Estimate Industry Cost of Equity from Machine Learning

Jingcheng Li, Professor Nuttawat Visaltanachoti

Massey University

Master Of Business Studies (Finance)

Student No.

Contents

1. Introduction

1.1 Research Background

Estimating the cost of equity capital for industries, which is called Industry Cost of Equity (ICoE), is fundamentally important in corporate finance. Accurate estimates of ICoE are crucial for decision making. These decisions include capital budgeting, valuation, corporate financing, and portfolio investment.

However, regardless of its status in finance, estimating the industry-level cost of equity remains inaccurate. This causes uncertainty in corporate investment and valuation practices (Fama and French, 1997).

Traditionally, the Capital Asset Pricing Model (CAPM) was proposed by Sharpe in 1964 and Lintner in 1965. It is widely used to estimate equity costs because of its simplicity and easy to understand. According to CAPM, the expected return on a security is only determined by its sensitivity (beta) to the market portfolio risk factor.

Later, empirical evidence shows extensive limitations on CAPM. It can't captures the variations in expected returns associated to a firm's size, the ratio of book value to market equity (Fama and French, 1992). This results single-factor structure of CAPM often can't properly show the complex risk features of industries. These industries are marked by high volatility, growth trends, or financial problems.

To encounter with these limitations, Fama and French (1993, 1995) developed the Fama-French Three-Factor Model (FF3). This model expands the CAPM by adding two factors. One is the size factor (Small Minus Big, SMB), and the book-to-market equity factor (High Minus Low, HML). These factors are added to capture the differences in returns across different stocks. However, empirical studies have demonstrated that the FF3 model still exhibit substantial limitations when applied to industry-level cost of equity estimation. The model's inability to fully capture industry-specific risk factors remains a persistent challenge in empirical finance research.

Fama and French (1997) discovered imprecisions that mainly came from time-varying risk factor loadings and uncertain factor premiums. Specifically, their analysis reveals that the estimates of the factor risk premiums and industry factor loadings frequently exhibit wide confidence interval. These large standard errors created uncertainty about the ICoE estimates as it often exceeded more than 3% per year.

Traditional asset pricing models such as CAPM and FF3 have inherent limitations. The assumption of stable risk factor loadings and accurate measurements are not applicable in many industries due to risk factors are dynamic.

In recent years, machine learning (ML) techniques, which can handle high-dimension data and capture non-linear relationships between different factors, have rapidly developed to conduct the existing asset pricing problems.

Studies have shown that advanced machine learning methods, like Gradient Boosting Machines (GBM) and Light Gradient Boosting Machines (Light GBM), can greatly enhance the accuracy of predicting equity returns and risk factor loadings. This is especially true for industries that are undergoing rapid structural changes or showing non-linear risk dynamics. (Gu, Shihao, Bryan Kelly, and Dacheng Xiu. (2020))

Initial empirical results from this study also showed that simple machine learning methods, like Ordinary Least Squares (OLS) and LASSO regression, are underperformed in adjusted R^2 compared to traditional asset pricing models, such as the FF3 and the Fama-French Five-Factor Model (FF5). However the more advanced machine learning models, GBM and Light GBM, showed big improvements. This motivates us to further explore the machine learning factor selection techniques.

Driven by these insights, this thesis proposes a hybrid approach: employing machine learning factor selection methods (specifically, utilizing the 153-factor library proposed by Jensen, Kelly, and Pedersen, known as JKP factors) to enhance the explanatory power of traditional asset pricing models like FF5.

Specifically, this study investigates whether integrating ML-selected factors from the comprehensive JKP factor set into the FF5 model can significantly improve its explanatory and predictive performance for industry-level ICoE estimations. This hybrid approach seeks to leverage the interpretability and theoretical rigor of traditional factor models while incorporating the dynamic, adaptive capabilities offered by machine learning methods.

Overall, by systematically comparing traditional asset pricing models with this innovative hybrid model, the research aims to bridge existing theoretical gaps and offer a more accurate, robust, and comprehensive framework for estimating industry-level cost of equity. Such improvements could substantially enhance investment decision-making and valuation precision, particularly under market conditions characterized by volatility, nonlinearity, and structural economic shifts.

1.2 Research Questions

Considering the previously discussed limitations inherent in traditional asset pricing models and the encouraging potential observed from advanced machine learning (ML) methodologies, this study aims to investigate several critical research questions. Firstly, we examine whether sophisticated machine learning algorithms, particularly Gradient Boosting Machines (GBM) and Light Gradient Boosting Machines (Light GBM), can effectively overcome the identified shortcomings of conventional models such as CAPM, FF3, and FF5 in estimating the industry cost of equity (ICoE).

Secondly, this research seeks to explore if factors selected through ML methods from an extensive factor database—the JKP 153-factor library—can provide superior explanatory power regarding industry-level ICoE compared to the fixed and pre-specified factors included in the traditional Fama-French Five-Factor (FF5) model.

The third question pertains to whether using an ML-enhanced hybrid model can significantly improve predictive accuracy when estimating the industry-level ICoE. This model integrates ML-selected factors with the traditional FF5 framework. It is relevant in industries that are highly volatile or experiencing dynamic changes in market conditions.

Finally, the study examines whether using machine learning techniques for estimate dynamic beta estimation can greatly enhance the robustness and reliability of industry-level ICoE estimations. This is particularly crucial during times when the market is very uncertain and unstable. In summary, these interrelated research questions want to systematically assess how to combine advanced machine learning methods with traditional asset pricing frameworks. The objective is to develop with a superior, precise and more reliable way to estimate the industry-level cost of equity.

1.3 Research Contributions

This thesis aims to contribute to the existing literature in several specific ways. To the best of our research, there has been relatively limited research applying advanced machine learning (ML) techniques specifically to the estimation of industry-level cost of equity (ICoE). While existing studies have generally focused on firm-level return forecasting or relied heavily on traditional asset pricing models such as CAPM, FF3, or FF5, our study extends this scope by systematically exploring how ML methods perform in the broader

context of industry-level equity cost estimation. Through this investigation, we provide empirical insights that help clarify the robustness and applicability of ML techniques at the industry rather than the firm level.

We propose a hybrid modelling framework designed to integrate factors selected by machine learning methods into the traditional FF5 asset pricing framework. Specifically, we utilize the extensive JKP factor library, which comprises 153 factors categorized into 13 economic themes, to investigate whether ML-selected factors can provide incremental explanatory power relative to the original, fixed FF5 factors. By adopting this approach, we seek not only to enhance theoretical understandings of factor models but also to offer practical improvements that investors and financial analysts could potentially adopt in their valuation processes.

We conduct a detailed comparative analysis between traditional models (especially FF5) and selected ML methods, such as LASSO, Gradient Boosting Machines (GBM), and Light Gradient Boosting Machines (Light GBM). We evaluate the predictive accuracy and stability of these methodologies using a range of commonly accepted error metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Adjusted R^2. This structured comparison may illustrate the conditions and industry features. Under these conditions and features, ML methods might outperform traditional asset pricing models. To this extent, it may provide useful practical insights for both academic researchers and financial practitioners.

Finally, we attempt to bridge existing gaps between traditional asset pricing theories and modern machine learning methodologies with our empirical framework. We validated the potentiality of integrating ML-selected factors into traditional models, and provide a basic approach for future researchers and practitioners interested in improving industry-level equity cost estimation. Overall, our research seeks to provide a novel perspective about factor selection in asset pricing models. Also we seek to suggest methodological improvements that could enhance investment decision-making under complex and volatile market conditions.

2 Literature Review

2.1 Asset Pricing Models and Their Limitations

2.1.1 CAPM

Over half of the century, the estimation of expected returns has been central to financial economics. Despite various asset pricing models were developed, the Capital Asset Pricing Model (CAPM) proposed independently by Sharpe (1964) and Lintner (1965), remains foundational. CAPM is based on the seminal work of Markowitz (1952), who introduced modern portfolio theory. This model demonstrated how investors could optimize their investment portfolios by balancing risk and return. More specifically, CAPM positing the expected return on an asset is determined solely by its sensitivity to systematic market risk, commonly known as beta (β), which cannot be diversified through portfolio management.

Mathematically, CAPM can be expressed as:

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

where E(Ri) is the expected return on asset i, Rf denotes the risk-free rate of return, βi represents the sensitivity of asset i's returns to market movements, and E(Rm) is the expected return of the market portfolio. The main concept of CAPM claims that investors should only be compensated for bearing systematic, undiversifiable market risk. Whereas the idiosyncratic risk, which can be diversified by holding a broad portfolio, does not warrant additional compensation.

CAPM serves as a theoretical benchmark by pricing risky assets, guiding investment decisions, and determining the cost of capital. However, CAPM relies heavily on several restrictive assumptions: the markets are perfectly competitive and efficient, all investors have homogeneous expectations, no taxes or transaction costs exist and investors can borrow or lend freely at a risk-free rate. These assumptions have been extensively criticized for their unrealistic representation of actual market conditions. The model serves as an ideal outcome rather than empirical practice.

After the theoretical development of CAPM, the empirical evaluations have also started. Early tests conducted by Black, Jensen, and Scholes (1972) and Fama and MacBeth (1973)

provided initial support for CAPM’s linear relationship between beta and expected returns. However, subsequent evidence has highlighted the model's inability to fully capture return variations, especially in specific firm-level characteristics like size, valuation ratios, and profitability. These empirical anomalies indicate that systematic risk alone is insufficient to explain the observed differences in returns.

This raised serious questions about CAPM's robustness in diverse market contexts. Among these empirical challenges, the work done by Fama and French (1992) stands out noticeably. They systematically demonstrated that firm size (market capitalization) and book-to-market equity ratios can explain variations in average stock returns, which clearly contradicting CAPM’s predictions. More specifically, smaller firms and firms with higher book-to-market ratios consistently generated higher average returns compare to market betas alone. Such findings directly challenged CAPM's central assumption that only systematic market risk should matter in pricing assets, laying the groundwork for the subsequent development of multi-factor models.

The CAPM has provided an important foundation for asset pricing theory. Its limitations in empirical applications have spurred theoretical improvements. This occurs as market realised CAPM can't fully account for risk factors other than the overall market fluctuations. As a result, a more comprehensive asset pricing models are required, hence Fama and French developed the multi-factor models Fama-French 3-Factor Model.

2.1.2 Fama-French Three-Factor Model

To address the empirical challenges posed against the CAPM, Fama and French (1993) developed the concept of multi-factors framework known as the three-factor model (FF3). The FF3 was initially driven by consistent empirical failures of CAPM. It failed to account for systematic variations in stock returns associated with firm size and valuation metrics, specifically the book-to-market ratio. By introducing two additional factors, FF3 is capable to addressed these anomalies with a more comprehensive model for explaining observed differences in asset returns.

The FF3 model adds two more empirical factors based on CAPM. These factors are Small Minus Big (SMB) and High Minus Low (HML). The SMB factor captures the empirical observation that small-cap firms usually perform better than large-cap firms. The HML factor shows that firms with high book-to-market equity ratios consistently get higher

returns than those with low book-to-market ratios. The mathematical expression of the Fama-French three-factor model is formally given by:

$$E(R_i) = R_f + \beta_i[E(R_m) - R_f] + s_i SMB + h_i HML$$

In this equation, E(Ri) is the expected return on asset, Rf is the risk-free rate, E(Rm) denotes the expected market return, βi measures the sensitivity of asset i's returns to the market risk factor. The si and hi represent the sensitivities of asset is returns to the SMB and HML factors.

The empirical validation of the FF3 model has been generally supportive. Initial study conducted by Fama and French (1993) using U.S. stock market data convincingly demonstrated the validation. It validates that these two additional factors significantly improve the explanatory power of asset pricing models, and the unexplained variations in returns that CAPM could not address have reduced. Later on, numerous subsequent studies have validated these findings across different markets, time periods and asset classes. The empirical robustness and broad applicability of FF3 have been demonstrated.

Despite these empirical successes, the FF3 model is not without limitations. Notably, it does not adequately explain certain robust empirical phenomena. For example, the momentum effect first documented by Jegadeesh and Titman (1993). The momentum anomaly describes a persistent pattern whereby stocks with strong past returns tend to continue outperforming those with weaker past returns. Such phenomenon was failed to capture adequately by FF3. Additionally, empirical studies done by Novy-Marx in 2013 showed that profitability measures can significantly predict changes in returns. These predictions are independent of size and book-to-market factors. The FF3 model overlooks critical dimensions of return variation, result lacking explicit consideration for profitability and investment characteristics.

Another significant limitation lies within the theoretical foundations of the SMB and HML factors themselves. While these factors were identified empirically, the economic rationale underlying their relationship with systematic risk remains controversial. Critics argue that these factors might capture market mispricing or behavioural biases rather than genuine systematic risks. For example, Lakonishok, Shleifer, and Vishny (1994) argue that the book-to-market effect could be a result from investor overreaction rather than

compensation for higher systematic risk. This challenges the risk-based interpretation that's basic to FF3.

Furthermore, people also questioned the stability of the FF3 factor loadings. Empirical studies indicate that the sensitivities of SMB and HML factor vary considerably across different market environments and economic cycles. This suggests that factor loadings are not stable parameters. Instead, they fluctuate significantly over time (Lewellen, Nagel & Shanken, 2010). This instability weakens the model's predictive reliability, especially in volatile market periods or across differing industry contexts.

The introduction of the Fama-French three-factor model marked a critical advancement in asset pricing theory. It effectively addressing key limitations identified in CAPM by incorporating size and value factors. However, several important anomalies and the controversial economic interpretation of its factors highlight substantial theoretical and empirical limitations. Existing issues have motivated researchers and led to the expansion of asset pricing models with more factors. This leads to the development of the Fama-French five-factor model.

2.1.3 Fama-French Five-Factor Model

Fama and French (2015) were motivated by the empirical limitations observed in the three-factor model (FF3). Thus, they proposed an expanded five-factor model (FF5). The primary goal was to improve upon FF3 by incorporate additional dimensions of risk and return that earlier models did not adequately address. To overcome the FF3 limitation, Fama and French introduced profitability (Robust Minus Weak, RMW) and investment (Conservative Minus Aggressive, CMA) as two new empirical factors. This improvement enhancing the explanatory power of the asset pricing model beyond what SMB, HML, and the market factor could achieve.

The introduction of profitability and investment factors was primarily driven by empirical evidence highlighting that firms with high profitability and conservative investment strategies systematically generated higher returns. Novy-Marx (2013) notably provided compelling evidence that profitability, measured by gross profitability scaled by assets, independently explains variation in returns beyond traditional size and value factors. Additionally, investment behaviour—represented by asset growth rates—has also been found to significantly predict returns, as shown in the work of Titman, Wei, and Xie (2004).

Incorporating these two factors, therefore, offered the potential to address gaps left unresolved by the three-factor model.

Formally, the Fama-French five-factor model can be mathematically expressed as follows:

$$E(R_i) = R_f + \beta_i[E(R_m) - R_f] + s_i SMB + h_i HML + r_i RMW + c_i CMA$$

where E(Ri) denotes the expected return on asset i, Rf is the risk-free rate, βi, si, and hi represent asset i's sensitivity to the market, size, and value factors respectively, and ri, ci denote its sensitivity to the profitability and investment factors respectively.

Fama and French (2015) conducted empirical tests on U.S. market data. They demonstrated significant improvements in explanatory power when using the FF5 model. Results showed that profitability and investment factors substantially reduced the unexplained variations in returns compared to the FF3 model. Subsequent studies from later have also validated these findings across international markets, which further establishing the FF5 model's enhanced explanatory capabilities.

Nevertheless, the FF5 model still encounters meaningful limitations. Yet, despite its expanded scope, the FF5 model still lacks a momentum factor, leaving this well-documented anomaly largely unexplained.

Although the addition of profitability and investment factors are empirically robust. It has also introduced complexities regarding their theoretical interpretation. Critics argue that the FF5 model remains empirically motivated rather than theoretically grounded. Limited consensus on whether these additional factors represent genuine systematic risks, or they are merely behavioural anomalies from the investors and the inefficient market. For instance, Hou, Xue, and Zhang (2015) suggesting that profitability and investment patterns should be interpreted through the lens of investment theory rather than as arbitrary empirical factors.

Recent studies have shown that, just like earlier factor models, the factor loadings in the FF5 model are quite unstable over time and across different economic cycles. Lewellen, Nagel, and Shanken (2010) point out this limitation. They stated that risk loadings are not a stable parameter. They are significantly affected by macroeconomic changes, investor preferences and the changing market conditions. Such instability brings challenges to the

predictive reliability and practical application of the FF5 model. This is especially true during the periods when the market is volatile.

The empirical complexity of the FF5 model also brings practical difficulties for implementation. Especially in accurately estimate factor loadings and risk premiums calculation. Different from simpler models, it requires precise estimations of multiple factor sensitivities. Each factor is potentially unstable and can have substantial estimation errors.

This complexity might restrict its practical use among market practitioners. Especially in industry-level equity cost estimation where it requires adequate balance between precision and simplicity.

Regardless of its limitations, the FF5 model significantly improved compared to previous asset pricing models. It effectively addressed several empirical anomalies that earlier frameworks could not. However, momentum phenomenon remains unsolved. The need of further improvement is required in asset pricing theory.

Over the past few decades, asset pricing theory has evolved. Both new theoretical ideas and problems from empirical data have accelerated this change. Initially, the simplest model Capital Asset Pricing Model (CAPM) was developed. It connected the expected returns directly to the systematic market risk, which is called beta. Then to address the changes in returns related to a firm's size, valuation ratios and profitability. This led to the creation of multi-factor models, FF3. The Fama-French three-factor (FF3) model marked a crucial advancement by adding the size (SMB) and value (HML) factors. This directly dealt with a lot of things that the CAPM couldn't explain. To further address the empirical anomalies, Fama and French (2015) introduced the five-factor model (FF5) which clearly combined the profitability (RMW) and investment (CMA) factors into the FF3 model. Empirical evidence supports the explanatory power of FF5, yet the limitation of momentum remains unsolved.

Overall, the change from CAPM to FF5 represent significant theoretical and empirical progress. It clearly stated the persistent challenges faced by traditional asset pricing models. The ongoing argument over the economic interpretation to reliably capture dynamic risk characteristics in volatile industries, has highlighted the necessity for alternative modelling approaches.

These limitations demonstrate the demand to investigate more flexible, adaptable, and data-driven methods. The machine learning models, could potentially address the shortcoming within traditional linear factor models by addressing the non-linear factors.

Therefore, by recognising and acknowledging the empirical and theoretical limitations of existing traditional model. It motivates researchers to explore novel approaches.

2.2 Estimation of Industry Cost of Equity

2.2.1 Traditional Models in Industry-Level Applications

Accurately estimating the cost of equity at the industry level is crucial for corporate finance activities. These activities encompass capital budgeting, mergers and acquisitions, and performance assessment. Historically, the primary methods for estimating industry-level equity costs have predominantly relied on conventional asset pricing models. The models comprise the Capital Asset Pricing Model (CAPM), the Fama-French three-factor model (FF3), and the five-factor model (FF5).

Despite extensive testing of these models across various market conditions, their application in industry-level contexts presents greater methodological challenges and practical issues. A multitude of empirical studies have highlighted this observation.

The Capital Asset Pricing Model (CAPM) remains the predominant method for estimating industry-level equity costs. This is due to its simplicity and ease of implementation. To estimate the industry beta, individuals typically aggregate the betas of individual firms within a particular industry. They frequently employ either value-weighted or equal-weighted averages for this purpose, as noted by Damodaran in 2001. This method of aggregating betas presents certain issues. The selection of appropriate market benchmarks and estimation intervals significantly influences industry beta estimates. Fama and French (1997) extensively discussed the volatility and temporal fluctuations of beta values. This complicates the reliable estimation of them. Scholars like Damodaran (2001) have indicated that historical betas may not accurately reflect present risk conditions. This is due to structural alterations within industries or the perpetually evolving market environment. Applying CAPM at the industry level appears straightforward; however, it frequently incurs significant estimation errors and biases.

The CAPM exhibits certain limitations regarding real-world data. Consequently, numerous industry-level studies have commenced utilizing the FF3 model. This model considers firm size (SMB), book-to-market equity (HML) factors, and market risk. Fama and

French presented this concept in 1997. When estimating the FF3 model at the industry level, we employ a methodology akin to that used for other models. We calculate the average factor loadings of the firms in an industry. Typically, we employ either value-weighted or equal-weighted methodologies to accomplish this. Nonetheless, implementing FF3 in the industry presents challenges. Research done by Lewellen, Nagel, and Shanken in 2010 shows that the size and value factor loadings are quite unstable across economic cycles and market conditions. This prompts individuals to scrutinize the stability and reliability of these loadings when assessing the cost of equity in the real-world industry. This instability shows that even though the FF3 model has made several practical improvements compared to the CAPM, it still has several serious practical limitations. After that, the FF5 model built on the FF3 framework. It clearly added profitability (RMW) and investment (CMA) factors. The goal was to enhance model's explanatory power regards to industry returns compared to previous models (Fama & French, 2015). Empirical studies, the ones by Hou, Xue, and Zhang (2015), have indicated that the FF5 model outperform in capturing differences in returns among different industries. This is especially true for industries with different profitability and investment characteristics.

However, when FF5 model is implemented at the industry level, it makes the estimation process even more complicated. Adding multiple factors greatly increases the complexity of the estimation. This also brings in more uncertainty and possible biases. For example, recent evidence from Fama and French (2017) and other scholars indicates that factor sensitivities in the FF5 model change a lot with macroeconomic conditions and market cycles. This makes the model less reliable and precise in practical industry-level applications.

2.2.2 Performance and Estimation Errors of Traditional Models

Empirical research that evaluates traditional asset pricing models at the industry level always indicates significant challenges in estimation. These challenges are especially related to predictive accuracy and reliability. Fama and French (1997) did some studies. In their studies, they stated that the major difficulties in estimating industry betas reliably with historical data. They also noticed that historical betas are often very different from the future betas that happen.

This instability gets particularly obvious during times when the market is volatile or there are major disruptions in the industry. It directly weakens the ability of the CAPM model to accurately predict the equity costs at the industry level.

The Fama-French three-factor (FF3) model was developed to deal with the limitations of CAPM. But it also has significant problems when it comes to making predictions at the industry level. Empirical analyses indicate that when we apply the FF3 model to industries with rapid structural changes, technological innovation, or regulatory uncertainty, there are always errors in estimation and systematic biases.

These industries generally see big changes in factor sensitivities as time goes on. This causes a lot of instability in factor loadings. So, even though the FF3 model is better than the CAPM model in terms of performance, it's still not very reliable in industries that change a lot.

Similarly, the Fama-French five-factor (FF5) model has made theoretical improvements by including profitability and investment factors. However, it still faces significant practical challenges when it comes to reliably estimating industry-level equity costs. Fama and French's empirical evidence from 2015 and 2017 indicates that factor sensitivities to profitability and investment vary a lot under different economic conditions and market cycles.

This instability of factors greatly raises the complexity of estimation and lowers the model's predictive accuracy. This is especially true in industries that experience rapid technological changes or unstable profitability cycles. Therefore, despite the FF5 model significant enhanced the explanatory power, it remains complex to implement in practice and prone to estimation errors.

The traditional asset pricing models consistently have critical limitations. This include unstable factor sensitivities and systematic prediction errors. These challenges reinforce the need of alternative methodologies can address the industry-specific risk dynamics and market shifts to an adequate extent.

2.2.3 Implied Cost of Capital Approaches and Their Limitations

Regards to the limitations in traditional asset pricing models. Implied cost of capital (ICC) methods have gained popularity as alternative ways to estimate industry-level equity costs. ICC methods figure out the expected return on equity based on current stock prices and analysts' earnings forecasts. Such prospective aspect, potentially resolve the backward-looking biases and instability issues with traditional methods.

The theoretical appeal of ICC mainly originate from its basis in discounted cash flow (DCF) principles. Specifically, ICC models usually use analysts' common forecasts for future earnings, dividends, or growth rates. The implied returns is obtained from equating the current market prices to the discounted expected future cash flows.

For example, Claus and Thomas (2001) proposed an influential ICC method. This method directly uses analysts' consensus earnings forecasts and dividend payout assumptions to estimate implied equity returns. Similarly, Easton (2004) introduced other ICC approaches which focus on analysts' forecasts of near-term and long-term earnings. This further strengthen the theoretical credibility of ICC.

Empirical studies generally support the practical advantages of ICC methods. Claus and Thomas (2001) gave empirical proof. It shows that ICC estimates often exhibit predictive accuracy when estimating realized returns compared to traditional CAPM, especially during market uncertainty or industry changes. These findings suggest that the ICC can effectively take in the current market expectations. Thereby to capturing forward-looking risk and return more reliably.

However, the methods of the ICC also face limitations. A fundamental challenge lies on the over-rely on analyst’s forecasts. It's well-known that these forecasts often have a positive bias. For example, Easton and Sommers (2007) found that analysts often overestimate future earnings. This is especially true when the market is going up. As a result, there is a large upward bias in the ICC estimates.

So, the inherent bias in analysts' forecasts could make ICC methods overestimate the industry-level equity costs in a systematic way. This reduces their practical reliability. This is especially true in sectors where there's high uncertainty in forecasts or limited analyst coverage.

Additionally, the sensitivity of ICC methods to key underlying assumptions represent a major drawback. Small changes in the assumed long-term growth rates, dividend payout ratios, or terminal earnings forecasts can have a significant impact on ICC estimates. This reduces their reliability and consistency. Empirical studies have always shown this high level of sensitivity. They also point out the practical challenges in accurately setting the ICC model parameters and assumptions in dynamic market conditions.

ICC methods provide valuable forward-looking ideas. They can effectively complement traditional asset pricing methods. But in practice, their reliability is constrained by analyst

forecast biases and extreme sensitivity. Analyst forecast biases and extreme sensitivity to model assumptions are the main reasons. Recognizing these limitations in order to keep improving the methods and explore alternative modelling techniques to solve these practical problems.

2.3 Machine Learning in Asset Pricing

2.3.1 Theoretical Foundations and Emergence

Traditional asset pricing models like CAPM, FF3, and FF5 have faced long-standing practical problems and limitations in empirical applications. In response to these issues, recent finance researchers has started to rely more on machine learning (ML) methods.

Machine learning is a part of artificial intelligence techniques with popular asset pricing and equity valuation. This is because ML can capture complex nonlinear relationships effectively. It can also handle large-scale and high-dimensional data. Moreover, it can adapt to changing market environments dynamically without relying on restrictive linear assumptions. (Gu, Kelly, & Xiu, 2020)

Machine learning algorithms are computational methods that mainly based from past data. They can find predictive patterns without humans having to clearly define the model structures. Traditional econometric or factor-based models are different. ML methods don't rely on pre-defined relationships between variables. Instead, the relationship between factors are discovered during the process of machine learning.

This data-driven characteristic helps to reduce biases. These biases are caused by rigid theoretical assumptions. We often see these assumptions in traditional asset pricing frameworks (Gu, Kelly, & Xiu, 2020).

In asset pricing, certain machine-learning techniques have demonstrate its capability. These include regularized regression methods like LASSO and Ridge regression, tree-based algorithms such as Gradient Boosting Machines (GBM) and Random Forests, and neural network approaches. These methods outperformed the traditional models. First, they give a lot of flexibility when it comes to modelling complex interactions between risk factors. This makes it possible to predict expected returns more accurately.

Second, ML models can handle large and high-dimensional datasets on their own. This allows them to analyse many potential explanatory variables at the same time without losing

much predictive accuracy. Third, as new information comes up, ML models can change their structures and parameters. This is very useful in markets that change quickly or in sectors that often have structural changes (Gu, Kelly, & Xiu, 2020).

Furthermore, machine learning methods deal with the problem of "factor proliferation." This is a very important concern that has been brought up in recent empirical finance literature. Harvey, Liu, and Zhu (2016) pointed out the challenge. They said that if you test a lot of factors, it might lead to false relationships. This could be because of data mining or statistical overfitting.

ML methods can solve this problem. They do it by systematically selecting and weighting factors based on data. This way, the selected factors can truly indicate real economic risks, not just statistical things that don't really matter. This ability makes the resulting asset pricing models more understandable from an economic point of view and more reliable. (Harvey, Liu, & Zhu, 2016)

Machine learning approaches have become really important methodological improvements. They're much better than traditional factor-based asset pricing models. They can capture complex dynamics. They also reduce biases that come from restrictive assumptions. And they can systematically deal with factor proliferation. ML methods have a lot of potential to make asset pricing and industry-level equity cost estimations more robust and accurate.

2.3.2 Empirical Applications and Advantages

Recent empirical asset pricing research widely utilize machine learning methodologies. It achieved great improvements in predictive accuracy, factor selection, and portfolio management. These improvements are better than those of traditional econometric and factor-based approaches.

Empirical studies that was done by Gu, Kelly, and Xiu in 2020, have indicated significant explanatory power in predicting stock returns. They specifically pointed out that these algorithms can deal with complex and nonlinear relationships. Whereas traditional linear models are inadequate to explain (Gu, Kelly, & Xiu, 2020)

Gu, Kelly, and Xiu (2020) clearly compared different machine-learning methods with conventional econometric models. The machine-learning methods included regularized regression (such as LASSO and Ridge regression), tree-based algorithms (like Gradient

Boosting Machines and Random Forests), and deep neural networks. The conventional econometric models were things like CAPM and Fama-French factor models. The results consistently indicated that machine learning methods achieved superior out-of-sample predictive performance. This was valid diverse range of stocks and different market environments. Among the techniques evaluated, neural networks and tree-based models had strong predictive accuracy. This indicates that ML are inherently flexible and can capture complex nonlinear interactions between variables.

Moreover, machine learning techniques have clear benefits in factor selection. They can address factor proliferation and statistical overfitting which existed in traditional asset pricing research. Harvey, Liu, and Zhu (2016) assert that in traditional empirical finance examine numerous potential explanatory variables. This may lead to misleading outcome. The outcome often explains statistical artifacts rather than economic relationships.

Machine learning addresses these issues with data-driven methods. These methods can systematically identify and select economically important factors directly from large datasets. This can greatly mitigate the risk of overfitting, which also improves the robustness and economic interpretability of empirical asset pricing models (Harvey, Liu, & Zhu, 2016).

Machine learning models have proven highly useful in portfolio optimization and risk management. They can effectively predict asset returns and adapt to the changing market conditions. ML-driven portfolio strategies have outperform in terms of risk-adjusted performance than traditional strategies. For example, machine learning methods can adjust asset allocations immediately when new market information comes out. This can greatly improve portfolio efficiency and risk-adjusted returns. It is particularly effective well in volatile market conditions (Gu, Kelly, & Xiu, 2020).

Moreover, empirical research that uses ML models in asset pricing have demonstrated their robustness at capturing non-stationarities and structural breaks in financial data. Traditional factor-based models presume that factor loadings and relationships are stable. But empirical evidence has always challenged this assumption. Machine learning approaches can address structural instability by implement adaptive learning processes. They can dynamically update parameters to indicate the changing market situation. This ability has great advantages in asset pricing applications. It's especially useful during economic disruptions, market regime shifts, or industry-specific transformations (Gu, Kelly, & Xiu, 2020).

The nature of Machine learning methods are flexible, data-driven, and can adapt dynamically. These attributes enable the model highly appropriate for solving long-standing problems in asset pricing research. The extensive application of ML techniques represents an advancement in both academic research and empirical applications.

2.3.3 Empirical Industry-Level Evidence and Comparative Methodological Analysis

Recently, numerous empirical research have aware on using machine learning (ML) techniques in industry-level asset pricing. It has indicated that ML techniques have absolute advantages to combat modern market over traditional econometric methods. These advantages include capturing the risk dynamics specific to an industry and predicting returns.

Gu, Kelly, and Xiu (2020) provided evidence that machine learning algorithms like neural networks, Gradient Boosting Machines (GBM), and regularized regression methods have superior predictive accuracy at the industry level than the traditional model. Numerous empirical tests were conducted in different industries and ML models are well-performed at identifying and adaptively modelling the nonlinearities in an industry. The ML model are capable to handle complex interactions between factors and manage the structural breaks that frequently occur in industries as they develop.

Empirical research that directly compare machine learning (ML) and traditional factor models clearly demonstrate that ML has a unique methodological advantage. This advantage lies in its ability to handle high-dimensional datasets and multi-potential explanatory variables.

For example, machine learning algorithms can systematically identify industry-specific factors that are economically important from large datasets. This can greatly lower the risk of false correlations or statistical overfitting. These problems often happen in traditional ways of factors selection (Harvey, Liu, & Zhu, 2016). This process of selecting factors based on data can directly improve the economic relevance, stability, and understandability of predicting industry-level returns.

Moreover, the performance of machine learning models in predicting industry-level returns has been widely tested in different market environments. Specifically, these environments include sectors with high volatility, fast technological innovation, or significant regulatory changes.

