

Stock Asset Pricing Based on the Integration of Multifactor Models and Machine Learning

Sihan Qin

*Irvine Valley College, Irvine, USA
kellyqin2006@gmail.com*

Abstract. This study examines whether combining traditional multifactor asset-pricing theory with machine-learning methods can improve the explanation and prediction of U.S. stock returns. The sample comprises ten large U.S.-listed firms across six sectors, with SPY as the market proxy. Monthly returns cover August 2023 through June 2026, providing 35 observations for the preliminary market-model analysis, while the underlying dataset also contains approximately three years of daily prices for the proposed machine-learning extension. Preliminary estimates reveal substantial heterogeneity: market betas range from 1.945 for NVIDIA to 0.117 for Johnson & Johnson, while R-squared ranges from 49.6% for Amazon to 0.6% for Johnson & Johnson. These differences indicate that a single market factor does not explain return variation equally across securities. The paper therefore develops a framework that uses the Fama-French five-factor model as an economically interpretable benchmark and Random Forest and Extreme Gradient Boosting (XGBoost) as nonlinear extensions. Factor selection and functional form are distinct: traditional factors capture economically meaningful risks, while machine learning tests nonlinear, interactive, and time-varying effects. However, additional complexity is justified only when it produces stable out-of-sample improvement. The study consequently treats interpretability, chronological validation, and resistance to overfitting as necessary conditions for accepting a hybrid asset-pricing model.

Keywords: asset pricing, Fama-French five-factor model, machine learning, Random Forest, stock returns

1. Introduction

Asset pricing tries to explain why different stocks earn different expected returns and what types of risk investors are compensated for bearing. The Capital Asset Pricing Model (CAPM) provides a simple answer by linking expected excess return to systematic market risk. However, later empirical research showed that market beta alone does not explain many persistent differences across stocks. This led to the development of multifactor models. The Fama-French framework, for example, expands the market factor by including size, value, profitability, and investment. The development of additional pricing factors shows that asset pricing is still an evolving research problem rather than a closed model with a fixed set of variables [1].

A larger number of factors, however, does not automatically produce a better model. Evidence across developed and emerging markets shows that the Fama-French five-factor model does not consistently dominate simpler three or four-factor alternatives [2]. Factor effectiveness can depend on the market, the test assets, and the sample period. New factors can also overlap with existing information. Chen et al. show that housing mortgage leverage contains information about cross-sectional returns in the Chinese A-share market, but they also find that simply adding it to an existing specification may create redundancy when the new factor is highly correlated with established factors [1]. This suggests that the real problem is not how many factors can be added, but whether they provide independent and economically meaningful information.

A second problem concerns functional form. Traditional factor models normally assume that returns are a linear combination of factor exposures. Xu argues that empirical asset-pricing data may contain nonlinear relationships, interactions among characteristics, and high-dimensional dependence that a standard linear model handles only imperfectly [3]. Liu et al. also show that statistically extracted factors can provide additional explanatory information beyond established Fama-French factors [4]. These findings raise an important possibility: a traditional factor model may underperform not only because it is missing a factor, but also because it combines existing information too rigidly.

Machine learning offers one way to relax this restriction. Random Forest and Extreme Gradient Boosting (XGBoost) can represent thresholds, nonlinear effects, and interactions without requiring every relationship to be specified in advance. Recent financial studies have found useful predictive information in such methods, although the best algorithm varies across stocks and datasets [5, 6]. Xue et al. go further by combining multifactor information with a dynamic attention network, allowing factor importance to change through time [7]. Yet machine learning creates its own problems. Financial data are noisy, complex models can overfit, and a high training accuracy does not necessarily represent stable information about future returns [8]. A recent hybrid study also finds that combining more technical and fundamental variables does not automatically improve generalization [9].

This paper therefore takes a middle position. It does not treat machine learning as a replacement for asset-pricing theory. Instead, economically meaningful factors provide the structure, while machine learning is used to examine whether the relationship between those factors and returns needs to be more flexible.

Three questions guide the analysis. First, does a market-based model explain returns equally well across different stocks? Second, is there a theoretical and empirical reason to move from a single market factor to a broader factor structure? Third, under what conditions would adding machine-learning flexibility represent a genuine improvement rather than overfitting?

