Common Stock Returns: Empirical Evidence.” *Journal of Finance* 43(2): 507–528.

Black, Fischer. 1972. “Capital Market Equilibrium With Restricted Borrowing.” *Journal of Business* 45(3): 444–454.

Brandt, Michael W, and Christopher S Jones. 2012. “Volatility Forecasting With Range-Based EGARCH Models.” *Journal of Business & Economic Statistics* 24(4): 470–486.

Cakici, Nusret. 2015. “The Five-Factor Fama-French Model: International Evidence.” Fordham University.

Cerqueira, Vitor, Luis Torgo, and Igor Mozetič. 2020. “Evaluating Time Series Forecasting Models: An Empirical Study on Performance Estimation Methods.” *Machine Learning* 109: 1997–2028.

Chan, Louis, Yasushi Hamao, and Josef Lakonishok. 1991. “Fundamentals and Stock Returns in Japan.” *Journal of Finance* 46(5): 1739–1789.

Chatterjee, Ananda, Hrisav Bhowmick, Jaydip Sen. 2022. “Stock Volatility Prediction using Time Series and Deep Learning Approach.” 2022 IEEE 2nd Mysore Sub Section International Conference.

Danial, Kent D, and Sheridan Titman. 1997. “Evidence on the Characteristics of Cross Sectional Variation in Stock Returns.” *Journal of Finance* 52(1): 1–33.

Diallo, Boubacar, Aliyu Bagudu, and Qi Zhang. 2023. “Fama-French Three Versus Five, Which Model is Better? A Machine Learning Approach.” *Journal of Forecasting* 42(6): 1461–1475.

Dittmar, Robert F. 2002. “Nonlinear Pricing Kernels, Kurtosis Preference, and Evidence from the Cross Section of Equity Returns.” *Journal of Finance* 57(1): 369–403.

Ezzat, Hassan. 2012. “The Application of GARCH and EGARCH in Modeling the Volatility of Daily Stock Returns During Massive Shocks: The Empirical Case of Egypt.” MPRA Paper 50530, University Library of Munich.

Fama, Eugene F. and Kenneth R. French. 1992. “The Cross-Section of Expected Stock Returns.” *Journal of Finance* 47(2): 427–465.

_________. 1993. “Common Risk Factors in the Returns on Stocks and Bonds.” *Journal of Financial Economics* 33: 3–56.

_________. 2003. “The Capital Asset Pricing Model: Theory and Evidence.” *Journal of Economic Perspectives* 18(3): 25–46.

_________. 2012. “Size, Value, and Momentum in International Stock Returns.” *Journal of Financial Economics* 105: 457–472.

_________. 2015. “A Five-Factor Asset Pricing Model.” *Journal of Financial Economics* 116: 1–22.

Filipović, Damir and Amir Khalilzadeh. 2021. “Machine Learning for Predicting Stock Return Volatility.” Swiss Finance Institute Research Paper No. 21–95.

Gibbons, Michael R., Stephen A. Ross, and Jay Shanken. 1989. “A Test of the Efficiency of a Given Portfolio.” *Econometrica* 57(5): 1121–1152.

Gogas, Periklis, Theofilos Papadimitriou, and Dimitrios Karagkiozis. 2018. “The Fama 3 and Fama 5 Factor Models Under a Machine Learning Framework.” Working Paper Series 18–05, Rimini Centre for Economic Analysis.

Griffin, John M., and Michael Lemmon. 2002. “Book-to-Market Equity, Distress Risk, and Stock Returns.” *Journal of Financial Economics* 57(5): 2317–2336.

Gu, Shihao, Bryan T Kelly, and Dacheng Xiu. 2018. “Empirical Asset Pricing Via Machine Learning.” Chicago Booth Research Paper No. 18–04, 31st Australasian Finance and Banking Conference 2018, Yale ICF Working Paper No. 2018–09.

Hyndman, Rob J, and George Athanasopoulos. 2018. “Forecasting: Principles and Practice.” OTexts.

Lemieux, Victoria, Payam S. Rahmdel, Rick Walker, B. L. William Wong, and Mark Flood. 2015. “Clustering Techniques And their Effect on Portfolio Formation and Risk Analysis.” Office of Financial Research (OFR) Staff Discussion Paper No. 2015–01.

León, Deigo, Arbey Aragón, Javier Sandoval, Germán Hernández, Andrés Arévalo, and Jaime Niño. 2017. “Clustering Algorithms for Risk-Adjusted Portfolio Construction.” *Procedia Computer Science* 108: 1334–1343.

Lintner, John. 1965. “The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets.” *Review of Economics and Statistics* 47(1): 13–37.

Markowitz, Harry. 1952. “Portfolio Selection.” *Journal of Finance* 7(1): 77–91.

Rossi, Matteo. 2016. “The Capital Asset Pricing Model: A Critical Literature Review.” *Global Business and Economics Review* 18(5): 604–617.

Sharpe, William F. 1964. “Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk”. *Journal of Finance* 19(3): 425–442.

Sinha, Bhaskar. 2012. “On Historical Volatility in Emerging Markets Using Advanced GARCH Models”. SSRN: http://dx.doi.org/10.2139/ssrn.3093469.

Wu, Dingming, Xiaolong Wang, and Shaocong Wu. 2022 “Construction of Stock Portfolios Based on K-Means Clustering of Continuous Trend Features”. *Knowledge-Based Systems* 252: 109358.