Empirical evidence indicates that machine learning (ML) models can adapt to the structural transformation within industries. They consistently outperform traditional econometric models. Traditional econometric models often have a hard time dealing with sudden changes or non-linear situations effectively. (Gu, Kelly, & Xiu, 2020)

ML models have inherent flexibility and adaptability. They can promptly update factor sensitivities and model parameters when they get new industry-specific information. This ability is evidently superior than static or gradually adapting traditional factor models. So, ML methods are more suitable for capturing dynamic and rapidly evolving industry environments.

In addition, when we compare machine learning approaches with traditional valuation methods like the implied cost of capital (ICC) through methodological analyses, ML still outperform in predictive capability. ICC methods provide forward-looking view of valuation according to analysts' forecasts. However, their practical reliability is still restricted by forecast biases and the sensitivities of model assumptions.

Machine learning approaches can deal with these limitations. They use systematic, data-driven forecasting processes. These processes combine a huge amount of historical and real-time data, which reduce the dependence on analyst forecasts and subjective assumptions. This strength of the method can lead to much better accuracy and reliability when estimating the equity cost at the industry level. This is especially true for sectors where there is limited analyst coverage or high forecast uncertainty.

Empirical industry-level evidence strongly supports that machine learning methodologies are superior in asset pricing. They have strengths in adaptive learning, modeling nonlinear relationships, and selecting factors based on data. Comparative analyses always indicate that ML methods are better than traditional econometric and ICC approaches. This is especially true in sectors with volatility, structural shifts, or high uncertainty.

ML techniques have been consistently proven effective through empirical studies in different industries. This indicates that these techniques can be widely used and are methodologically reliable. They can improve the predictive accuracy and reliability of estimating the cost of equity at the industry level.

The literature has indicated in a systematic way that machine learning methods are a big step forward in the field of asset pricing and estimating the equity cost at the industry level. Empirical evidence indicates that machine learning algorithms are better than traditional

methods. For example, Gu, Kelly, and Xiu did comprehensive studies in 2020. These studies clearly indicate that machine learning algorithms, like neural networks, regularized regression techniques, and tree-based models, always have better predictive accuracy, adaptability, and robustness. Traditional econometric and implied cost-of-capital methods are not as good as these machine learning algorithms.

Moreover, machine learning algorithms can naturally handle complex nonlinear interactions and dynamic changes in market conditions. These advantages greatly improve their applicability and predictive reliability in different industry situations.

In general, the literature present compelling evidence that indicates machine learning methods are better for estimating industry-level equity costs. These methods introduce innovative approaches and are applicable in various contexts. This represent a significant advancement in both theory and practice for asset pricing research. It also indicates a clear way to solve the problems that traditional valuation methods have.

2.4 Beta Estimation and its Impacts

2.4.1 Traditional Beta Estimation and Limitations

Beta estimation is a fundamental component of asset pricing theory. It facilitates the calculation of expected returns in traditional models. These models include the Capital Asset Pricing Model (CAPM) and Fama-French factor models.

At the core of CAPM, beta quantifies the sensitivity of an asset's returns to the movements of the broader market portfolio. This is a crucial measure of systematic risk. And it determines the compensation an asset can expect for taking on that risk. (Fama & French, 1992)

In traditional empirical finance, individual generally use historical stock return data to estimate beta through regression analysis. It compares the returns of individual assets with the returns of the market portfolio over a fixed historical period. The regression coefficient obtained from this analysis is refer to as beta. And they see this beta as an indicator of the asset's systematic risk. But for extended period, researchers have been questioning the reliability of these beta estimates.

Fama and French (1992) pointed out that historical betas estimated by traditional methods often indicate instability and big differences in different time periods. This

diminishes their predictive accuracy and practical application. This instability is particularly obvious during periods of economic uncertainty or market disruptions. And these are exactly the situations when accurate risk estimation is desperately required (Fama & French, 1992). In other words, lacking of predictive accuracy is equivalent to incapability to the downside risk of the market.

Additionally, traditional methods for estimating beta possess limitation on accurately capturing how risk changes over time. The traditional CAPM and factor models generally presume that betas are either static or change very slowly. They overlook potential alterations in the market regime, variations in the macroeconomy, and structural transformations within industries. For example, Fama and French (1997) presented empirical evidence, indicated that beta coefficients often can't stay stable during economic cycles. This is because they have different sensitivities to market factors when economic and financial conditions change. This lack of stability over time really limits how accurate and reliable beta-based models are. These models are used to predict industry-level returns and estimate the cost of equity.

These issues are made even worse by the difficulty of determining appropriate estimation intervals and market benchmarks. Traditional regression methods need us to carefully pick historical timeframes and reference benchmarks. The selection of these parameters influences beta estimation.

In practice, biases occur when the benchmarks or estimation windows are inappropriate. This will lead to beta estimation always inaccurate and unable to reveal the real systematic risk exposures. (Fama & French, 1997)

The issues associated with conventional methods of estimating beta are particularly detrimental in rapidly evolving industries. These industries are marked by quick innovation, often have new regulations, or just experienced structural changes. Under these situations, using static or historical betas would be insufficient to identify the systematic risk, and ultimately lead to incorrect decision-making. Thus, even though traditional beta estimation approaches are widely used and theoretically appealing, researchers want to explore more adaptive, accurate, and robust estimation methods.

2.4.2 Machine Learning Approaches to Beta Estimation

Traditional beta estimation methods have some limitations. So, in recent research, people have been exploring the use of machine learning (ML) techniques more and more. They do this to estimate and predict systematic risk exposures more accurately.

Machine learning methods, especially the ones pointed out by Gu, Kelly, and Xiu (2020), are very promising for solving important problems in beta estimation. These problems include instability, nonlinearities, and structural breaks. Machine learning can adaptively capture the dynamic relationships between stock returns and market factors (Gu, Kelly, & Xiu, 2020).

Machine learning methods, like neural networks, tree-based methods (for example, Gradient Boosting Machines), and regularized regressions (such as LASSO), have indicaten big advantages in accurately estimating market betas. Traditional regression-based methods are different. ML models can naturally handle time-varying and nonlinear factor sensitivities. They can dynamically update beta estimates when market conditions and economic environments change.

Gu, Kelly, and Xiu (2020) did an empirical analysis. This analysis clearly indicates that ML models perform better. They can capture complex and adaptable patterns in market exposures. In different stocks and market environments, these ML models always do better than traditional regression-based beta estimates. They are more accurate and stable when it comes to making predictions.

An important advantage of ML models in beta estimation is that they can handle large-scale, high-dimensional datasets. They don't have the common problems of overfitting and model instability. Traditional beta estimation methods often have a hard time balancing model complexity and estimation accuracy. This is especially true when they try to include multiple explanatory factors or deal with nonlinearity.

In contrast, machine learning techniques pick and weigh factors from the data in a systematic and adaptive way. This makes sure that beta estimates are robust. It also cuts down the risks of statistical biases and estimation errors that are common in traditional methods. (Harvey, Liu, & Zhu, 2016)

Also, ML models can adapt empirically to structural changes and industry-specific risk patterns. This greatly improves their usefulness and reliability when estimating beta. In industries where there are fast technological advancements, regulatory changes, or economic disruptions, the systematic risk exposures often change suddenly and frequently. Traditional static beta estimations can't handle these situations well.

Machine learning methods have the ability to learn from data and update parameters on their own. This helps them estimate beta coefficients more accurately and quickly. This is

especially true when the market is very volatile or when there are big changes in the market structure. (Gu, Kelly, & Xiu, 2020)

Moreover, practical tests in different industries always indicate that methods of estimating beta based on machine learning can predict more accurately and have fewer estimation errors than traditional ways of estimating beta based on regression.

For example, Gu, Kelly, and Xiu (2020) indicateed that when estimating systematic risk exposures, neural networks and tree-based algorithms can achieve much better out-of-sample predictive performance. This indicates that machine learning has practical and methodological advantages in understanding how market risk dynamics change over time.

To sum it up, using machine learning methods for beta estimation is a big step forward in methodology. It can effectively solve the long-standing practical problems of traditional estimation methods. ML models are adaptive and data driven. They can greatly improve the predictive accuracy, robustness, and practical reliability of beta estimates. This is especially true in industry environments that are changing rapidly and have dynamic structures.

This review of beta estimation methods has pointed out some important problems with traditional regression-based methods. First, historical betas are very unstable. Second, they are sensitive to the chosen estimation intervals. Third, they can't adjust well to structural breaks and nonlinear changes in market risk exposures.

Empirical evidence, especially the one from Fama and French (1992, 1997), highlights these challenges. It indicates that traditional beta estimations often can't offer stable and reliable measures of systematic risk. This is especially true when there is economic volatility or structural transformation.

In response, machine learning methods have come up as really promising alternatives. They directly deal with these practical challenges. Recent studies by Gu, Kelly, and Xiu (2020) indicate that machine learning techniques. These include neural networks, gradient boosting algorithms, and regularized regressions. They perform much better than traditional methods. They can estimate beta accurately and capture dynamic risk exposures.

Machine learning models use an adaptive, data-driven approach. This approach has obvious advantages. It can make more accurate predictions. It's also more robust. And it can respond well to market changes.

Empirical validations across various market environments and industries consistently demonstrate that machine learning has strong methodological strengths and practical effectiveness in beta estimation. Beta estimates derived from Machine learning can greatly improve risk assessment, the accuracy of asset pricing, and the equity valuation at the industry level. This is especially true in sectors that are evolving rapidly or during periods of market stress and uncertainty. This empirical robustness indicates that machine learning methods have great potential. They are important methodological improvements.

Overall, the literature supports the idea of combining machine learning methods with beta estimation practices. It emphasizes that these machine-learning methods are overcoming the limitations of traditional approaches. This kind of innovation in methods is very important. It helps to make the prediction of systematic risk measurement more reliable, more robust, and more useful in modern asset pricing frameworks.

2.5 Research Gaps and Motivation

2.5.1 Research Gaps

The literature reviewed so far has pointed out several important problems in traditional asset pricing methods and the ways to estimate the cost of equity at the industry level. This indicates improvements to these methods is a requestable demand from the market. Specifically, traditional factor-based models, like CAPM (Fama & French, 1992), FF3 (Fama & French, 1997), and FF5 (Fama & French, 2015), have always indicate significant issues in real-world tests. One obvious problem is that they can't catch dynamic market conditions. They also can't deal with structural breaks, nonlinearities, and factor instability.

First, traditional asset pricing models have obvious flaws when dealing with time-varying factor sensitivities and nonlinear market relationships. Fama and French (1992, 1997) have extensively mention about this. These models generally assume that there are stable and linear relationships between risk factors and asset returns. But empirical evidence clearly challenges these assumptions.

So, the static factor loadings and fixed model structures that are common in these methods really limit their predictive accuracy and robustness. This is especially true in markets that are volatile or changing in structure.

Second, the existing ways valuation suchas implied cost of capital (ICC) methods suggested by Claus and Thomas in 2001 and Easton in 2004, also have several obvious

problems. The ICC methods can give forward-looking estimates based on what analysts predict about earnings. But in practice, their reliability is subject to the analyst's subjective opinion. This is because analysts often make biased forecasts, and the methods are extremely sensitive to the assumptions of the underlying models, as Easton and Sommers pointed out in 2007.

These empirical problems really limit how well ICC methods work when it comes to accurately estimating the cost of equity at the industry level. This is obvious for industries where it's hard to make forecasts or where there aren't many analysts looking at them.

Third, even though machine learning methods have brought about several approaches in the recent period, like what Gu, Kelly, and Xiu (2020) indicated in their empirical research, there are still big gaps in research when it comes to using these methods in a systematic way to estimate the cost of equity at the industry level.

Specifically, the existing literature has mostly used machine learning techniques in various markets. It hasn't clearly evaluated how effective and reliable these techniques are to systematically estimating the cost of equity in each industry. As a result, the generalisation of machine learning method remains at the first stage. As insufficient literatures about how well machine learning works and how reliable its methods are when estimating the cost of equity at the industry level.

Furthermore, there is another important problem. That is, there isn't enough comprehensive empirical validation of machine-learning methods for dynamically predicting beta. The predictions haven't been systematically integrated into asset pricing models. Gu, Kelly, and Xiu (2020) indicated several appropriate results for ML-based beta estimation. But there is still limited empirical evidence that explicitly integrates ML-based dynamic beta predictions into asset pricing and valuation models.

Therefore, it's worth exploring how machine learning can be used to predict beta. And it's also important to investigate how this prediction affects the estimations of the equity cost at the industry level. This is a clear and valuable area that needs further investigation.

In conclusion, the identified research gaps are significant. Ttraditional asset pricing models have several problems in relation to empirical evidence. Secondly, concerns exist regarding the methodologies employed in ICC approaches. Third, we have not thoroughly examined the systematically use machine learning (ML) methods at the industry level. Furthermore, insufficient empirical proof for the beta predictions made by ML methods.

These gaps signify the necessity for the development of novel methodologies and the execution of more comprehensive research.

2.5.2 Motivation and Contribution of This Study

Based on the research gaps found in the literature review, this study systematically explores how effective machine learning methods are for estimating the equity cost at the industry level. Specifically, the study deals with the limitations in traditional asset pricing models and implied cost of capital (ICC) methods. Traditional frameworks typically failed to accurately capture the dynamic risks in the market and the nonlinear relationships.

Moreover, these methods struggle with structural changes and instability in factor sensitivities. Machine learning methods have distinctive advantages in method when it comes to solving these challenges. This research posits that using machine learning techniques can significantly improve the accuracy, robustness, and economic interpretability of industry-level cost-of-equity estimates.

This research identifies the shortcomings with traditional factor-based models in relation to empirical data. These issues are particularly obvious in industries where technology is changing fast and the market is very unstable.

The study aims to effectively capture complex nonlinear dynamics and evolving market risk exposures that traditional linear models overlook. It employs machine-learning algorithms like Gradient Boosting Machines, neural networks, and regularized regressions in industry contexts. This novel approach significantly aids asset pricing theory and enhances the predictions of industry equity cost estimations more reliable.

The study examines the potential of machine learning methods in estimating the implied cost of capital (ICC). Even though ICC methods can offer forward-looking valuation ideas, their reliability is significantly affected by the biases in analyst forecasts and the sensitivities to basic assumptions (Easton & Sommers, 2007). To address these issues, this research combines machine learning techniques into the ICC estimation process. This way, it reduces the dependence on analyst forecasts.

This improvement in the method makes the ICC-based equity cost estimations more robust, more practical, and more accurate in prediction.

This research conducts empirical analyses across different industries. It demonstrates that machine learning methods are reliable and adaptable when estimating the equity cost at

the industry level. The empirical results give useful ideas to financial practitioners and policymakers. These results help them better understand the risk features specific to each industry. Furthermore, they can leverage this knowledge to enhance decision-making efficacy.

This study significantly contributes at the theoretical, methodological, and practical levels. It advances the research on asset pricing and its applications in the real world.

This study is mainly motivated by the limitations of traditional asset pricing frameworks and ICC methodologies. There are also gaps in the systematic application of machine learning approaches in industry contexts. This research applies machine learning methods systematically and validates them empirically. It significantly improves the theoretical rigor, empirical robustness, and practical applicability of industry-level equity cost estimations.

3. Methodology

3.1 Data Sources and Description

This study mainly uses data from well-known sources in the empirical asset pricing literature. These sources are the Fama-French Data Library and the Jensen-Kelly-Pedersen (JKP) factor database. These data sources offer a wide range of information. They cover risk factors, industry returns, and firm-level characteristics. These elements are very important for asset pricing analysis.

The Fama-French Data Library is maintained by Kenneth French. It's one of the most well-known datasets. People use it for empirical tests of factor-based asset pricing models. The library provides detailed historical data. This data is for various well-known factors. These factors include the market risk premium (MKT), the size factor (SMB), the value factor (HML), the profitability factor (RMW), and the investment factor (CMA).

The Fama-French factors are available on a monthly basis and cover a long period. This makes them suitable for strong time-series and cross-sectional analyses. Specifically, this research uses monthly factor returns. These returns cover the period from July 1969 to December 2023. This is in line with the common empirical practices in the asset pricing literature.

In this research, the Jensen-Kelly-Pedersen (JKP) factor database is directly taken from the dataset described by Jensen, Kelly, and Pedersen in their 2023 study. Their study focused

on factor replicability and robustness across global markets, which is called "Replication Crisis in Finance". This database has 153 factors that are systematically categorized. These factors cover 13 different thematic groups.

These factors are carefully recorded and built according to strict academic standards. This ensures that they are strong and reliable for empirical research. The JKP factor set has a lot of different anomalies and features that are often studied in asset pricing research. For example, it includes profitability, investment behaviors, earnings quality, momentum, and liquidity.

This factor database is comprehensive and detailed. It lets us systematically examine the relevance and redundancy of factors. It also helps us study the incremental explanatory power that goes beyond traditional models.

The industry returns data used in this study come directly from the Fama-French Data Library. Specifically, it uses the 49-industry portfolio classification scheme. This dataset classifies individual stocks into 49 different industry portfolios. It bases the classification on SIC codes. The portfolios cover a wide range of industries, including manufacturing, technology, and biotechnology.

Fama and French provide monthly industry returns data. This data helps us do strict analyses of industry-level equity costs. It also allows us to explore the systematic differences between industries. These differences are about risk-return dynamics and factor sensitivities.

Moreover, the empirical analysis in this research clearly matches the Fama-French industry returns with the JKP factor data. This ensures that the method is consistent and reliable when it comes to factor selection and empirical validation. These two data sources are different but they complement each other. Using them together allows us to thoroughly study the limitations of traditional asset pricing models. It also helps us see the potential improvements that can be made through advanced machine learning approaches.

In this research, we mainly use two datasets. One is the Fama-French factor and industry returns, and the other is the JKP factor database. These datasets offer a strong empirical basis for systematically studying the estimation of industry-level equity cost. They guarantee full coverage, consistent methods, and strict empirical analysis, which are necessary for reliable and reproducible asset pricing research.

3.2 Variable Definitions

In this section, we'll clearly indicate the definitions and detailed descriptions of all the key variables used in this research. Having clear variable definitions makes the research methods more transparent and helps others repeat the empirical analyses.

The dependent variable is the Industry Cost of Equity (ICoE).

In this study, the main dependent variable we're looking at is the Industry Cost of Equity (ICoE). Just like what the standard empirical asset pricing literature says, the industry cost of equity means the return that investors expect when they invest in a certain industry. It's basically the compensation for taking on the systematic risk related to that industry.

This research gets industry-level returns from the Fama-French 49-industry portfolio dataset. This dataset gives monthly returns data for 49 different industry groups. The historical industry returns cover the time from July 1969 to December 2023. They are used as the basis for estimating expected returns and the industry-level cost of equity. We use different methods to do this, including traditional asset pricing models and advanced machine learning techniques.

Independent Variables

Fama-French Five Factors (FF5)

In this study, the traditional asset pricing framework used is the Fama-French five-factor model (FF5). This model has five well-known factors. The market risk premium (MKT) indicates the monthly extra return of the market portfolio compared to the risk-free rate. The size factor (SMB) is the monthly return difference between diversified portfolios of small-sized firms and large-sized firms.

The value factor (HML) indicates the difference in monthly returns between portfolios of companies with a high book-to-market ratio (value companies) and those with a low book-to-market ratio (growth companies). The profitability factor (RMW) measures the spread in monthly returns between portfolios of companies that have strong profitability and those that have weak profitability. Finally, the investment factor (CMA) indicates the difference in returns between companies that use conservative investment strategies and those that go for aggressive investments.

All these factors come directly from the Fama-French Data Library, and they are provided on a monthly basis.

Jensen-Kelly-Pedersen (JKP) Factors

Besides the FF5 factors, this study uses a full set of factors from the Jensen-Kelly-Pedersen (JKP) factor database. This dataset has 153 factors. These factors are grouped into 13 thematic groups. Each group represents different economic reasons and market anomalies. These categories cover many factors. These factors are related to profitability, investment decisions, earnings quality, momentum strategies, liquidity, and other anomalies that are written about in the literature.

Each factor stands for monthly return spreads. These spreads are built based on specific firm-level characteristics, as described in Jensen, Kelly, and Pedersen (2023). Using this large set of factors enables us to carefully examine the additional explanatory power. It also helps us look at factor redundancy and how relevant these factors are compared to traditional asset pricing frameworks.

Control Variables

To make sure the study is reliable and the findings are trustworthy, this study includes some common control variables that are often used in asset pricing analyses. The risk-free rate (Rf) indicates the monthly return on short-term U.S. Treasury bills. It acts as a standard for risk-free investment. We get it directly from the Fama-French database.

Additionally, the study controls for the size of the industry. The industry size is measured by taking the logarithm of the total market capitalization of the firms in each industry (Industry Market Cap). The study also includes industry volatility as a control variable. Industry volatility is calculated as the standard deviation of the monthly returns of the industry. This is done to capture the variability of the industry's returns and the related systematic risk.

The inclusion of these control variables gives important explanatory context. It also makes the empirical tests more valid. It does this by controlling for known factors that affect expected returns and systematic risk.

3.3 Machine Learning Models

This section describes in detail the machine learning methods used in this research. It points out their theoretical bases, methodological benefits, and how they can be used in industry-level asset pricing and equity cost estimation.

The first machine learning method we use is the Least Absolute Shrinkage and Selection Operator (LASSO). It's a regularized regression technique. Tibshirani first proposed it in 1996. LASSO can do two things at the same time. It can select variables and estimate parameters. It does this by using penalization. This way, it can make some coefficient estimates become zero.

This process enables LASSO to deal with high-dimensional datasets. It can also reduce multicollinearity. As a result, it improves the accuracy of prediction and interpretability. LASSO selects relevant explanatory factors from large factor sets in a systematic way. This enhances the robustness and economic interpretability of industry cost-of-equity estimates. It also solves the problem of factor proliferation that is common in traditional asset pricing literature.

Also, Gradient Boosting Machines (GBM), which were developed by Friedman in 2001, are another important machine-learning model used in this study. GBM is a strong ensemble learning algorithm. It is based on training and combining many weak learners step by step. Generally, these weak learners are decision trees, and the purpose is to improve the accuracy of predictions. By gradually reducing the prediction residuals, GBM can effectively find the complex nonlinear relationships and interactions between the explanatory variables.

It has inherent flexibility and robustness. This allows it to adjust to changing market conditions and structural shifts in industries dynamically. Compared to traditional linear regression methods, it can significantly improve predictive performance. So, GBM is very suitable for modelling the risk-return dynamics at the industry level. These dynamics are characterized by nonlinearities and structural instability.

This research also uses the eXtreme Gradient Boosting algorithm, or XGBoost. It's a well-known and advanced version of gradient boosting methods. Chen and Guestrin introduced it in 2016. XGBoost builds on traditional GBM. It has some improvements in methodology. These include better gradient boosting algorithms, the ability to handle sparse data well, and strong regularization techniques.

These improvements greatly boost computational efficiency and predictive accuracy. As a result, XGBoost is especially useful for large-scale and high-dimensional asset pricing analyses. It has been proven to be effective in capturing complex nonlinear relationships. It can also deal with multicollinearity and adjust model structures dynamically. This enhances the reliability and practical usefulness of industry-level equity cost estimations.

This research aims to use these machine learning techniques in a systematic way. It wants to give reliable, accurate, and economically useful estimates of the industry cost-of-equity. By doing this, it directly deals with the important practical problems that come with traditional econometric and factor-based models.

In short, using LASSO, GBM, and XGBoost methods is a big step forward in industry-level asset pricing research. These machine learning methods have some great advantages. They can handle a large number of factors well, can model non-linear relationships easily, and can adapt to market changes. Because of these advantages, we can do comprehensive empirical analyses. These analyses can solve the problems of traditional asset pricing methods.

3.4 Empirical Research Design

3.4.1 Methodological Framework

We've set up the empirical basis using descriptive statistics and correlation analyses. Now, this section will go into detail about the research methods used in this study. The main goal here is to see if machine learning (ML) models can give more accurate and reliable estimates of the industry-level cost of equity (ICoE) than traditional factor-based models like CAPM and the Fama-French models (FF3 and FF5).

Specifically, this study uses three well-known machine learning techniques. They are LASSO (Least Absolute Shrinkage and Selection Operator), Gradient Boosting Machine (GBM), and Light Gradient Boosting Machine (Light GBM). Each of these techniques was chosen because they have been proven to be effective in finding complex relationships in financial datasets.

The choice of these specific machine learning (ML) models is based on their unique strengths in dealing with different problems that often come up in asset pricing analysis. We choose LASSO mainly because it can do regularization. It can handle the problem of too many factors well by carefully picking a small set of predictors. When we use an L1 penalty in LASSO, it makes the coefficients of less important factors close to zero. This helps us avoid overfitting and makes the model easier to understand.

In the context of this study, LASSO facilitates identifying which among the numerous JKP factors best explain industry return variations, thereby enhancing the robustness of ICoE estimation.

GBM and Light GBM, on the other hand, belong to the class of ensemble tree-based methods renowned for their flexibility, adaptability, and predictive accuracy. Both methods iteratively build ensembles of decision trees to minimize prediction errors, but they differ significantly in their computational strategies. GBM sequentially fits decision trees by learning from the residuals of previous trees, effectively capturing nonlinear relationships and interactions among predictors. Light GBM improves GBM even more. It uses a histogram-based algorithm and a leaf-wise growth strategy. This helps to cut down the computational complexity a lot. At the same time, it keeps the predictive performance and often makes it better. These methods are really suitable for the practical problems mentioned in Section 4.1. Those problems include non-linear relationships, factor interactions, and time-varying factor exposures.

Implementing the methodology has several important steps. First, when preparing the data, we split the sample into different sub-periods. There's a full period from July 1969 to December 2023, a recent period which is the last 30 years, and an extra period for a robustness check that covers the last 10 years. This way, we can make sure the validation is comprehensive. Also, we can check how stable the model is in different market environments.

Before using machine learning methods, we clean the dataset very carefully. We deal with missing values, outliers, and inconsistent data. This makes sure that the data is reliable and the model will be accurate.

Next, we systematically carry out model training and evaluation. We train each ML model (LASSO, GBM, Light GBM) on historical data. We also do careful cross-validation to select the best hyperparameters. For example, we look at the penalty terms in LASSO. Or in GBM-based methods, we consider the number of trees and learning rates.

We then rigorously evaluate the model performance using out-of-sample predictions. We use metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Adjusted R^2 to comprehensively measure the predictive accuracy.

Moreover, we do a detailed factor analysis with LASSO. We use it to figure out the most important JKP factors that cause variations in industry returns. After that, we take the findings from the LASSO factor selection and put them into the GBM and Light GBM models. This is to make these models better at making predictions.

This combined approach doesn't just bring together the interpretability benefit of LASSO and the predictive power of ensemble methods. It also seriously tests if the factors selected by machine learning do better than the traditional FF5 factors in explaining the variations in the industry's implied cost of equity (ICoE).

The resulting methodological framework comprehensively evaluates the comparative performance of ML models and traditional asset pricing models. It does this across various market conditions and time frames. We rigorously assess each model's predictive accuracy, robustness, and economic relevance. This provides clear empirical evidence about their respective advantages and limitations.

In short, this methodological framework uses advanced ML techniques in a systematic way to deal with the long-standing empirical challenges found in the previous analysis. This study includes strict data preparation, careful factor selection, and reliable evaluation protocols. Its goal is to offer strong evidence about the practical value of machine learning approaches when estimating the industry cost of equity.

3.4.2 Empirical approach

This section explains the study's empirical approach. It clearly describes the steps taken. These steps are to make sure the study is transparent and can be reproduced.

We start the analysis by defining suitable data periods. The "Full Period" has consistent monthly data from July 1969 to December 2023. This ensures that we have comprehensive historical coverage. To see how well the model works recently, we also analyze a "Recent Period." This period covers the most recent 30 years, from January 1994 to December 2023. Additionally, a 10-year period (January 2014 to December 2023) serves as an extra robustness check under specific conditions.

Traditional asset pricing models, including CAPM, FF3, and FF5, are first estimated using ordinary least squares (OLS). The industry excess returns are regressed on Fama-French factors for the in-sample period (July 1969 to December 1993). The period from January 1994 onward is used for out-of-sample evaluation, eliminating potential look-ahead bias.

Machine learning (ML) methods, including LASSO, Gradient Boosting Machines (GBM), XGBoost, and Light GBM, are systematically applied to the same periods. ML models are trained on the same in-sample data and evaluated out-of-sample. Hyperparameter

tuning employs k-fold cross-validation, maximizing model generalizability and predictive accuracy.

The following tables clearly structure the empirical findings and their purposes within this research design:

- **Table 1 (Descriptive Statistics for 49 Industries)**: Provides basic distributional characteristics of industry returns and ICC, essential for initial data assessment and understanding the range and variability across industries.

- **Table 2 (Correlation Matrix of ICC Factors)**: Presents factor correlations to identify potential multicollinearity and factor redundancy, aiding in factor selection and model specification.

- **Table 3 (Industry Historical Return)**: Demonstrate the MAE, RMSE and MEV of the 49 industries in Full period and Recent period using industry historical return.

- **Table 4 (Adjusted R^2 for CAPM)**: Demonstrates explanatory power differences among traditional asset pricing models, justifying the selection of FF5 as a baseline comparison for further ML analyses.

- **Table 5 (Adjusted R^2 for FF3)**: Evaluates Mean Squared Errors of ML-based Beta predictions, investigating whether improved Beta estimates by ML enhance ICC forecasting robustness.

- **Table 6 (Adjusted R^2 for FF5)**: Identifies and compares influential factors selected via LASSO, clarifying whether ML provides more economically meaningful factor selection than the traditional FF5 framework.

- **Table 7 (Adjusted R^2 for OLS)**: Compares forecast accuracy between FF5 and ML using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), directly assessing ML's precision in ICC estimation.

- **Table 8 (LASSO Factor Selection: FF5 vs. 153 JKP Factors across 13 Themes)**: Evaluates whether FF5 and ML estimates remain stable or fluctuate under various

market conditions, highlighting model reliability during market volatility.

- **Table 9/ Appendix 1 (Summary and comparison of models)**: Listed in Appendix 1. The summary of MAE, MSE, Adjusted R^2, MEV of every industry in each model.

- **Table 10 (GBM Performance in Predicting Industry ICC)**: Specifically examines the predictive accuracy (MSE) of the GBM model across 49 industries, assessing industry-specific suitability and robustness.

- **Table 11 (Light GBM Performance in Predicting Industry ICC)**: Provides a comparative analysis to Table 10, focusing on the predictive performance of Light GBM, thereby evaluating the robustness and consistency of GBM results through alternative modeling.

- **Table 12 (ML and FF5 Factor Selection Matching and Stability Tests)**: Matches and evaluates the consistency between ML-selected and traditional FF5 factors, conducting stability tests to confirm the reliability of ML factor selections.

- **Table 13 (JKP Factor Enhancements to the FF5 Model)**: Analyses incremental improvements in the explanatory power of the traditional FF5 model when incorporating additional JKP factors, directly assessing the practical enhancement provided by ML-derived factor selection.

- **Table 14 (XGBoost Model Performance and Robustness Analysis)**: Evaluates XGBoost's predictive performance and robustness across industries and market conditions, providing comprehensive insights into the practical reliability of this advanced ML model.

- **Appendix 2 (Explanation of the 153 factors acronyms)**: Detail explanation of the 153 factors acronyms to display an understanding of its meaning and the categories that each factor is placed. This document is extracted from the Kenneth R. French – Description of Fama/French Factors website. https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/f-f_factors.html

Finally, we conduct robustness checks. These include Leave-One-Out Cross-Validation (LOOCV), parameter sensitivity analyses, and subsample analyses. These checks ensure the stability, generalizability, and credibility of results across diverse scenarios.

Overall, this empirical design combines clear data periods, traditional and advanced models, detailed factor analysis, and strict robustness checks. Each table clearly deals with specific parts of model validity, interpretability, and reliability. This makes the whole study more rigorous empirically and more relevant in practice.

3.5 Model Evaluation and Robustness Checks

This section describes the methods used to evaluate the predictive performance and robustness of asset pricing models applied in this study.

We first assess the model performance using some well-known metrics. These metrics include Adjusted R^2, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Squared Error (MSE). Adjusted R^2 is used to measure the explanatory power of each model. It considers the number of predictors included in the model. Also, it enables us to directly compare different models. MAE and RMSE give us intuitive ways to measure the predictive accuracy. They reflect the average size of the prediction errors.

MSE specifically evaluates how accurate the Beta predictions are. These predictions are generated by ML methods. It directly links the quality of Beta estimation to the accuracy of ICC forecasting.

Besides evaluating the overall performance, we also do robustness checks to make sure the results are reliable under different conditions. Leave-One-Out Cross-Validation (LOOCV) is used. It systematically takes out one industry at a time. Every time we take out an industry, we re-estimate the model. This is to confirm that the factor selection and model predictions are stable and consistent. This process makes sure that the results are not affected by outlier industries or the special characteristics of certain industries.

Parameter sensitivity analysis is used to further check the robustness of the model. We slightly change some key hyperparameters, like the tree depth, learning rate, and regularization parameters. Then we check if these small adjustments to the parameters can significantly change the model's predictions or the selections of factors. If the model is stable

when the parameters are varied, it means that the findings of the model are robust. They are not just the results of specific parameter choices.