The central argument of this paper is that factor selection and functional form are two different problems. Traditional finance theory is useful for deciding what information should matter. Machine learning may be useful for deciding how that information is related to returns. A hybrid model is valuable only if the additional flexibility survives out-of-sample testing and can still be connected to an interpretable economic mechanism.

2. Literature review

2.1. Development and limits of multifactor asset pricing

CAPM reduces priced risk to exposure to systematic market risk, implying that expected excess return is determined by an asset's sensitivity to movements in the market portfolio [10]. Multifactor

models expand this framework because empirical returns do not appear to be driven by market beta alone. Fama and French introduced a three-factor framework that adds size (SMB) and book-to-market (HML) factors to the market factor [11]. They later extended this structure to a five-factor model by adding profitability (RMW) and investment (CMA) factors [12]. This expansion is economically useful because each factor represents a specific dimension of return variation rather than an arbitrary statistical predictor.

However, evidence on the superiority of larger factor models is mixed. Lalwani and Chakraborty compare CAPM, three-factor, four-factor, five-factor, and augmented specifications across five developed and five emerging markets [2]. Their results show that five factors can improve pricing in some markets, including China and the United States, but the improvement is not universal. They therefore evaluate models using not only adjusted R-squared, but also pricing errors and parsimony. This leads to an important principle for the present study: a more complicated model should be preferred only when its additional complexity produces measurable improvement.

The identity of relevant risk factors may also depend on the economic environment. Mateev and Videv examine the Bulgarian equity market and find that macroeconomic variables have weak explanatory power at the individual-stock level, although the results improve after stocks are grouped into portfolios and a two-pass regression procedure is used [13]. They also emphasize problems such as thin trading, institutional conditions, and data quality in transition markets [13]. The implication is that factor performance cannot be separated completely from market structure.

Chen et al. provide another example. They identify housing mortgage leverage as a relevant pricing variable in the Chinese A-share market and connect it to the size effect [1]. At the same time, they show that directly inserting the variable into canonical models can create redundancy because the proposed factor overlaps with market and size information [1]. This result weakens the simple idea that asset-pricing progress comes from continuously adding new variables.

The broader lesson is that the factor zoo is partly a model-selection problem. An apparently significant factor may be economically meaningful, statistically redundant, sample-specific, or a combination of all three. A useful extension should therefore improve explanatory or predictive performance after existing information has already been controlled for.

2.2. From fixed factors to data-driven information

Liu et al. provide a useful bridge between traditional factor models and modern data-driven methods [4]. Instead of replacing financial factors, they construct statistical factors from a large panel of U.S. stock returns and test whether those factors add information beyond market, size, value, and momentum. Their results indicate incremental explanatory power from the newly constructed statistical factors.

This matters because it changes the debate. The choice is not necessarily financial theory or data science. A more useful structure is to allow theory to define economically motivated variables while statistical methods search for information that the original specification does not capture.

The same idea can be seen more broadly in financial pricing. Li et al. study European options under a multifactor volatility framework and argue that constant volatility is too restrictive because volatility changes over time and across market conditions [14]. Their study concerns derivatives rather than stock cross-sectional returns, so it cannot be used as direct evidence for equity factor pricing. Its relevance is conceptual: fixed-parameter financial models can miss dynamic risk structures.

2.3. Machine learning in financial prediction

Machine-learning research provides strong evidence that model specification matters. Rai and Soltanisehat compare nine algorithms using 36 technical indicators for Apple, Tesla, and NVIDIA. They find that the best model differs by company, while momentum indicators are generally influential and volatility measures matter especially for Tesla [5]. This indicates that a single fixed mapping from financial information to returns may not work equally well across securities.

Ensemble methods attempt to deal with this heterogeneity. Kayit and Ismail combine classifiers with different learning structures and argue that hybrid models can benefit from the complementary strengths of individual algorithms [6]. Their study also emphasizes feature selection and hyperparameter control rather than assuming that more complexity is automatically useful [6].