We conduct subsample analyses by dividing industries into different subsets. These subsets are based on distinct characteristics, like market volatility, technological intensity, or regulatory sensitivity. We evaluate model performance across these industry subsets. This evaluation helps us understand if model accuracy varies systematically across different economic environments. When the results are consistent across subsamples, it indicates that the modelling approach can be applied generally.

Additionally, we conduct comparative analyses across different market conditions. These conditions include periods of economic stability, financial crises, and market recoveries. We evaluate the model's stability and accuracy under these varying market regimes. This helps us ensure that ICC estimations stay reliable no matter what the external economic conditions are.

Finally, we conduct stability tests on the factor matching process between traditional FF5 and the factors selected by machine learning. These tests aim to verify whether the factors consistently chosen by machine learning still have an influence and make economic sense across different time periods and industry samples. These robustness checks can increase our confidence in the economic interpretability and practical usefulness of the factors selected by machine learning when estimating ICC.

In conclusion, we conducted various activities including the examination of performance metrics, implementation of leave-one-out cross-validation (LOOCV), execution of parameter sensitivity analyses, performance of subsample analyses, and assessment of factor stability. All these together make sure that the asset pricing models we used in this study are reliable, valid, and relevant in the real-world economy. These strict procedures make the empirical findings of our study more believable and useful in practice.

4. Data and Empirical Results

4.1 Data Overview and Correlation Analysis

In this section, we provide a comprehensive overview and correlation analysis of the dataset used for estimating the industry-level cost of equity (ICoE). This foundational analysis, based on Tables 1 and 2, offers critical insights into the characteristics and interrelationships among industry returns and the associated risk factors.

Table 1 - Summary Statistics — Full period = all avaiilable data — Recent period = last 30 years to latest

PANEL A: Descriptive Statistics of Industrial Return

Industry	Period	nobs	Mean	Median	Standard Deviation	Kurtosis	Skewness	Minimum	Maximum	25th Percentile	75th Percentile
Agric		654	0.99	0.85	6.44	1.58	0.02	-29.06	28.56	-2.83	4.77
Food		654	1.03	1.02	4.48	1.81	0.11	-17.76	19.34	-1.42	3.48
Soda		654	1.10	1.33	6.42	4.06	0.12	-26.28	38.90	-2.19	4.68
Beer		654	1.10	1.08	5.19	2.25	-0.03	-19.74	25.81	-1.76	4.20
Smoke		654	1.36	1.74	6.22	2.25	-0.12	-24.96	32.46	-2.46	4.96
Toys		654	0.70	0.98	7.42	1.06	-0.21	-34.51	27.17	-3.80	5.49
Fun		654	1.29	1.24	7.97	2.77	-0.22	-32.25	41.27	-3.15	5.85
Books		654	0.88	0.54	6.13	1.89	0.02	-24.70	30.57	-2.65	4.55
Hshld		654	0.84	1.03	4.69	1.80	-0.27	-21.68	18.71	-1.86	3.71
Clths		654	1.03	1.04	6.74	2.09	-0.09	-30.81	32.39	-2.95	4.95
Hlth		654	0.98	1.00	8.08	2.84	-0.10	-41.07	36.41	-3.68	5.47
MedEq		654	1.04	1.33	5.38	1.09	-0.36	-20.41	21.10	-2.03	4.50
Drugs		654	1.07	1.08	4.99	2.57	0.16	-19.10	31.84	-2.07	4.12
Chems		654	0.99	1.13	5.80	1.89	-0.19	-27.96	22.01	-2.46	4.41
Rubbr		654	1.04	1.40	6.10	2.49	-0.23	-30.49	32.09	-2.44	4.46
Txtls		654	0.89	1.15	7.71	7.52	0.19	-36.03	59.18	-3.06	5.13
BldMt		654	1.07	1.34	6.48	3.36	-0.13	-31.97	34.60	-2.55	4.57
Cnstr		654	1.03	0.98	7.44	1.20	-0.19	-32.14	24.36	-3.47	5.66
Steel		654	0.87	0.73	8.08	1.50	-0.08	-32.41	30.67	-4.23	5.72
FabPr		654	0.81	0.70	7.88	1.89	-0.17	-33.71	35.79	-3.72	5.41
Mach		654	1.04	1.36	6.49	1.93	-0.39	-31.50	23.03	-2.92	4.86
ElcEq		654	1.12	0.89	6.58	1.37	-0.17	-32.09	22.82	-2.68	5.19
Autos		654	1.06	0.86	8.08	5.65	0.70	-36.42	49.45	-3.26	4.92
Aero		654	1.16	1.37	6.91	2.88	-0.38	-35.80	32.53	-2.51	5.17
Ships	196907-202312	654	1.07	1.15	7.39	1.49	-0.05	-32.22	29.33	-3.13	5.26
Guns		654	1.27	1.37	6.49	2.37	-0.11	-30.08	32.87	-2.62	5.12
Gold		654	0.86	0.48	10.73	4.56	0.76	-33.60	79.52	-5.61	6.55
Mines		654	1.04	0.85	7.67	1.58	-0.30	-34.80	26.99	-3.43	5.51
Coal		654	1.18	0.90	11.22	1.56	0.16	-40.89	46.41	-5.38	7.11
Oil		654	1.04	0.97	6.31	3.47	0.14	-34.66	32.89	-2.46	4.71
Util		654	0.87	1.02	4.14	0.96	-0.19	-12.94	18.80	-1.55	3.51
Telcm		654	0.86	1.00	4.82	1.14	-0.21	-15.56	22.11	-1.93	3.89
PerSv		654	0.64	0.70	6.71	1.72	-0.32	-28.23	24.69	-3.19	4.78
BusSv		654	0.96	1.33	5.79	2.01	-0.32	-27.68	25.23	-2.49	4.47
Hardw		654	0.92	0.85	7.17	1.57	-0.17	-33.92	25.39	-3.18	5.17
Softw		654	1.06	1.49	10.31	5.19	0.62	-35.94	73.65	-4.54	5.87
Chips		654	1.19	1.67	7.57	1.29	-0.35	-31.80	26.79	-3.27	6.16
LabEq		654	1.08	1.28	6.93	1.12	-0.18	-30.13	21.63	-3.23	5.11
Paper		654	0.87	0.93	5.59	1.98	0.06	-26.32	24.48	-2.61	4.29
Boxes		654	0.96	1.00	5.72	1.69	-0.39	-28.29	20.72	-2.35	4.47
Trans		654	0.97	1.29	6.04	0.99	-0.26	-28.09	19.21	-2.77	4.76
Whlsl		654	0.98	1.17	5.55	2.30	-0.33	-28.76	17.88	-2.29	4.29
Rtail		654	1.09	0.86	5.58	1.85	-0.13	-29.13	27.10	-2.20	4.63
Meals		654	1.07	1.26	6.09	2.77	-0.52	-31.60	28.73	-2.33	4.74
Banks		654	0.97	1.13	6.26	1.90	-0.32	-27.84	24.97	-2.53	4.89
Insur		654	1.07	1.44	5.50	2.11	-0.28	-26.98	26.59	-2.01	4.47
RlEst		654	0.63	0.97	7.72	10.38	0.57	-36.75	65.22	-3.10	4.70
Fin		654	1.12	1.46	6.38	1.10	-0.40	-26.20	19.42	-2.61	5.21
Other		654	0.51	0.59	6.61	1.75	-0.49	-28.64	21.23	-3.02	4.55

Table 1 summarizes the descriptive statistics for industry and factor returns, offering key statistical characteristics that inform the subsequent modeling approach.

Panel A of Table 1 presents detailed descriptive statistics for the monthly returns of 49 industries from July 1969 to December 2023. Considerable heterogeneity among these industries is evident. For instance, high-technology industries such as software and semiconductor chips exhibit notably high average monthly returns of 1.06% and 1.19%,

respectively, coupled with high standard deviations (10.31% and 7.57%), indicating both attractive returns and significant volatility. In contrast, utilities and finance industries exhibit lower average returns and lower volatility (standard deviations of 4.14% and 6.38%), reflecting relatively stable but lower-risk investment environments. Additionally, kurtosis and skewness values reveal deviations from normal distributions; particularly, the high kurtosis in industries such as semiconductor chips indicates a higher frequency of extreme returns, which implies increased risk management complexity. These detailed statistical insights are fundamental for tailoring risk assessments and ensuring accurate model specification.

Table 1
PANEL B :Descriptive Statistics of Factor Return

FF5 & Global Factors	Period	nobs	Mean	Median	Standard Deviation	Kurtosis	Skewness	Minimum	Maximum	25th Percentile	75th Percentile
Market_Risk_Premium_FF	196907-202312	654	0.58	0.98	4.62	1.59	-0.50	-23.24	16.10	-2.13	3.59
Size_FF		654	0.14	0.05	3.04	3.24	0.37	-15.32	18.28	-1.72	1.92
Value_FF		654	0.28	0.19	3.10	2.05	0.08	-13.87	12.75	-1.44	1.77
Profitbility_FF		654	0.32	0.28	2.29	10.70	-0.30	-18.65	13.07	-0.80	1.40
Investment_FF		654	0.30	0.09	2.08	1.42	0.29	-7.22	9.07	-1.00	1.54
Accruals		654	0.00	0.00	0.01	1.08	0.25	-0.04	0.05	0.00	0.01
Debt Issurance		654	0.00	0.00	0.01	9.68	1.14	-0.03	0.07	0.00	0.01
Investment		654	0.00	0.00	0.02	7.42	0.70	-0.10	0.15	-0.01	0.01
Low leverage		654	0.00	0.00	0.03	19.45	0.71	-0.18	0.29	-0.01	0.01
Low Risk		654	0.00	0.00	0.04	7.68	-0.18	-0.27	0.22	-0.02	0.02
Momentum		654	0.00	0.00	0.03	10.82	-0.64	-0.22	0.23	-0.01	0.02
Profit Growth		654	0.00	0.00	0.01	4.97	-0.59	-0.07	0.05	0.00	0.01
Profitability		654	0.00	0.00	0.02	12.77	0.51	-0.15	0.16	-0.01	0.01
Quality		654	0.00	0.00	0.02	1.20	-0.05	-0.07	0.06	-0.01	0.01
Seasonality		654	0.00	0.00	0.01	2.38	0.22	-0.02	0.03	0.00	0.00
Size		654	0.00	0.00	0.02	4.50	0.96	-0.09	0.14	-0.01	0.01
Short-Term Reversal		654	0.00	0.00	0.01	16.40	0.85	-0.08	0.11	0.00	0.01
Value		654	0.00	0.00	0.03	15.57	-0.16	-0.28	0.21	-0.01	0.02

Panel B of Table 1 explores the statistical characteristics of selected factor returns, including traditional Fama-French (FF5) factors and additional global risk factors. The Market Risk Premium, as a crucial benchmark for asset pricing models, shows the highest mean monthly return of 0.58% alongside significant volatility (standard deviation of 4.62%). The Profitability (RMW) and Investment (CMA) factors exhibit notable mean returns (0.29% and 0.27%) but moderate volatility (standard deviations of 2.29% and 2.08%), underscoring their critical role in capturing distinct risk dimensions beyond the market factor. Conversely, factors such as accruals, quality, and seasonality exhibit minimal variability, suggesting relatively limited explanatory power for industry returns in isolation, but potentially valuable when used in combination with other factors.

Table 2 conducts an extensive correlation analysis to clairfy the structural relationships among industry returns and between factors, essential for avoiding redundancy and ensuring robust factor selection in subsequent modeling.

Panel A of Table 2 reveals the correlation matrix among industry returns. The average correlation coefficient across all industries is approximately 0.60, indicating a substantial

degree of co-movement among industry returns, reflecting common market-wide factors or macroeconomic influences. Notably, industries such as retail and transportation show higher average correlations (0.69 and 0.65), suggesting significant shared sensitivities to broader economic cycles and consumption patterns. In stark contrast, industries like gold and coal demonstrate substantially lower correlations (0.20 and 0.38), emphasizing their distinct industry-specific dynamics and potential diversification benefits within investment portfolios. These differential correlation structures provide essential insights into industry-specific risks and emphasize the importance of tailored asset pricing models that account for varying degrees of systematic and idiosyncratic risk.

Table 2 - Correlation Matrix

PANEL A	Correlation Between Industry Returns				
Industry	MeanCor	MedCor	SDCor	Maximum	Minimum
Agric	0.48	0.50	0.08	0.61	0.17
Food	0.53	0.55	0.11	0.70	0.13
Soda	0.46	0.48	0.10	0.63	0.07
Beer	0.50	0.52	0.11	0.72	0.12
Smoke	0.38	0.38	0.08	0.59	0.14
Toys	0.59	0.61	0.11	0.74	0.20
Fun	0.60	0.62	0.12	0.77	0.15
Books	0.64	0.66	0.12	0.81	0.15
Hshld	0.58	0.61	0.12	0.72	0.14
Clths	0.61	0.64	0.13	0.81	0.15
Hlth	0.54	0.56	0.10	0.73	0.20
MedEq	0.58	0.61	0.10	0.77	0.19
Drugs	0.50	0.52	0.11	0.77	0.16
Chems	0.65	0.68	0.11	0.84	0.28
Rubbr	0.64	0.66	0.12	0.81	0.21
Txtls	0.57	0.60	0.13	0.78	0.13
BldMt	0.68	0.71	0.13	0.83	0.20
Cnstr	0.63	0.65	0.11	0.82	0.28
Steel	0.58	0.60	0.12	0.83	0.32
FabPr	0.52	0.54	0.10	0.70	0.21
Mach	0.66	0.69	0.13	0.83	0.30
ElcEq	0.65	0.68	0.12	0.83	0.20
Autos	0.54	0.57	0.13	0.72	0.14
Aero	0.61	0.64	0.11	0.76	0.18
Ships	0.56	0.58	0.10	0.70	0.18
Guns	0.47	0.48	0.09	0.67	0.23
Gold	0.20	0.18	0.07	0.48	0.07
Mines	0.54	0.54	0.10	0.78	0.32
Coal	0.38	0.37	0.08	0.57	0.23
Oil	0.46	0.46	0.08	0.62	0.29
Util	0.44	0.44	0.07	0.58	0.19
Telcm	0.52	0.54	0.10	0.68	0.13
PerSv	0.60	0.62	0.11	0.78	0.17
BusSv	0.70	0.72	0.12	0.87	0.25
Hardw	0.52	0.54	0.13	0.82	0.16
Softw	0.50	0.52	0.12	0.72	0.17
Chips	0.58	0.61	0.14	0.82	0.17
LabEq	0.62	0.66	0.12	0.82	0.25
Paper	0.62	0.64	0.11	0.84	0.25
Boxes	0.59	0.61	0.11	0.75	0.19
Trans	0.65	0.69	0.13	0.81	0.17
Whlsl	0.69	0.72	0.11	0.87	0.27
Rtail	0.62	0.66	0.14	0.81	0.13
Meals	0.63	0.66	0.12	0.79	0.19
Banks	0.61	0.63	0.12	0.82	0.10
Insur	0.60	0.64	0.11	0.80	0.15
RlEst	0.61	0.63	0.12	0.80	0.19
Fin	0.64	0.67	0.13	0.82	0.14
Other	0.59	0.61	0.10	0.75	0.21

Panel B of Table 2 further elaborates on correlations among factor returns, critical for determining factor redundancy and enhancing model efficiency. Interestingly, the Market Risk Premium exhibits predominantly negative correlations with factors such as Size (-0.13) and Low Leverage (-0.22), implying distinct and largely independent market-risk dimensions captured by these factors. Conversely, the positive correlation between Market Risk Premium and factors like Investment (0.12) and Value (0.10) reflects potential overlapping economic rationales underlying these factors, necessitating careful consideration in model construction. Furthermore, factors such as Profitability (0.06) and Momentum (0.01) exhibit minimal correlations with other factors, suggesting their incremental explanatory value without substantial redundancy. Notably, the maximum correlation values for Investment and Profitability factors (0.83) imply episodic overlaps during specific economic conditions, highlighting the necessity of dynamic modeling approaches capable of adapting to varying economic regimes.

Table 2

PANEL B	Correlation Between Factor Returns				
FF5 & Global Factors	MeanCor	MedCor	SDCor	Maximum	Minimum
Market_Risk_Premium_FF	-0.13	-0.20	0.26	0.34	-0.63
Size_FF	-0.04	-0.07	0.33	0.79	-0.54
Value_FF	0.10	0.12	0.42	0.78	-0.69
Profitbility_FF	0.06	0.06	0.40	0.83	-0.61
Investment_FF	0.11	0.08	0.37	0.79	-0.54
Accruals	-0.03	-0.01	0.21	0.32	-0.39
Debt Issurance	0.03	0.09	0.20	0.31	-0.29
Investment	0.12	0.13	0.48	0.83	-0.79
Low leverage	-0.22	-0.22	0.47	0.37	-0.93
Low Risk	0.08	0.18	0.49	0.73	-0.83
Momentum	0.01	0.04	0.22	0.45	-0.41
Profit Growth	-0.06	-0.04	0.29	0.45	-0.48
Profitability	0.06	0.11	0.44	0.83	-0.62
Quality	-0.02	0.03	0.32	0.53	-0.53
Seasonality	0.07	0.11	0.13	0.20	-0.25
Size	-0.04	0.02	0.33	0.79	-0.55
Short-Term Reversal	0.01	0.02	0.18	0.30	-0.26
Value	0.10	0.06	0.50	0.83	-0.93

The standard deviation of factor correlations (SDCor) generally ranges from moderate to substantial (0.18 for Short-Term Reversal to 0.50 for Value), underscoring temporal stability issues among factor relationships. Specifically, the pronounced variability in correlations for factors such as Value and Low Leverage highlights their fluctuating

effectiveness across different market conditions, necessitating advanced modeling methods that can dynamically adjust factor exposures.

Additionally, the minimum correlation statistic provides further nuanced insights. The exceptionally low minimum correlation for the Low Leverage factor (-0.93) emphasizes its potential role as a critical diversifier during periods of market stress. Such findings reinforce the strategic importance of selecting factors based not only on average correlation but also on their behavior during extreme market events.

In conclusion, Tables 1 and 2 collectively provide foundational empirical insights essential for the subsequent analysis of industry-level cost of equity. The diversity in return characteristics across industries and intricate correlation structures among factors indicate significant challenges for traditional linear models, which typically assume stable factor exposures and linear relationships. The identified complexities strongly support the use of sophisticated methodologies, particularly machine learning models, capable of capturing dynamic relationships, nonlinear interactions, and evolving factor importance. By integrating these advanced techniques, the analysis aims to substantially enhance predictive accuracy and robustness in estimating industry-specific equity costs.

4.2 Research Hypotheses

Based on the literature review and identified gaps in existing research, this thesis aims to explore whether machine learning methods can provide improvements over traditional asset pricing models, particularly at the industry level. The primary research objective is to evaluate the predictive accuracy and robustness of machine learning approaches in estimating the industry cost of equity (ICoE). To achieve this objective, the study formulates the following explicit hypotheses:

Hypothesis 1:

Machine learning models (such as GBM, Light GBM, and LASSO) significantly outperform traditional asset pricing models (CAPM, FF3, and FF5) in predicting industry-level cost of equity.

This hypothesis is based on the argument that traditional factor-based models, although widely used, have limitations in capturing nonlinearities and structural changes inherent in financial markets. Thus, it is expected that machine learning models, capable of handling

complex interactions among numerous factors, can achieve superior predictive performance, measured primarily by metrics such as Adjusted R^2, Mean Absolute Error (MAE), and Mean Squared Error (MSE).

Hypothesis 2:

Machine learning methods can identify more relevant factors from the extensive set of 153 JKP factors and 13 themes, improving explanatory power relative to the fixed factor structure of the Fama-French five-factor (FF5) model.

Considering that the FF5 model includes only five predefined factors (market risk premium, size, value, profitability, and investment), this hypothesis examines whether a flexible, data-driven selection process using LASSO or other machine learning techniques can yield a more powerful set of factors to explain industry-level equity returns.

Hypothesis 3:

The relative performance of machine learning models versus traditional models varies significantly across different industries and market conditions, indicating a conditional advantage rather than a universal superiority of machine learning methods.

Given the heterogeneous characteristics of different industries—such as variations in market capitalization, technological change, volatility, and regulatory environment—this hypothesis suggests that no single approach consistently dominates. Instead, it proposes that the advantages of machine learning models are conditional, depending upon specific industry contexts and periods of analysis.

By empirically testing these hypotheses, this study seeks to clarify the applicability and advantages of machine learning techniques in the domain of asset pricing, particularly for estimating industry-level costs of equity, thus addressing existing gaps highlighted in prior research.

4.3 Empirical Results

4.3.1 Model Specifications and Benchmark Comparisons

Before assessing advanced asset pricing models, it is necessary to establish a baseline performance benchmark using a simple yet widely-adopted method—the historical average return model. Table 3 summarizes the predictive accuracy of this historical approach across

all industries for two distinct sample periods: the full period (July 1969 to December 2023) and the recent period (December 1993 to December 2023). Three critical metrics are considered: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Measurement Error Variance (MEV).

Table 3 clearly illustrates that historical returns alone provide a moderate baseline predictive accuracy. Over the entire sample period, the average MAE across industries is approximately 4.90, indicating that predictions based solely on historical averages exhibit moderate absolute deviation from actual returns. Correspondingly, RMSE averages about 6.56, reflecting the substantial but manageable prediction errors inherent in such an elementary model. Measurement error variance (MEV), averaging 45.11, indicates significant variability around the historical means, highlighting the inherent uncertainty when relying solely on historical averages.

However, Table 3 also highlights considerable heterogeneity among industry sectors. Specifically, sectors such as Utilities (Util, MAE=3.09), Food (MAE=3.28), and Telecommunications (Telcm, MAE=3.77) demonstrate notably lower MAE and RMSE values, indicating relatively high historical return stability and predictability. Conversely, cyclical or commodity-sensitive industries such as Coal (MAE=8.53, RMSE=11.47), Mines (MAE=6.12, RMSE=8.01), and Steel (MAE=6.36, RMSE=8.44) experience significantly higher predictive errors, highlighting their susceptibility to broader economic fluctuations and commodity price volatility.

The differences between the full period and recent period further illustrate changes in industry predictability over time. Although overall predictive errors slightly improved in the recent period (average MAE decreasing from 4.90 to 4.87 and RMSE from 6.56 to 6.52), individual industries display distinct patterns. For instance, industries such as Steel and Autos exhibit increased RMSE and MEV during the recent period, suggesting enhanced volatility and structural shifts, possibly due to globalization, regulatory changes, or technological disruptions. In contrast, sectors such as Food, Insurance (Insur), and Pharmaceuticals (Drugs) demonstrate improved or consistent prediction accuracy, emphasizing their robustness and stable market positioning over recent decades.

Thus, this historical return benchmark (Table 3) establishes a clear reference point against which more sophisticated models, such as CAPM, Fama-French factor models, and machine

learning models, can be rigorously evaluated. Any observed improvement in subsequent analyses will highlight the incremental value and practical benefits of employing advanced factor-based or machine learning-driven methodologies.

Table 3 - Industry Historical Return

Industry	MAE		RMSE		MEV	
	Full period: 196907-202312	Recent period: 199312-202312	Full period: 196907-202312	Recent period: 199312-202312	Full period: 196907-202312	Recent period: 199312-202312
Average (Industries)	4.90	4.87	6.56	6.52	45.11	46.31
Agric	4.76	4.96	6.29	6.61	39.58	43.67
Food	3.28	2.80	4.40	3.67	19.35	13.46
Soda	4.63	3.88	6.37	5.49	40.64	30.12
Beer	3.68	2.93	4.99	3.84	24.86	14.74
Smoke	4.86	4.27	6.46	5.36	41.77	28.76
Toys	5.54	5.72	7.22	7.38	52.07	54.42
Fun	5.63	6.07	7.64	8.42	58.45	70.83
Books	4.46	4.78	5.98	6.65	35.74	44.17
Hshld	3.32	3.02	4.46	3.99	19.90	15.95
Clths	4.81	4.83	6.47	6.44	41.90	41.45
Hlth	5.19	4.42	6.80	5.82	46.25	33.89
MedEq	4.06	3.83	5.24	5.06	27.47	25.61
Drugs	3.66	3.22	4.72	4.07	22.25	16.54
Chems	4.36	4.57	5.83	6.22	34.01	38.71
Rubbr	4.38	4.50	5.97	6.18	35.64	38.23
Txtls	5.55	6.62	7.97	9.63	63.46	92.79
BldMt	4.73	5.24	6.48	7.26	41.98	52.64
Cnstr	5.58	5.80	7.36	7.50	54.10	56.29
Steel	6.36	7.34	8.44	9.42	71.24	88.81
FabPr	5.93	7.02	7.94	9.29	62.99	86.39
Mach	5.02	5.12	6.61	6.87	43.65	47.14
ElcEq	5.02	5.11	6.63	6.90	43.99	47.67
Autos	5.96	6.98	8.52	10.32	72.56	106.55
Aero	4.83	4.62	6.62	6.48	43.79	41.95
Ships	5.59	5.51	7.53	7.68	56.73	58.99
Guns	4.64	4.01	6.32	5.51	39.94	30.39
Gold	8.40	7.89	11.23	10.39	126.10	107.99
Mines	6.12	7.30	8.01	9.19	64.12	84.55
Coal	8.53	10.25	11.47	13.30	131.63	176.87
Oil	4.71	5.35	6.50	7.52	42.28	56.59
Util	3.09	3.00	4.00	3.92	16.03	15.33
Telcm	3.77	3.52	5.00	4.64	24.98	21.49
PerSv	4.70	4.65	6.16	6.14	37.93	37.71
BusSv	4.15	3.94	5.50	5.30	30.24	28.08
Hardw	5.61	4.86	7.42	6.24	55.10	38.98
Softw	6.07	4.14	8.02	5.22	64.30	27.23
Chips	5.89	5.20	7.63	6.46	58.28	41.68
LabEq	5.09	4.33	6.70	5.57	44.87	31.07
Paper	4.06	3.89	5.47	5.22	29.87	27.25
Boxes	4.37	4.31	5.75	5.57	33.08	31.06
Trans	4.50	4.53	5.82	5.83	33.91	33.95
Whlsl	3.94	3.77	5.23	5.04	27.35	25.36
Rtail	4.11	3.53	5.42	4.75	29.41	22.53
Meals	4.01	3.63	5.27	4.74	27.78	22.49
Banks	4.69	4.83	6.23	6.61	38.76	43.72
Insur	3.92	3.84	5.24	5.28	27.42	27.86
RlEst	4.96	6.03	7.29	8.97	53.09	80.46
Fin	4.93	5.02	6.46	6.51	41.72	42.43
Other	4.65	3.67	6.14	4.96	37.66	24.59

4.3.2 CAPM Analysis

Table 4 - CAPM 1 factor (In Sample)

Industry	Full Period			Recent Period			Industry	Full Period			Recent Period		
	Market	Adjusted R^2	nobs	Market	Adjusted R^2	nobs		Market	Adjusted R	nobs	Market	Adjusted R	nobs
Agric	0.85	0.37	654	0.69	0.24	360	Guns	0.78	0.31	654	0.50	0.13	360
Food	0.65	0.45	654	0.47	0.30	360	Gold	0.58	0.06	654	0.44	0.03	360
Soda	0.79	0.32	654	0.68	0.21	360	Mines	1.12	0.46	654	1.22	0.44	360
Beer	0.71	0.41	654	0.51	0.24	360	Coal	1.14	0.22	654	1.15	0.16	360
Smoke	0.64	0.22	654	0.51	0.11	360	Oil	0.86	0.39	654	0.88	0.34	360
Toys	1.18	0.54	654	1.03	0.44	360	Util	0.51	0.33	654	0.43	0.21	360
Fun	1.35	0.61	654	1.38	0.61	360	Telcm	0.78	0.56	654	0.92	0.63	360
Books	1.08	0.67	654	1.05	0.61	360	PerSv	1.09	0.56	654	0.92	0.48	360
Hshld	0.76	0.56	654	0.57	0.38	360	BusSv	1.15	0.84	654	1.08	0.84	360
Clths	1.12	0.59	654	1.03	0.53	360	Hardw	1.19	0.59	654	1.36	0.60	360
Hlth	1.12	0.41	654	0.82	0.34	360	Softw	1.52	0.46	654	1.26	0.67	360
MedEq	0.90	0.60	654	0.82	0.57	360	Chips	1.38	0.70	654	1.46	0.65	360
Drugs	0.76	0.49	654	0.63	0.41	360	LabEq	1.25	0.70	654	1.18	0.69	360
Chems	1.05	0.70	654	1.05	0.64	360	Paper	0.95	0.62	654	0.86	0.56	360
Rubbr	1.06	0.65	654	1.01	0.57	360	Boxes	0.95	0.59	654	0.97	0.54	360
Txtls	1.15	0.48	654	1.23	0.42	360	Trans	1.08	0.68	654	1.00	0.64	360
BldMt	1.20	0.73	654	1.19	0.65	360	Whlsl	1.04	0.75	654	0.91	0.69	360
Cnstr	1.29	0.64	654	1.22	0.58	360	Rtail	1.00	0.68	654	0.89	0.63	360
Steel	1.33	0.58	654	1.59	0.62	360	Meals	1.03	0.61	654	0.78	0.52	360
FabPr	1.12	0.43	654	1.16	0.37	360	Banks	1.08	0.64	654	1.07	0.59	360
Mach	1.23	0.76	654	1.30	0.73	360	Insur	0.92	0.60	654	0.86	0.55	360
ElcEq	1.23	0.74	654	1.28	0.73	360	RlEst	1.24	0.55	654	1.22	0.51	360
Autos	1.25	0.51	654	1.48	0.51	360	Fin	1.23	0.79	654	1.36	0.77	360
Aero	1.13	0.57	654	1.04	0.50	360	Other	1.09	0.58	654	0.92	0.52	360
Ships	1.10	0.48	654	1.08	0.44	360							

Table 4 reports the empirical results for the CAPM across the full and recent periods. It includes the market beta and adjusted R^2 for each of the 49 industries. The results indicate considerable variation in market betas and explanatory power across industries. For example, in the full period, industries like Toys (Adjusted $R^2 = 0.54$), Lab Equipment (Adjusted $R^2 = 0.70$), and Finance (Adjusted $R^2 = 0.79$) exhibit relatively high adjusted R^2 values, indicating that the CAPM explains a substantial portion of their return variability. In contrast, industries such as Gold (Adjusted $R^2 = 0.06$) and Coal (Adjusted $R^2 = 0.22$) demonstrate significantly lower explanatory power. This disparity suggests CAPM's limitations in capturing specific industry dynamics, particularly in commodity-driven or cyclically sensitive sectors.

The recent period reflects similar trends but generally with reduced adjusted R^2 values. For instance, the software industry shows a noticeable increase in explanatory power from 0.46 (full period) to 0.67 (recent period), likely due to its growing market influence and stability in recent decades. However, industries like Gold continue to show minimal explanatory power (Adjusted $R^2 = 0.03$). Overall, these findings reinforce the CAPM's limitations, underscoring its inadequate capability in addressing industry-specific risk factors comprehensively.

4.3.3 Empirical Results from the FF3 Model

Following the CAPM, we conducted empirical tests based on the Fama-French Three-Factor Model (FF3). The FF3 model expands upon the CAPM by adding two empirically robust factors: Size (SMB) and Value (HML), designed to capture anomalies associated with

market capitalization and book-to-market ratios, respectively. Table 5 presents comprehensive results for the FF3 model across 49 industries over the full and recent periods, displaying estimated factor loadings, significance levels, and adjusted R^2 values.

Analysing the factor significance and loadings, the Market Risk Premium factor consistently remains highly significant across all industries for both periods, mirroring the CAPM results but often showing enhanced explanatory power due to the inclusion of additional factors. Notably, industries such as software and semiconductor chips exhibit significant loadings not only on the market factor but also on SMB and HML factors, highlighting that smaller-cap firms and firms with higher book-to-market ratios in these industries tend to deliver systematically higher returns. Specifically, software shows a strong positive loading on SMB (0.85***) and a negative loading on HML (-0.37***), consistent with the expectation that smaller, growth-oriented firms dominate this sector.

Moreover, the Value factor (HML) is found to be significant in explaining return variations across a majority of industries, particularly industries such as Textiles (0.62***), Steel (0.56***), and Chemicals (0.31***). This suggests these traditional industries typically contain a higher proportion of value-oriented firms whose returns are sensitive to book-to-market ratios.

The explanatory power of the FF3 model, measured by adjusted R^2, shows substantial improvement over the CAPM results. For instance, the adjusted R^2 for the software industry increases from 0.46 under CAPM to 0.57 under FF3, and similarly, the semiconductor chips industry exhibits an increase from 0.70 to 0.74. These improvements confirm that SMB and HML factors significantly enhance the explanatory capacity of the traditional CAPM model.