Xue et al. are especially relevant to the current paper because they combine multifactor information with a dynamic attention mechanism [7]. Their model allows the importance of different factors to change before using recurrent learning to capture temporal dependence. Their reported prediction accuracy of 90.42% exceeds the comparison models in their own sample [7]. The most important implication, however, is not the numerical accuracy itself. It is the idea that factor relevance may be dynamic rather than constant.

There is also strong reason for caution. Zhang et al. show that machine-learning and deep-learning models that perform well in other domains may fail to extract stable future information from real financial price series [8]. They identify black-box behavior, overfitting, parameter sensitivity, and noise as major concerns [8]. Abd-Elwahab et al. similarly find that combining technical and fundamental information does not consistently improve out-of-sample performance; feature relevance and stability can matter more than the amount of data supplied to the model [9].

Meera et al. demonstrate a standard forecasting pipeline in which historical prices and volume are preprocessed, engineered into features, and evaluated using regression and long short-term memory (LSTM) models [15]. Their framework is useful methodologically, but it also illustrates a distinction important to this paper: forecasting prices from past market data is not the same question as explaining why investors earn compensation for risk.

The research gap therefore lies between two literatures. Traditional factor models offer economic interpretation but impose relatively restrictive structures. Machine learning offers flexible functions but can lose economic meaning. The present study argues that the useful question is not whether one approach can defeat the other. It is whether nonlinear methods can extract additional information from economically motivated factors without destroying interpretability.

3. Theoretical foundation

The theoretical foundation begins with the distinction between total and systematic risk. Diversification can reduce firm-specific risk, so investors should not automatically receive higher expected returns simply because an individual stock is more volatile. CAPM instead links expected excess return to systematic market exposure [10]:

$$E(R_i) - R_f = \beta_i[E(R_m) - R_f] \quad (1)$$

The strength of CAPM is that it establishes a clear pricing principle: only risk that cannot be eliminated through diversification should command a systematic risk premium. Its main restriction for the present study is that market exposure is treated as the central source of priced risk.

Fama and French extended this framework by introducing size and value factors in addition to the market factor [11]. Their later five-factor model further incorporates profitability and investment [12]:

$$R_{i,t} - R_{f,t} = \alpha_i + \beta_M(MKT_t - RF_t) + \beta_S SMB_t + \beta_H HML_t + \beta_R RMW_t + \beta_C CMA_t + \varepsilon_{i,t} \quad (2)$$

Here, SMB captures the size premium, HML captures the value premium, RMW captures the profitability premium, and CMA captures the investment premium [12]. The empirical factor series used in this study are obtained from the Kenneth R. French Data Library, which constructs the five research factors from value-weighted portfolios formed on size and book-to-market, operating profitability, and investment [16].

The key theoretical extension in this paper concerns the function connecting these factors to returns. A standard linear factor model assumes a fixed additive relationship:

$$E(R) = \alpha + \sum_{k=1}^K \beta_k X_k \quad (3)$$

A nonlinear model instead allows:

$$E(R) = f(X_1, \dots, X_K) \quad (4)$$

This distinction matters because a factor may not have the same marginal effect under every market condition. For example, value exposure may have a stronger relationship with returns among smaller firms, or profitability may become more relevant under particular market conditions. A linear specification can capture such relationships only when the relevant interaction or nonlinear term is specified in advance. Tree-based machine-learning models provide a way to test these relationships without manually imposing every possible functional form.

The argument of this study is therefore that poor linear performance does not immediately imply that the underlying factors are useless. The information may be relevant while the assumed functional relationship is wrong.

Random Forest is used because decision trees can capture thresholds and interactions, while aggregation across trees reduces dependence on a single tree. XGBoost provides a second nonlinear specification in which trees are added sequentially and regularization limits unnecessary complexity. These models are not selected because machine learning is assumed to be superior. They are selected because they provide two different ways to test whether nonlinear structure adds value beyond the linear factor benchmark.

A meaningful performance gain must persist out of sample, and the source of that gain should also be interpretable. Feature importance alone is insufficient because it primarily ranks the contribution of variables without fully describing the direction and shape of their effects. This study therefore proposes SHAP (SHapley Additive exPlanations), introduced by Lundberg and Lee, to interpret nonlinear model predictions [17]. SHAP assigns each feature a contribution to an individual prediction, allowing the analysis to examine whether a factor raises or lowers predicted excess return at different values and whether its effect changes in the presence of other factors [17].