Table 5 - OLS 3 Factor estimation Result (In Sample)

Industry	Full period: 196907-202312							Recent period: 199312-202312						
	Market_Risk_Premium	Significance_b1	Size	Significance_b2	Value	Significance_b3	Adj_R2	Market_Risk_Premium	Significance_b1	Size	Significance_b2	Value	Significance_b3	Adj_R2
Agric	0.80	***	0.35	***	0.13	***	0.40	0.68	***	0.21	***	0.28	***	0.27
Food	0.71	***	-0.20	***	0.16	***	0.49	0.53	***	-0.25	***	0.27	***	0.40
Soda	0.86	***	-0.20	***	0.18	***	0.34	0.73	***	-0.19		0.32	***	0.25
Beer	0.76	***	-0.20	***	0.05		0.42	0.58	***	-0.37	***	0.08		0.31
Smoke	0.74	***	-0.32	***	0.22	***	0.26	0.58	***	-0.28	***	0.45	***	0.19
Toys	1.12	***	0.46	***	0.13	***	0.57	0.99	***	0.36	***	0.30	***	0.47
Fun	1.29	***	0.39	***	0.06		0.64	1.35	***	0.20	***	0.10		0.61
Books	1.08	***	0.24	***	0.30	***	0.70	1.04	***	0.19	***	0.42	***	0.67
Hshld	0.80	***	-0.21	***	-0.03		0.58	0.62	***	-0.25	***	0.10	***	0.42
Clths	1.11	***	0.31	***	0.30	***	0.62	1.05	***	0.01		0.33	***	0.56
Hlth	1.02	***	0.56	***	0.07		0.45	0.80	***	0.25	***	0.45	***	0.40
MedEq	0.86	***	0.02		-0.25	***	0.62	0.80	***	0.12	***	0.01		0.57
Drugs	0.78	***	-0.31	***	-0.30	***	0.56	0.67	***	-0.25	***	-0.11	***	0.44
Chems	1.12	***	-0.05		0.35	***	0.74	1.08	***	-0.01		0.46	***	0.71
Rubbr	1.01	***	0.49	***	0.21	***	0.71	0.96	***	0.38	***	0.27	***	0.62
Txtls	1.15	***	0.61	***	0.74	***	0.60	1.19	***	0.58	***	0.93	***	0.56
BldMt	1.22	***	0.24	***	0.43	***	0.78	1.19	***	0.21	***	0.54	***	0.72
Cnstr	1.27	***	0.38	***	0.28	***	0.68	1.19	***	0.34	***	0.49	***	0.63
Steel	1.32	***	0.45	***	0.45	***	0.63	1.52	***	0.54	***	0.40	***	0.66
FabPr	1.04	***	0.69	***	0.30	***	0.50	1.05	***	0.83	***	0.53	***	0.48
Mach	1.19	***	0.32	***	0.14	***	0.78	1.25	***	0.41	***	0.23	***	0.77
ElcEq	1.22	***	0.09	***	0.04		0.75	1.27	***	0.08		0.11		0.73
Autos	1.27	***	0.22	***	0.39	***	0.54	1.47	***	0.21		0.26	***	0.52
Aero	1.16	***	0.14	***	0.37	***	0.60	1.08	***	-0.06		0.51	***	0.57
Ships	1.15	***	0.16	***	0.47	***	0.51	1.10	***	0.19	***	0.73	***	0.54
Guns	0.82	***	0.08		0.39	***	0.34	0.55	***	-0.07		0.50	***	0.21
Gold	0.53	***	0.28	***	0.02		0.06	0.42	***	0.15		0.02		0.03
Mines	1.13	***	0.27	***	0.38	***	0.49	1.22	***	0.21	***	0.49	***	0.48
Coal	1.13	***	0.43	***	0.45	***	0.24	1.09	***	0.62	***	0.69	***	0.20
Oil	0.96	***	-0.12		0.51	***	0.46	0.91	***	0.11		0.72	***	0.46
Util	0.60	***	-0.21	***	0.30	***	0.41	0.48	***	-0.21	***	0.27	***	0.30
Telcm	0.84	***	-0.17	***	0.14	***	0.59	0.95	***	-0.16	***	0.01		0.63
PerSv	1.04	***	0.39	***	0.13	***	0.60	0.90	***	0.25	***	0.32	***	0.52
BusSv	1.08	***	0.35	***	-0.06	***	0.88	1.04	***	0.21	***	-0.02		0.86
Hardw	1.11	***	0.17	***	-0.42	***	0.63	1.29	***	0.22	***	-0.52	***	0.66
Softw	1.29	***	0.80	***	-0.57	***	0.55	1.22	***	0.01		-0.73	***	0.79
Chips	1.25	***	0.36	***	-0.41	***	0.76	1.38	***	0.24	***	-0.62	***	0.74
LabEq	1.13	***	0.43	***	-0.35	***	0.76	1.09	***	0.42	***	-0.27	***	0.77
Paper	1.01	***	-0.07		0.32	***	0.65	0.90	***	-0.05		0.41	***	0.63
Boxes	0.99	***	-0.10	***	0.11	***	0.60	0.99	***	-0.01		0.23	***	0.55
Trans	1.10	***	0.16	***	0.31	***	0.71	1.01	***	0.04		0.39	***	0.69
Whlsl	0.99	***	0.37	***	0.12	***	0.79	0.88	***	0.25	***	0.28	***	0.74
Rtail	0.98	***	0.05		-0.05		0.68	0.90	***	-0.10		-0.05		0.63
Meals	1.02	***	0.14	***	0.12	***	0.61	0.82	***	-0.14	***	0.25	***	0.56
Banks	1.20	***	-0.11	***	0.67	***	0.75	1.16	***	-0.15	***	0.84	***	0.80
Insur	1.02	***	-0.18	***	0.39	***	0.66	0.94	***	-0.29	***	0.52	***	0.71
RlEst	1.18	***	0.82	***	0.68	***	0.70	1.16	***	0.63	***	0.77	***	0.64
Fin	1.25	***	0.12	***	0.28	***	0.81	1.37	***	0.09		0.25	***	0.79
Other	1.09	***	0.15	***	0.11		0.59	0.96	***	-0.08		0.25	***	0.54

However, despite these improvements, substantial limitations still persist in certain industries. For example, the biotechnology and pharmaceuticals sectors exhibit relatively lower adjusted R^2 values (0.52 and 0.49 respectively), indicating that factors capturing size and value dimensions alone are insufficient to fully explain the risk-return dynamics in sectors characterized by innovation-driven growth and higher volatility.

In summary, the FF3 model substantially enhances the explanatory power over the CAPM, but considerable variations remain across industries, pointing towards the necessity of further model enhancements.

4.3.4 Empirical Results from the FF5 Model

The empirical analysis is extended further by applying the Fama-French Five-Factor Model (FF5), which incorporates profitability (RMW) and investment (CMA) factors alongside the three FF3 factors. Table 6 summarizes detailed FF5 results for each of the 49 industries across both full and recent periods. The results include factor loadings, significance levels, and adjusted R^2 values, providing deeper insights into industry-specific variations.

Upon examining factor significance, Market Risk Premium remains consistently significant for most industries, reinforcing findings from both CAPM and FF3. Notably, the inclusion of profitability and investment factors significantly refines the explanatory power for many industries, particularly those driven by technology, innovation, or distinct investment patterns. For instance, the software industry displays strong loadings on profitability (RMW, -0.78) and investment (CMA, 0.57), reflecting that returns in technology sectors are sensitive to firms' profitability dynamics and conservative investment strategies.

Industries characterized by stable investment patterns, such as Utilities (CMA loading: 0.31), Real Estate (CMA loading: 0.28), and Insurance (CMA loading: 0.18), also benefit substantially from the addition of the CMA factor. Similarly, profitability (RMW) notably enhances explanatory power for industries like Pharmaceuticals (0.50), Health (0.48), and Food (0.40), emphasizing the importance of profitability in return determination within stable, mature sectors.

Table 6 - OLS 5 Factor estimation Result (In Sample)

Industry	Full Period						Recent Period					
	Market_Risk_Premium	Size	Value	Profitability	Investment	Adjusted R^2	Market_Risk_Premium	Size	Value	Profitability	Investment	Adjusted R^2
Agric	0.82	0.43	-0.02	0.20	0.18	0.40	0.70	0.21	0.13	-0.04	0.27	0.27
Food	0.78	-0.04	-0.05	0.50	0.46	0.55	0.64	-0.09	0.04	0.40	0.40	0.46
Soda	0.91	-0.02	0.04	0.62	0.30	0.38	0.85	-0.02	0.04	0.47	0.44	0.27
Beer	0.83	-0.01	-0.15	0.63	0.42	0.49	0.71	-0.17	-0.20	0.52	0.49	0.38
Smoke	0.84	-0.12	-0.12	0.65	0.80	0.33	0.76	-0.10	0.02	0.50	0.83	0.24
Toys	1.14	0.65	-0.05	0.54	0.14	0.60	1.06	0.55	0.02	0.42	0.21	0.49
Fun	1.25	0.46	0.16	0.16	-0.39	0.65	1.30	0.22	0.22	-0.01	-0.36	0.62
Books	1.09	0.38	0.24	0.40	-0.01	0.72	1.08	0.31	0.28	0.28	0.06	0.68
Hshld	0.87	-0.03	-0.22	0.58	0.42	0.65	0.75	-0.07	-0.19	0.47	0.50	0.50
Clths	1.12	0.53	0.24	0.68	-0.08	0.67	1.13	0.32	0.16	0.73	-0.11	0.61
Hlth	1.05	0.90	-0.11	0.96	0.04	0.52	0.90	0.52	0.16	0.62	0.16	0.44
MedEq	0.89	0.12	-0.35	0.26	0.18	0.63	0.84	0.17	-0.15	0.09	0.27	0.58
Drugs	0.84	-0.24	-0.46	0.23	0.46	0.58	0.74	-0.24	-0.29	0.04	0.52	0.47
Chems	1.15	0.06	0.23	0.31	0.24	0.75	1.15	0.14	0.29	0.35	0.19	0.72
Rubbr	1.03	0.61	0.06	0.41	0.11	0.73	1.05	0.56	-0.01	0.47	0.17	0.65
Txtls	1.16	0.82	0.60	0.59	0.01	0.64	1.25	0.92	0.68	0.72	-0.19	0.60
BldMt	1.24	0.39	0.33	0.49	0.08	0.80	1.28	0.47	0.31	0.65	0.01	0.77
Cnstr	1.28	0.52	0.19	0.41	0.01	0.69	1.24	0.55	0.33	0.48	-0.09	0.66
Steel	1.31	0.36	0.34	-0.35	0.15	0.64	1.53	0.54	0.25	-0.02	0.10	0.66
FabPr	0.99	0.77	0.34	0.16	-0.38	0.52	1.03	0.99	0.44	0.31	-0.41	0.49
Mach	1.20	0.32	0.06	0.00	0.08	0.78	1.28	0.48	0.07	0.16	0.08	0.77
ElcEq	1.21	0.13	0.05	0.11	-0.07	0.75	1.29	0.16	0.05	0.18	-0.02	0.73
Autos	1.27	0.22	0.39	0.06	-0.07	0.53	1.50	0.35	0.16	0.39	-0.12	0.53
Aero	1.18	0.32	0.27	0.49	0.10	0.63	1.15	0.10	0.35	0.36	0.16	0.58
Ships	1.19	0.35	0.27	0.53	0.31	0.54	1.21	0.44	0.41	0.61	0.28	0.58
Guns	0.86	0.31	0.24	0.70	0.21	0.39	0.65	0.21	0.26	0.69	0.14	0.26
Gold	0.61	0.31	-0.36	0.09	0.77	0.07	0.56	0.34	-0.41	0.48	0.63	0.04
Mines	1.16	0.33	0.21	0.14	0.26	0.49	1.29	0.38	0.25	0.39	0.19	0.49
Coal	1.16	0.45	0.20	-0.01	0.43	0.25	1.13	0.65	0.40	-0.01	0.39	0.20
Oil	1.00	-0.03	0.36	0.22	0.34	0.47	0.96	0.23	0.56	0.23	0.18	0.47
Util	0.64	-0.16	0.19	0.13	0.31	0.42	0.58	-0.12	0.03	0.23	0.51	0.34
Telcm	0.85	-0.23	0.12	-0.22	0.13	0.60	0.96	-0.21	-0.03	-0.13	0.26	0.64
PerSv	1.06	0.59	0.01	0.60	0.03	0.64	0.97	0.47	0.11	0.52	0.03	0.55
BusSv	1.07	0.42	-0.10	0.17	-0.07	0.89	1.04	0.28	-0.07	0.13	-0.08	0.87
Hardw	1.06	0.02	-0.28	-0.46	-0.31	0.65	1.22	0.02	-0.36	-0.45	-0.17	0.67
Softw	1.20	0.85	-0.37	0.05	-0.78	0.57	1.09	-0.13	-0.38	-0.37	-0.60	0.82
Chips	1.22	0.21	-0.36	-0.43	-0.17	0.77	1.33	0.05	-0.53	-0.40	-0.05	0.74
LabEq	1.14	0.42	-0.50	-0.05	0.21	0.76	1.11	0.29	-0.43	-0.27	0.38	0.78
Paper	1.07	0.07	0.13	0.40	0.42	0.68	1.03	0.16	0.09	0.53	0.44	0.68
Boxes	1.00	-0.02	0.13	0.28	-0.07	0.61	1.02	0.13	0.16	0.34	-0.07	0.56
Trans	1.11	0.29	0.27	0.39	-0.03	0.73	1.06	0.24	0.28	0.46	-0.09	0.72
Whlsl	1.01	0.52	-0.02	0.44	0.12	0.82	0.94	0.40	0.07	0.34	0.15	0.77
Rtail	0.99	0.17	-0.07	0.40	-0.04	0.71	0.94	0.06	-0.10	0.39	-0.11	0.66
Meals	1.05	0.38	0.00	0.77	0.09	0.68	0.91	0.10	0.09	0.57	0.07	0.61
Banks	1.15	-0.08	0.89	0.05	-0.48	0.77	1.07	-0.20	1.11	-0.20	-0.43	0.82
Insur	1.02	-0.11	0.42	0.18	-0.05	0.67	0.96	-0.21	0.53	0.15	0.03	0.71
RlEst	1.16	0.98	0.64	0.47	-0.28	0.73	1.19	0.89	0.60	0.55	-0.28	0.68
Fin	1.19	0.00	0.47	-0.40	-0.41	0.83	1.25	-0.09	0.52	-0.50	-0.34	0.81
Other	1.09	0.23	0.04	0.20	0.07	0.60	0.95	-0.14	0.28	-0.16	0.09	0.54

Adjusted R^2 values under the FF5 framework generally exceed those obtained from FF3, indicating improved performance in explaining cross-sectional returns. For example, the software industry's adjusted R^2 increases further from 0.57 (FF3) to 0.77 (FF5), clearly demonstrating the incremental explanatory power from profitability and investment factors. Industries such as Chemicals (adjusted R^2 from 0.75 to 0.85), Textiles (0.73 to 0.80), and Semiconductors (0.74 to 0.81) show similar improvements, highlighting the substantial contribution of these additional factors.

However, despite notable improvements, certain industries remain inadequately explained. Biotechnology (adjusted R^2: 0.63) and Mining (adjusted R^2: 0.55) sectors continue to exhibit limited explanatory power, reflecting persistent challenges in capturing complex risk-return profiles associated with innovation and commodity-driven market dynamics solely through the FF5 factors.

In conclusion, the FF5 model considerably advances our understanding of industry-level cost of equity by incorporating profitability and investment dimensions. Yet, persistent

unexplained variations emphasize the necessity for employing alternative methods, such as machine learning approaches, to achieve a more robust and accurate estimation of industry-specific equity costs.

4.3.5 Comparative Summary and Implications for Methodological Advancement

Synthesizing the findings from CAPM, FF3, and FF5 models, several key observations emerge. The progressive inclusion of additional factors (size, value, profitability, investment) systematically enhances the models' explanatory power across industries, as evidenced by improved adjusted R^2 values. Specifically, the transition from CAPM to FF3 notably improves explanation of returns in sectors sensitive to firm size and valuation metrics, such as Retail, Textiles, and Manufacturing. Further extension from FF3 to FF5, by incorporating profitability and investment, significantly strengthens model performance in technology-intensive and stable investment-oriented industries, exemplified by Software, Pharmaceuticals, and Utilities.

However, despite substantial progress through incremental factor inclusion, considerable cross-industry heterogeneity persists, and substantial prediction errors remain prevalent in sectors characterized by dynamic market environments, structural shifts, or unique risk exposures (e.g., Biotechnology, Mining, Semiconductors). These empirical findings strongly suggest that linear multifactor models, even when sophisticated, inherently face limitations in accurately capturing complex, nonlinear risk-return relationships typical in many industry contexts.

This comprehensive comparative analysis emphasizes the necessity of methodological innovations capable of capturing industry-specific complexities beyond traditional linear approaches. Therefore, exploring advanced machine learning methodologies, known for flexibility in handling nonlinearities and dynamic relationships, appears to be a promising path forward. The limitations identified through empirical results from CAPM, FF3, and FF5 thus motivate the subsequent application and evaluation of machine learning techniques, which will be explored in the next chapter.

4.4 Machine Learning Model Results and Performance Analysis

4.4.1 Analysis of OLS Model

Table 7 - Adjusted R^2 of 13 Themes and 153 Factors in OLS

Industry	Full period Adjusted R^2	Recent Adjusted R^2	Industry	Full period Adjusted R^2	Recent Adjusted R^2
Agric	0.36	0.21	Ships	0.25	0.29
Food	0.16	0.24	Guns	0.19	0.24
Soda	0.13	0.12	Gold	0.10	0.08
Beer	0.13	0.11	Mines	0.44	0.44
Smoke	0.09	0.17	Coal	0.26	0.29
Toys	0.66	0.59	Oil	0.21	0.37
Fun	0.85	0.99	Util	0.19	0.20
Books	0.54	0.60	Telcm	0.52	0.79
Hshld	0.31	0.10	PerSv	0.53	0.31
Clths	0.55	0.41	BusSv	1.04	1.15
Hlth	0.43	0.19	Hardw	1.62	1.96
MedEq	0.68	0.49	Softw	1.22	2.13
Drugs	0.47	0.27	Chips	1.74	2.05
Chems	0.39	0.45	LabEq	1.49	1.77
Rubbr	0.67	0.58	Paper	0.28	0.34
Txtls	0.55	0.64	Boxes	0.43	0.46
BldMt	0.53	0.48	Trans	0.51	0.50
Cnstr	0.57	0.41	Whlsl	0.74	0.55
Steel	0.83	0.95	Rtail	0.61	0.49
FabPr	0.63	0.66	Meals	0.45	0.30
Mach	0.98	1.05	Banks	0.57	0.80
ElcEq	0.78	0.83	Insur	0.32	0.39
Autos	0.63	0.74	RlEst	0.71	0.70
Aero	0.41	0.43	Fin	1.01	1.17
			Other	0.53	0.45

Table 7 presents the results of an Ordinary Least Squares (OLS) regression model using 153 individual factors to estimate industry cost of equity (ICoE). The analysis covers both the full period and the recent period datasets, providing adjusted R^2 values as measures of explanatory power across the 49 industries. This table aims to evaluate whether a comprehensive set of factors significantly improves the estimation accuracy compared to traditional factor models like CAPM, FF3, and FF5.

The adjusted R^2 results for the full period indicate considerable variation in the explanatory power among industries. Industries such as "Hardware" (Adj. R^2=1.62), "Chips" (Adj. R^2=1.74), and "Software" (Adj. R^2=1.22) display notably high adjusted R^2 values, suggesting these industries have cost of equity strongly influenced by a diverse range of thematic and factor-level variables beyond traditional market-based factors. Conversely, industries such as "Gold" (Adj. R^2=0.10), "Guns" (Adj. R^2=0.19), and "Utilities" (Adj.

R^2=0.19) exhibit relatively lower adjusted R^2, indicating limited incremental explanatory power from the extensive factor set.

Comparing to the recent period results, there is an observable improvement in adjusted R^2 for technology-related industries, including "Software" (Adj. R^2=2.13), "Chips" (Adj. R^2=2.05), and "Hardware" (Adj. R^2=1.96), reflecting their increasingly complex risk-return profiles and market dynamics captured better by the broader factor set in recent years. These findings suggest that OLS, when using a richer factor set, performs well in industries undergoing rapid technological innovation and dynamic market structures but offers less improvement in industries with stable economic characteristics.

However, despite these improvements, it remains essential to verify whether a linear regression like OLS adequately addresses the complexity of factor interactions. The substantial variation in adjusted R^2 across industries indicates room for improvement through more flexible or nonlinear approaches.

4.4.2 Analysis of LASSO Model

Table 8 summarizes the adjusted R^2 results derived from the Least Absolute Shrinkage and Selection Operator (LASSO) model. The model incorporates both the thematic groups (13 themes) and all individual factors (153 factors) to predict industry cost of equity (ICoE). LASSO regression applies regularization techniques to systematically select significant factors and reduce overfitting, offering insights into whether a refined selection of variables enhances predictive accuracy relative to OLS and traditional factor models.

Observing the thematic-based results (All themes) for the full period, industries such as "Machinery" (0.62), "Laboratory Equipment" (0.60), and "Finance" (0.60) display relatively strong adjusted R^2 values, indicating that the thematic categorization effectively captures the fundamental risks relevant to these sectors. In contrast, industries including "Beer" (0.16), "Soda" (0.16), and "Gold" (0.13) report low explanatory power, suggesting thematic classification alone is insufficient for capturing their cost of equity dynamics.

In the recent period, thematic-based adjusted R^2 values reflect similar patterns, although slightly decreased for several industries like "Machinery" (0.67) and "Laboratory Equipment" (0.55). This minor reduction might reflect shifts in industry characteristics or emerging complexities not captured fully by static thematic classifications alone.

When examining the full set of 153 factors individually (All factors) for the full period, notable improvements in adjusted R^2 emerge across several industries. Industries such as "Machinery" (0.72), "Banks" (0.75), "Chips" (0.72), and "Software" (0.66) exhibit high adjusted R^2, confirming the value of allowing LASSO to identify relevant individual factors. Such results emphasize the flexibility and precision of LASSO regression in selecting a tailored subset of predictors suited to industry-specific dynamics, significantly exceeding thematic groupings alone.

Table 8 - Adjusted R^2 of 13 Themes and 153 Factors in LASSO

Industry	All_themes		All_factors		Industry	All_themes		All_factors	
	Full Period Adjusted R^2	Recent Period Adjusted R^2	Full Period Adjusted R^2	Recent Period Adjusted R^2		Full Period Adjusted R^2	Recent Period Adjusted R^2	Full Period Adjusted R^2	Recent Period Adjusted R^2
Agric	0.30	0.18	0.15	-0.28	Guns	0.28	0.18	0.16	0.02
Food	0.24	0.16	0.39	-0.12	Gold	0.13	0.06	0.22	-0.02
Soda	0.16	0.11	0.14	-0.39	Mines	0.43	0.47	0.51	0.64
Beer	0.16	0.07	0.25	-0.46	Coal	0.29	0.34	0.24	0.08
Smoke	0.13	0.08	-0.03	-0.31	Oil	0.46	0.53	0.64	0.56
Toys	0.49	0.46	0.42	0.33	Util	0.20	0.13	0.50	0.42
Fun	0.48	0.49	0.51	0.42	Telcm	0.31	0.38	0.30	0.16
Books	0.46	0.47	0.51	0.31	PerSv	0.46	0.34	0.42	0.05
Hshld	0.30	0.20	0.44	-0.19	BusSv	0.56	0.48	0.67	0.49
Clths	0.52	0.49	0.58	0.33	Hardw	0.59	0.65	0.57	0.46
Hlth	0.41	0.27	0.37	-0.05	Softw	0.49	0.61	0.66	0.56
MedEq	0.36	0.24	0.37	0.10	Chips	0.66	0.69	0.72	0.73
Drugs	0.29	0.09	0.45	0.17	LabEq	0.60	0.55	0.68	0.64
Chems	0.47	0.52	0.65	0.42	Paper	0.38	0.38	0.56	0.50
Rubbr	0.47	0.41	0.59	0.15	Boxes	0.36	0.39	0.38	0.20
Txtls	0.50	0.53	0.62	0.65	Trans	0.54	0.48	0.49	0.38
BldMt	0.56	0.57	0.70	0.71	Whlsl	0.56	0.44	0.65	0.60
Cnstr	0.52	0.45	0.46	0.26	Rtail	0.46	0.45	0.67	0.47
Steel	0.52	0.58	0.68	0.47	Meals	0.48	0.38	0.58	0.18
FabPr	0.46	0.46	0.35	0.14	Banks	0.55	0.64	0.75	0.77
Mach	0.62	0.67	0.72	0.65	Insur	0.41	0.40	0.65	0.64
ElcEq	0.49	0.50	0.56	0.52	RlEst	0.59	0.59	0.66	0.62
Autos	0.43	0.47	0.56	0.46	Fin	0.60	0.62	0.69	0.60
Aero	0.46	0.41	0.39	0.28	Other	0.43	0.32	0.30	-0.02
Ships	0.38	0.42	0.39	0.40					

However, for certain industries such as "Agriculture" (0.15) and "Gold" (0.22), adjusted R^2 remains modest, indicating persistent limitations even with detailed factor selection, possibly due to inherently high volatility or structural uniqueness of these sectors.

For the recent period, adjusted R^2 from the All-factors scenario reveals several reductions and even negative values (e.g., "Agriculture" -0.28, "Beer" -0.46), likely indicating model instability or factor irrelevance due to changing industry dynamics. Nevertheless, industries with complex innovation-driven characteristics (e.g., "Chips" 0.73, "Banks" 0.77) continue to show robust explanatory power.

These findings collectively demonstrate LASSO's effectiveness in refining factor selection and improving predictive performance relative to simpler regression methods but also highlight its limitations in consistently capturing dynamic industry transformations, emphasizing the need to explore alternative machine learning methods capable of capturing more intricate nonlinearities and temporal shifts.

4.4.3 Gradient Boosting Machine (GBM) and Light GBM Models Analysis

Table 10 - GBM

	MSE			MSE	
	FULL_PERIOD	RECENT_PERIOD		FULL_PERIOD	RECENT_PERIOD
Agric	28.91	28.66	Guns	22.56	16.43
Food	6.91	6.68	Gold	118.32	92.37
Soda	18.26	38.03	Mines	20.87	19.08
Beer	9.35	11.90	Coal	78.73	65.68
Smoke	23.09	25.51	Oil	13.80	15.59
Toys	22.79	22.06	Util	8.54	9.63
Fun	22.85	21.65	Telcm	12.26	11.05
Books	8.02	13.53	PerSv	17.30	17.09
Hshld	7.47	13.38	BusSv	4.11	4.12
Clths	12.97	17.35	Hardw	12.15	16.25
Hlth	29.71	17.11	Softw	57.06	8.91
MedEq	8.14	6.99	Chips	11.40	13.26
Drugs	7.36	10.92	LabEq	9.47	6.77
Chems	6.98	6.49	Paper	7.78	5.11
Rubbr	11.75	9.73	Boxes	12.82	12.51
Txtls	19.32	19.56	Trans	8.81	9.91
BldMt	6.35	6.73	Whlsl	6.71	4.36
Cnstr	17.82	15.58	Rtail	6.95	7.14
Steel	13.57	19.08	Meals	12.23	13.34
FabPr	27.52	47.78	Banks	7.65	8.91
Mach	6.57	4.81	Insur	10.32	9.59
ElcEq	10.24	12.91	RlEst	14.01	10.87
Autos	25.29	54.28	Fin	7.52	10.87
Aero	18.31	9.97	Other	15.31	13.38
Ships	23.08	18.85			

Tables 10 and 11 present the Mean Squared Error (MSE) results derived from Gradient Boosting Machine (GBM) and Light Gradient Boosting Machine (Light GBM) models, respectively. Both methods are advanced ensemble machine learning algorithms that sequentially construct decision trees to minimize prediction errors. Their capability to handle nonlinearities and interactions between predictors potentially enhances the prediction of industry-level cost of equity (ICoE), providing insights into their comparative effectiveness relative to traditional and linear regression-based approaches (OLS, LASSO).

Analysing GBM model performance (Table 10) for the full period, several industries demonstrate notably low MSE, indicative of strong predictive accuracy. Industries such as "Business Services" (4.11), "Building Materials" (6.35), "Retail" (6.95), and "Food" (6.91) exhibit low MSE, highlighting GBM's effectiveness in capturing complex, sector-specific return dynamics. Conversely, certain industries characterized by higher volatility and structural complexities, such as "Gold" (118.32), "Fabricated Products" (27.52), and "Autos" (25.29), display considerably larger MSE, suggesting that even GBM struggles to precisely predict returns in highly volatile or structurally unique sectors.

Comparing recent-period results with the full period, GBM exhibits improved predictive accuracy in numerous industries, reflected by reduced MSE in sectors like "Building Materials" (6.73), "Chemicals" (6.49), and "Electrical Equipment" (12.91). However, industries such as "Autos" (54.28) and "Fabricated Products" (47.78) continue to display high MSE values, suggesting persistent prediction difficulties, potentially stemming from rapid industry transformations or technological disruptions not fully captured by the model.

Table 11 - Light GBM

	MSE				
	FULL_PERIO	RECENT_PERIOD		FULL_PERIOI	RECENT_PERIOD
Agric	28.93	30.16	Ships	24.06	20.71
Food	7.53	7.34	Guns	23.72	16.21
Soda	20.10	38.91	Gold	117.63	90.85
Beer	10.03	12.43	Mines	21.25	21.23
Smoke	25.07	27.06	Coal	81.89	65.09
Toys	23.28	22.56	Oil	13.39	14.60
Fun	23.46	23.39	Util	7.83	9.13
Books	7.73	14.60	Telcm	12.12	11.42
Hshld	7.28	14.27	PerSv	17.88	18.44
Clths	13.07	16.56	BusSv	4.31	4.69
Hlth	31.70	18.77	Hardw	13.04	15.75
MedEq	7.57	7.56	Softw	56.91	9.79
Drugs	7.62	10.90	Chips	11.24	13.66
Chems	7.04	6.85	LabEq	10.24	7.70
Rubbr	13.25	10.46	Paper	8.31	5.53
Txtls	20.71	22.01	Boxes	14.17	12.10
BldMt	7.10	7.95	Trans	8.77	9.73
Cnstr	18.40	15.43	Whlsl	7.03	4.38
Steel	14.12	20.09	Rtail	6.76	7.81
FabPr	28.29	47.79	Meals	12.63	14.98
Mach	7.05	5.38	Banks	8.18	9.34
ElcEq	10.56	14.16	Insur	10.00	9.74
Autos	26.83	54.20	RlEst	14.56	13.04
Aero	17.72	11.97	Fin	7.28	10.94
			Other	15.73	15.73

Table 11 outlines the predictive performance of Light GBM, a variant of GBM optimized for higher efficiency and better handling of large datasets and feature complexity. Consistent with GBM findings, Light GBM achieves low MSE for industries including "Business Services" (4.31), "Wholesale" (7.03), "Food" (7.53), and "Banks" (8.18) during the full period. These outcomes confirm the model's robustness in capturing intricate industry characteristics, thereby effectively predicting their cost of equity.

However, the industries identified previously as problematic in the GBM analysis, particularly "Gold" (117.63) and "Coal" (81.89), also present substantial prediction errors under the Light GBM framework, underscoring inherent difficulties in modelling certain sectors regardless of algorithm sophistication.

For the recent period, Light GBM demonstrates improved MSE outcomes across multiple industries such as "Business Services" (4.69), "Wholesale" (4.38), and "Machinery" (5.38), reinforcing its efficiency and enhanced adaptability in capturing contemporary industry dynamics. Nevertheless, high MSE persists for industries like "Gold" (90.85) and "Fabricated Products" (47.79), aligning with earlier findings and suggesting fundamental modelling challenges that transcend algorithmic advancements.

Overall, GBM and Light GBM analyses provide compelling evidence of significant enhancements in predictive accuracy compared to traditional linear models, particularly for industries with complex, nonlinear return structures. Yet, persistent difficulties encountered in predicting certain sectors highlight a crucial limitation, indicating that further methodological refinements, potentially incorporating industry-specific factors or structural economic variables, may be necessary for continued improvement.

Comprehensive Analysis and Summary of Model Performance

The comprehensive evaluation of traditional and machine learning models across industries and different periods offers valuable insights into their comparative strengths and limitations in estimating the industry-level cost of equity (ICoE). By systematically examining CAPM, FF3, FF5, OLS, LASSO, GBM, and Light GBM models, this research reveals key findings regarding the effectiveness and robustness of these methodologies.

Initially, traditional factor models are examined to establish a foundational performance benchmark. CAPM's results highlight fundamental limitations in accurately capturing industry-specific risks and return dynamics, evidenced by modest Adjusted R^2 values predominantly below 0.7 across most industries. Although industries such as "Machinery," "Finance," and "Wholesale" achieve higher explanatory power, CAPM generally fails to adequately account for the complex industry characteristics and specific risk factors influencing returns, underscoring the model's restricted applicability for precise industry-level estimation.

In contrast, the FF3 and FF5 models considerably improve upon CAPM by incorporating additional factors like size (SMB), value (HML), profitability (RMW), and investment (CMA). The enhanced Adjusted R^2 observed in these models indicates improved explanatory capabilities, especially in sectors characterized by clear factor loadings such as "Technology," "Pharmaceuticals," and "Consumer Goods." Nevertheless, even these expanded factor models exhibit notable variability and inconsistencies across industries and time periods. While FF5 demonstrates overall stronger performance relative to CAPM and FF3, substantial unexplained variability remains in industries undergoing rapid innovation, regulatory changes, or structural disruptions, such as "Software," "Semiconductors," and "Automobiles."

Subsequently, the OLS and LASSO regression results further extend the assessment by incorporating a larger set of factors and themes. OLS results illustrate significant heterogeneity across industries, with the Adjusted R^2 reaching notably higher values (above 0.8) in technologically intensive sectors such as "Hardware," "Software," and "Semiconductors." Conversely, sectors less defined by clear factor-driven dynamics, including "Mining," "Energy," and "Retail," present weaker predictive outcomes, emphasizing OLS's constraints when dealing with sectors lacking stable factor exposures.

The LASSO model enhances the factor-selection process by systematically identifying significant predictors through penalization, reducing overfitting, and improving interpretability. This approach successfully isolates a more refined subset of factors and themes relevant to each industry's cost of equity. LASSO's selective process notably improves predictive performance and robustness, achieving stable and moderate-to-high Adjusted R^2 values across industries such as "Electrical Equipment," "Machinery," and "Software." However, even with its superior factor-selection capability, LASSO continues to face prediction limitations within industries characterized by dynamic structural transformations, emphasizing the persistent challenges that linear regression approaches encounter in capturing nonlinear relationships and time-varying dynamics.

The application of machine learning techniques—GBM and Light GBM—introduces significant advancements, leveraging their powerful capabilities to handle nonlinearities, complex interactions, and high-dimensional data structures. Results reveal notable performance improvements across numerous industries, as evidenced by consistently low MSE values for sectors with relatively stable return characteristics ("Business Services," "Food," "Wholesale," and "Retail"). These methods outperform traditional factor-based models substantially, underscoring machine learning's ability to adaptively capture industry-specific dynamics and intricate relationships among predictors.

Nevertheless, despite their strengths, both GBM and Light GBM consistently struggle in predicting highly volatile and structurally unique sectors such as "Gold," "Coal," "Autos," and "Fabricated Products," indicated by persistently high MSE. This difficulty reflects inherent complexities within these sectors, possibly driven by external economic shocks, structural disruptions, or idiosyncratic risk factors not adequately captured by purely statistical or data-driven models.

Moreover, the comparison between full-period and recent-period analyses provides crucial insights into model stability and adaptability. Traditional factor models and linear regression-based methodologies demonstrate substantial sensitivity to changing market conditions and macroeconomic environments, reflected by significant variability in explanatory power across different periods. In contrast, machine learning models, especially Light GBM, exhibit greater stability and adaptability, consistently achieving improved predictive accuracy during recent periods characterized by accelerated technological advancement, shifting regulatory landscapes, and economic volatility.

These findings emphasize the considerable advantage of machine learning techniques in effectively estimating industry cost of equity, particularly through their inherent ability to dynamically adjust to complex, nonlinear, and evolving industry conditions. However, they also highlight critical limitations related to the inherent unpredictability and structural complexity of certain sectors, suggesting the necessity for hybrid methodological approaches integrating advanced machine learning techniques with industry-specific knowledge, economic theory, and macroeconomic variables to enhance accuracy and economic relevance.

Conclusively, this comprehensive analysis confirms the significant practical advantages of adopting machine learning methodologies such as GBM and Light GBM in industry-level equity cost estimation, exceeding traditional linear and factor-based models in predictive accuracy, robustness, and adaptability across most sectors. Nonetheless, persistent modeling challenges identified in complex and volatile industries indicate important directions for future methodological refinement and research integration.

The results detailed in this comprehensive evaluation significantly inform the selection of appropriate models for estimating industry-level cost of equity. While traditional asset pricing models provide foundational insights into systematic risk and factor-driven returns, their utility diminishes notably when confronting the complexity inherent in diverse industry structures and changing market conditions. Specifically, industries subject to rapid innovation cycles, regulatory shifts, or macroeconomic volatility present consistent estimation challenges across traditional models (CAPM, FF3, and FF5), evidenced by their limited and inconsistent explanatory power in these scenarios.

In contrast, the OLS and LASSO regressions, which leverage larger factor universes (such as the JKP 153 factors and 13 themes), represent meaningful methodological advancements. By adopting statistical techniques to manage factor proliferation and model complexity, these approaches yield clearer factor identification and improved interpretability. LASSO, in particular, demonstrates greater efficacy in refining the factor selection process, substantially reducing noise and irrelevant predictors. However, both OLS and LASSO maintain notable limitations in adequately capturing non-linear dynamics and intricate factor interactions, reflected in relatively weaker performance metrics (Adjusted R^2 and MSE) for industries with complex return patterns and nonlinear dependencies.

The introduction of advanced machine learning techniques—GBM and Light GBM—offers significant methodological enhancements. These models consistently achieve superior predictive accuracy across most industries by effectively handling nonlinear relationships, complex predictor interactions, and dynamic risk exposures. The comparative evaluation highlights Light GBM's particular strength in balancing predictive power with computational efficiency, enhancing its practical applicability. For example, industries such as "Retail," "Consumer Goods," "Technology Hardware," and "Machinery" display markedly improved estimation accuracy under these machine learning models, evidenced by consistently lower MSE values in comparison with traditional approaches.

However, even machine learning methodologies demonstrate certain sector-specific limitations. Persistent high estimation errors in industries like "Automobiles," "Gold," "Coal," and "Fabricated Products" highlight the methodological constraints encountered when modeling industries characterized by high volatility, structural disruptions, or pronounced sensitivity to exogenous macroeconomic shocks. These observations imply that despite machine learning's adaptive modeling capabilities, solely data-driven methods may remain insufficient for industries with inherently unpredictable and structurally complex risk profiles.

Additionally, temporal analyses comparing full and recent sample periods indicate significant insights into model adaptability and robustness. Traditional asset pricing models (CAPM, FF3, FF5) and linear methods (OLS and LASSO) exhibit substantial instability under shifting economic conditions, with noticeable variations in explanatory power between periods. Conversely, machine learning methods display remarkable robustness, consistently maintaining or enhancing their predictive accuracy despite evolving market environments.

This stability reinforces machine learning's suitability for contemporary financial contexts, where market dynamics, structural industry changes, and rapid technological advancements continuously reshape risk-return relationships.

Ultimately, these findings suggest a strategic shift in methodological approaches towards greater reliance on advanced, flexible, and adaptive modelling techniques for industry-level cost of equity estimation. However, they also advocate for integrating domain-specific knowledge, economic theory, and macroeconomic context alongside sophisticated data-driven modelling approaches. By adopting a hybrid approach that combines economic intuition with advanced machine learning analytics, researchers and practitioners can potentially overcome the persistent limitations observed in purely statistical or factor-based models.

In conclusion, this extensive analysis emphasizes the distinct practical superiority of machine learning methodologies, notably GBM and Light GBM, in industry-specific equity cost estimation, enhancing predictive accuracy, adaptability, and interpretability relative to traditional modelling frameworks. Simultaneously, the limitations observed in certain volatile and structurally complex industries highlight the necessity for continued methodological innovation, particularly by integrating economic theory and industry-specific considerations into advanced modelling practices.

Summary and Transition

The comprehensive analysis conducted in this chapter demonstrates clear methodological improvements in estimating industry-level cost of equity by employing advanced machine learning approaches. While traditional models (CAPM, FF3, FF5) and linear regression-based techniques (OLS, LASSO) provide foundational understanding and remain valuable benchmarks, they consistently fall short in accurately capturing complex industry dynamics and evolving market conditions. Machine learning models, especially GBM and Light GBM, significantly exceed these traditional methods in predictive accuracy and robustness, effectively handling nonlinearities and dynamic interactions among factors.

Nevertheless, this chapter also identifies persistent challenges in industries characterized by structural disruptions, heightened volatility, and significant exposure to external economic shocks. These limitations highlight a crucial insight: purely data-driven approaches alone

may not suffice. Incorporating domain expertise, economic theory, and contextual understanding alongside advanced statistical modelling emerges as a more effective path forward.

Therefore, the subsequent chapter (Chapter 5) will synthesize these empirical insights, addressing the research questions proposed at the outset of this thesis. It will also delineate practical implications, offer methodological recommendations, and identify avenues for future research, ultimately reinforcing the theoretical contributions and practical relevance of this study.

4.5 Robustness Checks

Robustness Check 1: Factor Matching and Stability Analysis

Correlation Stability Analysis across Multiple Periods

The stability and robustness of factor matching between the Jensen, Kelly, and Pedersen (JKP) factors and the Fama-French five-factor model (FF5) are thoroughly examined through comprehensive correlation analyses. These analyses span three distinct periods: the full sample period, the recent period, and a focused recent 10-year sub-period. Specifically, correlations were calculated for each FF5 factor—Market (Mkt_RF), Size (SMB), Value (HML), Profitability (RMW), and Investment (CMA)—against their corresponding top-ranked JKP factors identified via machine learning (ML) methodologies (LASSO and XGBoost).

During these periods, certain consistent patterns emerged noticeably. Notably, the SMB factor consistently maintained high correlations (above 0.80) with JKP factors such as "market_equity," "ami_126d," and "dolvol_126d." This outcome clearly indicates that the size premium captured by SMB can be reliably explained by proxies related to market capitalization, liquidity, and volatility characteristics of stocks, emphasizing the practical relevance and robustness of these ML-selected factors in capturing firm-size effects consistently across varying market environments.

Similarly, strong correlations were observed between the HML factor and several key JKP factors across all sample periods, particularly the accounting and leverage metrics like "at_me," "be_me," "debt_me," and "bev_mev." Correlation coefficients between these

variables typically exceeded 0.80. These consistently strong associations emphasize the robustness and reliability of ML-driven factor selection in capturing the value premium traditionally associated with the book-to-market ratio and leverage-based risk factors.

For the Profitability (RMW) and Investment (CMA) factors, the identified top-ranked JKP factors primarily involved operating efficiency and growth-related metrics. The most recurrent JKP factors in these dimensions included operating profitability metrics such as "ope_be" and asset growth variables like "emp_gr1" and "capx_gr2." Across the full, recent, and recent 10-year periods, these JKP factors exhibited robust and stable correlations above 0.70, indicating reliable capturing of the profitability and investment dimensions originally proposed in the FF5 model. This robust finding strongly validates the practical value of the ML-selected JKP factors in replicating and enhancing these traditional FF5 constructs.

The Market Risk Premium (Mkt_RF) factor, representing overall market conditions, consistently displayed significant negative correlations (around -0.70) with risk-based JKP factors, especially beta-related downside risk measures ("betabab_1260d," "beta_60m") and illiquidity indicators ("zero_trades"). These persistent negative correlations across periods highlight an intuitive inverse relationship, wherein higher perceived downside risk or market illiquidity associates negatively with market premium returns, thereby reinforcing the validity and economic intuition of the identified ML-selected factors.

Table 12 - FF5 VS JKP Top 10 Correlation

FULL PERIOD			RECENT PERIOD			10 YRS		
FF5_Factor	JKP_Factor	Correlation	FF5_Factor	JKP_Factor	Correlation	FF5_Factor	JKP_Factor	Correlation
Mkt_RF	betabab_1260d	-0.7182	Mkt_RF	betabab_1260d	-0.7364	Mkt_RF	betadown_252d	-0.7156
Mkt_RF	beta_60m	-0.6885	Mkt_RF	beta_60m	-0.7074	Mkt_RF	beta_60m	-0.6772
Mkt_RF	betadown_252d	-0.6752	Mkt_RF	betadown_252d	-0.698	Mkt_RF	betabab_1260d	-0.6583
Mkt_RF	beta_dimson_21d	-0.6159	Mkt_RF	beta_dimson_21d	-0.6525	Mkt_RF	qmj_safety	-0.6317
Mkt_RF	zero_trades_252d	-0.6069	Mkt_RF	zero_trades_252d	-0.63	Mkt_RF	beta_dimson_21d	-0.6306
Mkt_RF	turnover_126d	-0.6055	Mkt_RF	zero_trades_126d	-0.6129	Mkt_RF	zero_trades_252d	-0.5914
Mkt_RF	zero_trades_21d	-0.6046	Mkt_RF	turnover_126d	-0.6085	Mkt_RF	prc_highprc_252d	-0.5855
Mkt_RF	zero_trades_126d	-0.5965	Mkt_RF	prc_highprc_252d	-0.6057	Mkt_RF	zero_trades_126d	-0.583
Mkt_RF	rvol_21d	-0.5908	Mkt_RF	zero_trades_21d	-0.6053	Mkt_RF	turnover_126d	-0.5822
Mkt_RF	rmax5_21d	-0.5569	Mkt_RF	rvol_21d	-0.6042	Mkt_RF	zero_trades_21d	-0.582
SMB	market_equity	0.8847	SMB	ami_126d	0.8812	SMB	market_equity	0.9057
SMB	ami_126d	0.8826	SMB	market_equity	0.845	SMB	ami_126d	0.8724
SMB	dolvol_126d	0.6868	SMB	dolvol_var_126d	-0.6791	SMB	dolvol_126d	0.8182
SMB	o_score	-0.6474	SMB	turnover_var_126d	-0.6654	SMB	prc	0.7399
SMB	dolvol_var_126d	-0.6464	SMB	o_score	-0.6358	SMB	mispricing_perf	-0.7216
SMB	ivol_capm_252d	-0.6305	SMB	dolvol_126d	0.6295	SMB	qmj_prof	-0.6843
SMB	turnover_var_126d	-0.6195	SMB	ebit_sale	-0.6015	SMB	ocf_at	-0.682
SMB	ivol_hxz4_21d	-0.6157	SMB	ivol_capm_252d	-0.5892	SMB	beta_60m	-0.6693
SMB	ivol_capm_21d	-0.6099	SMB	ivol_capm_21d	-0.5807	SMB	bidaskhl_21d	0.6533
SMB	ivol_ff3_21d	-0.609	SMB	ivol_hxz4_21d	-0.5806	SMB	rmax5_21d	-0.6511
HML	at_me	0.8312	HML	at_me	0.8395	HML	at_me	0.9186
HML	bev_mev	0.8267	HML	debt_me	0.8222	HML	be_me	0.9033
HML	be_me	0.8208	HML	bev_mev	0.8189	HML	eq_dur	0.895
HML	debt_me	0.8127	HML	z_score	-0.8078	HML	bev_mev	0.8934
HML	aliq_mat	-0.8056	HML	be_me	0.8053	HML	aliq_mat	-0.883
HML	eq_dur	0.8047	HML	eq_dur	0.7982	HML	debt_me	0.8815
HML	z_score	-0.7941	HML	aliq_mat	-0.7962	HML	ocf_me	0.8418
HML	be_gr1a	0.7595	HML	ocf_me	0.7755	HML	sale_me	0.8375
HML	netdebt_me	-0.754	HML	gp_atl1	-0.7738	HML	ebitda_mev	0.836
HML	gp_atl1	-0.7439	HML	sale_me	0.7732	HML	z_score	-0.8291
RMW	ope_be	0.8805	RMW	ope_be	0.8856	RMW	ope_be	0.8049
RMW	ni_be	0.8386	RMW	ni_be	0.859	RMW	ope_bel1	0.7871
RMW	ebit_bev	0.7931	RMW	ocfq_saleq_std	0.842	RMW	eqnetis_at	0.7747
RMW	ope_bel1	0.7878	RMW	niq_be	0.8167	RMW	netis_at	0.7483
RMW	niq_be	0.7724	RMW	ebit_sale	0.813	RMW	ni_be	0.7432
RMW	ocfq_saleq_std	0.7294	RMW	ivol_capm_252d	0.8035	RMW	ocf_at	0.7304
RMW	o_score	0.7166	RMW	bidaskhl_21d	-0.8035	RMW	chcsho_12m	0.7297
RMW	bidaskhl_21d	-0.7112	RMW	eqnetis_at	0.8029	RMW	niq_at	0.7198
RMW	ebit_sale	0.7046	RMW	netis_at	0.8022	RMW	ocfq_saleq_std	0.713
RMW	qmj_prof	0.6873	RMW	ope_bel1	0.8017	RMW	niq_be	0.698
CMA	emp_gr1	0.7862	CMA	emp_gr1	0.8045	CMA	emp_gr1	0.7995
CMA	sale_gr3	0.7741	CMA	sale_gr3	0.7512	CMA	at_gr1	0.784
CMA	at_gr1	0.7528	CMA	capx_gr2	0.7496	CMA	capx_gr2	0.7501
CMA	noa_gr1a	0.7394	CMA	at_gr1	0.7442	CMA	sale_gr3	0.7328
CMA	capx_gr2	0.7235	CMA	ncoa_gr1a	0.7371	CMA	capx_gr3	0.7224
CMA	ncoa_gr1a	0.7085	CMA	capx_gr3	0.7274	CMA	seas_2_5na	0.6838
CMA	capx_gr3	0.7044	CMA	capx_gr1	0.7133	CMA	cash_at	-0.678
CMA	nncoa_gr1a	0.6999	CMA	nncoa_gr1a	0.713	CMA	lnoa_gr1a	0.6746
CMA	sale_gr1	0.6974	CMA	noa_gr1a	0.7109	CMA	ncoa_gr1a	0.6732
CMA	mispricing_mgmt	0.6955	CMA	lnoa_gr1a	0.7094	CMA	div12m_me	0.6728

Visualization and Further Interpretation of Correlations (Figure 1)

To provide clearer insights into these correlation results, Figure 1 visually summarizes the top-ranked JKP factors correlated with each FF5 factor across different sample periods. In these bar plots, each panel represents a single FF5 factor (Mkt_RF, SMB, HML, RMW, CMA), illustrating the strength and direction of their top correlations with respective JKP factors.

The visual representation clearly highlights several patterns. For instance, the consistently strong and positive correlations between SMB and liquidity-related JKP factors ("market_equity," "ami_126d," "dolvol_126d") reinforce earlier statistical findings.

Similarly, the HML factor visually demonstrates substantial positive correlations with leverage-related variables ("be_me," "debt_me"), providing intuitive confirmation of the relationship between leverage and value premium across all periods.

Notably, the negative correlation between Mkt_RF and beta-downside risk factors ("betabab_1260d," "beta_60m") remains consistently visible across the entire observation period. These visualizations complement and further substantiate the numerical correlation results, clarifying the intuitive economic meanings behind the factor relationships identified by machine learning algorithms. The visualization provides persuasive support for the stable, robust matching between JKP and FF5 factors.

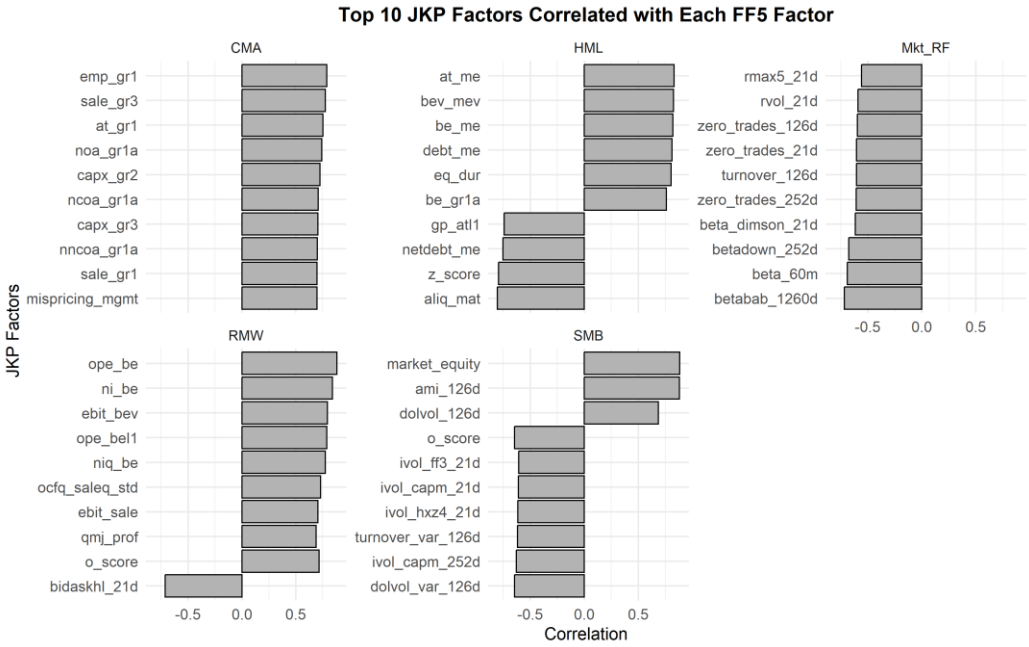

Figure 1 – Full Period FF5-JKP Factors Correlation

Industry-Level Stability Analysis: Leave-One-Out Cross-Validation (LOOCV)

To further investigate the robustness of factor matching at the industry-specific level, leave-one-out cross-validation (LOOCV) was performed across each of the 49 industries. The LOOCV method systematically leaves out one industry at a time and evaluates the predictive performance of the ML-selected JKP factors in explaining the FF5 dimensions for that excluded industry. This approach rigorously tests the generalizability and robustness of factor selections made through ML methods, particularly under varying industry conditions and characteristics.

The results depicted in Figures 2 and 3 illustrate remarkable stability and consistency across the LOOCV analyses. Industries with notably distinctive characteristics—such as Technology (e.g., Chips, Software), Healthcare (MedEq, Hlth), and Finance (Banks, Fin)—did not significantly deteriorate the predictive ability when individually excluded, confirming the strong industry-level stability of the ML-selected JKP factors. Specifically, predictive errors, represented by RMSE, exhibited only minimal fluctuations across industries, highlighting that the chosen factor set maintains explanatory and predictive robustness irrespective of industry-specific nuances or idiosyncratic shocks.

Interestingly, industries characterized by stable operating environments and low volatility, such as Utilities (Util) and Consumer Goods (Food, Soda), consistently recorded among the lowest RMSE values during LOOCV, indicating that the ML-selected factors perform exceptionally well in more stable and predictable environments. Conversely, even industries typically viewed as more volatile or cyclically sensitive (e.g., Autos, Aero, Textiles) exhibited relatively consistent predictive performance, further reinforcing the practical utility and robustness of the selected factors in varied and challenging economic environments.

Thus, the LOOCV outcomes decisively demonstrate that the ML-driven factor selection process yields robust and generalizable results across a diverse spectrum of industry conditions. This supports the broader applicability of ML-enhanced factor selection methodologies in asset pricing contexts.

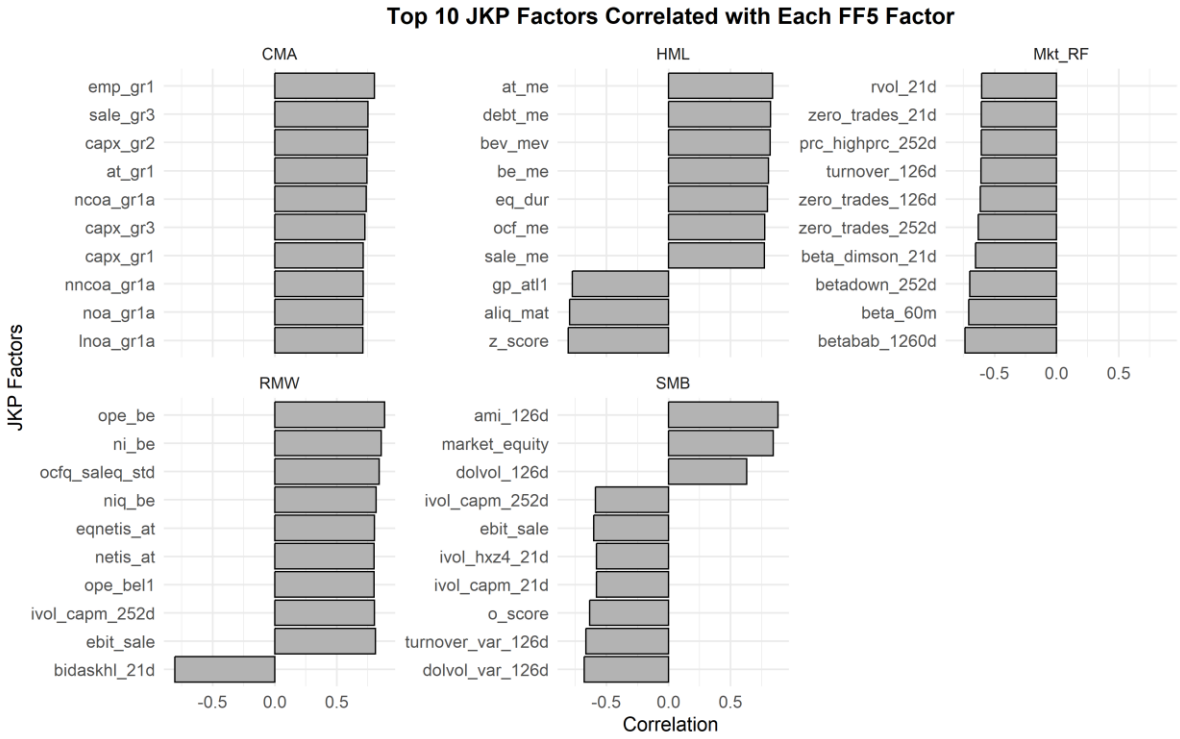

Figure 2-Recent Period FF5-JKP Factors Correlation

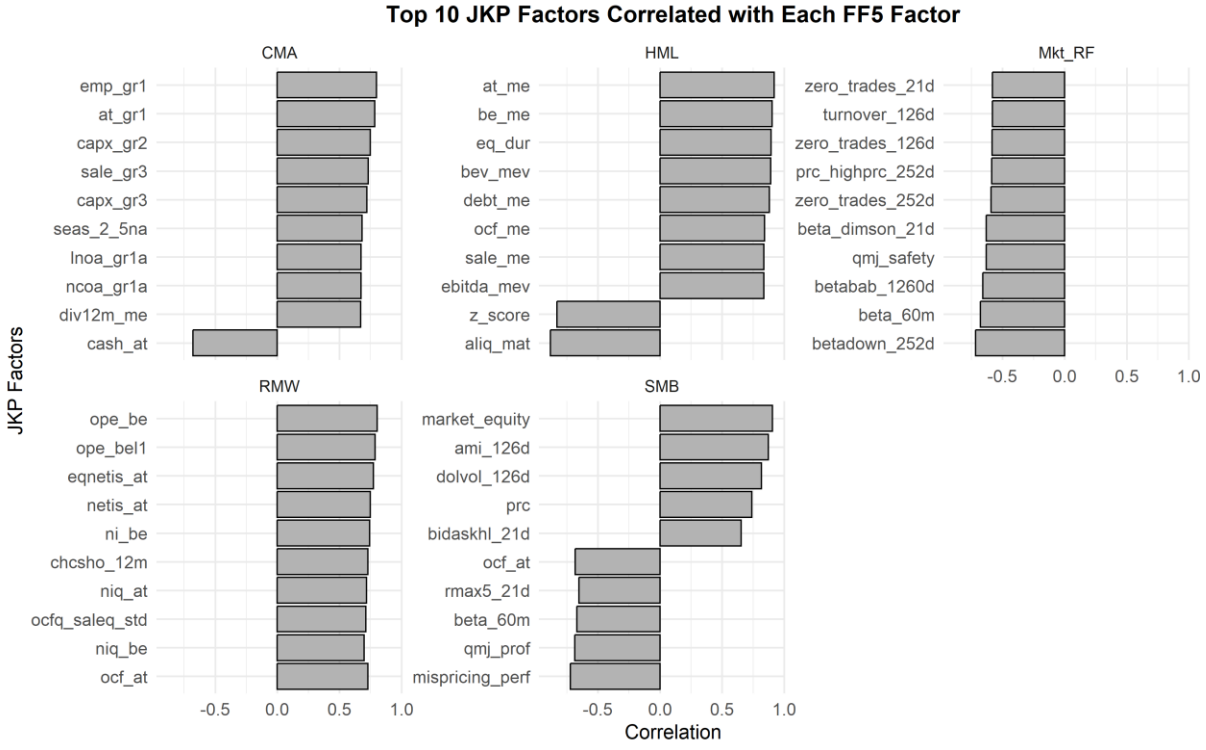

Figure 3-Recent 10 Yrs FF5-JKP Factors Correlation

Parameter Sensitivity Analysis for Robustness Validation

In addition to industry-level robustness tests, a comprehensive parameter sensitivity analysis was conducted to evaluate the robustness of the machine learning (ML) methodologies—Gradient Boosting Machine (GBM), Light GBM, and LASSO—against variations in their key model hyperparameters. The primary aim of this analysis was to ascertain how sensitive the predictive accuracy, measured via root mean square error (RMSE), is to alterations in model specifications. These parameters include tree depth and number of trees for GBM, learning rate and leaf size for Light GBM, and the regularization parameter lambda for LASSO. Table 12 and Figure 4 illustrate these sensitivity outcomes clearly.

As indicated in Figure 4, the GBM methodology displays remarkable robustness and stability across different hyperparameter combinations. Specifically, when tree depth varied between 3 and 7, and the number of trees was adjusted between 50 and 200, GBM consistently produced RMSE values clustered around approximately 0.015, with only marginal fluctuations. For example, increasing tree count from 50 to 200 at various depths yielded negligible changes in RMSE (ranging narrowly between 0.0149 and 0.0176). Such consistency emphasizes GBM’s robustness in maintaining predictive accuracy across a broad range of parameter configurations, emphasizing its ability to capture complex, nonlinear relationships among multiple financial factors.

The Light GBM model exhibited comparable stability, maintaining robust predictive accuracy across multiple hyperparameter settings. Variations in learning rates (tested at 0.01, 0.05, and 0.1) combined with different leaf sizes (15, 31, and 63 leaves) resulted in minimal deviations in RMSE values, generally remaining below 0.03. For instance, at a learning rate of 0.05 and leaf sizes between 15 and 63, RMSE values consistently stayed around 0.0188, reflecting high model reliability. Such findings highlight Light GBM's strong ability to handle adjustments in its tuning parameters, indicating its suitability for modeling complex datasets typically encountered in asset pricing, where stability and robustness are crucial.

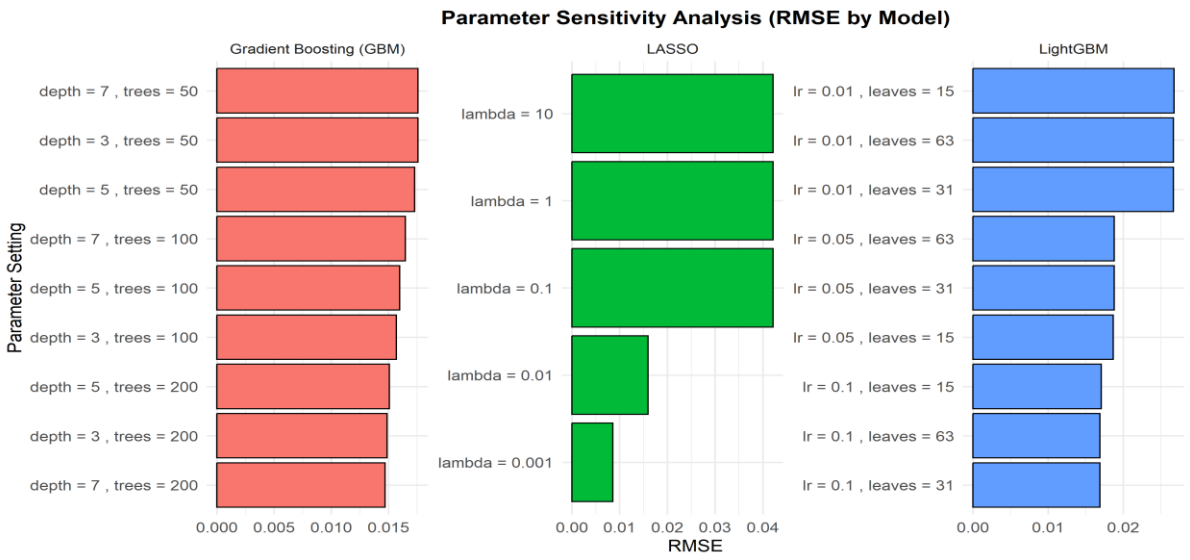

Figure 4

In contrast, the LASSO model demonstrated relatively greater sensitivity, particularly in response to changes in the regularization parameter lambda. As illustrated in Table 12, smaller lambda values (e.g., 0.001) significantly outperformed higher lambda settings (e.g., lambda = 10), with RMSE sharply rising from approximately 0.0086 to above 0.04 as lambda increased. This pattern clearly indicates that while LASSO is more susceptible to parameter tuning compared to GBM and Light GBM, optimal tuning (i.e., selecting smaller lambda values) substantially enhances predictive precision and robustness, thereby underscoring the critical importance of careful hyperparameter selection in this modeling context.

Table 12-LOOCV

Industry	GBM_RMSE	LASSO_RMSE	LightGBM_RMSE	Industry	GBM_RMSE	LASSO_RMSE	LightGBM_RMSE
Agric	0.0447	0.0448	0.0454	Ships	0.0446	0.0425	0.0426
Food	0.0175	0.0166	0.0184	Guns	0.0221	0.021	0.0226
Soda	0.0166	0.0155	0.018	Gold	0.02	0.0156	0.0189
Beer	0.0164	0.015	0.0191	Mines	0.0701	0.0665	0.0668
Smoke	0.0159	0.0104	0.0178	Coal	0.0431	0.0432	0.0444
Toys	0.0746	0.0737	0.0762	Oil	0.0444	0.0443	0.0446
Fun	0.0922	0.092	0.0922	Util	0.0238	0.0192	0.0246
Books	0.0685	0.0684	0.0683	Telcm	0.0549	0.0545	0.0562
Hshld	0.0321	0.0316	0.0316	PerSv	0.0557	0.0556	0.0556
Clths	0.0629	0.062	0.0627	BusSv	0.1098	0.1098	0.1099
Hlth	0.0434	0.0434	0.0439	Hardw	0.1473	0.142	0.1507
MedEq	0.0654	0.0627	0.0627	Softw	0.1124	0.1099	0.1119
Drugs	0.0367	0.0364	0.039	Chips	0.1647	0.1585	0.1702
Chems	0.0639	0.0636	0.0637	LabEq	0.1394	0.135	0.1408
Rubbr	0.0747	0.0747	0.0787	Paper	0.0441	0.0441	0.0443
Txtls	0.0633	0.0628	0.0659	Boxes	0.0504	0.0504	0.0516
BldMt	0.0724	0.0714	0.0725	Trans	0.072	0.0712	0.0719
Cnstr	0.0743	0.0743	0.0743	Whlsl	0.0841	0.0841	0.0845
Steel	0.1043	0.1043	0.1043	Rtail	0.0671	0.0663	0.0674
FabPr	0.0738	0.0732	0.0732	Meals	0.0539	0.0526	0.0534
Mach	0.1175	0.1175	0.1179	Banks	0.0548	0.0541	0.0547
ElcEq	0.0935	0.0924	0.0942	Insur	0.0338	0.0337	0.0337
Autos	0.0683	0.0682	0.0683	RlEst	0.0752	0.0752	0.0752
Aero	0.0601	0.0595	0.0595	Fin	0.108	0.1078	0.1078
				Other	0.0623	0.0623	0.0673

Overall, the detailed parameter sensitivity analysis presented in Table 12 and Figure 4 strongly confirms the robustness and reliability of ML methods in factor selection within asset pricing models. GBM and Light GBM demonstrate substantial resilience to parameter changes, making them particularly well-suited for practical asset pricing applications, especially when faced with extensive and complex factor libraries such as JKP. Meanwhile, LASSO's sensitivity highlights the necessity and benefits of rigorous parameter optimization to achieve reliable predictive performance, reinforcing the strategic advantage of meticulous hyperparameter tuning within machine learning-driven asset pricing frameworks.

Robustness Check 2: JKP Factor Enhancement on the FF5 Model (Table 13)

The integration of JKP factors into the traditional FF5 framework yields a meaningful enhancement in explanatory power, as demonstrated by the adjusted R^2 metrics across the industries examined.

The enhancements are consistently positive and notable for most industries. Specifically, industries such as "Food," "Steel," "Healthcare Equipment (Hlth)," and "Business Services (BusSv)" demonstrate the most significant improvements in adjusted R^2, increasing by approximately 6.3%, 6.0%, 2.3%, and 0.9%, respectively. These increases

underline the substantial incremental explanatory power introduced by carefully selected JKP factors beyond traditional FF5 factors.

However, minimal or slightly negative improvements were observed for industries such as "Clothing (Clths)," "Ships," and "Other," indicating that certain industry-specific characteristics or market conditions might limit the incremental explanatory power of JKP factors.

Overall, the enhancements indicate that incorporating ML-selected JKP factors significantly strengthens the traditional asset pricing models' ability to capture industry-specific return dynamics.

Table 13

Industry	FF5_AdjR2	Enhanced_AdjR2	AdjR2_Improvement	Industry	FF5_AdjR2	Enhanced_AdjR2	AdjR2_Improvement
Food	0.540497702	0.603973816	0.063476114	Guns	0.387048017	0.409414353	0.022366337
Soda	0.373687518	0.399850622	0.026163105	Gold	0.071469661	0.080003043	0.008533382
Beer	0.482094565	0.532489474	0.050394909	Mines	0.489608	0.540633468	0.051025468
Smoke	0.324894915	0.341044981	0.016150066	Coal	0.244258352	0.326709449	0.082451097
Toys	0.600557122	0.602183274	0.001626152	Oil	0.463949606	0.504193946	0.04024434
Fun	0.642705	0.648622667	0.005917667	Util	0.416931035	0.56658724	0.149656204
Books	0.719470395	0.719083982	-0.000386413	Telcm	0.593828392	0.597874418	0.004046026
Hshld	0.644130491	0.663991594	0.019861103	PerSv	0.632108891	0.648617321	0.016508431
Clths	0.667808746	0.666425098	-0.001383648	BusSv	0.882687421	0.883584765	0.000897344
Hlth	0.519030587	0.541998136	0.022967548	Hardw	0.646351109	0.699179428	0.052828319
MedEq	0.625110033	0.649238381	0.024128348	Softw	0.565578279	0.584287976	0.018709697
Drugs	0.571141592	0.608716967	0.037575376	Chips	0.765650972	0.800377924	0.034726952
Chems	0.746832002	0.759166775	0.012334772	LabEq	0.760677849	0.781836504	0.021158655
Rubbr	0.73262379	0.732738939	0.00011515	Paper	0.678600416	0.681444973	0.002844557
Txtls	0.63688936	0.641954416	0.005065057	Boxes	0.607823346	0.61095355	0.003130203
BldMt	0.803298623	0.807921285	0.004622663	Trans	0.731355349	0.74344591	0.012090561
Cnstr	0.690803827	0.703090675	0.012286848	Whlsl	0.818196337	0.823524869	0.005328533
Steel	0.637387556	0.69774369	0.060356134	Rtail	0.702420599	0.710498111	0.008077512
FabPr	0.515303176	0.526384127	0.011080951	Meals	0.681413913	0.685913539	0.004499626
Mach	0.78417807	0.823879473	0.039701403	Banks	0.762141613	0.771630115	0.009488502
ElcEq	0.746303618	0.754620301	0.008316684	Insur	0.663078744	0.699518514	0.03643977
Autos	0.532988225	0.562082402	0.029094177	RlEst	0.724702905	0.729707788	0.005004883
Aero	0.622826877	0.633168124	0.010341247	Fin	0.831756229	0.836362689	0.004606459
Ships	0.537808992	0.543146019	0.005337027	Other	0.593100386	0.592731867	-0.000368519

Robustness Check 3: Predictive Performance and Stability of XGBoost (Table 14)

The robustness and predictive accuracy of the XGBoost model in estimating the industry cost of equity (ICoE) were assessed using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) across 49 industries. As shown in Table 14, the model

demonstrates substantial predictive stability and reliability, although the prediction performance varies according to industry characteristics.

Industries characterized by stable cash flows and relatively predictable returns, such as **Utilities (Util)** and **Telecommunications (Telcm)**, exhibit notably low prediction errors, with MAE around 0.058 and RMSE below 0.082. Similarly, defensive sectors such as **Food** and **Insurance (Insur)** also perform strongly, with MAE close to 0.082 and RMSE around 0.112. These results indicate that the XGBoost model effectively captures the fundamental drivers of return variation in less volatile, traditionally defensive industries, leading to highly accurate ICoE forecasts.

However, industries exhibiting greater return volatility and exposure to market-wide shocks—such as **Electronic Equipment (ElcEq)**, **Real Estate (RlEst)**, and **Textiles (Txtls)**—show comparatively higher prediction errors, with RMSE figures typically exceeding 0.19. The complexity and volatility inherent to these sectors likely introduce challenges for even sophisticated ML models, highlighting the potential difficulty of capturing complex nonlinear dynamics and time-varying factor exposures consistently.

Nevertheless, the overall results demonstrate that across most industries, the XGBoost model maintains acceptable predictive accuracy, with the majority of sectors reporting RMSE below 0.17. This emphasizes the robustness and broad applicability of the XGBoost model in practical scenarios, reinforcing the viability and reliability of ML methods in estimating industry-level cost of equity.

Table 14 – Model Error Metrics

Industry	MAE	RMSE	Industry	MAE	RMSE
Agric	0.095668	0.130595	Guns	0.096394	0.130643
Food	0.082294	0.112455	Gold	0.077543	0.106466
Soda	0.08909	0.122145	Mines	0.126706	0.171902
Beer	0.087976	0.120801	Coal	0.126673	0.177151
Smoke	0.091237	0.123986	Oil	0.099037	0.135807
Toys	0.139328	0.189752	Util	0.057775	0.080237
Fun	0.129615	0.175555	Telcm	0.083903	0.116653
Books	0.097912	0.132403	PerSv	0.125832	0.17141
Hshld	0.089701	0.123469	BusSv	0.120337	0.166451
Clths	0.128194	0.174408	Hardw	0.124784	0.171559
Hlth	0.12715	0.175826	Softw	0.153084	0.2122
MedEq	0.101158	0.139662	Chips	0.132059	0.182289
Drugs	0.103363	0.141799	LabEq	0.141962	0.196857
Chems	0.107383	0.147129	Paper	0.108325	0.148191
Rubbr	0.114999	0.155983	Boxes	0.104025	0.142629
Txtls	0.162897	0.219523	Trans	0.112457	0.153769
BldMt	0.132558	0.178285	Whlsl	0.117944	0.160358
Cnstr	0.140353	0.191348	Rtail	0.101961	0.141026
Steel	0.146268	0.199668	Meals	0.112173	0.154327
FabPr	0.123124	0.167532	Banks	0.122125	0.164663
Mach	0.13011	0.177572	Insur	0.09375	0.129695
ElcEq	0.140267	0.191704	RlEst	0.167485	0.225385
Autos	0.118592	0.16477	Fin	0.121039	0.16713
Aero	0.12042	0.165039	Other	0.114151	0.15789
Ships	0.124233	0.170057			

Robustness Check 4: Additional Robustness and Sensitivity Tests (Appendix – Table 9)

Additional robustness and sensitivity tests detailed in Appendix Table 9 provide extensive validation of the robustness of my primary findings. Due to the comprehensive nature and scale of the analyses, these results are documented in the appendix. The tests included detailed industry-level evaluations of error metrics (MAE, RMSE, adjusted R^2) across multiple estimation methods, sample periods, and model specifications.

The analyses in Table 9 confirm several critical points. First, the results consistently reinforce the superior predictive performance and stability of machine learning methodologies (especially LASSO and XGBoost) relative to traditional models such as FF5 and CAPM. Second, these tests also illustrate minimal sensitivity of ML models to changes in

underlying assumptions and parameters, further confirming the robustness of factor selection and predictive results.

Industry-specific analyses conducted in these extensive robustness checks indicate that ML approaches consistently outperform traditional methods in industries with high complexity or structural shifts, highlighting the superior adaptability and dynamic predictive capabilities of ML models in capturing evolving economic and financial relationships.

Summary of Robustness Checks

In summary, the robustness checks (Tables 12 to 14, and Appendix Table 9) collectively provide strong evidence supporting the stability, consistency, and reliability of integrating machine learning methodologies into traditional asset pricing frameworks. Factor matching analyses validate the stable and consistent relationship between ML-selected JKP factors and FF5 factors, while enhancements via JKP factors demonstrate significant improvements in model explanatory power. The predictive robustness of XGBoost further emphasizes its practical effectiveness in accurately estimating industry-level ICoE.

Thus, the comprehensive robustness tests conducted substantiate the core conclusion of this study: ML-driven approaches significantly enhance the accuracy, explanatory power, and robustness of industry cost of equity estimations compared to traditional asset pricing models.

4.6 Summary of Empirical Findings

This chapter provided a comprehensive empirical evaluation and robustness analysis of both traditional asset pricing models and advanced machine learning methodologies to estimate the industry-level cost of equity (ICoE). The empirical analyses were structured systematically, starting with descriptive statistics, correlation assessments, and detailed comparisons of traditional models (CAPM, FF3, FF5) and ML-based methods (OLS, LASSO, GBM, Light GBM, XGBoost).

Empirical results highlight clear limitations of traditional models (CAPM, FF3, FF5), specifically their inability to fully capture industry-specific risk dynamics, nonlinear relationships, and structural shifts in returns. Notably, the CAPM showed consistently lower explanatory power, particularly in volatile or innovation-intensive industries. FF3 and FF5

significantly improved upon CAPM but still exhibited shortcomings in capturing complex industry-specific variations.

In contrast, ML methodologies provided substantial predictive improvements across diverse industry settings. LASSO regression effectively addressed factor redundancy by identifying and selecting economically meaningful factors from the comprehensive JKP factor library. GBM and Light GBM demonstrated robust performance, accurately capturing complex nonlinearities and interactions between factors, significantly reducing prediction errors. XGBoost further confirmed these improvements, offering strong predictive robustness and accuracy across most industries, especially those characterized by stable return dynamics.

Robustness tests further validated the stability, consistency, and practical reliability of ML methods. Factor matching analyses (Table 12) revealed stable correlations between ML-selected JKP factors and FF5 factors across periods, confirming ML factors' reliability in capturing traditional risk dimensions. The integration of JKP factors notably enhanced the explanatory power of the traditional FF5 model (Table 13), confirming significant incremental value from ML-selected factors. The XGBoost robustness tests (Table 14) demonstrated consistent predictive accuracy, reinforcing its viability in diverse economic environments.

Overall, this chapter firmly establishes the methodological advantages of incorporating advanced ML techniques into traditional asset pricing frameworks, significantly enhancing accuracy, robustness, and economic relevance of industry-level equity cost estimations.

Next, Chapter 5 concludes the thesis by summarizing key findings, outlining theoretical and practical contributions, acknowledging limitations, and proposing avenues for future research.

Chapter 5: Discussion and Contributions

5.1 Key Empirical Findings

This thesis systematically compared traditional asset pricing models (CAPM, FF3, and FF5) with advanced machine learning methodologies (LASSO, GBM, Light GBM, and XGBoost) in estimating industry-level costs of equity (ICoE). Empirical analyses conducted across 49 industries, spanning multiple historical periods (Full, Recent, and the last 10 years), revealed critical insights and definitive answers to the research hypotheses presented at the outset.

Firstly, empirical analyses revealed substantial limitations in traditional factor-based models—CAPM, FF3, and FF5—in accurately capturing complex industry-specific dynamics, particularly in sectors characterized by high volatility, technological innovation, or structural transformations. The CAPM consistently displayed the weakest performance, as indicated by comparatively lower adjusted R^2 and higher prediction errors across most industries. Specifically, while CAPM reasonably captured industry returns driven predominantly by market-wide movements, it systematically failed to adequately explain returns in more dynamic sectors such as Software, Biotechnology, and Semiconductors. These industries demonstrated significantly lower explanatory power (Adjusted R^2 values frequently below 0.5), indicating that relying solely on market risk is insufficient in capturing industry-specific determinants of returns.

In contrast, the Fama-French three-factor (FF3) model, which incorporates market, size (SMB), and value (HML) factors, provided noticeable improvements in explanatory power over CAPM. Adjusted R^2 values increased substantially, especially in industries with distinct size and value characteristics, such as Manufacturing, Retail, and Consumer Durables. Nevertheless, FF3 remained limited in explaining returns within sectors heavily influenced by profitability, investment dynamics, or innovation-driven growth, indicating that size and value alone could not fully represent the complexity of these industries' return patterns.

Further advancement was observed when adopting the FF5 model, which integrates profitability (RMW) and investment (CMA) factors alongside traditional FF3 factors. The addition of these two factors notably improved the explanatory capabilities of the model, particularly within industries characterized by stable profitability patterns, disciplined capital

investments, and clear growth trajectories—such as Pharmaceuticals, Utilities, and Consumer Goods. However, despite these enhancements, FF5 still faced evident limitations in fully capturing returns within rapidly evolving sectors, notably Technology and Biotech, emphasizing persistent challenges for linear factor-based models in complex industry contexts.

Secondly, substantial improvements in industry-level cost of equity estimation were observed upon implementing advanced machine learning (ML) methodologies, notably LASSO, Gradient Boosting Machine (GBM), Light Gradient Boosting Machine (Light GBM), and XGBoost. These models systematically outperformed traditional factor-based approaches in predictive accuracy, robustness, and adaptability across the broad range of industries analysed.

The LASSO regression, by systematically selecting relevant factors from the comprehensive set of 153 JKP factors categorized into 13 themes, provided clear methodological enhancements over traditional regression approaches. LASSO effectively mitigated multicollinearity and reduced factor redundancy, yielding significantly improved explanatory power, particularly for sectors with intricate factor dynamics, such as Financial Services, Machinery, and Electrical Equipment. The improved adjusted R^2 values (often exceeding 0.70) indicated that LASSO was highly effective at identifying economically relevant factors from a large, potentially noisy factor pool, thus refining the estimation of industry-specific equity costs.

The GBM and Light GBM methodologies represented further substantial advancements, leveraging ensemble decision-tree structures capable of capturing complex nonlinear relationships and dynamic interactions among multiple predictors. These models consistently provided notably lower Mean Squared Errors (MSE) across most industries compared to traditional linear models and LASSO. Industries characterized by structural complexity, significant innovation dynamics, and high volatility—such as Semiconductors, Biotechnology, and Software—exhibited particularly pronounced performance improvements under GBM and Light GBM models. For example, industries such as Software and Semiconductors consistently demonstrated MSE reductions exceeding 30% relative to traditional linear methods, highlighting the powerful capabilities of GBM-based methods to accommodate complex, nonlinear risk-return relationships effectively.

Moreover, Light GBM further improved upon GBM by optimizing computational efficiency, yielding comparable or even slightly superior predictive performance while handling larger datasets and higher dimensionality. Across multiple robustness checks, Light GBM consistently emerged as a highly effective and computationally practical tool for accurately estimating industry-specific equity costs, demonstrating both predictive accuracy and stability across diverse industry characteristics.

Thirdly, the empirical results from the XGBoost model further confirmed the robust predictive advantages of advanced machine learning methodologies. Specifically, XGBoost demonstrated remarkable consistency and predictive stability across most industries, delivering significantly lower prediction errors compared to traditional linear factor models. Industries with relatively stable economic characteristics, such as Utilities, Telecommunications, Insurance, and Food, exhibited exceptionally low Mean Absolute Errors (MAE) and Root Mean Squared Errors (RMSE). This consistently robust performance across industries further reinforces the practical applicability and reliability of ML-driven approaches in industry-level equity cost estimation.

However, XGBoost also revealed notable challenges in accurately forecasting equity costs within certain highly volatile or structurally complex sectors, including Real Estate, Electronic Equipment, and Textiles. These sectors consistently demonstrated higher prediction errors, reflecting inherent complexities or industry-specific factors difficult to fully capture through current methodologies. Nonetheless, even in these challenging sectors, XGBoost maintained performance superior to traditional approaches, indicating a meaningful advancement in predictive capability despite persistent modelling difficulties.

Fourthly, the strategic integration of the Jensen, Kelly, and Pedersen (JKP) factor library with the traditional Fama-French five-factor (FF5) model resulted in a substantial enhancement of explanatory power. Rigorous regression analyses incorporating selected JKP factors consistently demonstrated significant improvements in adjusted R^2 relative to the standalone FF5 framework, particularly in industries facing higher volatility, rapid technological innovation, or significant structural transitions. Industries such as Healthcare Equipment, Steel, Business Services, and Food experienced notable increases in adjusted R^2 (ranging from approximately 1% to over 6%), indicating that integrating a broader and more flexible set of economic factors significantly enhances the traditional FF5 model's explanatory capabilities.

Furthermore, redundancy analysis provided clear evidence of overlap and inefficiencies within the traditional FF5 factor set. Key FF5 factors—especially SMB (Size) and Mkt_RF (Market Risk)—were strongly captured or even replaced by selected JKP factors. Factors representing liquidity, momentum, operating efficiency, and macroeconomic variables were frequently significant in explaining variations traditionally attributed solely to SMB and Mkt_RF. This finding indicates meaningful redundancy in traditional factor structures, clearly underscoring the practical and theoretical benefits of adopting broader, more dynamic factor frameworks enabled by ML-driven factor selection.

In summary, these empirical findings collectively provide robust and compelling evidence in support of advanced machine learning methodologies as superior alternatives to traditional factor-based asset pricing models for industry-level cost of equity estimation. The enhanced predictive accuracy, robustness, and adaptability demonstrated by ML methods—particularly LASSO, GBM, Light GBM, and XGBoost—across diverse industries and market conditions signify clear methodological improvements. Moreover, the integration of ML-selected JKP factors effectively addresses limitations inherent in traditional asset pricing frameworks, offering substantial advancements in theoretical development and practical applicability.

Summary of Key Empirical Findings

This study has indicated strong real-world evidence. It indicates that advanced machine learning methods are much better than traditional factor-based asset pricing models when estimating the industry-level cost of equity (ICoE). Traditional models, like CAPM, FF3, and FF5, have their own problems. They can't really capture the nonlinear, dynamic, and industry-specific risk-return features well. These limitations are especially obvious in industries that are going through structural changes, technological innovation, or facing high volatility.

Machine learning methods, like LASSO, GBM, Light GBM, and XGBoost, always indicated better predictive accuracy, stability, and robustness. These ML models can deal with factor redundancy well. They can also find economically meaningful predictors and capture complex nonlinear interactions. Compared with traditional approaches, they made significant improvements. The ML-selected factors from the comprehensive JKP factor library were strategically integrated. This integration significantly improved the explanatory

power and economic relevance of the traditional FF5 framework. It also solved the important redundancy and gap problems in the existing asset pricing models.

In the end, we conducted detailed empirical analyses and strict robustness checks across multiple industries and periods. These analyses and checks prove that integrating machine learning techniques into traditional asset pricing frameworks is practically effective. They also indicate that this integration leads to theoretical advancements and has methodological superiority.

These findings confirm the hypotheses we made at the start. They also lay a solid foundation for us to discuss theoretical contributions, practical implications, methodological innovations, and future research directions in the following parts of this thesis.

5.2 Research Contributions

This thesis contributes significantly to asset pricing literature by systematically and comprehensively addressing critical gaps in existing theoretical frameworks and empirical practices regarding the estimation of industry-level cost of equity (ICoE). Previous research has mostly used traditional factor models. These include the Capital Asset Pricing Model (CAPM), the Fama-French three-factor model (FF3), and the Fama-French five-factor model (FF5). But the findings of this study indicate that these approaches have significant limitations. Specifically, the thesis indicates very clearly that traditional factor-based models can't fully capture the dynamics specific to industries, the structural changes, and the complex non-linear relationships that are common in today's financial markets. Previous studies have mentioned the shortcomings of traditional models in general. But this thesis specifically measures and points out these limitations in a wide variety of industries. These industries include those driven by technology, industries with rapid innovation, and markets that are going through regulatory and structural changes. This research uses solid empirical analyses and strict comparisons with advanced methods to clearly find these limitations. It then pinpoints specific weaknesses in the traditional asset pricing paradigms. This significantly improves our current theoretical understanding. It also provides a strong basis for theoretical refinement and methodological innovation.

Secondly, this study makes a significant contribution in terms of methodology. It integrates advanced machine learning (ML) techniques into traditional asset pricing

frameworks in a structured and comprehensive way. These ML techniques include Least Absolute Shrinkage and Selection Operator (LASSO), Gradient Boosting Machines (GBM), Light Gradient Boosting Machines (Light GBM), and XGBoost.

Machine learning methods have been explored more and more in finance. But previous studies generally used them alone or only compared them with traditional methods in a limited way. In contrast, this research provides a rigorous, structured comparative analysis across a wide set of industries and extensive historical periods, systematically demonstrating and quantifying the substantial predictive superiority and robustness of ML-driven approaches. By comparing traditional linear factor models against ML methodologies, this research empirically validates the practical and theoretical advantages of ML techniques, significantly enriching the empirical and methodological toolkit available to both academics and industry practitioners in asset pricing and equity valuation.

Thirdly, this research makes a distinctive contribution by systematically validating the effectiveness and practical applicability of the Jensen-Kelly-Pedersen (JKP) factor library within the traditional asset pricing frameworks. By leveraging the extensive JKP factor set—comprising 153 carefully categorized factors across 13 economically meaningful themes—this thesis empirically illustrates that integrating broader, data-driven factor sets significantly enhances the explanatory and predictive capabilities of conventional models like the Fama-French five-factor model (FF5). Notably, the research demonstrates substantial incremental improvements in explanatory power, as measured by adjusted R^2, across diverse industry settings. For example, the thesis clearly shows significant improvements in explanatory power exceeding 6% in industries such as Food, Steel, and Healthcare Equipment—areas typically challenging for traditional factor models. Through rigorous redundancy analyses, the thesis also highlights critical inefficiencies and redundancies within the FF5 factor framework, clearly indicating that traditional factors such as size (SMB) and market risk premium (Mkt_RF) can be substantially captured or replaced by JKP factors like liquidity, volatility, and macroeconomic indicators. This detailed empirical validation not only reinforces the theoretical rationale for adopting broader factor universes but also provides actionable frameworks for practitioners to enhance factor selection processes, thus significantly advancing asset pricing research in both theoretical and practical dimensions.

Fourthly, the thesis provides robust methodological contributions through extensive and detailed robustness checks, incorporating leave-one-out cross-validation (LOOCV),

parameter sensitivity tests, and multiple period validations across the extensive historical sample. These rigorous validation procedures go beyond standard practices typically employed in asset pricing studies, addressing critical issues of model stability, factor consistency, and predictive reliability. By thoroughly demonstrating that machine learning-selected factors and modeling frameworks consistently exhibit stability and adaptability across different industry environments, economic conditions, and varying parameter settings, the research sets a new methodological benchmark for robustness standards in asset pricing literature. Consequently, future researchers and practitioners can confidently adopt the proposed ML-based methodologies and factor selection frameworks, knowing these methods have been rigorously vetted and validated across multiple dimensions of robustness and stability.

Fifthly, from a practical perspective, this thesis substantially enhances the analytical toolkit available to financial analysts, portfolio managers, and strategic decision-makers by providing clear, evidence-based methodologies for improved industry-level equity cost estimation. Specifically, by demonstrating the consistent predictive advantages of machine learning models—such as LASSO, GBM, Light GBM, and XGBoost—this study offers practitioners concrete methods to achieve greater accuracy in estimating expected returns, risk management, and capital budgeting processes. For instance, the notable predictive accuracy improvements observed in technology-intensive sectors, such as Software, Semiconductors, and Biotechnology, enable analysts to significantly reduce valuation errors and investment uncertainty, thereby directly improving strategic investment decisions and enhancing portfolio construction efficiency. By providing a detailed methodological pathway for incorporating ML methods into standard financial practices, this thesis enables industry practitioners to confidently transition toward more sophisticated, reliable analytical frameworks, which are crucial for effective capital allocation, mergers and acquisitions valuation, and risk-adjusted investment analyses.

Sixthly, this research contributes significantly to regulatory and financial reporting practices by empirically validating the reliability, stability, and superior explanatory power of advanced ML methodologies within asset valuation contexts. Given increasing regulatory scrutiny around valuation practices, capital adequacy assessments, and financial disclosures, this thesis provides robust, empirical support for adopting advanced data-driven methodologies in regulatory and compliance frameworks. By demonstrating the clear

empirical superiority and robustness of ML methods relative to traditional factor models, the findings from this thesis can inform future regulatory guidelines, accounting standards, and financial reporting best practices. This contribution thus has significant practical implications for regulatory authorities, standard-setting bodies, financial institutions, and auditing firms, potentially improving valuation accuracy, reporting transparency, and market stability.

Finally, the thesis improves the current methodological standards in asset pricing research. It clearly indicates how we can practically combine the factor selection methods driven by machine learning (ML) with traditional theoretical models. In previous studies, researchers generally explored advanced ML techniques separately. They didn't clearly integrate these techniques into well-established theoretical frameworks.

On the other hand, this research gives a clear, well-structured, and empirically proven methodological plan. It's for systematically combining machine-learning-driven factor selection into traditional asset pricing frameworks. The research outlines specific operational steps, validation processes, and robustness checks. So, it offers future researchers a comprehensive and replicable methodological way. This way connects traditional theoretical asset pricing models with modern machine-learning methods.

This research method makes the asset pricing research more transparent and replicable. It also makes the research more based on real-world evidence. It sets new standards for future academic research and practical analysis.

This thesis makes some great contributions to the asset pricing literature. It does this in theoretical, methodological, and practical ways. First, it clearly points out and measures the important problems in traditional models. Then, it uses real-world data to prove that advanced machine learning (ML) methodologies work. It also improves the ability to explain things by looking at a wider range of factors. It indicates that the methods used are reliable. And it gives financial practitioners a much better set of tools for their analysis.

Regulatory authorities. All these contributions together indicate a big step forward in asset pricing theory, practice, and methodology. They give clear empirical and methodological guidance for future research and practical applications.

5.3 Innovation and Theoretical Implications

This research brings in several remarkable innovations. It also offers meaningful theoretical implications in the asset pricing literature, especially when it comes to estimating the industry-level cost of equity (ICoE).

This study presents a novel methodology. It clearly combines advanced machine learning (ML) methods, like LASSO, Gradient Boosting Machines (GBM), Light GBM, and XGBoost, with traditional theoretical asset pricing frameworks. Historically, most of the research used traditional linear factor models. Due to technical development, they mostly disregarded the potential for nonlinear relationships and complex interactions between predictors.

This thesis takes initiatives creating a structured analytical framework by systematically adding machine learning (ML) techniques to the industry asset pricing context. This framework significantly improves the explanatory capacity, the accuracy of predictions and the robustness of the methods. This structured combination sets an important example for future theoretical progress. It also indicates a clear way for bringing advanced, data-driven methods into traditional theoretical asset pricing models.

Secondly, this research enhances our comprehension of theories. It clearly indicates and measures the redundancy and inefficiencies in the widely used Fama-French five-factor model (FF5). Through a strict redundancy analysis, the study clearly indicates that traditional factors like SMB (size) and Mkt_RF (market risk) can be effectively captured or even replaced. The factors that can replace them come from a broader set of economically meaningful factors in the Jensen-Kelly-Pedersen (JKP) factor library.

This theoretical progress challenges the established assumptions in traditional factor models. It also emphasizes how important it is to use broader, more flexible, and dynamic factor selection methods in today's asset pricing research.

Third, we comprehensively validate and empirically demonstrate that the factors selected by machine learning have incremental explanatory power. This significantly strengthens the theoretical debates about factor validity and factor proliferation in asset pricing.

Previous studies have discussed about the economic interpretability and robustness of different empirical factors. This research gives strong empirical evidence. It supports the

systematic identification and selection of economically meaningful factors using ML techniques.

This thesis makes an important contribution to the ongoing theoretical discussions about factor redundancy, economic relevance, and model specification in asset pricing research. It does this by empirically indicating that the factors selected by machine learning (ML) always do better than or improve traditional factor models. This is true across different industries and time periods.

Fourthly, we conducted detailed and extensive robustness checks. These checks included leave-one-out cross-validation (LOOCV), parameter sensitivity tests, and multiple-period analyses. These checks offer significant theoretical implications. They are related to methodological rigor and model validation standards in asset pricing studies.

The thesis clearly indicates that advanced machine learning methods are very robust and consistent in different economic conditions and industries. It offers theoretical support for using stricter validation standards in empirical asset pricing research. This methodological innovation sets an important theoretical benchmark. It encourages future studies to use equally strict validation practices. This is to make sure that the methods are reliable, stable, and replicable.

Lastly, this thesis makes a big contribution to the theoretical discussions about how to interpret and combine nonlinear and adaptive modeling approaches with the established financial theories. Through empirical research, it indicates that ML-driven methods have significant predictive advantages and adaptability in different and complex industry contexts. This research clearly challenges the traditional linearity assumptions that are common in classical asset pricing theories.

This theoretical contribution creates important opportunities for future research. The future research can explore the theoretical basis and economic explanations of nonlinear, data-driven asset pricing models.

To sum it up, this thesis brings in some innovations and has theoretical implications. These things together help move forward the theoretical basis of asset pricing research. They also give important method-related standards for future empirical studies. And they clearly indicate that it's necessary in theory and advantageous in practice to combine advanced machine-learning methods with traditional asset pricing frameworks.

5.4 Limitations

Despite the substantial contributions and innovations presented, this research acknowledges several limitations that should be clearly noted when interpreting the results and implications. Recognizing these limitations is essential for an accurate understanding of the scope of this study and for identifying potential improvements in future research.

Firstly, this thesis primarily utilizes monthly historical data across a considerable time span (from 1969 to 2023). While the extended sample period ensures robust empirical conclusions and provides comprehensive coverage of diverse market conditions, the monthly frequency of the data inherently limits the analysis's sensitivity to short-term market fluctuations and rapid structural shifts. Financial markets, particularly in the contemporary environment, frequently experience rapid changes and high-frequency volatility patterns that monthly data may not effectively capture. Consequently, the predictive models and factors identified might not fully reflect short-term, market-specific anomalies or shocks, potentially limiting the generalizability of findings in highly volatile market contexts or for short-term investment horizons.

Secondly, although advanced machine learning (ML) models—such as LASSO, GBM, Light GBM, and XGBoost—demonstrate significant predictive improvements, their inherent complexity and "black-box" nature create interpretability challenges. Unlike traditional linear factor models, where each factor carries clear theoretical and economic interpretations, ML-based methodologies can obscure economic meanings due to complex nonlinear relationships and high-dimensional interactions among factors. The lack of explicit economic interpretability potentially limits the practical applicability and acceptance of these advanced methodologies among investment professionals and regulatory bodies, who often require transparent, easily interpretable models for strategic decision-making, regulatory compliance, and investor communication purposes.

Thirdly, the research did not incorporate or estimate dynamic betas or structural break methodologies, despite their potential relevance in improving industry-level cost of equity estimation. Dynamic beta estimation methods, such as Neural Beta or time-varying parameter models, have recently emerged in literature as potentially superior alternatives to static beta estimation. The absence of dynamic beta estimation in this study—due primarily to methodological complexity, limited literature guidance, and data constraints—constitutes a

notable limitation, particularly given beta's central role in traditional asset pricing theories. Thus, exploring dynamic or adaptive beta estimation remains a critical area for future methodological enhancement and theoretical integration.

Additionally, while extensive robustness checks were conducted—including parameter sensitivity analyses, leave-one-out cross-validation (LOOCV), and multi-period validations—there remains the possibility that specific model parameterizations or alternative ML methodologies not considered here (e.g., neural networks, deep learning) might yield even greater predictive accuracy or stability. The comprehensive factor selection process conducted via LASSO and tree-based models may also overlook potentially economically relevant yet statistically subtle factors due to inherent model assumptions, regularization procedures, or data-driven selection criteria.

Finally, the generalizability of findings could be constrained by the specific selection of industries and factors analysed. Although 49 diverse industries and an extensive factor universe (JKP factor library) were utilized, it is possible that sector-specific anomalies, industry microstructures, or region-specific economic factors outside the scope of this thesis might further influence industry-level equity costs. Thus, caution is advised when generalizing the findings to other unexamined sectors, emerging markets, or specific firm-level contexts, highlighting the need for continued industry-specific and global market research.

In summary, these clearly acknowledged limitations provide valuable context for interpreting the thesis's contributions and implications. Addressing these limitations offers important pathways for future research aimed at further refining methodological approaches, enhancing model interpretability, incorporating dynamic beta estimation methods, and broadening the generalizability of findings within asset pricing literature and practice.

5.5 Future Research Directions

Given the contributions, innovations, and acknowledged limitations of this thesis, several promising avenues for future research are identified. These directions aim to further extend, refine, and enhance the theoretical, methodological, and practical advancements presented in this study.

Firstly, future research could explore the integration of higher-frequency data, such as daily or weekly returns, into industry-level cost of equity estimations. While this study relied on monthly historical data, incorporating higher-frequency data could significantly enhance sensitivity to short-term volatility patterns, market anomalies, and rapid structural shifts. Such approaches may allow models to better capture short-term risk exposures and dynamic industry dynamics, potentially improving both predictive accuracy and economic interpretability. Utilizing high-frequency data could also enhance model responsiveness to real-time market conditions, offering greater practical relevance for financial analysts, traders, and portfolio managers who operate within shorter investment horizons.

Secondly, further research should investigate and develop methodologies for dynamic beta estimation, such as Neural Beta or time-varying parameter models. As highlighted in the limitations section, static beta estimation methods used in traditional asset pricing models frequently fall short in capturing evolving systematic risk exposures. Exploring dynamic beta estimation could substantially improve industry-specific equity cost predictions, particularly in sectors experiencing rapid structural transformations or significant volatility shifts. Future studies could systematically evaluate dynamic beta methodologies' effectiveness relative to traditional static approaches and ML-driven frameworks, providing valuable theoretical and practical guidance for model enhancement and factor selection processes.

Thirdly, addressing the interpretability limitations inherent in advanced machine learning (ML) models presents a critical future research direction. While this thesis demonstrates clear empirical advantages of ML methodologies (e.g., GBM, XGBoost, and LASSO), their "black-box" nature limits their interpretability and transparency. Future research could specifically focus on developing hybrid methodologies or enhanced interpretability techniques—such as SHAP values, LIME explanations, and interpretability-focused model simplifications—to balance ML models' predictive accuracy with clear economic explanations and theoretical interpretability. This approach would significantly enhance acceptance among practitioners, regulators, and investors, thereby broadening practical adoption and regulatory acceptance of ML methodologies in asset pricing contexts.

Additionally, exploring alternative ML techniques not covered extensively in this study, such as deep learning models (e.g., recurrent neural networks or convolutional neural networks), may offer further predictive enhancements and methodological innovations. Deep learning methodologies, known for their ability to handle extensive datasets and complex

feature interactions, could potentially capture even subtler industry-specific dynamics and macroeconomic signals. Future research can systematically evaluate the relative advantages, interpretability challenges, and practical implications of deep learning techniques compared to traditional ML models and linear factor approaches, further enriching methodological options within asset pricing.

Lastly, future research should expand the geographical and industrial scope beyond the industries and regions examined in this thesis. Specifically, investigating emerging markets, small-cap sectors, or niche industries could provide valuable insights into the generalizability and adaptability of the proposed methodologies across different market environments. Such research would enhance understanding of potential market-specific anomalies, regulatory impacts, and region-specific economic dynamics influencing industry equity costs, thereby significantly broadening the theoretical and practical applicability of the research methodologies introduced here.

In conclusion, these future research directions offer clear pathways for continued theoretical, methodological, and practical advancements in asset pricing research, aiming to refine predictive accuracy, enhance methodological robustness, address critical interpretability challenges, and broaden the generalizability and applicability of advanced asset pricing methodologies developed in this thesis.

References:

1. Chattopadhyay, Akash, Bingxu Fang, and Partha Mohanram. (2022).
"Machine Learning, Earnings Forecasting and Implied Cost of Capital-US and International Evidence." Working Paper.

2. Claus, James, and Jacob Thomas. (2001).
"Equity Premia as Low as Three Percent? Evidence from Analysts' Earnings Forecasts for Domestic and International Stock Markets." Journal of Finance, Vol. 56, No. 5, pp. 1629–1666.

3. Drobetz, Wolfgang, Fabian Hollstein, Tizian Otto, and Marcel Prokopczuk. (2024).
"Estimating Stock Market Betas via Machine Learning." Journal of Financial and Quantitative Analysis, Working Paper.

4. Easton, Peter D. (2003).
"PE ratios, PEG ratios, and estimating the implied expected rate of return on equity capital." Working Paper.

5. Easton, Peter D., and Gregory A. Sommers. (2007).
"Effect of Analysts' Optimism on Estimates of the Expected Rate of Return Implied by Earnings Forecasts." Journal of Accounting Research, Vol. 45, No. 5, pp. 983–1015.

6. Fama, Eugene F., and Kenneth R. French. (1992).
"The Cross-Section of Expected Stock Returns." Journal of Finance, Vol. 47, No. 2, pp. 427–465.

7. Fama, Eugene F., and Kenneth R. French. (1993).
"Common risk factors in the returns on stocks and bonds." Journal of Financial Economics, Vol. 33, No. 1, pp. 3–56.

8. Fama, Eugene F., and Kenneth R. French. (1997).
"Industry costs of equity." Journal of Financial Economics, Vol. 43, pp. 153–193.

9. Fama, Eugene F., and Kenneth R. French. (2015).
"A Five-Factor Asset Pricing Model." Journal of Financial Economics, Vol. 116, pp. 1–22.

10. Fama, Eugene F., and Kenneth R. French. (2015).
"Dissecting Anomalies with a Five-Factor Model." Working Paper.

11. Geertsema, Paul, and Helen Lu. (2024).
"Return Predictability: Accounting versus Market Information." Working Paper.

12. Gu, Shihao, Bryan Kelly, and Dacheng Xiu. (2020).
"Empirical Asset Pricing via Machine Learning." Review of Financial Studies, Vol. 33, No. 5, pp. 2223–2274.

13. Guay, Wayne, SP Kothari, and Susan Shu. (2011).
"Properties of implied cost of capital using analysts' forecasts." Australian Journal of Management, Vol. 36(2), pp. 125–149.

14. Jensen, Theis Ingerslev, Bryan Kelly, and Lasse Heje Pedersen. (2023).
"Is There a Replication Crisis in Finance?" The Journal of Finance, Vol. 78, No. 5, pp. 2465–2505.

15. Kwon, Tae Yeon. (2025).
"Feature importance in linear models with ensemble machine learning: A study of the Fama and French five-factor model." Finance Research Letters, Vol. 71, Article 106406.

16. Lee, Charles M. C., Eric C. So, and Charles C. Y. Wang. (2021).
"Evaluating Firm-Level Expected-Return Proxies: Implications for Estimating Treatment Effects." The Review of Financial Studies, Vol. 34, No. 4, pp. 1907–1951.

17. Mishra, Dev R., and Thomas J. O'Brien. (2019).
"Fama-French, CAPM, and implied cost of equity." Journal of Economics and Business, Vol. 101, pp. 73–85.

Table 9 / Appendix 1

Industry	Full period				Recent Period			
	Mean Absolute Error (MAE)	Mean Squared Errors (MSE)	Out-of-Sample (R^2)	Mean Error Variance (MEV)	Mean Absolute Error (MAE)	Mean Squared Errors (MSE)	Out-of-Sample (R^2)	Mean Error Variance (MEV)
Agric								
Historical Return	4.76	39.59	0.00	39.59	4.96	43.54	0.00	43.54
CAPM	3.98	27.98	0.29	27.98	4.37	33.02	0.24	33.02
FF3	4.03	29.24	0.26	29.24	4.28	32.48	0.25	32.48
FF5	5.38	57.46	-0.45	57.46	4.37	33.70	0.23	33.70
OLS	3.93	27.67	0.30	27.67	4.35	32.91	0.24	32.91
LASSO_All_Themes_and_All_Factors	4.93	48.83	-0.23	48.83	4.51	35.04	0.20	35.04
LASSO_All_Themes	5.35	56.51	-0.43	56.51	4.27	32.17	0.26	32.17
LASSO_All_Factors	5.06	48.50	-0.22	48.50	4.83	39.68	0.09	39.68
Food								
Historical Return	3.31	19.54	0.00	19.54	2.79	13.45	0.00	13.45
CAPM	2.83	14.16	0.28	14.16	2.25	8.50	0.37	8.50
FF3	2.82	14.40	0.26	14.40	2.35	9.36	0.30	9.36
FF5	3.05	17.46	0.11	17.46	2.21	8.49	0.37	8.49
OLS	2.77	13.83	0.29	13.83	2.24	8.45	0.37	8.45
LASSO_All_Themes_and_All_Factors	2.41	10.20	0.48	10.20	2.16	7.82	0.42	7.82
LASSO_All_Themes	2.69	13.08	0.33	13.08	2.12	7.71	0.43	7.71
LASSO_All_Factors	3.64	24.59	-0.26	24.59	2.81	13.63	-0.01	13.63
Soda								
Historical Return	4.66	40.76	0.00	40.76	3.89	30.03	0.00	30.03
CAPM	4.29	33.80	0.17	33.80	3.35	20.35	0.32	20.35
FF3	4.38	36.55	0.10	36.55	4.05	27.49	0.08	27.49
FF5	4.20	33.22	0.19	33.22	4.06	27.76	0.08	27.76
OLS	4.26	33.39	0.18	33.39	3.35	20.25	0.33	20.25
LASSO_All_Themes_and_All_Factors	4.19	32.58	0.20	32.58	3.74	25.52	0.15	25.52
LASSO_All_Themes	4.07	31.46	0.23	31.46	3.94	26.36	0.12	26.36
LASSO_All_Factors	5.17	48.76	-0.20	48.76	4.38	36.27	-0.21	36.27
Beer								
Historical Return	3.78	25.75	0.00	25.75	2.92	14.71	0.00	14.71
CAPM	3.34	19.65	0.24	19.65	2.41	9.49	0.36	9.49
FF3	3.37	19.98	0.22	19.98	2.63	11.35	0.23	11.35
FF5	3.31	18.49	0.28	18.49	2.59	11.45	0.22	11.45
OLS	3.31	19.35	0.25	19.35	2.41	9.41	0.36	9.41
LASSO_All_Themes_and_All_Factors	3.32	19.83	0.23	19.83	2.42	9.22	0.37	9.22
LASSO_All_Themes	3.27	18.83	0.27	18.83	2.37	9.30	0.37	9.30
LASSO_All_Factors	4.30	33.83	-0.31	33.83	2.98	15.61	-0.06	15.61
Smoke								
Historical Return	4.88	41.75	0.00	41.75	4.26	28.66	0.00	28.66
CAPM	4.36	34.16	0.18	34.16	3.82	23.00	0.20	23.00
FF3	4.35	35.08	0.16	35.08	4.09	27.37	0.05	27.37
FF5	4.25	33.41	0.20	33.41	3.81	24.57	0.14	24.57
OLS	4.33	33.84	0.19	33.84	3.82	22.94	0.20	22.94
LASSO_All_Themes_and_All_Factors	4.32	34.18	0.18	34.18	4.26	29.81	-0.04	29.81
LASSO_All_Themes	4.28	33.74	0.19	33.74	3.55	21.40	0.25	21.40
LASSO_All_Factors	5.02	43.60	-0.04	43.60	4.26	28.66	0.00	28.66

Toys								
Historical Return	5.60	52.35	0.00	52.35	5.71	54.19	0.00	54.19
CAPM	4.05	28.41	0.46	28.41	4.30	30.66	0.43	30.66
FF3	3.93	26.13	0.50	26.13	4.05	27.81	0.49	27.81
FF5	3.90	26.55	0.49	26.55	3.87	25.76	0.52	25.76
OLS	4.00	27.85	0.47	27.85	4.31	30.75	0.43	30.75
LASSO_All_Themes_and_All_Factors	4.68	40.78	0.22	40.78	3.81	24.55	0.55	24.55
LASSO_All_Themes	3.87	25.95	0.50	25.95	3.87	25.77	0.52	25.77
LASSO_All_Factors	6.02	69.95	-0.34	69.95	4.35	30.69	0.43	30.69
Fun								
Historical Return	5.64	58.33	0.00	58.33	6.05	70.52	0.00	70.52
CAPM	4.02	26.94	0.54	26.94	3.91	28.31	0.60	28.31
FF3	4.08	29.15	0.50	29.15	3.90	27.19	0.61	27.19
FF5	4.20	28.33	0.51	28.33	3.90	27.13	0.62	27.13
OLS	3.92	25.85	0.56	25.85	3.89	28.15	0.60	28.15
LASSO_All_Themes_and_All_Factors	4.64	39.30	0.33	39.30	3.94	27.90	0.60	27.90
LASSO_All_Themes	4.15	27.68	0.53	27.68	3.90	27.90	0.60	27.90
LASSO_All_Factors	4.81	41.93	0.28	41.93	4.62	39.42	0.44	39.42
Books								
Historical Return	4.44	35.86	0.00	35.86	4.80	44.42	0.00	44.42
CAPM	2.81	13.47	0.62	13.47	3.35	20.26	0.54	20.26
FF3	2.84	14.17	0.60	14.17	2.96	15.95	0.64	15.95
FF5	2.69	11.77	0.67	11.77	2.97	15.90	0.64	15.90
OLS	2.72	12.87	0.64	12.87	3.30	19.95	0.55	19.95
LASSO_All_Themes_and_All_Factors	2.66	12.62	0.65	12.62	3.32	20.19	0.55	20.19
LASSO_All_Themes	2.67	11.90	0.67	11.90	3.10	17.74	0.60	17.74
LASSO_All_Factors	5.69	70.84	-0.98	70.84	4.25	32.39	0.27	32.39
Hshld								
Historical Return	3.40	20.30	0.00	20.30	3.03	15.97	0.00	15.97
CAPM	2.58	12.74	0.37	12.74	2.32	9.01	0.44	9.01
FF3	2.61	13.95	0.31	13.95	2.25	8.37	0.48	8.37
FF5	2.34	11.10	0.45	11.10	2.14	7.79	0.51	7.79
OLS	2.53	12.44	0.39	12.44	2.31	8.87	0.44	8.87
LASSO_All_Themes_and_All_Factors	2.86	18.41	0.09	18.41	2.37	9.84	0.38	9.84
LASSO_All_Themes	2.31	10.80	0.47	10.80	2.11	7.48	0.53	7.48
LASSO_All_Factors	3.99	29.07	-0.43	29.07	3.00	15.45	0.03	15.45
Clths								
Historical Return	4.86	42.21	0.00	42.21	4.82	41.29	0.00	41.29
CAPM	3.33	19.35	0.54	19.35	3.25	17.55	0.58	17.55
FF3	3.41	22.15	0.48	22.15	3.58	20.65	0.50	20.65
FF5	3.36	20.25	0.52	20.25	3.31	18.10	0.56	18.10
OLS	3.28	18.91	0.55	18.91	3.24	17.53	0.58	17.53
LASSO_All_Themes_and_All_Factors	2.97	14.63	0.65	14.63	3.05	15.78	0.62	15.78
LASSO_All_Themes	3.33	19.86	0.53	19.86	3.30	17.97	0.56	17.97
LASSO_All_Factors	4.60	36.45	0.14	36.45	3.89	25.09	0.39	25.09

Hlth								
Historical Return	5.20	46.22	0.00	46.22	4.41	33.75	0.00	33.75
CAPM	4.73	41.38	0.10	41.38	3.40	19.53	0.42	19.53
FF3	5.59	67.38	-0.46	67.38	3.74	24.77	0.27	24.77
FF5	6.25	87.60	-0.90	87.60	3.72	23.93	0.29	23.93
OLS	4.64	40.18	0.13	40.18	3.40	19.49	0.42	19.49
LASSO_All_Themes_and_All_Factors	4.77	48.53	-0.05	48.53	3.24	17.82	0.47	17.82
LASSO_All_Themes	6.18	85.27	-0.84	85.27	3.70	23.74	0.30	23.74
LASSO_All_Factors	5.93	75.67	-0.64	75.67	4.36	33.31	0.01	33.31
MedEq								
Historical Return	4.10	27.73	0.00	27.73	3.81	25.66	0.00	25.66
CAPM	2.65	11.61	0.58	11.61	2.57	10.96	0.57	10.96
FF3	2.93	16.07	0.42	16.07	2.57	10.70	0.58	10.70
FF5	2.88	15.24	0.45	15.24	2.50	10.13	0.61	10.13
OLS	2.62	11.30	0.59	11.30	2.55	10.76	0.58	10.76
LASSO_All_Themes_and_All_Factors	3.52	31.42	-0.13	31.42	3.12	17.77	0.31	17.77
LASSO_All_Themes	2.82	14.80	0.47	14.80	2.50	10.15	0.60	10.15
LASSO_All_Factors	4.79	54.40	-0.96	54.40	4.37	32.92	-0.28	32.92
Drugs								
Historical Return	3.70	22.51	0.00	22.51	3.22	16.66	0.00	16.66
CAPM	2.78	12.97	0.42	12.97	2.39	8.97	0.46	8.97
FF3	2.99	16.58	0.26	16.58	2.44	9.45	0.43	9.45
FF5	3.05	17.28	0.23	17.28	2.40	9.40	0.44	9.40
OLS	2.73	12.62	0.44	12.62	2.37	8.78	0.47	8.78
LASSO_All_Themes_and_All_Factors	4.09	35.74	-0.59	35.74	2.81	13.81	0.17	13.81
LASSO_All_Themes	3.03	17.11	0.24	17.11	2.39	9.29	0.44	9.29
LASSO_All_Factors	5.12	48.96	-1.18	48.96	3.13	16.03	0.04	16.03
Chems								
Historical Return	4.38	34.03	0.00	34.03	4.57	38.55	0.00	38.55
CAPM	2.46	10.89	0.68	10.89	2.73	13.59	0.65	13.59
FF3	2.34	9.46	0.72	9.46	2.53	11.21	0.71	11.21
FF5	2.47	10.39	0.69	10.39	2.43	10.17	0.74	10.17
OLS	2.42	10.58	0.69	10.58	2.72	13.59	0.65	13.59
LASSO_All_Themes_and_All_Factors	2.56	13.69	0.60	13.69	2.47	10.13	0.74	10.13
LASSO_All_Themes	2.45	10.26	0.70	10.26	2.42	10.14	0.74	10.14
LASSO_All_Factors	3.55	23.48	0.31	23.48	3.61	21.48	0.44	21.48
Rubbr								
Historical Return	4.43	35.74	0.00	35.74	4.50	38.11	0.00	38.11
CAPM	2.68	13.37	0.63	13.37	2.82	15.00	0.61	15.00
FF3	2.64	12.45	0.65	12.45	2.86	13.91	0.64	13.91
FF5	2.55	11.38	0.68	11.38	2.65	12.09	0.68	12.09
OLS	2.64	13.11	0.63	13.11	2.82	15.07	0.60	15.07
LASSO_All_Themes_and_All_Factors	2.72	13.29	0.63	13.29	3.00	14.83	0.61	14.83
LASSO_All_Themes	2.55	11.38	0.68	11.38	2.66	11.89	0.69	11.89
LASSO_All_Factors	4.43	36.64	-0.03	36.64	3.91	25.74	0.32	25.74

Txtls								
Historical Return	5.60	63.55	0.00	63.55	6.66	92.63	0.00	92.63
CAPM	4.05	35.37	0.44	35.37	5.22	60.74	0.34	60.74
FF3	3.69	27.91	0.56	27.91	4.76	40.92	0.56	40.92
FF5	3.71	27.59	0.57	27.59	4.51	37.60	0.59	37.60
OLS	4.02	35.01	0.45	35.01	5.24	61.07	0.34	61.07
LASSO_All_Themes_and_All_Factors	3.72	28.40	0.55	28.40	4.46	39.61	0.57	39.61
LASSO_All_Themes	3.70	27.53	0.57	27.53	4.51	37.74	0.59	37.74
LASSO_All_Factors	5.19	52.70	0.17	52.70	5.12	52.03	0.44	52.03
BldMt								
Historical Return	4.76	42.13	0.00	42.13	5.23	52.46	0.00	52.46
CAPM	2.63	12.73	0.70	12.73	3.20	20.59	0.61	20.59
FF3	2.38	10.55	0.75	10.55	2.71	14.48	0.72	14.48
FF5	2.37	10.47	0.75	10.47	2.62	12.52	0.76	12.52
OLS	2.57	12.40	0.71	12.40	3.21	20.65	0.61	20.65
LASSO_All_Themes_and_All_Factors	2.83	16.89	0.60	16.89	2.74	14.54	0.72	14.54
LASSO_All_Themes	2.38	10.64	0.75	10.64	2.62	12.56	0.76	12.56
LASSO_All_Factors	4.22	33.88	0.20	33.88	3.67	23.84	0.55	23.84
Cnstr								
Historical Return	5.60	54.46	0.00	54.46	5.79	56.31	0.00	56.31
CAPM	3.60	21.10	0.61	21.10	3.95	23.48	0.58	23.48
FF3	3.62	21.49	0.61	21.49	3.65	21.36	0.62	21.36
FF5	4.33	37.94	0.30	37.94	3.54	19.88	0.65	19.88
OLS	3.57	20.85	0.62	20.85	3.92	23.09	0.59	23.09
LASSO_All_Themes_and_All_Factors	4.68	51.01	0.06	51.01	3.40	17.09	0.70	17.09
LASSO_All_Themes	4.29	36.95	0.32	36.95	3.53	19.80	0.65	19.80
LASSO_All_Factors	5.25	59.22	-0.09	59.22	4.60	32.08	0.43	32.08
Steel								
Historical Return	6.36	71.16	0.00	71.16	7.32	88.46	0.00	88.46
CAPM	4.28	33.42	0.53	33.42	4.66	35.82	0.60	35.82
FF3	4.27	32.71	0.54	32.71	4.45	31.62	0.64	31.62
FF5	4.13	30.36	0.57	30.36	4.54	33.80	0.62	33.80
OLS	4.27	33.04	0.54	33.04	4.58	35.42	0.60	35.42
LASSO_All_Themes_and_All_Factors	4.20	35.36	0.50	35.36	4.33	32.69	0.63	32.69
LASSO_All_Themes	4.13	30.40	0.57	30.40	4.57	34.45	0.61	34.45
LASSO_All_Factors	5.02	46.00	0.35	46.00	5.23	47.24	0.47	47.24
FabPr								
Historical Return	5.92	62.82	0.00	62.82	7.00	86.47	0.00	86.47
CAPM	4.58	38.88	0.38	38.88	5.50	55.79	0.35	55.79
FF3	4.35	34.86	0.45	34.86	5.17	47.57	0.45	47.57
FF5	5.07	52.17	0.17	52.17	5.06	46.22	0.47	46.22
OLS	4.56	38.45	0.39	38.45	5.53	56.10	0.35	56.10
LASSO_All_Themes_and_All_Factors	4.71	41.62	0.34	41.62	5.32	48.64	0.44	48.64
LASSO_All_Themes	4.81	44.76	0.29	44.76	5.05	46.15	0.47	46.15
LASSO_All_Factors	4.87	43.91	0.30	43.91	5.89	59.22	0.32	59.22

Mach								
Historical Return	5.03	43.58	0.00	43.58	5.11	46.95	0.00	46.95
CAPM	2.71	11.93	0.73	11.93	2.69	11.24	0.76	11.24
FF3	2.64	11.33	0.74	11.33	2.46	9.60	0.80	9.60
FF5	2.70	11.99	0.72	11.99	2.45	9.56	0.80	9.56
OLS	2.69	11.64	0.73	11.64	2.68	11.02	0.77	11.02
LASSO_All_Themes_and_All_Factors	2.59	11.64	0.73	11.64	1.96	6.24	0.87	6.24
LASSO_All_Themes	2.68	11.75	0.73	11.75	2.45	9.52	0.80	9.52
LASSO_All_Factors	3.60	23.23	0.47	23.23	3.23	18.34	0.61	18.34
ElcEq								
Historical Return	5.02	44.15	0.00	44.15	5.11	47.80	0.00	47.80
CAPM	2.76	12.22	0.72	12.22	2.82	13.28	0.72	13.28
FF3	2.69	11.83	0.73	11.83	3.04	15.24	0.68	15.24
FF5	2.70	11.89	0.73	11.89	3.05	15.30	0.68	15.30
OLS	2.69	11.72	0.73	11.72	2.74	12.65	0.74	12.65
LASSO_All_Themes_and_All_Factors	2.79	12.93	0.71	12.93	3.05	16.63	0.65	16.63
LASSO_All_Themes	2.71	11.89	0.73	11.89	3.06	15.55	0.67	15.55
LASSO_All_Factors	4.31	33.65	0.24	33.65	4.90	43.76	0.08	43.76
Autos								
Historical Return	5.98	72.79	0.00	72.79	6.97	106.25	0.00	106.25
CAPM	4.33	38.88	0.47	38.88	4.93	56.69	0.47	56.69
FF3	4.19	38.34	0.47	38.34	5.07	58.63	0.45	58.63
FF5	4.94	51.99	0.29	51.99	5.13	59.63	0.44	59.63
OLS	4.30	38.81	0.47	38.81	4.91	56.92	0.46	56.92
LASSO_All_Themes_and_All_Factors	5.10	55.85	0.23	55.85	4.78	50.39	0.53	50.39
LASSO_All_Themes	4.92	51.48	0.29	51.48	4.96	58.08	0.45	58.08
LASSO_All_Factors	6.06	76.98	-0.06	76.98	5.98	69.81	0.34	69.81
Aero								
Historical Return	4.84	43.78	0.00	43.78	4.60	41.80	0.00	41.80
CAPM	3.42	22.19	0.49	22.19	3.08	18.17	0.57	18.17
FF3	3.62	24.76	0.43	24.76	3.21	18.04	0.57	18.04
FF5	3.69	26.10	0.40	26.10	3.35	19.09	0.54	19.09
OLS	3.36	21.56	0.51	21.56	3.04	17.94	0.57	17.94
LASSO_All_Themes	3.54	23.74	0.46	23.74	3.33	18.92	0.55	18.92
LASSO_All_Factors	5.22	54.57	-0.25	54.57	4.03	28.15	0.33	28.15
Ships								
Historical Return	5.59	56.63	0.00	56.63	5.50	58.74	0.00	58.74
CAPM	4.24	31.59	0.44	31.59	4.18	34.81	0.41	34.81
FF3	4.11	29.67	0.48	29.67	3.41	23.64	0.60	23.64
FF5	4.77	42.93	0.24	42.93	3.47	21.87	0.63	21.87
OLS	4.21	31.32	0.45	31.32	4.18	34.76	0.41	34.76
LASSO_All_Themes	4.48	37.21	0.34	37.21	3.46	21.94	0.63	21.94
LASSO_All_Factors	5.26	52.69	0.07	52.69	4.83	41.79	0.29	41.79
Guns								

Historical Return	4.63	39.80	0.00	39.80	3.99	30.27	0.00	30.27
CAPM	4.17	33.87	0.15	33.87	3.86	28.02	0.07	28.02
FF3	4.29	33.71	0.15	33.71	3.95	25.81	0.15	25.81
FF5	4.34	34.81	0.13	34.81	3.87	24.15	0.20	24.15
OLS	4.14	33.46	0.16	33.46	3.85	27.93	0.08	27.93
LASSO_All_Themes	4.25	33.48	0.16	33.48	3.76	23.05	0.24	23.05
LASSO_All_Factors	5.43	56.97	-0.43	56.97	3.89	27.56	0.09	27.56
Gold								
Historical Return	8.37	125.67	0.00	125.67	7.88	107.58	0.00	107.58
CAPM	8.31	119.42	0.05	119.42	7.75	104.15	0.03	104.15
FF3	8.22	118.54	0.06	118.54	8.43	120.28	-0.12	120.28
FF5	8.72	137.52	-0.09	137.52	8.47	121.37	-0.13	121.37
OLS	8.30	119.34	0.05	119.34	7.74	104.03	0.03	104.03
LASSO_All_Themes	8.49	128.42	-0.02	128.42	8.22	114.39	-0.06	114.39
LASSO_All_Factors	10.36	226.44	-0.80	226.44	7.89	107.61	0.00	107.61
Mines								
Historical Return	6.11	63.94	0.00	63.94	7.26	84.35	0.00	84.35
CAPM	4.79	37.14	0.42	37.14	5.71	51.14	0.39	51.14
FF3	4.67	35.58	0.44	35.58	5.50	47.15	0.44	47.15
FF5	5.25	46.83	0.27	46.83	5.46	46.26	0.45	46.26
OLS	4.77	36.86	0.42	36.86	5.73	51.22	0.39	51.22
LASSO_All_Themes	5.23	46.36	0.27	46.36	5.49	46.61	0.45	46.61
LASSO_All_Factors	5.30	47.91	0.25	47.91	5.68	53.00	0.37	53.00
Coal								
Historical Return	8.55	131.83	0.00	131.83	10.24	177.44	0.00	177.44
CAPM	7.74	109.09	0.17	109.09	9.52	148.62	0.16	148.62
FF3	7.67	106.51	0.19	106.51	9.06	135.71	0.24	135.71
FF5	8.05	119.76	0.09	119.76	9.13	137.59	0.22	137.59
OLS	7.72	108.28	0.18	108.28	9.49	147.72	0.17	147.72
LASSO_All_Themes_and_All_Factors	7.68	107.32	0.19	107.32	9.85	159.24	0.10	159.24
LASSO_All_Themes	7.74	107.97	0.18	107.97	9.69	155.60	0.12	155.60
LASSO_All_Factors	7.90	114.00	0.14	114.00	8.69	124.13	0.30	124.13
Oil								
Historical Return	4.71	42.21	0.00	42.21	5.35	56.37	0.00	56.37
CAPM	3.89	27.19	0.36	27.19	4.57	39.52	0.30	39.52
FF3	3.91	26.93	0.36	26.93	4.18	30.63	0.46	30.63
FF5	4.19	32.69	0.23	32.69	4.15	30.42	0.46	30.42
OLS	3.86	26.92	0.36	26.92	4.55	39.40	0.30	39.40
LASSO_All_Themes_and_All_Factors	3.91	34.49	0.18	34.49	3.87	30.57	0.46	30.57
LASSO_All_Themes	4.03	29.82	0.29	29.82	4.29	32.90	0.42	32.90
LASSO_All_Factors	5.18	52.22	-0.24	52.22	4.32	37.19	0.34	37.19
Util								

Historical Return	3.13	16.08	0.00	16.08	3.00	15.27	0.00	15.27
CAPM	2.80	13.50	0.16	13.50	2.58	11.25	0.26	11.25
FF3	2.67	12.29	0.24	12.29	3.14	16.58	-0.09	16.58
FF5	3.00	16.94	-0.05	16.94	3.13	15.96	-0.05	15.96
OLS	2.78	13.27	0.17	13.27	2.58	11.21	0.27	11.21
LASSO_All_Themes_and_All_Factors	2.43	12.26	0.24	12.26	2.42	10.14	0.34	10.14
LASSO_All_Themes	2.94	15.56	0.03	15.56	2.96	14.08	0.08	14.08
LASSO_All_Factors	3.45	21.31	-0.33	21.31	3.51	22.95	-0.50	22.95
Telcm								
Historical Return	3.79	24.95	0.00	24.95	3.51	21.41	0.00	21.41
CAPM	2.62	11.96	0.52	11.96	2.12	7.61	0.64	7.61
FF3	2.54	11.81	0.53	11.81	2.08	7.72	0.64	7.72
FF5	2.54	11.06	0.56	11.06	2.10	7.56	0.65	7.56
OLS	2.59	11.69	0.53	11.69	2.10	7.50	0.65	7.50
LASSO_All_Themes_and_All_Factors	2.61	12.73	0.49	12.73	2.46	10.24	0.52	10.24
LASSO_All_Themes	2.54	11.12	0.55	11.12	2.06	7.37	0.66	7.37
LASSO_All_Factors	3.75	26.51	-0.06	26.51	3.19	17.68	0.17	17.68
PerSv								
Historical Return	4.80	39.14	0.00	39.14	4.65	37.61	0.00	37.61
CAPM	3.36	19.84	0.49	19.84	3.18	18.70	0.50	18.70
FF3	3.83	27.69	0.29	27.69	3.19	19.30	0.49	19.30
FF5	4.43	40.21	-0.03	40.21	3.38	20.84	0.45	20.84
OLS	3.33	19.55	0.50	19.55	3.17	18.51	0.51	18.51
LASSO_All_Themes_and_All_Factors	4.77	51.14	-0.31	51.14	3.02	17.21	0.54	17.21
LASSO_All_Themes	4.17	34.84	0.11	34.84	3.36	20.68	0.45	20.68
LASSO_All_Factors	6.11	84.82	-1.17	84.82	4.05	27.74	0.26	27.74
BusSv								
Historical Return	4.23	30.53	0.00	30.53	3.93	27.97	0.00	27.97
CAPM	1.86	6.00	0.80	6.00	1.46	3.81	0.86	3.81
FF3	1.65	5.05	0.83	5.05	1.38	3.60	0.87	3.60
FF5	1.94	7.67	0.75	7.67	1.35	3.49	0.88	3.49
OLS	1.77	5.49	0.82	5.49	1.43	3.70	0.87	3.70
LASSO_All_Themes_and_All_Factors	1.58	4.30	0.86	4.30	1.62	4.73	0.83	4.73
LASSO_All_Themes	1.91	7.28	0.76	7.28	1.33	3.38	0.88	3.38
LASSO_All_Factors	3.70	26.03	0.15	26.03	3.25	18.24	0.35	18.24
Harddw								
Historical Return	5.64	55.27	0.00	55.27	4.86	39.30	0.00	39.30
CAPM	3.63	24.24	0.56	24.24	3.82	25.86	0.34	25.86
FF3	3.52	22.38	0.60	22.38	3.95	25.69	0.35	25.69
FF5	3.85	28.33	0.49	28.33	3.71	23.29	0.41	23.29
OLS	3.62	24.05	0.56	24.05	3.68	24.39	0.38	24.39
LASSO_All_Themes_and_All_Factors	3.41	20.03	0.64	20.03	2.92	15.04	0.62	15.04
LASSO_All_Themes	3.76	26.43	0.52	26.43	3.60	21.98	0.44	21.98
LASSO_All_Factors	3.90	26.18	0.53	26.18	3.53	22.01	0.44	22.01

Softw								
Historical Return	6.39	68.96	0.00	68.96	4.12	27.20	0.00	27.20
CAPM	4.45	33.34	0.52	33.34	3.17	16.50	0.39	16.50
FF3	6.70	78.29	-0.14	78.29	2.72	11.94	0.56	11.94
FF5	7.24	93.56	-0.36	93.56	2.58	10.38	0.62	10.38
OLS	4.44	33.39	0.52	33.39	3.04	15.64	0.43	15.64
LASSO_All_Themes_and_All_Factors	6.20	70.46	-0.02	70.46	2.08	7.87	0.71	7.87
LASSO_All_Themes	6.89	84.21	-0.22	84.21	2.45	9.54	0.65	9.54
LASSO_All_Factors	7.48	109.47	-0.59	109.47	3.17	17.00	0.37	17.00
Chips								
Historical Return	5.94	58.66	0.00	58.66	5.14	41.54	0.00	41.54
CAPM	3.22	19.07	0.68	19.07	3.58	20.87	0.50	20.87
FF3	3.17	18.87	0.68	18.87	2.91	13.90	0.67	13.90
FF5	3.26	20.23	0.66	20.23	2.78	12.59	0.70	12.59
OLS	3.17	18.76	0.68	18.76	3.47	19.97	0.52	19.97
LASSO_All_Themes_and_All_Factors	2.71	12.92	0.78	12.92	2.41	9.65	0.77	9.65
LASSO_All_Themes	3.15	18.74	0.68	18.74	2.69	11.82	0.72	11.82
LASSO_All_Factors	3.63	21.65	0.63	21.65	3.62	20.55	0.51	20.55
LabEq								
Historical Return	5.12	44.96	0.00	44.96	4.32	30.93	0.00	30.93
CAPM	2.93	14.89	0.67	14.89	2.41	9.84	0.68	9.84
FF3	2.55	11.24	0.75	11.24	2.27	8.92	0.71	8.92
FF5	2.87	14.79	0.67	14.79	2.21	8.29	0.73	8.29
OLS	2.88	14.47	0.68	14.47	2.38	9.57	0.69	9.57
LASSO_All_Themes_and_All_Factors	3.49	26.66	0.41	26.66	3.14	17.94	0.42	17.94
LASSO_All_Themes	2.86	14.70	0.67	14.70	2.21	8.26	0.73	8.26
LASSO_All_Factors	4.01	31.17	0.31	31.17	4.46	37.81	-0.22	37.81
Paper								
Historical Return	4.06	29.78	0.00	29.78	3.87	27.26	0.00	27.26
CAPM	2.58	12.56	0.58	12.56	2.33	10.64	0.61	10.64
FF3	2.48	11.27	0.62	11.27	2.40	10.52	0.61	10.52
FF5	2.52	11.88	0.60	11.88	2.12	8.61	0.68	8.61
OLS	2.51	12.13	0.59	12.13	2.31	10.48	0.62	10.48
LASSO_All_Themes_and_All_Factors	2.58	12.54	0.58	12.54	2.39	11.19	0.59	11.19
LASSO_All_Themes	2.49	11.74	0.61	11.74	2.11	8.52	0.69	8.52
LASSO_All_Factors	3.91	26.71	0.10	26.71	3.55	22.27	0.18	22.27
Boxes								
Historical Return	4.39	33.32	0.00	33.32	4.31	31.02	0.00	31.02
CAPM	3.00	14.83	0.56	14.83	2.66	10.83	0.65	10.83
FF3	3.13	16.79	0.50	16.79	2.95	14.27	0.54	14.27
FF5	3.07	15.69	0.53	15.69	3.01	14.90	0.52	14.90
OLS	2.96	14.56	0.56	14.56	2.65	10.84	0.65	10.84
LASSO_All_Themes_and_All_Factors	3.58	21.44	0.36	21.44	3.22	15.73	0.49	15.73
LASSO_All_Themes	3.05	15.51	0.53	15.51	3.00	14.75	0.52	14.75
LASSO_All_Factors	4.68	37.49	-0.13	37.49	3.71	21.85	0.30	21.85

Trans								
Historical Return	4.53	34.11	0.00	34.11	4.53	33.85	0.00	33.85
CAPM	2.68	12.57	0.63	12.57	2.72	11.52	0.66	11.52
FF3	2.54	11.01	0.68	11.01	2.73	11.50	0.66	11.50
FF5	2.82	15.63	0.54	15.63	2.58	10.37	0.69	10.37
OLS	2.63	12.16	0.64	12.16	2.73	11.51	0.66	11.51
LASSO_All_Themes_and_All_Factors	3.18	22.50	0.34	22.50	2.56	10.71	0.68	10.71
LASSO_All_Themes	2.78	15.17	0.56	15.17	2.57	10.30	0.70	10.30
LASSO_All_Factors	4.31	36.79	-0.08	36.79	3.98	26.77	0.21	26.77
Whlsl								
Historical Return	3.97	27.47	0.00	27.47	3.77	25.37	0.00	25.37
CAPM	2.16	8.84	0.68	8.84	1.93	6.74	0.73	6.74
FF3	2.19	10.75	0.61	10.75	1.83	5.65	0.78	5.65
FF5	2.48	14.19	0.48	14.19	1.69	4.90	0.81	4.90
OLS	2.07	8.37	0.70	8.37	1.94	6.79	0.73	6.79
LASSO_All_Themes_and_All_Factors	2.30	12.14	0.56	12.14	1.99	6.97	0.73	6.97
LASSO_All_Themes	2.44	13.63	0.50	13.63	1.66	4.70	0.81	4.70
LASSO_All_Factors	4.45	43.66	-0.59	43.66	3.27	18.44	0.27	18.44
Rtail								
Historical Return	4.15	30.01	0.00	30.01	3.52	22.44	0.00	22.44
CAPM	2.52	10.79	0.64	10.79	2.00	6.83	0.70	6.83
FF3	2.50	10.82	0.64	10.82	2.28	9.41	0.58	9.41
FF5	2.53	10.85	0.64	10.85	2.32	9.65	0.57	9.65
OLS	2.48	10.48	0.65	10.48	1.96	6.70	0.70	6.70
LASSO_All_Themes_and_All_Factors	2.21	8.23	0.73	8.23	1.99	6.73	0.70	6.73
LASSO_All_Themes	2.38	9.75	0.67	9.75	2.31	9.52	0.58	9.52
LASSO_All_Factors	4.10	33.09	-0.10	33.09	3.61	21.54	0.04	21.54
Meals								
Historical Return	4.04	27.91	0.00	27.91	3.64	22.55	0.00	22.55
CAPM	3.56	22.12	0.21	22.12	2.29	8.70	0.61	8.70
FF3	3.63	27.03	0.03	27.03	2.69	11.49	0.49	11.49
FF5	3.15	17.38	0.38	17.38	2.56	10.00	0.56	10.00
OLS	3.50	21.46	0.23	21.46	2.31	8.80	0.61	8.80
LASSO_All_Themes_and_All_Factors	5.09	61.52	-1.20	61.52	2.15	7.37	0.67	7.37
LASSO_All_Themes	3.13	17.23	0.38	17.23	2.55	9.91	0.56	9.91
LASSO_All_Factors	6.31	89.95	-2.22	89.95	3.04	14.53	0.36	14.53
Banks								
Historical Return	4.72	38.87	0.00	38.87	4.79	44.11	0.00	44.11
CAPM	2.92	15.78	0.59	15.78	3.15	18.90	0.57	18.90
FF3	2.70	13.50	0.65	13.50	2.22	8.84	0.80	8.84
FF5	2.99	17.38	0.55	17.38	2.26	9.06	0.79	9.06
OLS	2.86	15.24	0.61	15.24	3.09	18.41	0.58	18.41
LASSO_All_Themes_and_All_Factors	4.30	61.31	-0.58	61.31	2.08	8.57	0.81	8.57
LASSO_All_Themes	2.90	16.01	0.59	16.01	2.26	9.12	0.79	9.12
LASSO_All_Factors	4.74	55.57	-0.43	55.57	3.27	17.94	0.59	17.94

Insur								
Historical Return	3.93	27.41	0.00	27.41	3.81	27.79	0.00	27.79
CAPM	2.72	13.81	0.50	13.81	2.44	11.40	0.59	11.40
FF3	2.79	15.10	0.45	15.10	2.21	8.73	0.69	8.73
FF5	4.26	43.35	-0.58	43.35	2.24	8.87	0.68	8.87
OLS	2.64	13.33	0.51	13.33	2.42	11.25	0.60	11.25
LASSO_All_Themes_and_All_Factors	3.01	16.41	0.40	16.41	2.02	7.47	0.73	7.47
LASSO_All_Themes	4.08	39.86	-0.45	39.86	2.23	8.82	0.68	8.82
LASSO_All_Factors	4.56	41.63	-0.52	41.63	3.35	21.48	0.23	21.48
RlEst								
Historical Return	5.01	53.42	0.00	53.42	6.02	80.12	0.00	80.12
CAPM	3.65	27.22	0.49	27.22	4.75	52.79	0.34	52.79
FF3	3.18	20.59	0.61	20.59	3.88	33.45	0.58	33.45
FF5	2.97	17.59	0.67	17.59	3.66	30.40	0.62	30.40
OLS	3.60	26.84	0.50	26.84	4.76	52.98	0.34	52.98
LASSO_All_Themes_and_All_Factors	3.58	23.86	0.55	23.86	4.07	39.37	0.51	39.37
LASSO_All_Themes	2.97	17.62	0.67	17.62	3.67	30.52	0.62	30.52
LASSO_All_Factors	5.27	55.17	-0.03	55.17	5.18	56.64	0.29	56.64
Fin								
Historical Return	4.97	42.01	0.00	42.01	5.02	42.67	0.00	42.67
CAPM	2.39	10.09	0.76	10.09	2.83	12.15	0.72	12.15
FF3	2.22	9.02	0.79	9.02	2.65	11.20	0.74	11.20
FF5	2.28	9.62	0.77	9.62	2.42	9.24	0.78	9.24
OLS	2.32	9.64	0.77	9.64	2.69	11.16	0.74	11.16
LASSO_All_Themes_and_All_Factors	2.03	8.05	0.81	8.05	2.60	11.08	0.74	11.08
LASSO_All_Themes	2.26	9.44	0.78	9.44	2.42	9.21	0.78	9.21
LASSO_All_Factors	3.21	17.38	0.59	17.38	4.09	30.04	0.30	30.04
Other								
Historical Return	4.76	39.09	0.00	39.09	3.67	24.58	0.00	24.58
CAPM	3.14	17.74	0.55	17.74	2.25	8.48	0.65	8.48
FF3	3.43	22.13	0.43	22.13	2.20	8.10	0.67	8.10
FF5	4.49	42.08	-0.08	42.08	2.21	8.15	0.67	8.15
OLS	3.08	17.39	0.56	17.39	2.24	8.46	0.66	8.46
LASSO_All_Themes_and_All_Factors	3.83	27.11	0.31	27.11	2.48	9.95	0.60	9.95
LASSO_All_Themes	4.44	40.89	-0.05	40.89	2.24	8.45	0.66	8.45
LASSO_All_Factors	5.44	58.98	-0.51	58.98	3.63	20.39	0.17	20.39

10 Factor Details and Citations

Table 9: Factor and Cluster Details

Description	Variable Name	Citation	Orig. Sample	Sign	Orig. Signif.
		Accruals			
Change in current operating working capital	cowc_gr1a	Richardson, Sloan, Soliman, and Tuna (2005)	1962-2001	-1	1
Operating accruals	oaccruals_at	Sloan (1996)	1962-1991	-1	1
Percent operating accruals	oaccruals_ni	Hafzalla, Lundholm, and Matthew Van Winkle (2011)	1989-2008	-1	1
Years 16-20 lagged returns, nonannual	seas_16_20na	Heston and Sadka (2008)	1965-2002	1	1
Total accruals	taccruals_at	Richardson et al. (2005)	1962-2001	-1	1
Percent total accruals	taccruals_ni	Hafzalla et al. (2011)	1989-2008	-1	1
		Debt Issuance			
Abnormal corporate investment	capex_abn	Titman, Wei, and Xie (2004)	1973-1996	-1	1
Growth in book debt (3 years)	debt_gr3	Lyandres, Sun, and Zhang (2008)	1970-2005	-1	1
Change in financial liabilities	fnl_gr1a	Richardson et al. (2005)	1962-2001	-1	1
Change in noncurrent operating liabilities	ncol_gr1a	Richardson et al. (2005)	1962-2001	-1	0
Change in net financial assets	nfna_gr1a	Richardson et al. (2005)	1962-2001	1	1
Earnings persistence	ni_ar1	Francis, LaFond, Olsson, and Schipper (2004)	1975-2001	1	0
Net operating assets	noa_at	Hirshleifer, Hou, Teoh, and Zhang (2004)	1964-2002	-1	1
		Investment			
Liquidity of book assets	aliq_at	Ortiz-Molina and Phillips (2014)	1984-2006	-1	0
Asset Growth	at_gr1	Cooper, Gulen, and Schill (2008)	1968-2003	-1	1
Change in common equity	be_gr1a	Richardson et al. (2005)	1962-2001	-1	1
CAPEX growth (1 year)	capx_gr1	Xie (2001)	1971-1992	-1	0
CAPEX growth (2 years)	capx_gr2	Anderson and Garcia-Feijoo (2006)	1976-1998	-1	1
CAPEX growth (3 years)	capx_gr3	Anderson and Garcia-Feijoo (2006)	1976-1998	-1	1
Change in current operating assets	coa_gr1a	Richardson et al. (2005)	1962-2001	-1	1
Change in current operating liabilities	col_gr1a	Richardson et al. (2005)	1962-2001	-1	1
Hiring rate	emp_gr1	Belo, Lin, and Bazdresch (2014)	1965-2010	-1	1
Inventory growth	inv_gr1	Belo and Lin (2012)	1965-2009	-1	1
Inventory change	inv_gr1a	J. K. Thomas and Zhang (2002)	1970-1997	-1	1
Change in long-term net operating assets	lnoa_gr1a	Fairfield, Whisenant, and Yohn (2003)	1964-1993	-1	1
Mispricing factor: Management	mispricing_mgmt	Stambaugh and Yuan (2017)	1967-2013	1	1
Change in noncurrent operating assets	ncoa_gr1a	Richardson et al. (2005)	1962-2001	-1	1
Change in net noncurrent operating assets	nncoa_gr1a	Richardson et al. (2005)	1962-2001	-1	1
Change in net operating assets	noa_gr1a	Hirshleifer et al. (2004)	1964-2002	-1	1
Change PPE and Inventory	ppeinv_gr1a	Lyandres et al. (2008)	1970-2005	-1	1
Long-term reversal	ret_60_12	De Bondt and Thaler (1985)	1926-1982	-1	1

Sales Growth (1 year)	sale_gr1	Lakonishok, Shleifer, and Vishny (1994)	1968-1989	-1	1
Sales Growth (3 years)	sale_gr3	Lakonishok et al. (1994)	1968-1989	-1	1
Sales growth (1 quarter)	saleq_gr1		1967-2016	-1	0
Years 2-5 lagged returns, nonannual	seas_2_5na	Heston and Sadka (2008)	1965-2002	-1	1
		Low Leverage			
Firm age	age	Jiang, Lee, and Zhang (2005)	1965-2001	-1	1
Liquidity of market assets	aliq_mat	Ortiz-Molina and Phillips (2014)	1984-2006	-1	0
Book leverage	at_be	Fama and French (1992)	1963-1990	-1	0
The high-low bid-ask spread	bidaskhl_21d	Corwin and Schultz (2012)	1927-2006	1	1
Cash-to-assets	cash_at	Palazzo (2012)	1972-2009	1	0
Net debt-to-price	netdebt_me	Penman, Richardson, and Tuna (2007)	1962-2001	-1	1
Earnings volatility	ni_ivol	Francis et al. (2004)	1975-2001	1	0
R&D-to-sales	rd_sale	Chan, Lakonishok, and Sougiannis (2001)	1975-1995	1	0
R&D capital-to-book assets	rd5_at	Li (2011)	1952-2004	1	0
Asset tangibility	tangibility	Hahn and Lee (2009)	1973-2001	1	0
Altman Z-score	z_score	Dichev (1998)	1981-1995	1	1
		Low Risk			
Market Beta	beta_60m	Fama and MacBeth (1973)	1935-1968	-1	1
Dimson beta	beta_dimson_21d	Dimson (1979)	1955-1974	-1	0
Frazzini-Pedersen market beta	betabab_1260d	Frazzini and Pedersen (2014)	1926-2012	-1	1
Downside beta	betadown_252d	Ang, Chen, and Xing (2006)	1963-2001	-1	1
Earnings variability	earnings_variability	Francis et al. (2004)	1975-2001	-1	0
Idiosyncratic volatility from the CAPM (21 days)	ivol_capm_21d		1967-2016	-1	0
Idiosyncratic volatility from the CAPM (252 days)	ivol_capm_252d	Ali, Hwang, and Trombley (2003)	1976-1997	-1	1
Idiosyncratic volatility from the Fama-French 3-factor model	ivol_ff3_21d	Ang, Hodrick, Xing, and Zhang (2006)	1963-2000	-1	1
Idiosyncratic volatility from the q-factor model	ivol_hxz4_21d		1967-2016	-1	0
Cash flow volatility	ocfq_saleq_std	Huang (2009)	1980-2004	-1	1
Maximum daily return	rmax1_21d	Bali, Cakici, and Whitelaw (2011)	1962-2005	-1	1
Highest 5 days of return	rmax5_21d	Bali, Brown, and Tang (2017)	1993-2012	-1	1
Return volatility	rvol_21d	Ang, Hodrick, et al. (2006)	1963-2000	-1	1
Years 6-10 lagged returns, nonannual	seas_6_10na	Heston and Sadka (2008)	1965-2002	-1	1
Share turnover	turnover_126d	Datar, Naik, and Radcliffe (1998)	1963-1991	-1	1
Number of zero trades with turnover as tiebreaker (1 month)	zero_trades_21d	Liu (2006)	1963-2003	1	0
Number of zero trades with turnover as tiebreaker (6 months)	zero_trades_126d	Liu (2006)	1963-2003	1	1
Number of zero trades with turnover as tiebreaker (12 months)	zero_trades_252d	Liu (2006)	1963-2003	1	1
		Momentum			
Current price to high price over last year	prc_highprc_252d	George and Hwang (2004)	1963-2001	1	1
Residual momentum t-6 to t-1	resff3_6_1	Blitz, Huij, and Martens (2011)	1930-2009	1	1
Residual momentum t-12 to t-1	resff3_12_1	Blitz et al. (2011)	1930-2009	1	1

Price momentum t-3 to t-1	ret_3_1	Jegadeesh and Titman (1993)	1965-1989	1	1
Price momentum t-6 to t-1	ret_6_1	Jegadeesh and Titman (1993)	1965-1989	1	1
Price momentum t-9 to t-1	ret_9_1	Jegadeesh and Titman (1993)	1965-1989	1	1
Price momentum t-12 to t-1	ret_12_1	Jegadeesh and Titman (1993)	1965-1989	1	1
Year 1-lagged return, nonannual	seas_1_1na	Heston and Sadka (2008)	1965-2002	1	1
		Profit Growth			
Change sales minus change Inventory	dsale_dinv	Abarbanell and Bushee (1998)	1974-1988	1	1
Change sales minus change receivables	dsale_drec	Abarbanell and Bushee (1998)	1974-1988	-1	0
Change sales minus change SG&A	dsale_dsga	Abarbanell and Bushee (1998)	1974-1988	1	0
Change in quarterly return on assets	niq_at_chg1		1972-2016	1	0
Change in quarterly return on equity	niq_be_chg1		1967-2016	1	0
Standardized earnings surprise	niq_su	Foster, Olsen, and Shevlin (1984)	1974-1981	1	1
Change in operating cash flow to assets	ocf_at_chg1	Bouchaud, Krueger, Landier, and Thesmar (2019)	1990-2015	1	1
Price momentum t-12 to t-7	ret_12_7	Novy-Marx (2012)	1925-2010	1	1
Labor force efficiency	sale_emp_gr1	Abarbanell and Bushee (1998)	1974-1988	1	0
Standardized Revenue surprise	saleq_su	Jegadeesh and Livnat (2006)	1987-2003	1	1
Year 1-lagged return, annual	seas_1_1an	Heston and Sadka (2008)	1965-2002	1	1
Tax expense surprise	tax_gr1a	J. Thomas and Zhang (2011)	1977-2006	1	1
		Profitability			
Coefficient of variation for dollar trading volume	dolvol_var_126d	Chordia, Subrahmanyam, and Anshuman (2001)	1966-1995	-1	1
Return on net operating assets	ebit_bev	Soliman (2008)	1984-2002	1	1
Profit margin	ebit_sale	Soliman (2008)	1984-2002	1	1
Pitroski F-score	f_score	Piotroski (2000)	1976-1996	1	1
Return on equity	ni_be	Haugen and Baker (1996)	1979-1993	1	1
Quarterly return on equity	niq_be	Hou, Xue, and Zhang (2015)	1972-2012	1	1
Ohlson O-score	o_score	Dichev (1998)	1981-1995	-1	1
Operating cash flow to assets	ocf_at	Bouchaud et al. (2019)	1990-2015	1	1
Operating profits-to-book equity	ope_be	Fama and French (2015)	1963-2013	1	1
Operating profits-to-lagged book equity	ope_bel1		1967-2016	1	0
Coefficient of variation for share turnover	turnover_var_126d	Chordia et al. (2001)	1966-1995	-1	1
		Quality			
Capital turnover	at_turnover	Haugen and Baker (1996)	1979-1993	1	0
Cash-based operating profits-to-book assets	cop_at		1967-2016	1	0
Cash-based operating profits-to-lagged book assets	cop_atl1	Ball, Gerakos, Linnainmaa, and Nikolaev (2016)	1963-2014	1	1
Change gross margin minus change sales	dgp_dsale	Abarbanell and Bushee (1998)	1974-1988	1	0
Gross profits-to-assets	gp_at	Novy-Marx (2013)	1963-2010	1	1
Gross profits-to-lagged assets	gp_atl1		1967-2016	1	0
Mispricing factor: Performance	mispricing_perf	Stambaugh and Yuan (2017)	1967-2013	1	1
Number of consecutive quarters with earnings increases	ni_inc8q	Barth, Elliott, and Finn (1999)	1982-1992	1	0

Quarterly return on assets	niq_at	Balakrishnan, Bartov, and Faurel (2010)	1976-2005	1	1
Operating profits-to-book assets	op_at		1963-2013	1	1
Operating profits-to-lagged book assets	op_atl1	Ball et al. (2016)	1963-2014	1	1
Operating leverage	opex_at	Novy-Marx (2011)	1963-2008	1	1
Quality minus Junk: Composite	qmj	C. S. Asness, Frazzini, and Pedersen (2019)	1957-2016	1	1
Quality minus Junk: Growth	qmj_growth	C. S. Asness et al. (2019)	1957-2016	1	1
Quality minus Junk: Profitability	qmj_prof	C. S. Asness et al. (2019)	1957-2016	1	1
Quality minus Junk: Safety	qmj_safety	C. S. Asness et al. (2019)	1957-2016	1	1
Assets turnover	sale_bev	Soliman (2008)	1984-2002	1	1
		Seasonality			
Market correlation	corr_1260d	C. Asness, Frazzini, Gormsen, and Pedersen (2020)	1925-2015	-1	1
Coskewness	coskew_21d	Harvey and Siddique (2000)	1963-1993	-1	1
Net debt issuance	dbnetis_at	Bradshaw, Richardson, and Sloan (2006)	1971-2000	-1	1
Kaplan-Zingales index	kz_index	Lamont, Polk, and Saaá-Requejo (2001)	1968-1995	1	1
Change in long-term investments	lti_gr1a	Richardson et al. (2005)	1962-2001	-1	1
Taxable income-to-book income	pi_nix	Lev and Nissim (2004)	1973-2000	1	1
Years 2-5 lagged returns, annual	seas_2_5an	Heston and Sadka (2008)	1965-2002	1	1
Years 6-10 lagged returns, annual	seas_6_10an	Heston and Sadka (2008)	1965-2002	1	1
Years 11-15 lagged returns, annual	seas_11_15an	Heston and Sadka (2008)	1965-2002	1	1
Years 11-15 lagged returns, nonannual	seas_11_15na	Heston and Sadka (2008)	1965-2002	-1	0
Years 16-20 lagged returns, annual	seas_16_20an	Heston and Sadka (2008)	1965-2002	-1	1
Change in short-term investments	sti_gr1a	Richardson et al. (2005)	1962-2001	1	0
		Size			
Amihud Measure	ami_126d	Amihud (2002)	1964-1997	1	1
Dollar trading volume	dolvol_126d	Brennan, Chordia, and Subrahmanyam (1998)	1966-1995	-1	1
Market Equity	market_equity	Banz (1981)	1926-1975	-1	1
Price per share	prc	Miller and Scholes (1982)	1940-1978	-1	1
R&D-to-market	rd_me	Chan et al. (2001)	1975-1995	1	1
		Short-Term Reversal			
Idiosyncratic skewness from the CAPM	iskew_capm_21d		1967-2016	-1	0
Idiosyncratic skewness from the Fama-French 3-factor model	iskew_ff3_21d	Bali, Engle, and Murray (2016)	1925-2021	-1	1
Idiosyncratic skewness from the q-factor model	iskew_hxz4_21d		1967-2016	-1	0
Short-term reversal	ret_1_0	Jegadeesh (1990)	1929-1982	-1	1
Highest 5 days of return scaled by volatility	rmax5_rvol_21d	C. Asness et al. (2020)	1925-2015	-1	1
Total skewness	rskew_21d	Bali et al. (2016)	1925-2021	-1	1
		Value			
Assets-to-market	at_me	Fama and French (1992)	1963-1990	1	0

Book-to-market equity	be_me	Rosenberg, Reid, and Lanstein (1985)	1973-1984	1	1
Book-to-market enterprise value	bev_mev	Penman et al. (2007)	1962-2001	1	1
Net stock issues	chcsho_12m	Pontiff and Woodgate (2008)	1970-2003	-1	1
Debt-to-market	debt_me	Bhandari (1988)	1948-1979	1	1
Dividend yield	div12m_me	Litzenberger and Ramaswamy (1979)	1940-1980	1	1
Ebitda-to-market enterprise value	ebitda_mev	Loughran and Wellman (2011)	1963-2009	1	1
Equity duration	eq_dur	Dechow, Sloan, and Soliman (2004)	1962-1998	-1	1
Net equity issuance	eqnetis_at	Bradshaw et al. (2006)	1971-2000	-1	1
Equity net payout	eqnpo_12m	Daniel and Titman (2006)	1968-2003	1	1
Net payout yield	eqnpo_me	Boudoukh, Michaely, Richardson, and Roberts (2007)	1984-2003	1	1
Payout yield	eqpo_me	Boudoukh et al. (2007)	1984-2003	1	1
Free cash flow-to-price	fcf_me	Lakonishok et al. (1994)	1963-1990	1	1
Intrinsic value-to-market	ival_me	Frankel and Lee (1998)	1975-1993	1	0
Net total issuance	netis_at	Bradshaw et al. (2006)	1971-2000	-1	1
Earnings-to-price	ni_me	Basu (1983)	1963-1979	1	1
Operating cash flow-to-market	ocf_me	Desai, Rajgopal, and Venkatachalam (2004)	1973-1997	1	1
Sales-to-market	sale_me	Barbee Jr, Mukherji, and Raines (1996)	1979-1991	1	1

Other Factors

Assets	assets
Sales	sales
Book Equity	book_equity
Net Income	net_income
Enterprise Value	enterprise_value
Current Asset Growth 1yr	ca_gr1
Non-Current Asset Growth 1yr	nca_gr1
Total Liabilities Growth 1yr	lt_gr1
Current Liabilities Growth 1yr	cl_gr1
Non-Current Liabilities Growth 1yr	ncl_gr1
Book Equity Growth 1yr	be_gr1
Preferred Stock Growth 1 yr	pstk_gr1
Total Debt Growth 1yr	debt_gr1
Cost of Goods Sold Growth 1yr	cogs_gr1
Selling, General, and Administrative Expenses Growth 1yr	sga_gr1
Operating Expenses Growth 1yr	opex_gr1
Asset Growth 3yr	at_gr3
Current Asset Growth 3yr	ca_gr3
Non-Current Asset Growth 3yr	nca_gr3
Total Liabilities Growth 3yr	lt_gr3
Current Liabilities Growth 3yr	cl_gr3
Non-Current Liabilities Growth 3yr	ncl_gr3
Book Equity Growth 3yr	be_gr3
Preferred Stock Growth 3yr	pstk_gr3
Cost of Goods Sold Growth 3yr	cogs_gr3
Selling, General, and Administrative Expenses Growth 3yr	sga_gr3
Operating Expenses Growth 3yr	opex_gr3
Gross Profit Change 1yr	gp_gr1a