This creates a falsifiable proposition: if nonlinear models do not produce consistent out-of-sample improvement over the linear factor benchmark, the additional complexity is not justified. In that case, the simpler factor model should be preferred. The purpose of machine learning in this

study is therefore not to replace established asset-pricing theory, but to test whether economically meaningful factors contain nonlinear or interactive information that a fixed linear specification fails to capture.

4. Model construction and empirical analysis

4.1. Data and sample

The stock dataset contains adjusted historical prices for ten U.S.-listed companies: Apple (AAPL), Microsoft (MSFT), NVIDIA (NVDA), Amazon (AMZN), Alphabet (GOOGL), Meta Platforms (META), JPMorgan Chase (JPM), Johnson & Johnson (JNJ), Exxon Mobil (XOM), and Procter & Gamble (PG). SPY is included as a broad U.S. market proxy. The underlying workbook records daily adjusted closing prices and monthly returns for the sample.

The daily dataset begins in July 2023 and contains about 750 trading-day price observations. Monthly returns are calculated from the last adjusted closing price of each complete month. The resulting monthly-return sample contains 35 observations from August 2023 through June 2026.

The sample is deliberately heterogeneous. It contains large technology and communication firms as well as financial, healthcare, energy, and defensive consumer companies. This matters because a model that works only for highly correlated growth stocks provides weak evidence of a general asset-pricing relationship.

The observed monthly statistics already show large differences. NVIDIA has the highest average monthly return, 4.82%, and the highest monthly standard deviation, 11.22%. Alphabet averages 3.28%, while JPM averages 2.44% with a substantially lower standard deviation of 5.81%. Apple averages 1.35% with a standard deviation of 6.42%, while PG has the lowest average return, 0.14%, and a standard deviation of 4.96%.

These values show why total volatility cannot be treated as a complete explanation of returns. For example, JPM has a higher realized average monthly return than Apple despite lower monthly volatility. This does not reject CAPM because realized sample returns are not identical to expected returns, but it shows that the simple statement that higher volatility produces higher return is inadequate.

The correlation structure also confirms that the sample contains different risk exposures. MSFT and AMZN have a return correlation of 0.661, and AMZN and GOOGL have a correlation of 0.643. In contrast, MSFT and XOM have a correlation of -0.421, while AMZN and JNJ have a correlation of -0.382. These differences provide a concrete reason to examine systematic factors rather than treating the ten securities as similar exposures.

4.2. Preliminary market-model evidence

Before estimating a larger factor structure, a preliminary market model is used:\

$$R_{i,t} = \alpha_i + \beta_i R_{SPY,t} + \varepsilon_{i,t} \quad (5)$$

This is deliberately described as a market-model regression rather than a final CAPM test, because SPY return rather than the official market excess return is used at this stage. That distinction avoids overstating what the regression proves.

Table 1. Preliminary market-model results based on 35 monthly observations (August 2023-June 2026)

Stock	Market Beta	R-squared	Beta p-value
AAPL	0.904	0.288	0.0009
MSFT	1.264	0.442	<0.001
NVDA	1.945	0.438	<0.001
AMZN	1.527	0.496	<0.001
GOOGL	1.250	0.268	0.0014
META	1.572	0.404	<0.001
JPM	0.969	0.405	<0.001
JNJ	0.117	0.006	0.645
XOM	-0.393	0.060	0.158
PG	0.392	0.091	0.078

Using the 35 monthly observations gives the results reported in Table 1. The most useful result is not simply that NVIDIA has the highest beta. The main finding is the large dispersion in explanatory power. SPY explains approximately 49.6% of monthly variation in Amazon, 44.2% in Microsoft, 43.8% in NVIDIA, and 40.4% in Meta. For Johnson & Johnson, the corresponding R-squared is only 0.6%, and for Exxon Mobil it is 6.0%.

This provides direct evidence for the first argument of the paper: the weakness of a single-factor structure is security-specific rather than uniform. Market movements contain substantial information for several growth-oriented firms, but almost none of the monthly variation in JNJ is explained by the same market proxy. A broader factor structure therefore has a clear empirical purpose: it must explain variation that market exposure leaves behind.

A second result is equally important. Realized average returns are not mechanically ordered by estimated beta. NVIDIA combines the highest return and beta, but JPM earns a 2.44% monthly average with a beta close to one, while JNJ earns 1.59% with an estimated beta of only 0.117. This is not evidence that CAPM is mathematically wrong; the sample is short and realized returns contain shocks. The correct conclusion is narrower: market beta alone does not produce a simple monotonic ordering of realized returns in this sample.

4.3. Multifactor and machine-learning design

For the multifactor extension, the Fama-French factor set contains MKT-RF, SMB, HML, RMW, CMA, together with the risk-free rate (RF). The correct dependent variable is excess return:

$$ER_{i,t} = R_{i,t} - RF_t \quad (6)$$

The linear benchmark is then the Fama-French five-factor (FF5) model, while Random Forest and XGBoost use the same factor information under a flexible function. Keeping the input information constant is essential. If the machine-learning model were given dozens of extra indicators while FF5 received only five factors, a higher predictive score would not identify a functional-form improvement; it could simply reflect additional information.

The data must also be split chronologically rather than randomly. Random mixing would allow later market conditions to enter the training set and would produce an unrealistically favorable test. Hyperparameters should therefore be selected only from earlier observations, with later dates reserved for evaluation.

The principal evaluation measures are root mean squared error (RMSE), mean absolute error (MAE), and out-of-sample R^2 . The decision rule should be set before looking at the results. If FF5 improves on the market benchmark but Random Forest and XGBoost do not improve on FF5 out of sample, the evidence supports additional factors but not additional nonlinearity. If the nonlinear models improve consistently and SHAP reveals stable threshold or interaction effects, the hybrid hypothesis receives support. If training performance improves while test performance deteriorates, the appropriate conclusion is overfitting.

Importantly, this paper does not report fabricated Random Forest or XGBoost accuracy values. A complete numerical comparison requires an exactly aligned factor-return panel and a reproducible training procedure. Reporting invented R-squared, RMSE, or SHAP values simply to make the hybrid model appear successful would directly contradict the methodological argument of this study. That limitation is itself informative because the literature repeatedly warns that machine-learning performance in finance can be unstable [8, 9]. A credible hybrid model has to earn its complexity through reproducible evidence.

5. Conclusion

This paper examines asset pricing from a different angle than a simple contest between traditional finance and machine learning. The main argument is that factor information and functional form should be treated separately. Traditional factor models provide economic discipline by identifying risk exposures that may justify return differences. Machine learning provides flexibility by allowing those exposures to affect returns nonlinearly or interact with one another.

The real stock data provide a clear reason to move beyond a single market factor. Across the ten-stock sample, market-model R-squared ranges from approximately 49.6% for Amazon to only 0.6% for Johnson & Johnson. Estimated market betas also range from 1.945 for NVIDIA to 0.117 for JNJ, with XOM producing a negative point estimate over the short sample. These differences show that common market exposure is important for some securities but far from sufficient for all of them.

The evidence does not justify the stronger claim that machine learning is automatically superior. That would require a properly aligned FF5 dataset, chronological validation, and reproducible out-of-sample comparison. The literature itself provides reasons not to make such an assumption [8, 9]. A model that achieves better training fit but loses that advantage in later observations has not improved asset pricing; it has simply used more flexibility to describe history.

The resulting conclusion is therefore more specific than saying that machine learning improves stock prediction. A factor-machine-learning hybrid is useful only when the economic information contained in traditional factors is genuine but the relationship between those factors and returns is too nonlinear, interactive, or time-dependent for a fixed linear specification. If that condition cannot be demonstrated out of sample, the simpler factor model should be preferred.

Future work should construct a fully aligned daily panel of stock excess returns and FF5 factors over a longer period, estimate the same information set using FF5, Random Forest, and XGBoost, and use SHAP to identify whether any out-of-sample gain comes from economically meaningful nonlinear factor effects. The contribution of machine learning should then be judged not by complexity or headline accuracy, but by whether it produces stable, interpretable, and genuinely incremental pricing information.

References

- [1] Chen, Q., Li, H., Lin, J., & Gao, Y. (2025). The role of housing mortgage leverage in stock asset pricing: Evidence from the Chinese A-share market. *The Journal of Real Estate Finance and Economics*, 71(2), 209–253. <https://doi.org/10.1007/s11146-023-09940-5>
- [2] Lalwani, V., & Chakraborty, M. (2020). Multi-factor asset pricing models in emerging and developed markets. *Managerial Finance*, 46(3), 360–380. <https://doi.org/10.1108/MF-12-2018-0607>
- [3] Xu, X. (2025). Empirical asset pricing based on network big data mining and privacy protection. *Neural Computing and Applications*, 37(12), 7841–7855. <https://doi.org/10.1007/s00521-024-10110-8>
- [4] Liu, X., Guerard, J., Chen, R., & Tsay, R. (2025). Improving estimation of portfolio risk using new statistical factors. *Annals of Operations Research*, 346(1), 245–261. <https://doi.org/10.1007/s10479-024-06307-8>
- [5] Rai, B., & Soltanisehat, L. (2025). A multi-model machine learning framework for daily stock price prediction. *Big Data and Cognitive Computing*, 9(10), Article 248. <https://doi.org/10.3390/bdcc9100248>
- [6] Kayit, A. D., & Ismail, M. T. (2025). Advancing stock price prediction through the development of hybrid ensembles: A comprehensive comparative analysis of machine learning approaches. *Journal of Big Data*, 12, Article 232. <https://doi.org/10.1186/s40537-025-01185-8>
- [7] Xue, Z., Zhang, L., Zhang, Y., & He, C. (2026). Forecasting stock return based on the multifactor dynamic attention network. *Discrete Dynamics in Nature and Society*, 2026, Article 6662757. <https://doi.org/10.1155/ddns/6662757>
- [8] Zhang, P., Yang, J., Zhu, H., Hou, Y., Liu, Y., & Zhou, C. (2021). Failure in stock price prediction: A comparison between the curve-shape-feature and non-curve-shape-feature modes of existing machine learning algorithms. *International Journal of Computers Communications & Control*, 16(6), Article 4549. <https://doi.org/10.15837/ijccc.2021.6.4549>
- [9] Abd-Elwahab, A. M., El-Horbaty, Y. S., Abd-Elfattah, M. H., & Bahloul, M. M. (2026). A hybrid machine learning framework integrating technical and fundamental analysis for stock market prediction. *Future Business Journal*, 12, Article 191. <https://doi.org/10.1186/s43093-026-00897-4>
- [10] Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19(3), 425–442. <https://doi.org/10.1111/j.1540-6261.1964.tb02865.x>
- [11] Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1), 3–56. [https://doi.org/10.1016/0304-405X\(93\)90023-5](https://doi.org/10.1016/0304-405X(93)90023-5)
- [12] Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1), 1–22. <https://doi.org/10.1016/j.jfineco.2014.10.010>
- [13] Mateev, M., & Videv, A. (2008). Multifactor asset pricing model and stock market in transition: New empirical tests. *Eastern Economic Journal*, 34(2), 223–237. <https://doi.org/10.1057/palgrave.eej.9050031>
- [14] Li, L., Shi, G., Zhou, L., & Sheng, Y. (2025). European option pricing under uncertain multifactor exponential Ornstein-Uhlenbeck volatility model. *Journal of Applied Mathematics and Computing*, 71(4), 5159–5184. <https://doi.org/10.1007/s12190-025-02434-y>
- [15] Nagul Meera, S. D., Khan, M. F., Mallikarjuna, M., Bramaiah, M., & Reddy, A. (2025). Predicting stock prices using advanced machine learning models. *International Journal of Communication Networks and Information Security*, 17(3), 160–165.
- [16] French, K. R. (n.d.). Description of Fama/French 5 factors (2x3). *Kenneth R. French Data Library*. Retrieved August 19, 2026, from https://mba.tuck.dartmouth.edu/pages/faculty/Ken.french/Data_Library/f_5_factors_2x3.html
- [17] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* 30 (pp. 4765–4774). <https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions>