Classifying Stock Outperformance:

A Machine Learning Framework for Dynamic Factor Allocation

Vrije Universiteit Amsterdam

Duisenberg Honours Programme in Finance and Technology

Student: Vasil Zhiliev (2815500)

Supervisor: Dr. Albert J. Menkveld

Date: 27.06.2025

Abstract

This paper investigates the potential for persistent equity investment outperformance using a dynamic, industry aware machine learning model trained on a subsample of factor-based signals. Drawing from the vast "Factor Zoo" literature, recurring signals across the major risk categories are identified. The most robust proxies from each factor category are selected to form the feature set of a unified ML model. The framework utilizes a rolling-window training design to generate probabilistic forecasts of stock outperformance relative to both, industry peers and a value-weighted benchmark index. The model allows for time-sensitive factor allocation, capturing evolving economic conditions and style rotations within U.S. equities. Results show that over the 44-year out-of-sample period (1981-2024), a long-only strategy based on 29 input variables achieved a compounded annual return of 20.6% with an annualized alpha of 7.3%, statistically significant at the 1% level.

Contents

1. Introduction 2
2. Literature Review 3
2.1. Overview of Traditional Investment Approaches and Product Strategies 3
2.2. Machine Learning in Asset Pricing 5
3. Data 6
4. Research Methodology 8
4.1. Risk Factor Dimensions and Proxies 8
4.2. Feature Implementation and Engineering 8
4.3. Model Design 10
4.4. Training and Validation 11
4.5. Portfolio Formation 12
4.5. Evaluation Framework 13
5. Results 14
5.1. Holistic Overview 14
5.1.1. Top Minus Bottom Deciles 15
5.1.2. Prediction Accuracy/Hit Rates 16
5.1.3. Portfolio Composition 17
5.1.4. Performance Breakdown and Robustness Checks 20
5.1.5. SHAP Values/Feature Importance 23
5.1.6. Factor Relevance Over Time 25
6. Conclusions, Limitations & Implications for future research 26
7. References 29
8. Appendices 33

1. Introduction

Over the past decades, there has been a growing body of literature identifying well-established predictors that form broad factor categories such as Value, Momentum, Profitability, and Investment, among others. The accumulation of such signals is now referred to as the "Factor Zoo," reflecting both the depth of reported findings, and the growing concern over their reliability and intuitiveness. Some factors have proven to be robust across time and different geographies, while also remaining economically grounded, but many others suffered from publication bias, had data mining concerns or experienced poor realized performance.

This increasing number of return predictors creates two important challenges. Firstly, the vast amount of proposed signals is accompanied by the complex task of determining which ones are reliable and economically meaningful. Second, and more importantly, even for those signals that have been validated, their implementation and degree of utility remains constrained by numerous structural limitations. Traditional factor-based frameworks/approaches are often built on rigid assumptions, include fixed, linear relationships and have static weighting schemes, among other shortcomings, resulting in a general inability to deal with broader changes in financial markets. Such models can struggle to reflect a time-varying nature of factors relevance and often operate without sufficient contextual awareness of sectoral or macroeconomic context.

When combined, the issues of factor relevance and successful implementation raise an important question for researchers and practitioners alike: whether an adaptive framework can systematically navigate the Factor Zoo, while preserving economic interpretability and delivering persistent outperformance.

This thesis investigates whether modern machine learning approaches can intelligently integrate traditional factor signals and overcome the limitations of static investment frameworks, with the ultimate objective of generating superior returns.

To support this overarching objective, the study explores several related sub-questions:

- Of the utilized signals, which are most predictive across decades, and how does their relevance evolve over time?
- Do firm groupings based on industry (GIND) economic activity provide strong context towards the model's predictions?
- Do model diagnostics, such as SHAP values, help reveal persistent economic structure and support interpretation of the model's internal logic?

In answering these questions, quarterly fundamental and price data for U.S. equities from 1961 to 2024 is used as the study's dataset. The feature set consists of 28 factor-based variables from Robeco's "Factor Zoo" subset. Alongside them, a categorical variable is created based on the GICS firm classification, (GIND), with firms being assigned to their respective industries. XGBoost is selected as the model architecture due to its favourable trade-off between efficiency and reported accuracy. A rolling window training design is used, re-fitting the model after each quarterly interval, with data oversampling of the last 4 years to emphasize recent market dynamics. The target is a binary indicator showing whether a stock's next-quarter returns exceed that of the industry it belongs to, alongside that of the market index.

Results show that the proposed framework consistently outperformed benchmarks for the entire out of sample period (Q1 1981 – Q4 2024). The Amalgam model achieved compounded annual returns of 20.6%, maintaining a Sharpe ratio of 1, while also having superior risk-adjusted returns. The model's average accuracy in identifying outperformers is aligned with similar findings, sitting at 51.3%. Despite the low accuracy, excess

performance persisted due to the positive asymmetry in the return distributions between winning and losing stocks (large gains in the top decile and substantial losses avoided in the bottom). Implementing trading costs results in a modest yearly average portfolio value decrease of 0.92%, which is attributable to the lower rebalancing frequency (quarterly as opposed to monthly). Fama French factor regressions display a strong annualized average gross alpha of 7.3%, significant at the 1% level, which decreases moderately to 4.95% if adjusted for transaction costs. Factor and feature importance analyses reveal "Value" and Industry belonging as the most consistent and informative towards classifying outperformance, with others such as "Leverage" and "Momentum" declining in importance over the decades.

Overall, this study addresses gaps in factor-based asset pricing by establishing, empirically validating and dissecting the results of a detailed machine learning framework. It contributes to the existing literature by combining economic intuition from multiple fields and implementing them within a data-driven method, thus proposing a robust, adaptable and transparent framework for quantitative investing.

2. Literature Review

The following section outlines common investment approaches found in academic literature and practice. Each is examined in terms of implementation logic, along with the main strengths and drawbacks. Placing them within the broader context of asset pricing establishes the groundwork for subsequent evaluation. The overarching aim is to assess whether a custom framework can deliver meaningful, lasting outperformance by preserving each strategy's strengths while mitigating its weaknesses.

2.1. Overview of Traditional Investment Approaches and Product Strategies

2.1.1. Single-Factor Strategies

These strategies utilize unidirectional (long-only), well-researched anomalies with robust historical premiums to form portfolios, one risk dimension at a time. No complex modelling is done beyond computing factor metrics and ranking stocks by their raw exposure to the selected dimension (Pimco, n.d). For example, value factor portfolios would rank stocks by high book-to-market or low P/E and invest in a value or equal weighted selection of the companies which have the highest ratios. While such strategies are widespread and are represented by many "smart beta" investment vehicles, they typically do not account for inter-sector differences which can lead to sectoral bias by the sheer nature of individual companies' business and accounting structure (e.g. Technology Sector stocks tend to always have high(er) book-to-market ratios than Consumer Discretionary stocks). Another feature that serves as a potential drawback for such unidirectional strategies lies within the sacrifice of depth. One-dimensional smart beta funds/products explicitly overweigh the top-ranked stocks and under weigh, exclude, or sometimes short bottom-ranked stocks (of that dimension) which creates a very narrow source of returns. They can also be seen as naive, meaning that expected returns are tightly associated with the factor risk premium, without explicit probability measurements or return forecasts. This serves as a downside because individual factors/anomalies can fade, have diminishing long-term returns, or be entirely cyclical. For example, a momentum factor portfolio can suffer severe drawdowns, upwards of 50% in recessionary environments, and be subdued in relation to other factors for a prolonged period of time.

2.1.2. Multi-Factor/Composite Strategies

Multi-Factor strategies expand on the previous section by combining several signals as a framework for stock-selection. There can be as little as two factors used simultaneously comprising such strategy. Their combinations differ depending on the product "style", and the formulation of such portfolios add a degree of rigor. Stocks are ranked by a composite score, which is an additive or weighted average combination of their exposure to different risk dimensions. Similar to single factors, these target medium-term expected returns and because multiple signals are combined, performance over time is expected to be more stable. Further, specific risk dimensions can be complementary when combined due to diverging risk sources. For example, value and momentum tend to be negatively correlated, so a stock that has a high composite score by ranking well on both can provide a powerful source of diversification (Barber et al., 2015). Therefore, if one factor is underperforming, the other(s) may be producing superior returns, which reduces the portfolio's holistic volatility and tracking error relative to broad-market indices, which makes for a robust strategy (Pimco, n.d[2]). While such strategies are a marked improvement over univariate factor strategies, several difficulties can arise during their implementation: A stock with a high composite score in practice, can sometimes be **driven by one factor** if others are underperforming during the same time. By design, the approach would choose the strongest value stocks if/when momentum score premiums are globally out of favor. In such extreme cases a multifactor portfolio is essentially truncated to a single-factor one, losing the intended diversification benefits. Moreover, composite ranks are often weighed on a fixed basis, which implicitly assumes that factor premiums are stable and **equally** important over-time. If practitioners want to alter this aspect, they can introduce bias or misjudge by manually assigning weights to each proxy comprising the composite score (e.g Composite Score = 0.7.*Value Rank + 0.3*Momentum Rank). This can prove to be difficult and is heavily liable if, and when, factor efficacy varies (Barber et al., 2015[2]).

2.1.3. Stock-Screening Rules Based Approaches

These strategies are perhaps one of the most wide-spread stylistic investment tools. They utilize academic or practitioner insights and are often propagated by funds and investing professionals/gurus. Rules- based approaches, similar to multi-factor strategies, include frameworks comprised of (broad) sets of fundamental and behavioral proxies in order to filter and rank stocks, to form a portfolio. Some notable examples include "The Magic Formula" by Joel Greenblatt, "Conservative Investing" by Blitz and van Vliet and other industry-based readily available products such as the RAFI Indices. They diverge from multifactor approaches in the manner by which they assess a stock's suitability. Oftentimes, custom proxies/ratios, thresholds and proprietary insights are utilized (e.g. Using Operating Cash Flow to Price Ratio as a value factor risk proxy instead of the academic standard of BtM ratio, or using Book-Value instead of Market Value weights for index construction, etc.). They vary widely in their implementation and "model" assumptions, which makes it hard to pin-point any one definitive downside. What permeates across the entire category, however, is that they wholeheartedly rely on practitioner judgement, and a continuity of performance for the specific framework post-publication. Some of these formulas/product indices have been documented to experience diminishing returns post-publication, with a strong downward trend in recent history, which is suggestive that whatever anomalies were present in the first place, are nearly exhausted for a variety of reasons, one being their popularity and potential widespread application (Schwartz & Hanauer, 2024). Moreover, many traditional approaches are inherently rigid. They assume a linear and additive relationship between risk-factor proxies and stock performance, often omit sector or industry context, and assess companies in absolute rather than relative feature ranks. This all serves as a potential blind-spot for bias, and an over-reliance on static frameworks. Similar to the previous approach, the scoring of individual features can be seen as a heuristic. For example, the Magic Formula equally weights two ratios as the *definitive* way to measure a stock by its value and quality rank, which can seem arbitrary. Finally, as a segue to the next methods, rule-based strategies lack a formal mechanism for

understanding correlations across inputs and offer no probabilistic guidance or confidence levels for performance, as they are either inclusive or exclusive of a stock based on the predefined narrow set of dimensions.

2.1.4. Linear Factor Models

Historically, the linear cross-sectional model can be seen as the "benchmark" approach for both academics practitioners alike (Fieberg et al., 2022). These models are largely unconstricted and allow for incorporation of a broad array of firm-level characteristics. Efficacy of individual risk dimensions can be assessed by testing dozens of factors into regressions or similar linear models to forecast stock-level returns more broadly. They allow for the estimation of marginal effects from each variable while controlling for the presence of remaining features. Sector or industry context can and is typically built-in through the inclusion of categorical sector dummies for instance. One of the major strengths of linear factor models lies in their interpretability and systematic structure. Unlike previously discussed heuristic/fixed-weight models, linear regressions find the statistically optimal combination of factors from historical data, account for multicollinearity among predictors, and provide interpretable signals. Because of their simplicity, transparency, and tractability, linear models have long been the cornerstone of quantitative alpha generation. However, these models are not without limitations. Perhaps their biggest flaw is that similar to previous methods, they assume that the relationship between factor exposures and expected returns is linear and additive. Furthermore, linear models are prone to overfitting, especially as the feature set expands, which is an issue when evaluating the robustness of individual signals from large "Factor Zoo" datasets. Furthermore, coefficient estimates can also be unstable over time, as factor premia shift across market regimes. This can cause the use of long historical windows to be unreliable in assessing premia, because they can dilute current relationships. In comparison, shorter timeframes on the other hand, can thus be overinfluenced by noise (Dujava, 2024). Most importantly, linear models do not inherently recognize nonlinearities or specific interactions unless they are explicitly engineered into the feature set. For instance, if a certain value signal only works in low-volatility environments, this nuance must be manually encoded through interaction terms. Accounting for such variability and "one offs" can be limited, as these models are moderately scalable.

2.2. Machine Learning in Asset Pricing

In the last decade or so, ML approaches have gained traction as vast financial datasets and cheaper compute resources enabled users to move from handcrafted, rule-based strategies to more flexible, granular and scalable alternatives. ML models can automatically process complex, nonlinear interactions to identify subtle patterns that linear or hierarchical models often miss. This intricacy makes ML thrive in big datasets containing hundreds of predictors, without having to pre-specify linear relationships or manually encode potential "quirks" between factor interactions.

Compared to the previously defined conventional approaches, ML models consider all inputs jointly and are especially good at filtering based on their observed robustness. Qualitative context can also be embedded explicitly through categorical inputs, allowing for "custom" logic across subgroups within the broad data. For example, tree-based models have the ability to discover that a given predictor, such as net income-to-market equity ($\frac{NI}{ME}$), carries varying importance depending on the time frame, the sector, or at the individual firm level. Frameworks like random forests (RF) or gradient boosting machines (GBM) have been documented to more than double the Sharpe ratios compared to linear factor models, namely due to such increased utility (Gu, Kelly, and Xiu, 2020). Studies explicitly on non-US markets also report similar results, with ML measures

outperforming linear benchmarks by ~0.6% per month, and without an increase in risk or turnover (Fieberg et al., 2022). Additionally, regularization through hyperparameter tuning and validation can help mitigate redundancy and noise brought on by countless predictors in the "Factor Zoo". Ensemble ML methods can offer further stability through techniques like bagging and shrinkage, with research suggesting this adds value during market downturns (Nti et al., 2020).

However, these benefits are not without serious caveats. The biggest drawback of ML models is that they are often considered "black boxes," due to their lack of transparency and unintuitive process when modelling. The added flexibility raises the risk of overfitting, particularly in high-dimensional datasets, such as financial ones, where signals can be weak and fleeting. Also, if not properly validated, models may inadvertently focus on transient data (noise) rather than real structure. Moreover, over-reliance on in-sample performance without an extended, well documented out-of-sample evaluation can lead to false confidence in the model's predictive capabilities. Additionally, particularly rigorous models, such neural networks (NN) with many nodes, can often underperform simpler models in financial applications due to cross sectional or time series limitations in the underlying data (Lekan et al., 2025). Ultimately, while ML can revolutionize quantitative asset pricing, its success is contingent on careful model selection, training design, regularization, and extensive post-hoc analyses that trace and validate the underlying framework's output. Without these measures, the same complexity that made ML appealing in the first place can end up undermining its utility altogether, as interpretability and practical application will be highly limited.

3. Data

The datasets used in this study are derived from two primary sources: the Center of Research in Security Prices (CRSP) and the CRSP/Compustat Merged Database. CRSP provides daily market data for all publicly listed US equities, including price, volume, and returns. The CRSP/Compustat Merged Database supplies all quarterly accounting data. In combination, these two sources provide complete firm-level coverage for both market-based and fundamental variables.

This sample universe consists of all US common shares, identified with CRSP codes 10 and 11. Securities such as private shares, preferred shares, hybrid shares, ETFs are excluded. Returns and fundamental data from all delisted firms are retained to counteract the impact of survivorship bias. Missing or undefined accounting items are manually imputed based on common frameworks that use derivations from adjacent line items (Jensen et al. ,2023), (Schwartz and Hanauer, 2024). For firms with different accounting structures, such as those in the Financials or Utilities sectors, industry-specific line-item equivalents are applied based on classification variables from CRSP/Compustat. This treatment is further detailed in the methodology section.

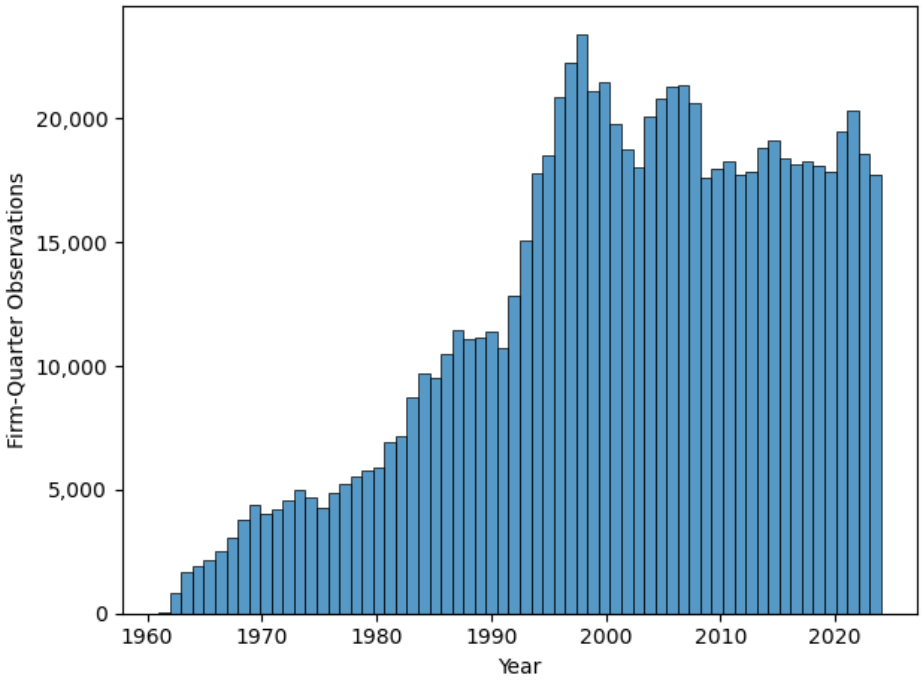
Furthermore, several adjustments are made for key variables. Firstly, observations with negative or zero book-to-market (BtM) values are set as missing (NaN) to avoid distortions from accounting anomalies or disproportionate leverage (Fama&French, 1992). Secondly, Research & Development (R&D) expenditures are capitalized and amortized over their estimated useful life, instead of being expensed immediately, following Damodaran's proposed framework (Damodaran, 2025). Table 1 describes all affected accounting variables and their adjustments. Table 2 provides the R&D amortization schedules by industry.

A notable limitation in the dataset is the lack of publication dates for quarterly financial statements. Approximately 89% of firm-quarters have no timestamps, therefore it is unknown whether accounting updates were released in time, before the end of a given fiscal quarter. For such instances, the academic convention is

to apply a 2-quarter lag (Fama&French, 1992). This ensures all accounting data had been publicly available to investors at the time of portfolio formation and addresses potential leakage of future information during model training. A one quarter lag is applied to the remaining 11% of firm-quarters with known publication dates that *precede* the last day of a fiscal quarter. The remaining portion of data is based on market variables and is uniformly lagged by one quarter, since there is no risk of lookahead bias.

Overall, the study spans from Q1 1961 to Q4 2024, representing *the maximal* period during which coverage of both accounting and market data is available in CRSP/Compustat. All firm quarters with market capitalizations (MC) below $50 million (in real terms) were excluded to reduce the impact of illiquid microcap investments. In total, the dataset covers 64 years, consists of 27,167 firms and contains approximately 1,249,397 firm-quarter observations (shown in Figure 1), offering a comprehensive view of firm performance and financial characteristics across varying market conditions.

Figure 1: Scope of the sample period



4. Research Methodology

The following section describes the implementation and rationale behind the core topics introduced in the literature review. It shows how feature selection, data engineering, and model architecture are operationalized, establishing a baseline for evaluation and discussion in later sections.

4.1. Risk Factor Dimensions and Proxies

This paper utilizes a condensed version of the "Factor Zoo" by replicating 28 out of the 30 most statistically robust proxies identified in Robeco's "Factor Zoo (.zip)", (Swade & Hanauer, 2024). These proxies span key risk dimensions such as value, profitability, quality, investment and low risk, among others. They were selected based on the demonstrated ability to explain the cross-section of expected returns and reduce alpha via a Gibbons, Ross and Shanken (GRS) test at the 5% and 1% significance levels across a broad universe of US stocks. This reduced proxy count not only offers a clear distinction between predictive signal and statistical noise, but it is also particularly suitable when modelling with limited compute power. For a complete overview of each feature, along with its formulaic expression, refer to Table 3.

In addition to using factor-based inputs, firm classification approaches have gained increasing attention in recent literature. Whether based on Fama-French industry definitions, SIC/NAICS codes, or related commercial taxonomies, the grouping of firms has proven to be non-trivial in evaluating both "within-group" and "across-group" performance (Hanauer et al., 2025). Fortunately, this importance is now empirically proven, not just based on intuition. Building on this, I extend on traditional approaches by adopting a more granular firm classification than is typically used. While prior studies often rely on 11 to 12 sector conventions, such groupings might obscure finer, within-sector, differences that would emerge only when categories are disaggregated further (Bagnara & Goodarzi, 2023).

To capture these potentially nuanced effects, the Global Industry Classification Standard (GICS) is utilized, focusing on the 85-category six-digit industry level (GIND). GIND offers a good medium since it is more granular than conventional 12-count sector classifications, while also being less scattered than SIC codes, which are outdated beyond 1987 and contain hundreds of sparsely populated/obsolete firm categories. Such intermediate level of detail can preserve essential within-sector variation, while also remaining feasible for implementation within a machine learning model that relies only on local compute power.

4.2. Feature Implementation and Engineering

Before calculating any of the financial ratios that form the feature set, the relevant line items from the three financial statements must be defined. The process follows the methodology of related implementations (Schwarz & Hanauer, 2024), providing formulas using the exact CRSP/Compustat variable names. However, two important additions to their approach are applied in this study.

First, the legacy accounting treatment of research and development (R&D) expenses, where R&D is fully expensed as part of general operating costs (SG&A), can distort *financial* analysis. Direct expensing assumes that R&D expenditures provide no economic benefit whatsoever, beyond the current period. This is inconsistent with the role R&D can have towards long-term value creation, particularly in innovation-intensive sectors (Information Technology and Healthcare, among others). Many studies have argued that R&D investments have lasting impact and contribute significantly to future firm valuation levels (Lev & Sougiannis,

1999) Penman & Zhang, 2000). Thus, expensing R&D can understate actual asset values and future earning potential. This biases performance comparisons negatively for firms/industries that rely heavily on re-investment as a primary source of growth. In this study, R&D is reclassified as a capital investment and amortized over its useful life to better reflect its role in future cash flow generation. This modified treatment alters key line items, such as cash profitability, long-term assets, and book equity. Further, the amortization schedules for these research assets follow a GICS industry classification logic, recognizing that useful lives differ for firms across different industries.

The second change concerns firms with SIC codes 6000–6999 (Financials) and 4900–4999 (Utilities), where general accounting fields are substituted with the relevant industry-specific variables provided by CRSP. This ensures that individual line items accurately reflect differences in accounting conventions for non-industrial firms (e.g. using "cheq" (cash) for industrial firms and "finchq" for financial firms). For a complete overview of the variable implementation logic, refer to Table 4. Once all accounting inputs are established, the 28 continuous feature proxies used for modelling are constructed, per the definitions mentioned in Table 3.

To embed industry awareness into the model, each firm-quarter observation is grouped by its corresponding GIND code, with the full 85 GICS industries encoded as categorical variables alongside all continuous feature proxies. These continuous features were initially expressed as raw values (e.g., a firm in GIND 551020 reports $\frac{NI}{ME} = 5.9\%$ in Q2 2010). While such format offers interpretability, it can be problematic. Most machine learning studies utilize raw-value inputs across observations, but this may not be optimal for financial ratio data.

It has been demonstrated that, all else equal, reducing the magnitude of a feature's variation can improve the signal-to-noise ratio (STN) within ML models during training, possibly resulting in better performance out of sample (Chen et al., 2024). Practitioners often normalize values by subtracting the group-level mean/median, and/or scale individual observations by their training period standard deviation. While this is a common and valid approach, its implementation given the context of this study can prove to be suboptimal for two reasons:

Firstly, normalizing via z-scores assumes an approximately symmetric distribution which is rarely true for financial ratio data, especially when computed on a quarterly basis. Many financial ratios exhibit high skewness caused by a small number of unusually distressed/profitable firms, which can disproportionately influence the group mean/median (e.g. GIND average $\frac{NI}{ME} = 5\%$ for Q2 2010, but 3 firms have $\frac{NI}{ME} = -8\%$).

Second, especially in earlier sample periods, GIND groups can contain fewer than 30 firms, causing a feature's parametric transformation to be statistically unstable/unreliable.

Nevertheless, there is a good medium that both, expresses raw values in lower magnitudes to increase STN, and is also consistent for firm-quarters with lower observations. To address these issues, percentile-based transformations are utilized for each feature within its GIND group, applied to each firm-quarter. With this approach the cross-sectional order of firms is preserved, while also compressing feature magnitudes to the "0-1" interval. For example, a firm with a moderate $\frac{NI}{ME}$ within its GIND for Q2 2010 is encoded as 0.5.

Unlike winsorization, this method does not omit the effect and validity of outlier firms, but it also does not allow their raw magnitude to dominate the model's learning process. Instead, the impact is implicitly capped to the bounded positions.
With this approach, the aim is to achieve a balance between interpretability, robustness to outliers and statistical consistency, even in GIND categories with consistently limited sample sizes. For a complete overview regarding the implementation of this process, refer to Table 5.

4.3. Model Design

To counteract the main limitations of traditional asset pricing approaches, being assumptions of linear and additive relationships and static forecasts based on heuristics, this study adopts a ML model with a high data-fitting capacity. Such models can "uncover" and utilize complex, nonlinear, time-varying interactions between firm characteristics and future returns, which are all elements that conventional models and investment formulas are likely to miss. This is particularly important given the fact that financial markets are non-stationary, can exhibit path dependency and overgo regime shifts/structural breaks often (Lo, 2004), (Ang & Timmermann, 2011). Because of this, the architecture must reflect market reality, which is ever evolving and chaotic in nature. In doing so, the model ought to be re-trained on a quarterly basis, which is the highest temporal frequency available in the dataset. This results in a high training cycle count (64 years; 256 quarters), imposing significant computational demands. Nevertheless, it is seen a **necessary** design choice that aims to minimize signal decay and ensures accurate responsiveness to broad shocks or trend reversals.

Moreover, given the long span of the data sample, it is also crucial to mitigate historical bias, while also not overfitting to short-term noise. Having this in mind, the training architecture prioritizes market performance in the latest 4 years, based on the assumption that the most recent history can be more representative of current and near-future market conditions. While it is just an assumption, it holds some theoretical grounds based on literature covering temporal relevance in financial modeling (Bengio et al, 2014), (Spiliotis, 2023). Therefore, in selecting a suitable framework, several practical constraints must be considered given the above arguments. The model should have proven strong performance via generalization, a high ability to detect nonlinear interactions and have high robustness to multicollinearity and noise, while maintaining feasible training times given the necessary condition of iterative re-fitting on a quarterly basis.

Table 6: Comparison of ML models on their fitting ability and training time

This table compares the fitting ability and training time of the some commonly used ML models. In data fit, 'low' means that the model assuming simple or linear relationships; 'medium' means that the model can capture non-linear relationships to a certain extent but is prone to overfitting; and 'high' means that the model can capture intricate and highly non-linear relationships and can avoid overfitting with appropriate setting of hyperparameters.

Model	Data Fit	Training Time
Logistic Regression	Low	Short
Decision Trees	Medium	Short
Random Forest	High	Medium
Support Vector Machines	High	Long
Neural Networks	High	Long
Gradient Boosting	High	Medium

These criteria narrow the model candidates to tree-based methods like Random Forest (RF) and Gradient Boosting Machines (GBM). While RF fit the criteria, prior literature has consistently shown GBMs to outperform RF accuracy-wise across a range of datasets (Bentéjac et al., 2021). Within the GBM range, XGBoost stands out as a powerful framework, as it combines forecasts from multiple "weak learning" shallow decision trees into a single strong learner. This adaptive learning structure results in detailed pattern detection, making it a good candidate for large-scale classification tasks, while also offering plenty of tuning options to counteract overfitting (Chen & Guestrin, 2016). While alternatives such as Support Vector Machines (SVM) and Neural Networks (NN) have demonstrated strong performance as well, their application is accompanied by substantially higher training times and RAM memory requirements (Nalepa & Kawulok, 2018) (Chen et al., 2016), (Almaspoor et al., 2021). Therefore, their scope makes implementation cumbersome, especially for large-scale, rolling-window datasets that are ran on the hardware of one personal computer, as is the case in this study.

4.4. Training and Validation

To ensure an appropriate amount of emphasis on the most recent market dynamics, the model is re-trained at each quarterly prediction step, using all available historical data up to, but not including, the prediction quarter. With such approach, the training window is anchored at the beginning of the dataset, expanding with each quarterly update, after predictions have been made. The end of the Training Set always includes the final quarter prior to the quarter being forecasted in the Test Set. For example, the outperformance predictions in Q1 1981 use data from Q1 1961 to Q4 1980, and projections in Q2 1981 include all data from Q1 1961 through Q1 1981, and so forth. This setup essentially accumulates training data with the progression of quarters, while preserving a strictly *out-of-sample* forecasting structure. For an overview of the timeline in the example, refer to Figure 2.

As part of the whole training set, observations in the last 4 years are reserved for validation and tuning. From the visual example, in the first fold, the validation set represents 20% of the full training set (Q1 1977 – Q4 1980). The aim during validation is to select a hyperparameter combinations that minimize log-loss function values, resulting in a more accurate model.

Table 7: Hyperparameters Ranges.

Hyperparameter in XGBoost classifier	Candidate values	Initial Values
max_depth	[3, 4, 5, 6, 7]	6
min_child_weight	[1, 3, 5]	1
gamma	[0, 1, 10, 20]	0
subsample	[0.6, 0.8, 1.0]	0.8
colsample_bytree	[0.6, 0.8, 1.0]	0.8
reg_alpha	[0, 0.1, 1, 5]	0
reg_lambda	[0, 1, 5]	0
learning_rate	[0.01, 0.05, 0.1]	0.1
n_estimators	[10, 20, 30, 50, 75, 100]	30

Rather than optimizing over all hyperparameters simultaneously, a four-tiered grid search is applied to prevent excessively long run-times and overfitting due to a large search-space. The hyperparameter types and their respective ranges shown in Table 7 are divided into four subsets and tuned sequentially, starting from the proposed list of Initial Values. The first combination includes "max_depth", "min_child_weight" and "gamma"; the second includes "subsample" and "colsample_bytree"; the third contains "reg_alpha" and "reg_lambda"; and lastly "learning_rate" and "n_estimators". This hierarchy is based on each hyperparameter's group influence on reported model performance (Putatunda and Rama, 2018). The candidate values are directly taken from commonly used ranges in relevant literature (Bentéjac et al., 2021), (Kavzoglu & Teke, 2022), (Putatunda & Rama, 2018). "n_estimators" are purposefully constrained to a lower range since the signal-to-noise ratio (STN) in financial data is relatively low, opening the possibility of overfitting.

Lastly, to account for the time-varying nature of return predictability, the last four years of each training set are oversampled. Oversampling is a common technique and can be implemented via replication and targeted sampling from a minority class (Mohammed et al., 2020). With this application of oversampling, it is assumed that recent data can better capture current and near-future market dynamics. In this study, the training set is divided into two segments: an older segment **"n"**, consisting of observations older than four years from the prediction fold, and a recent segment **"m"**, with observations *within* the last four years. By design, **"m"** is the minority class with less firm-quarter observations, and **"n"** is the majority, making $\boldsymbol{m < n}$ always true.

With oversampling, the aim is to equalize the number of observations between the two segments, making $\boldsymbol{m = n}$. This is how more recent market dynamics in **"m"** are emphasized equally with respect to older ones in **"n"**. Thus, observations from the recent segment are duplicated $\left\lfloor \frac{\boldsymbol{n}}{\boldsymbol{m}} \right\rfloor$ times, and **"m"** mod **"n"** additional observations are then randomly drawn from the recent segment, balancing the minority training set. This (oversampled) recent segment is then merged with the older segment, forming the final training set used by the model. Even though the oversampling process coincides with the validation set's timeframe (latest 4 years), oversampling is performed only **after** hyperparameter tuning, having no influence on the hyperparameter selection itself.

After tuning and oversampling, the optimal combination of hyperparameters is used to re-train the entire model on the balanced training set, and the model is then deployed on the test set. This process is repeated on a quarterly basis for the full duration of the back test.

4.5. Portfolio Formation

This study uses XGBoost for probabilistic classification with a dual criteria labelling scheme. In training, firms were assigned a label of "1" (outperformer) if they met two conditions within a given quarter:

1) their return percentile relative to firms in the same GICS industry group ("ret_pct_gind") exceded 0.5 (median)

2) This same return percentile also surpassed the value-weighted market index return ("vwretd_pct").

Otherwise, the firm was labelled "0" (underperformer). The dual condition ensures that firms are not only outperforming industry peers, but that they also exceed the benchmark set by the value-weighted index. This combines sector and market-relative criteria, making the model identify and learn from stocks with both, *localized* and *absolute* outperformance.

In addition to the dual label-scheme, this study adopts common and related methods in forming predictions for classification tasks (Rasekhschaffe & Jones, 2019). Each observation in the dataset represents a firm-quarter, where the current features are used to rank firms in the subsequent quarter. Once the binary labels are assigned, the following three-step process follows:

1) Train the model on the balanced historical data. Rather than assigning a discrete label, a continuous probability score (0 to 1) is the test-set's output, reflecting the estimated likelihood of outperformance in the next quarter.

2) After being fitted, the trained model generates probability scores for *all firms* in the test set. These scores are then ranked in descending order, reflecting the odds of outperformance from highest to lowest.

3) From the sorted predictions, the top and bottom deciles (top and bottom 10% of stocks for each quarter) with the highest and lowest predicted odds are selected to form the long and short portfolios used for subsequent analysis.

4.5. Evaluation Framework

Since the model is trained and provides output on a quarterly basis, stock selection and portfolio rebalancing are also at quarterly intervals. Each period, firms are sorted by their outperformance projections, with long and short portfolios reflecting the top and bottom deciles. The deciles are equally weighted, consistent with the convention in asset pricing studies. The model's effectiveness is then evaluated using a range of qualitative and quantitative diagnostics:

First, the holistic out-of-sample (OOS) performance is assessed, including benchmarks for cumulative returns and other relevant performance metrics, comparing the Amalgam model with broad market benchmark indices. In this section, an overview of the average industry composition and GIND concentration is included, along with hit rates/prediction accuracy for the long and short portfolios.

Second, the portfolios are regressed on a Fama-French five-factor model, including momentum (FF5+MOM), to analyze the sources of returns. This gives a detailed snapshot of the long, short and "long minus short" strategies, revealing their exposure to common risk factors, with the possibility of statistically significant alpha.

Third, one-sided t-tests are conducted to assess the significance levels of risk-adjusted returns from the top-decile portfolio, compared to benchmark indices.

Lastly, model insights are shown via SHAP values (SHapley Additive exPlanations), which quantify the absolute and signed, average contributions from each feature towards the probability of outperformance. The SHAP analysis is extended by tracking the evolution of factor importance across decades, offering insight into how the model's predictive focus shifted over time.

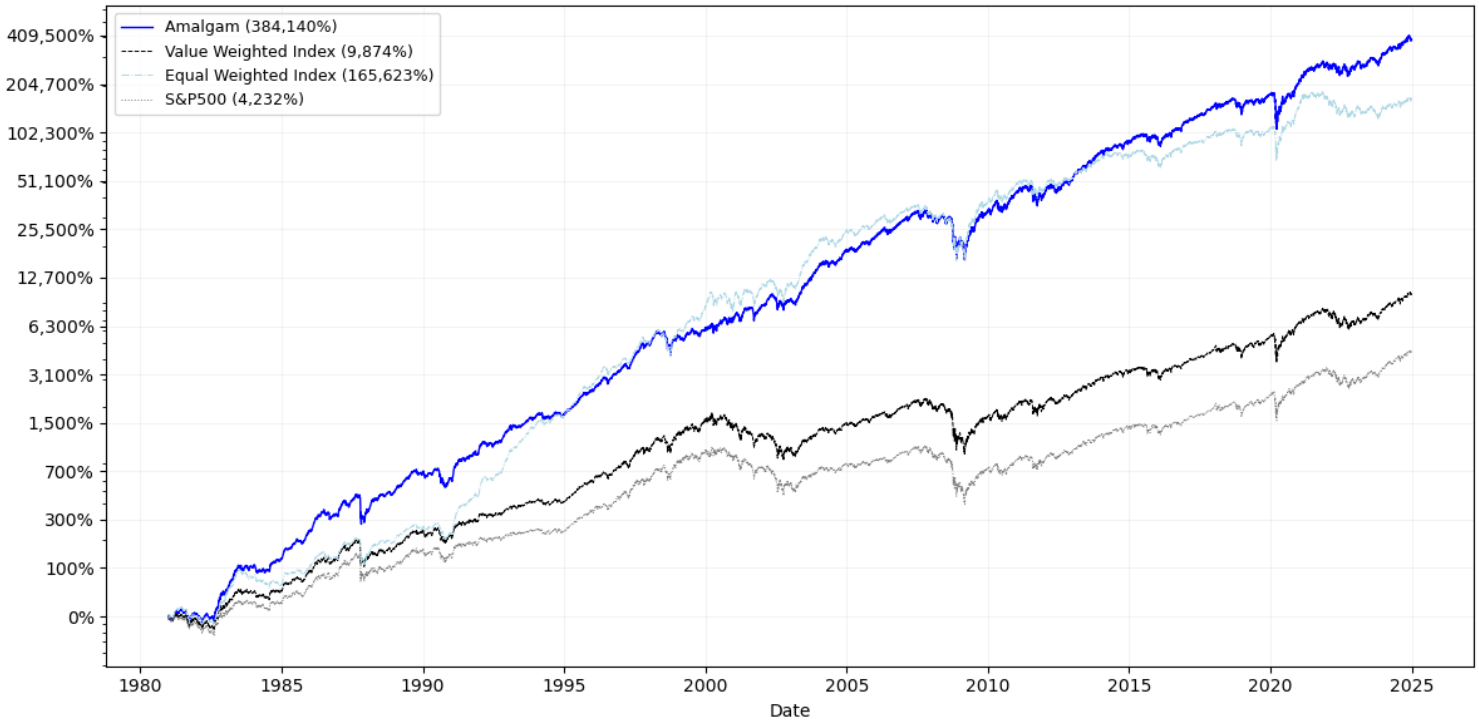
It is important to mention that while a long-short decile strategy is used for evaluation, it was motivated solely by the academic convention in empirical finance studies. The "short leg" is included mainly to contextualize and provide contrast to the top decile's return magnitude, with its real-world feasibility being limited, perhaps. Constraints such as regulatory short-sale bans, high borrowing costs and insufficient share float designated for short selling, among others, may have prevented the full implementation of the bottom decile portfolio in given

periods, or for certain stocks. Therefore, in practical terms, the most relevant version of this strategy is the long-only, buy-and-hold top-decile portfolio.

5. Results

5.1. Holistic Overview

Figure 3: Holistic Performance Comparison: Amalgam vs Broad Market Indices



*All return progressions include dividend distributions and reinvestment.

Q1 1981 – Q4 2024	Cumulative	CAGR	St. Dev.	Sharpe Ratio	M.D.D.
Amalgam Model	384,140%	20.6%	16.3%	1.0	(50.7%)
Equal Weighted Index	165,623%	18.4%	14.9%	0.95	(55.1%)
Value Weighted Index	9,874%	11.0%	17.3%	0.47	(55.5%)
S&P500	4,232%	8.9%	17.9%	0.36	(56.8%)
Q1 2020 – Q4 2024	**Cumulative**	**CAGR**	**St. Dev.**	**Sharpe Ratio**	**M.D.D.**
Amalgam Model	116.5%	16.7%	22.7%	0.69	(40.7%)
Equal Weighted Index	52.3%	8.8%	21.2%	0.39	(38.5%)
Value Weighted Index	81.4%	12.7%	21.1%	0.56	(34.7%)
S&P500	82.1%	12.7%	21.3%	0.56	(33.9%)

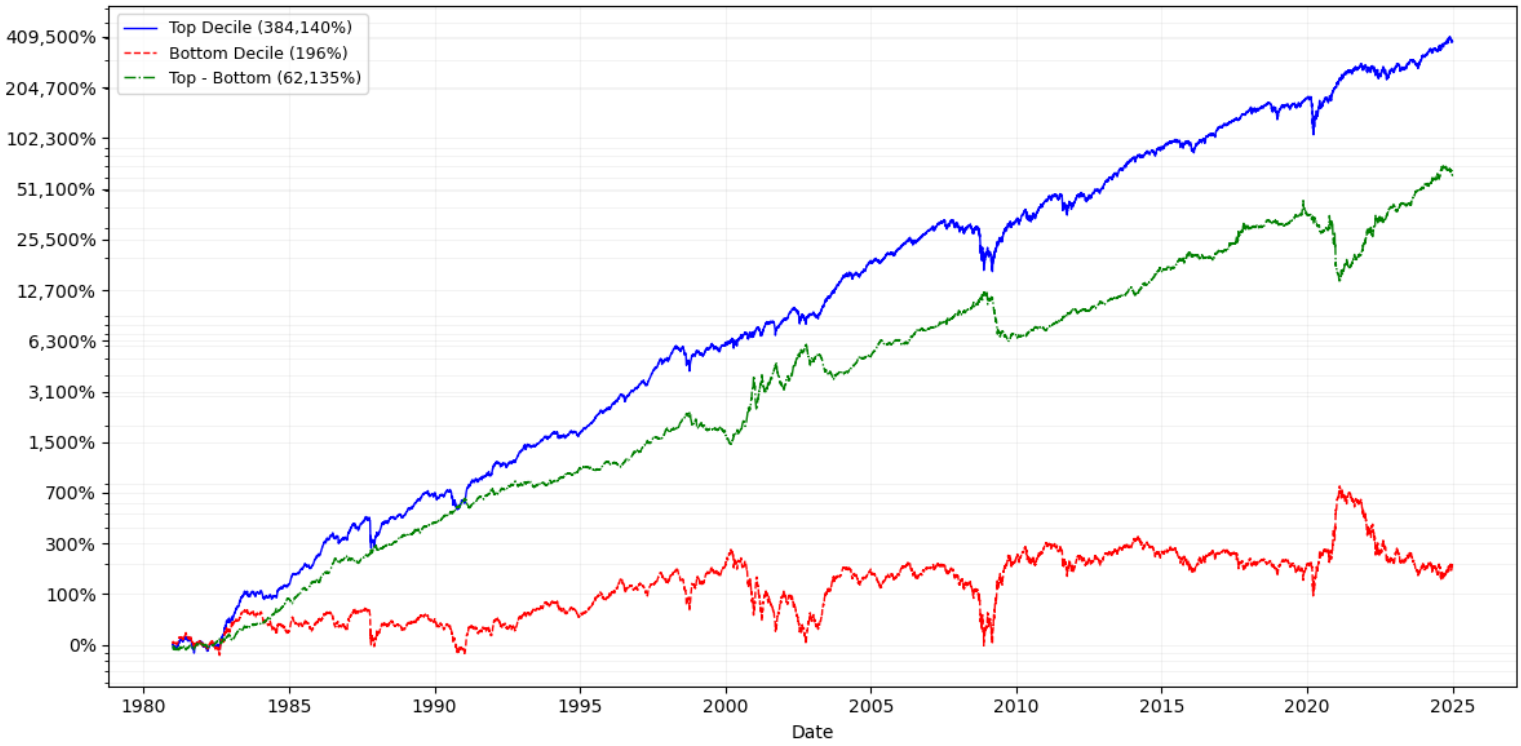
The Amalgam model had strong long-term outperformance, with a high cumulative return gap over all benchmarks in the full OOS period. It greatly outpaced the equal-weighted index (EWTD) during the 1980s and again after ~2014 as well, despite being trained relative to value-weighted (VWTD) benchmarks and

constrained to firms with market caps above $50 mil. (in real terms). Lower drawdowns and volatility are observed as well, relative to both VWTD indices, even though the top decile is equally weighted. The return sequence effectively mirrored indices during market-wide turmoil periods (Black Monday Crash in 1987, Global Financial Crisis in 2008, COVID in 2020).

Nevertheless, performance weakened in the most recent period, with the strategy returning only 4% CAGR more than the VWTD benchmarks, with slightly higher volatility and deeper drawdowns. The model's decay coincides with that of EWTD. Returns of the VWTD indices rose in the last 5 years in contrast to the entire OOS period, as the S&P500 effectively matches CRSP's VWTD universe of assets. This reflects increasing market concentration in large and mega cap stocks, likely due to the onset of ETFs and the rising popularity of passive, cap-weighted products as a whole (McNichols, 2025), (Kaczmarski, 2023). Even though the Amalgam still outperforms marginally, this structural shift can reduce the effectiveness of EWTD and/or mid-cap strategies, possibly reflecting decreased signal strength among the model's feature set.

5.1.1. Top Minus Bottom Deciles

Figure 4: Amalgam Top and Bottom Decile Portfolios



The return progression of the bottom decile acting as the short leg, or "avoided" portion of stocks, is shown alongside the cumulative spread. It is primarily used as a diagnostic, showcasing the model's ability to distinguish between winners and losers in a robust and consistent manner. Ideally, a perfectly calibrated strategy would yield spreads that surpass the top decile alone, driven by systematically negative returns in the bottom decile. This is not observed here, since in market rebound years (post-2003, ~2009, and ~2021), the bottom decile had sharp upswings that compressed the spread. The imperfection is common, and aligns with findings in ML asset pricing literature, showing that predictive power tends towards the upper quantiles/ranges

(Gu et al., 2018). After all, such models are optimized to identify outperformers, which doesn't necessarily translate to isolating **net** losers (Bryzgalova et al., 2019). For future research, it would be interesting to construct a long-short portfolio using a second model, deliberately trained to identify net negative underperformance. Nevertheless, consistent net losers were harder to capture, as poor stock performance in general, tends to be highly transient and perhaps more idiosyncratic in nature (Asness & Frazzini, 2012).

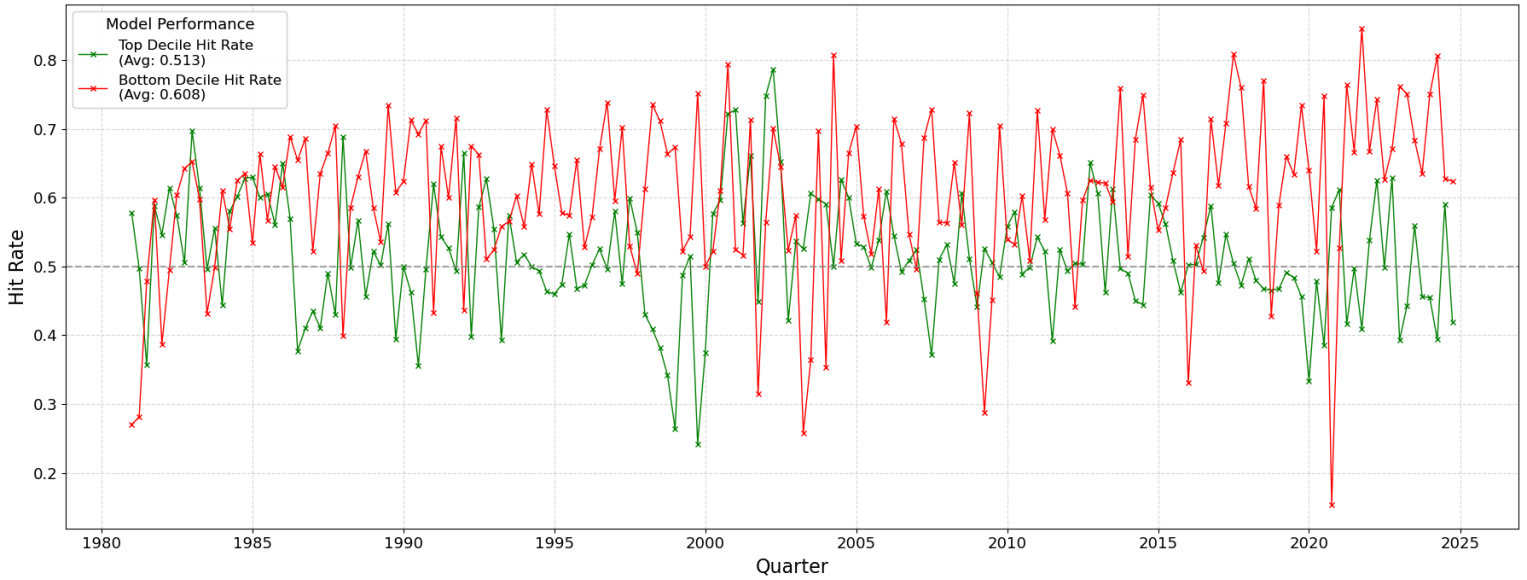
5.1.2. Prediction Accuracy/Hit Rates

Figure 5: Amalgam Top and Bottom Decile Hit Rates Per Quarter

This figure plots the quarterly hit rates of the top and bottom predicted deciles from Q1 1981 to Q4 2024. For the top decile, accuracy is defined as the proportion of firm projections correctly classified to outperform the VWTD benchmark in a given quarter (#ret>vwretd). For the bottom decile, accuracy is defined as the proportion of firm projections correctly classified to **underperform**. (#ret<vwretd)
The average decile hit rate for the entire period, shown in parentheses in the legend, is calculated by:

$$\mu_{accuracy} = \frac{\sum_t Hits_t}{\sum_t Decile\ Size_t}$$

Where $Hits_t$ represents the number of accurate classifications in a given quarter, and $Decile\ Size_t$ is the total number of stocks in the respective decile.



Decile	Positive %	Negative %	Median Positive Return	Median Negative Return
Top	59.7%	40.3%	12.0%	-9.4%
Bottom	47.0%	52.9%	12.7%	-17.4%

As part of the model's quantitative evaluation, it's important to report the full OOS average prediction accuracy for both deciles. On average, 51.3% of chosen stocks in the top outperformed the VWTD benchmark, and 60.8% of bottom-decile stocks underperformed it (meaning that in the remaining 39.2% of firm-quarter observations, the bottom decile contained outperformers). It's worth noting, that these results arise from the stricter dual-condition label used in training, where outperformance was defined as surpassing both the median returns within the industry group, and those of the VWTD benchmark.

While a 51.3% hit rate is punitive in absolute terms, returns compounding persisted even when accuracy was only marginally better than chance. To give insight to this dynamic, the realized return classification of both portfolios is shown. Across all top-decile firm-quarters, 59.7% had positive returns with a median gain of +12% and 40.3% were negative, but with a smaller loss of -9.4%. In contrast, the bottom decile had fewer firm-quarters with positive returns (47%), coupled with larger losses among the negatives (-17.4% per quarter). Even though raw accuracy leaves plenty of room for improvement, the directional return asymmetry of the deciles left the top with a net upside, and the bottom with a large net loss. This allowed the model to form portfolios that compounded positive returns consistently, explaining the high cumulative figure observed previously. It serves as an example that, while ML can be "disappointing" in absolute terms, financial success can rely more on avoiding large downsides over the long term, rather than maximizing precision (Buczynski et al., 2021).

5.1.3. Portfolio Composition

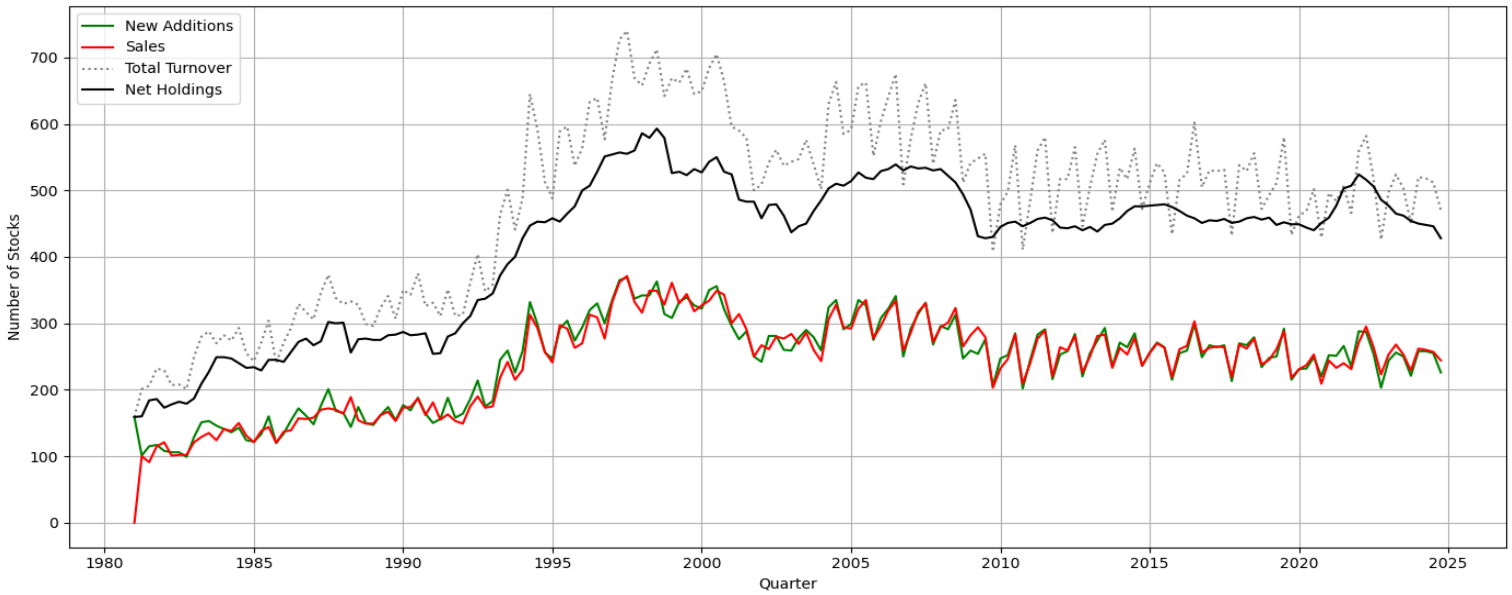
Figure 6: Amalgam Top Decile Quarterly Position Statistics

This figure displays the quarterly progression of the portfolio's holdings, along with stock purchases, sales and turnover. **"New Additions"** represent the number of stocks entering the portfolio, not held in the previous quarter.
"Sales" represent the number of stocks that were discarded from the portfolio from one quarter to the next.
"Total Turnover" represents the rebalancing activity in a given quarter, resulting from the sum of "New Additions" and "Sales".
"Net Holdings" reflect the number of unique stocks forming the portfolio in a given quarter, calculated as:

$$Net\ Holdings_t = Net\ Holdings_{t-1} + New\ Additions_t - Sales_t$$

"Turnover Ratio" represents the trading activity relative to the current portfolio size, calculated as:

$$Turnover\ Ratio_t = \frac{Total\ Turnover_t}{Net\ Holdings_t}$$



Quarter	New Additions	Sales	Total Turnover	Turnover Ratio	Net Holdings
Q1 1981	159	0	159	1.0	159
Q2 1981	101	100	201	1.26	160
Q3 1981	115	91	206	1.12	184

Machine learning strategies are reported to maintain high position counts, typically accompanied by high stock turnover across rebalancing periods as well (Azevedo et al., 2024), (Chin et al., 2022). Therefore, it's crucial to examine the evolution of the Amalgam portfolio's metrics in order to establish a baseline understanding for subsequent trading cost analysis. The progression begins in Q1 1981 with the starting portfolio comprising 159 unique stocks, which increase until the late 1990s. After peaking throughout 1998, the position size tapers down and stabilizes at 450 stocks on average, with a relatively high quarterly turnover of ~550. This trend aligns with the decreasing number of U.S. public listings after the early 2000s, as a consequence from the plateau in IPO activity (Kardashian, 2024).

Figure 7: Quarterly Transaction Costs

This figure displays the quarterly progression of the top decile portfolio's total trading costs, **reported in decimals**. The transaction costs reflect bilateral trading activity from the quarter's "New Additions" and "Sales", summed under "Total Turnover", and calculated as:

$$Total\ Cost_t = \sum_{i=1}^{N_t} w_i \cdot (\alpha + \beta_1 \cdot \frac{w_i}{MC_i})$$

where:

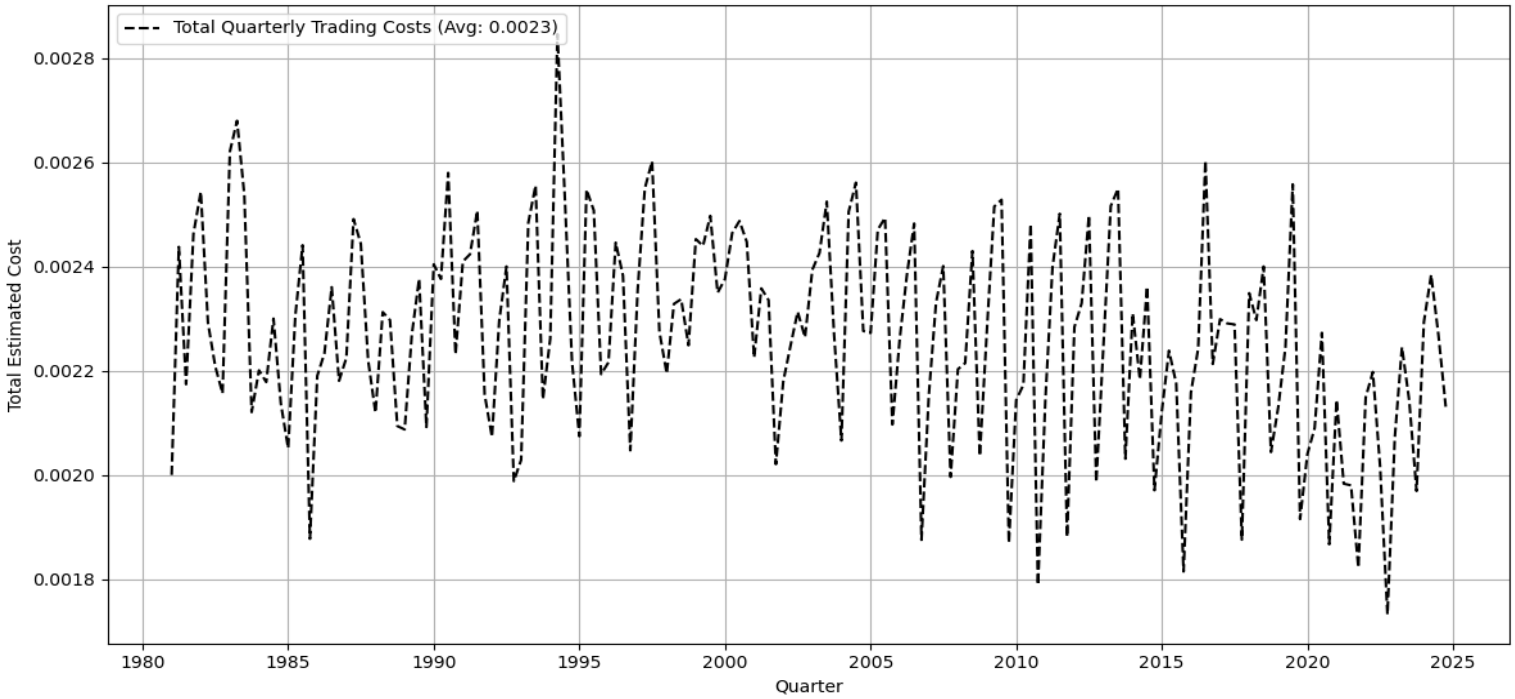
$w_i = \frac{1}{N_t}$ denotes the equal weight assigned to each of the traded stocks in a given quarter.

$\alpha = 0.002\ (20\ bps)$ denotes the baseline assumption for trading costs

$\beta_1 = 0.0025\ (25\ bps)$ is the secondary component, acting as a liquidity penalty that increases as firm size decreases

$\frac{w_i}{MC_i}$ is a given stock's weight in the portfolio, divided by its market cap (in millions), acting as the scalar for β_1

This formula applies a flexible cost penalty for illiquidity, reflecting real-world frictions that may arise when trading in small(er) firms. Total costs are computed across all "New Additions" and "Sales", reflecting the "round trip" during quarterly portfolio rebalancing.



Quarter	New Additions	Sales	Total Turnover	Mean Cost per Trade	Total Cost
Q1 1981	159	0	159	~0.000006	0.002000
Q2 1981	101	100	201	~0.000006	0.002438
Q3 1981	115	91	206	~0.000005	0.002174

While this study focuses on establishing a theoretical framework based on machine learning, the implications of trading costs are also included to assess the practical feasibility of the Amalgam model. Conventionally, periodic cost bases have been expressed as a linear function consisting of a flat cost component (α), ranging from 10 bps to 20 bps, and a flexible one which accounts for increased illiquidity relative to the firm's size (Avramov et al., 2019). A similar methodology is applied here, only excluding the NASDAQ exchange dummy variable due to insufficient data. The upper bound of 20 bps (0.002) was chosen to remain conservative, as the portfolio composition tended towards small and mid-cap stocks in earlier decades (≤$2 billion from 1981 to 2000, shown exclusively in the code appendix).

On average, the observed *quarterly* transactions were 0.23% of the portfolio's return, which amount to annualized costs of ~0.92%. These measures are very modest in contrast to the reported ~3 - 5% annual range for similar high-turnover ML strategies (Azevedo et al., 2024), (Kelly & Xiu, 2023), however, it's important to note that the relevant studies rebalanced portfolios on a *monthly* basis. If the Amalgam portfolio were to be rebalanced 12 times per year instead of 4, the average annualized costs would approximate 3%, which is in line with the literature. Furthermore, a mild downward trend in quarterly costs is observed over time, beginning after the year 2000. This occurrence is tied to the consistent increase in the portfolio's average market cap, rising from ~$5 billion in 2000 to ~$20 billion in 2024. This trend suggests a lasting shift away from small caps, and a preference towards mid and large-cap firms, resulting in slightly lower size-related illiquidity penalties.

Finally, these estimated costs adjust performance moderately, though the overall framework remains robust in outperforming the benchmarks. The total OOS period cumulative returns decline from 384,140% to 257,823% (-126,317%), top-decile average annualized alpha from 7.29% to 4.95% (-2.34%), and compounded annual returns from 20.65% to 19.6% (-1.05%). This is especially surprising since trading costs were not embedded into the model's classification logic during training/testing, meaning that outperformance projections were not optimized using this as an additional constraint.

For qualitative context showing the average industry composition of the top and bottom decile portfolios, refer to Figure 8.

5.1.4. Performance Breakdown and Robustness Checks

Table 8: Regression Results on Amalgam Model Decile Portfolios

This table reports regression results of equal-weighted returns for the top, bottom and "top minus bottom" long-short decile portfolios, regressed on the Fama-French five factors plus momentum (FF5+MOM). Results are based on the full out of sample period and for the most recent 5-year subsample. Regressions are estimated using daily returns, with annualized factor loadings. The t-stats are shown in parentheses and are based on Newey-West (HAC) standard errors with 1 lag. One-sided significance tests are used for **α,** with one, two, and three stars corresponding to statistical significance at the 10%, 5%, and 1% levels.

Full Out of Sample Period (Q1 1981 – Q4 2024)

	Amalgam Top Decile	Amalgam Bottom Decile	Top – Bottom
α	7.29%	-2.80%	10.10%
	(12.80***)	(-1.72**)	(6.16***)
RMRF	0.90	0.87	0.03
	(169.38)	(68.04)	(2.79)
SMB	0.52	0.64	-0.12
	(56.55)	(39.75)	(-8.46)
HML	0.15	-0.13	0.28
	(16.57)	(-6.85)	(15.00)
RMW	0.12	-0.47	0.59
	(13.34)	(-19.28)	(25.47)
CMA	-0.01	0.05	-0.06
	(-0.84)	(1.92)	(-2.37)
MOM	0.03	-0.25	0.29
	(5.58)	(-16.43)	(19.97)
R^2	0.95	0.76	0.27

Last 5 Years (Q1 2020 – Q4 2024)

	Amalgam Top Decile	Amalgam Bottom Decile	Top - Bottom
α	4.41%	-2.49%	7.26%
	(2.12**)	(-0.29)	(0.85)
RMRF	0.90	0.85	0.05
	(86.20)	(32.63)	(1.83)
SMB	0.48	0.72	-0.25
	(16.28)	(14.49)	(-4.73)
HML	0.20	-0.16	0.36
	(9.57)	(-3.61)	(8.20)
RMW	0.04	-0.74	0.78
	(2.28)	(-10.86)	(11.27)
CMA	-0.02	0.20	-0.22
	(-0.72)	(2.52)	(-2.78)
MOM	0.02	-0.15	0.16
	(1.17)	(-4.82)	(5.29)
R^2	0.96	0.66	0.27

These regression results offer good evidence of both economically and statistically significant alpha, demonstrating that excess returns can be captured without resorting to overly complex and opaque investment strategies. Even with a compact model comprising 29 features, this framework captured a long, **7.3%** annualized alpha, significant at the 1% level, in the full OOS period. Performance decayed moderately in the most recent 5-year subsample however, as the top decile dropped in magnitude and strength **(4.4%; 2.1 t-stat),** and the long-short spread became statistically insignificant. Still, the persistence of these anomalous returns highlights the lasting potential of dynamic, industry aware ML models, despite an increasingly competitive market landscape.

Moving onto common factor scores, it's worth noting that the Amalgam's alpha is not resulting from a consistently high market exposure in the top decile, as the long-short portfolio's **RMRF** is nearly neutral for both regression windows.

Size-wise, the model shows a moderate **SMB** score which is more pronounced in the bottom decile, suggesting that the long leg had more of a mid-cap structure on average. To reiterate from previous sections, the high **SMB** in the short decile may have been constrained and irreproducible in practice (short sell share availability, regulatory restrictions, etc), making the spread somewhat unreliable.

CMA and **HML** exposures are low to moderate and consistent in both instances, though this score is likely influenced by the added measure of R&D capitalization, which had a downstream effect on the accounting line items, and the model's features derived from them**.** Treating R&D expenses as an amortizable asset raised book-assets, equity values, profitability levels, and in turn compressed the FF Book-to-Market ratios and **decreased** traditional reinvestment levels (change in assets). As a result, small(er) firms with high(er) reinvestment under standard accounting measures were reclassified during the model's predictions and decile formation, potentially distorting CMA and HML exposures in the regressions.

Moreover, **RMW** scores offer an important insight. Even though the top decile is nearly neutral, the real effect lies in the strong negative **RMW** score of the short leg, implying the model consistently avoided "junkier" firms, even after increased profitability levels via R&D expense removal from SG&A. In recent years, RMW increased, and CMA decreased further, along with the observed decay in alpha. This may hint at a shift in market sentiment, where unprofitable and aggressive reinvestment firms are no longer penalized to the same degree as they were historically, making the model's performance worse (Asness et al., 2013), (Harvey et al., 2015).

Finally, **MOM** exposures are neutral in both regression timeframes. This is not necessarily an absence of holistic momentum, but perhaps a form of substitution. The model includes **"ivol_ff3_21d",** idiosyncratic volatility after adjusting for **RMRF, HML** and **SMB**, which carries strong modelling effect sizes in the **opposite** direction of the Jegadeesh & Titman **"mom_12_1"** feature. As a result, there might be a slight deviation away from traditional momentum, reflecting a preference towards volatility-induced idiosyncratic mispricing rather than pure returns persistence.

Table 9: T-test results: Difference between risk-adjusted returns

This table reports the one-sided t-test results for the difference between risk-adjusted daily average returns of the top decile portfolio and the broad-market indices:

$$H_0: \mu_{Amalgam} \leq \mu_{Benchmark}\ ; H_1: \mu_{Amalgam} > \mu_{Benchmark}$$

The risk-adjusted return is computed as $R_{adj,t} = \frac{R_t - r_{f,t}}{\sigma(R_t)}$.

Significance levels at 10%, 5%, and 1% are denoted by one, two, and three stars, respectively.

Full Out of Sample Period (Q1 1981 – Q4 2024)

Benchmark:	**EWTD**	**VWTD**	**S&P500**
t-statistic	1.59**	8.93***	8.60***

Last 5 Years (Q1 2020 – Q4 2024)

Benchmark:	**EWTD**	**VWTD**	**S&P500**
t-statistic	2.42**	0.84	0.69

In the full sample period, the Amalgam top decile significantly outperformed all benchmarks on an absolute and risk-adjusted basis with high certainty. As for the most recent historic window, risk-adjusted outperformance relative to the EWTD benchmark improved modestly, while becoming statistically indistinguishable from both VWTD indices.

This contrast continues to hint at a broader trend in the investment landscape. Namely, that cap-weighted strategies, despite historically lower risk-adjusted and absolute returns, have begun outperforming EWTD benchmarks. The previously observed narrowing in Sharpe ratios, combined with the current parity in risk-adjusted returns suggests that even systematic EWTD approaches, such as the Amalgam model, struggle to compete with the sheer momentum of cap-weighted indices. While declining efficacy of the model's features explain this, it's also plausible that the observed performance degradation is influenced by the recent cap-weighted phenomenon.

5.1.5. SHAP Values/Feature Importance

Figure 9: SHAP Summary Plot

This plot displays the top 12 most influential features in the model's prediction of next-quarter stock outperformance, ranked by their mean absolute SHAP value. Each color strip is formed from individual dots, corresponding to firm-quarter observations across the full sample (Q1 1981 – Q4 2024).

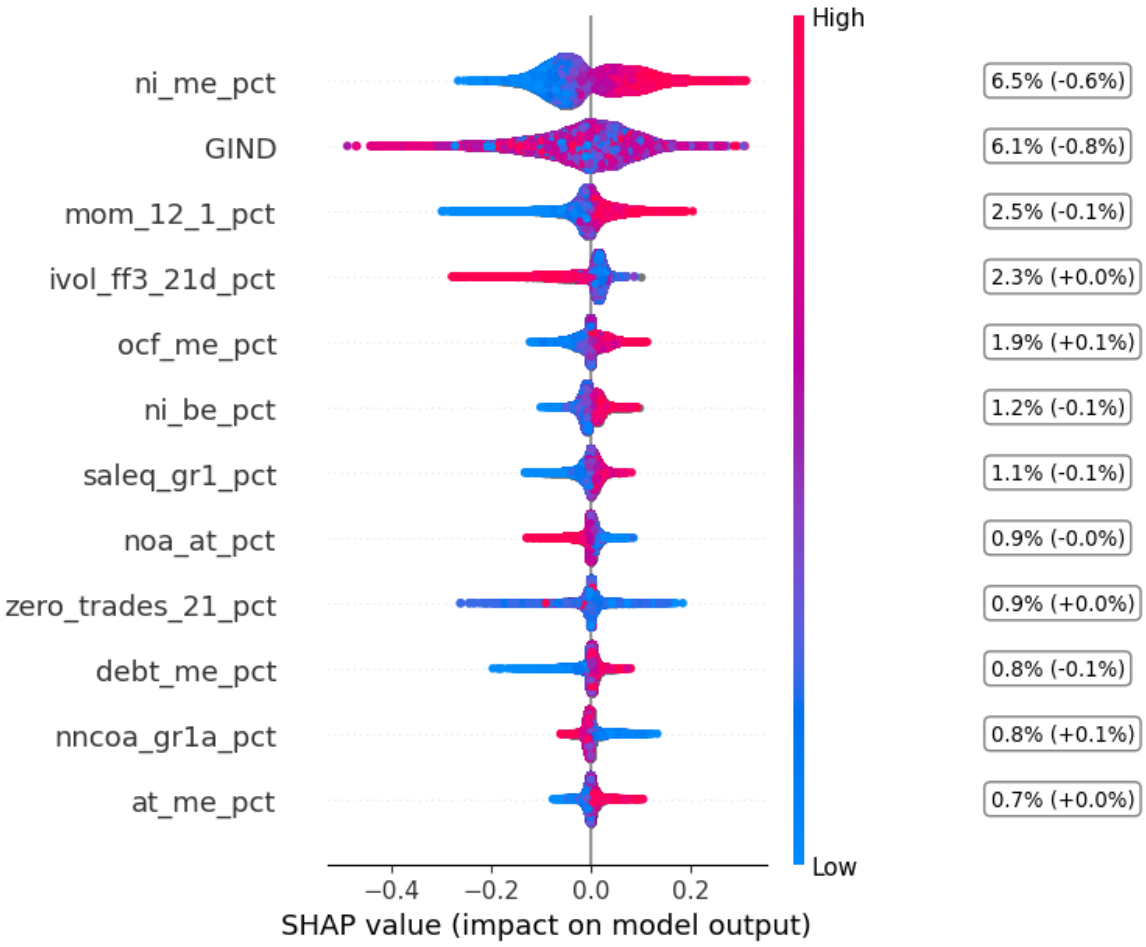
The hue in the strips reflects the raw feature value for a given observation. Red indicates high feature values, while blue indicates low ones. Dot saturation shows sample density, with more saturated regions reflecting a concentration of firm-quarters with similar SHAP outcomes.

SHAP values on the x-axis (-0.4 to +0.2) quantify each feature's marginal contribution to the model's predicted probability of outperformance. Both high and low (red and blue colored) feature values (dots) may appear on either side of x-axis, which occurs when a feature's effect is non-linear, or contextually dependent on other features.

The boxes on the right report the feature's average contribution to overall predictive power in two ways:

The **absolute** SHAP mean is the feature's total influence on the model, regardless of sign. For instance, when the model makes a prediction, ~**6.5%** of that decision is, on average, attributable to **"ni_me_pct"**, regardless of whether its effect is positive or negative.

The **signed** SHAP mean indicates the directionality of the average effect in each firm-quarter. For **"ni_me_pct"**, the signed mean of **-0.6%** suggests that, on average, higher net income-to-market ratios lower the log-odds of predicted outperformance by 0.006 below the baseline of 0. In probability terms, this translates to a decrease from 50% to approximately 49.85%, all other variables held constant[1].



[1]Conversion from log-odds to probabilities: $P = \frac{1}{1+e^{-(-.006)}} = \frac{1}{1+e^{0.006}} = \sim 49.85\%$

Most features have SHAP distributions centered around zero, and often with long, thin tails. This indicates that the majority of firm-quarter observations had mild shifts away from the 50% baseline probability of outperformance. On the other hand, only a minority of values in the tails caused large negative/positive shifts, likely being a result of extreme periods such as crises, unusual earnings cycles, regime changes, etc. In such atypical market conditions, features scores deviate largely and thus, may disproportionately drive model decisions. This is consistent with findings from related literature, stating that nonlinear ML models dynamically re-weight predictors based on time-varying conditional impacts, especially during periods of economic uncertainty (Gu, Kelly, and Xiu, 2020).

After "ni_me_pct", the categorical GIND feature which captures 85 distinct industries is the second most influential. Unlike the rest of the features, GIND carries no inherent ordinal or metric structure, and its predictive power comes from group-based interactions learned by the model, hence the visual heterogeneity in the coloured plot. The same industry (dot color) can appear on opposing sides of the axis, reflecting a contextual dependence. This occurs because a given industry's effect on outperformance varies and is based on the interaction with other features over all firm-quarter observations.

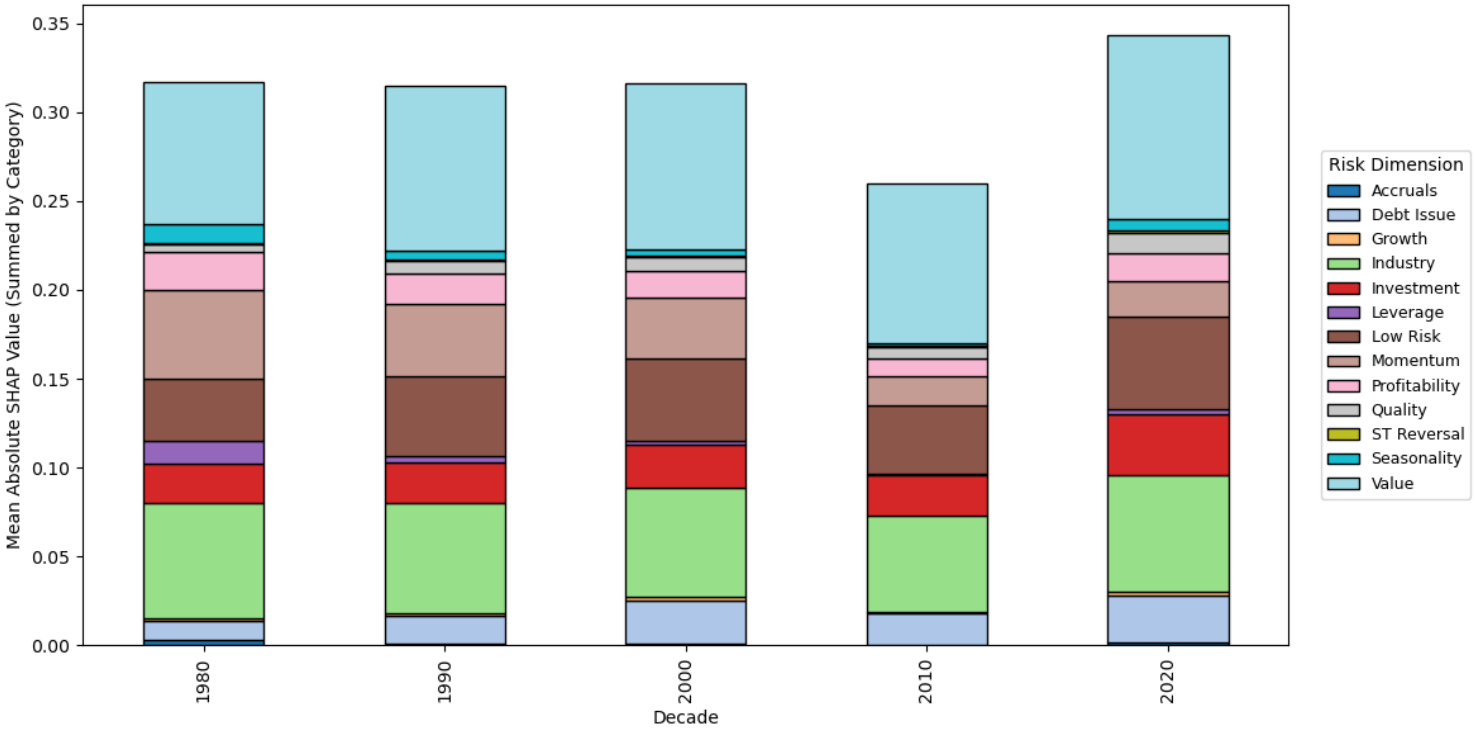
The signed GIND mean of –0.8% indicates that, on average, being in a given industry tends to *reduce* the predicted outperformance probability. This net negative effect shouldn't be misinterpreted as a universal penalty since the distribution is long tailed and highly left skewed. While most industries contributed neutrally or slightly positively (as seen in the higher density of positive SHAP dots on the right side), only a small handful of industry–time combinations exerted disproportionately negative effects, driving the average downward.

It suggests that in certain market environments, a few specific industries acted as strong detractors for outperformance. In practical terms, and consistent with recent history, hospitality and leisure experienced substantially larger declines relative to other industries during the COVID-19 crash (Mazur et al., 2021), and banks are often the first and most affected by rate shocks and liquidity strains (Acharya et al., 2024). In this sense, the GIND feature can reflect broader cycles and structural characteristics, allowing the model to continuously adjust expectations in a conditional and nonlinear way.

5.1.6. Factor Relevance Over Time

Figure 10: Cumulative SHAP Values by factor

This chart displays the sum of mean absolute SHAP values per factor for each decade, with the 2020s having only 4 full years of data. Each decade bar represents the cumulative average predictive contribution of all risk dimensions (e.g., Value, Momentum, Quality). The colored segments (factors) within each decade bar reflect the summed contribution of all features corresponding to that factor group. The height of each bar captures the model's total reliance on predictive signals that decade, and the distribution of colors across the bars show the changing importance of each factor.



The following analysis uses a per-decade decomposition of mean absolute SHAP values to trace how the model's reliance on key risk dimensions evolved. SHAP values do not establish economic causality, rather, their interpretation reflects the structure of the model and the design of the training and validation framework. Therefore, a higher SHAP value for a factor indicates that the model increasingly utilized that signal to classify outperformers during a given period, not that the corresponding factor premium was necessarily stronger or more persistent in realized returns. Nevertheless, when a factor exhibits *consistently* increasing or decreasing relative importance across decades, it provides robust evidence of enduring or fading predictive quality, which may *cautiously* suggest broader economic (ir)relevance in the cross-section.

The **Value** factor, comprised by the "debt_me", "ocf_me" and "ni_me" features, consistently carried the highest aggregate importance, increasing from ~8% in the 1980s to 10.4% in the 2020s. Notably, "ocf_me" rose from 0.4% to 2.7% across the period, suggesting that cash-flow-based valuation signals became increasingly relevant when estimating outperformance likelihoods.

In second place, **Industry** classification (GIND categorical variable) showed consistently high importance, with a mean absolute value of above 6% in every decade except the 2010s, when the predictive strength of the entire model experienced a decrease. Its persistent and high relative contribution validates the idea that firm

taxonomy meaningfully contributes to forecasting, however, it's important to note that the exact effect sizes are model-dependent.

The **Investment** and **Low Risk** factors followed closely, as they exhibited a lower but stable importance of above 2%, rising to 3.5% and 5.3% respectively in the 2020s. This suggests that asset growth and risk suppression especially gained more traction in classifying outperformance during the last years.

Moreover, **Debt Issue** captured by the "ni_arl", "noa_at" and "at_me" features, had the second highest consistent increase, growing from 1% in the 1980s to 2.6% in the 2020s. In contrast, **Leverage** ("nfna_grla"), showed a monotonic decline after the 1980s, implying that the contribution of this standalone feature faded over time. For an overview of all feature and factor SHAP values, refer to Table 10.

Finally, total SHAP importance peaked in the 2020s, even though returns and overall performance declined relative to the full OOS period. In essence, the model realized worse results despite being "more confident" in its signal use. This highlights that higher internal modelling "conviction", even after rigorous regularization and training, doesn't always translate into better realized performance. When return dynamics and feature values shift abruptly, as was the case with the post-2020 shock, certain signals can appear as more predictive in-sample, leading the model to mistakenly emphasize transient patterns (noise) rather than structural ones.

6. Conclusions, Limitations & Implications for future research

This thesis examined whether a custom-built classification framework could achieve persistent and interpretable stock outperformance. The proposed model was built using a curated subsample of 28 features, capturing 13 broad factor categories, with additional industry-based categorical signals which contextualized all feature scores by their respective industry. Data preprocessing included R&D capitalization for accounting metrics and percentile-based normalization of features and the label. Training utilized an expanding rolling-window design, with oversampling of the most recent 4 years to emphasize current market dynamics. This led to a statistically significant annualized alpha of 7.3% over the 44-year OOS period, confirming the model's ability to capture non-linear patterns and convert them into tractable, robust long-term gains.

The study's results align, contextualize, and extend several key insights from the literature:

1) Consistent with prior findings, tree-based models such as XGBoost can outperform traditional linear factor models and market benchmarks, offering superior OOS returns while being computationally efficient (Gu et al., 2020), (Hanauer et al., 2025).
2) Including R&D Amortization to adjust accounting-based features was motivated by prior literature, extending its application to a modelling setting. While the (potentially incremental) impact wasn't explicitly benchmarked, it ensures a theoretically informed way to evaluate high reinvestment firms.
3) Data normalization techniques shown to enhance the signal-to-noise ratio in ML (Chen et al., 2024), were implemented using percentile ranks that reduced the overall magnitude of feature and target values. Although the isolated effect wasn't explored here, the method is theoretically grounded and consistent with the strong observed outcomes.
4) Firm-group classification is a highly informative signal (Hanauer et al., 2025), (Bagnara & Goodarzi, 2023). Including industry-level categorical variables (as opposed to sector groupings) adds a degree of novelty to

the literature, and further reinforces these findings demonstrated by the relatively high and persistent mean absolute GIND SHAP values over time.

5) Oversampling more recent data is often used to emphasize contemporary market dynamics (Bengio et al., 2012), (Spiliotis, 2023). However, in this study it appeared to introduce overfitting risk in the last decade, as indicated by the elevated global SHAP values, without a simultaneous and notable increase in realized returns.

6) Model performance was concentrated in the top decile portfolio, a pattern that supports prior studies showing that ML techniques tend to capture signals more effectively for outperformers, rather than bottom range, underperforming firms (Bryzgalova et al., 2019).

7) Hit rate analysis revealed that low absolute classification accuracy could still lead to substantial excess returns, affirming that absolute accuracies substantially higher than chance are not a hard requirement for meaningful investment performance (Buczynski et al., 2021).

8) The evolution of feature relevance affirmed prior research documenting the fleeting nature of individual signals (Ilmanen et al., 2019). While relative strength varied across decades, the Value Factor as a whole, among others, remained consistently strong. This also aligns with the view that Value Investing is not structurally obsolete (Arnott et al., 2019), (Israel et al., 2020), (AQR, 2020), contrasting claims and observations of its diminished relevance (Lev & Srivastava, 2019), (Bezuidenhout & Vuuren, 2021).

9) Using a suite of diagnostics helped interpret the model's internal logic on multiple levels. These findings support recent literature showcasing their benefits in "demystifying" ML outputs (Tursunalieva et al., 2024).

Still, the documented findings are not without caveats. Overall performance declined meaningfully in the most recent 5-year window. While the model outperformed VWTD indices in absolute returns, it reached parity on a risk-adjusted basis. This likely reflects moderate decay in feature efficacy, alongside a market-wide shift towards large-cap equities and cap-weighted passive investment products. Moreover, while SHAP values offered interpretability, it's worth reiterating that they remain model-relative diagnostics and do not necessarily reflect causality or generalizable patterns across the cross-section. Lastly, the model's average accuracy was only 51.3%, which aligns with generally observed values. This leaves plenty of room for improvement, however, as observed ranges in the literature vary between 50.2% and 56.4%, depending on the timeframe and application specifics (Buczynski et al., 2021).

The study also contains several limitations that may impact generalizability and potentially reduce overall accuracy:

The geographic and asset scope focused exclusively on U.S. listed equities. As a result, the overall performance and findings and may not generalize to international markets, where return drivers and microstructure can differ greatly.

The model was built with only 29 features, representing one of the most informative subsamples in recent literature. Nevertheless, it still omits potentially valuable predictors. For example, sentiment signals derived from news outlets, earnings calls, social media, search engine results, etc. have gained significant traction in recent years and can be largely orthogonal to fundamental and market-based variables (Cookson et al., 2025). Additionally, analyst revisions and institutional coverage have also demonstrated strong explanatory power

(Jung et al., 2017). These were excluded for consistency with Robeco's baseline, but their inclusion could very likely improve performance.

Model run-time and memory limitations greatly hindered the scalability of the study. The current framework requires approximately five hours to process from start to finish, per iteration. These constraints prevented the use of ensemble methods/model stacking, which often lead to improved predictive performance by aggregating and balancing the outputs from multiple model architectures (Cheng & Zhao, 2022), (Nti et al., 2020).

Data limitations within the WRDS suite impacted the timeliness of accounting information. As mentioned previously, ~89% of firm-quarter observations lacked date coverage for earnings announcements. To avoid lookahead bias, accounting features were lagged by two quarters. This conservative adjustment likely weakened overall results, as model predictions were essentially based on information from 6 months in the past, rather than only 3.

Future research could explore whether augmenting the feature set with the proposed additions, among other relevant ones, mitigates the performance decay observed in the most recent 5-year period. The current framework can benefit from stacking, where the XGB model is paired with other algorithms (e.g. SVMs, RFs, NNs or related GBMs) to form a meta-model that balances predictions and improves overall accuracy. In addition, the entire study can be extended explicitly for practical use, by implementing constraints for trading costs that alter the model's internal logic. Lastly, applying the framework to international markets would offer insights into generalizability and robustness across different macroeconomic regimes, accounting standards, and institutional environments.

7. References

Acharya , Engle , Jager, & Steffen. (2024, September). *How credit line drawdowns and repayments affected bank performance during COVID-19*. Suerf. https://www.suerf.org/publications/suerf-policy-notes-and-briefs/how-credit-line-drawdowns-and-repayments-affected-bank-performance-during-covid-19/

Almaspoor, Safei, & Minaei. (2021, June). Support Vector Machines in Big Data Classification: A systematic literature review. https://www.researchgate.net/publication/353786557_Support_Vector_Machines_in_Big_Data_Classification_A_Systematic_Literature_Review

Arnott, R. D., Harvey, C. R., Kalesnik, V., & Linnainmaa, J. T. (2019, December 2). *Reports of value's death may be greatly exaggerated*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3488748

Asness, C. S., & Frazzini, A. (2012, May 9). *The devil in HML's details*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2054749

Asness, C. S., Frazzini, A., & Pedersen, L. H. (2013, August 19). *Quality minus junk*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2312432

Avramov, D., Cheng, S., & Metzker, L. (2019, September 18). *Machine learning vs. economic restrictions: Evidence from stock return predictability*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3450322

Azevedo, V., Hoegner, C., & Velikov, M. (2024, February 7). *The expected returns on machine-learning strategies*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4702406

Azevedo, V., Kaiser, S., & Müller, S. (2022, April 13). *Stock market anomalies and machine learning across the Globe*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4071852

Barber, Bennett, & Gvozdeva. (2015). *How to choose a strategic multi-factor equity portfolio*. Rusell Investments. https://russellinvestments.com/-/media/files/nz/insights/how-to-choose-a-strategic-multifactor-equity-portfolio.pdf

Bengio, Y., Courville, A., & Vincent, P. (2014, April 23). Representation learning: A review and new

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Bezuidenhout, J.-M., & Vuuren, G. van. (2021, January). *Spectral Analysis and the death of Value Investing*. Research Gate. https://www.researchgate.net/publication/356433581_Spectral_analysis_and_the_death_of_value_investing

Bryzgalova, S., Pelger, M., & Zhu, J. (2019, December 19). *Forest through the trees: Building cross-sections of stock returns*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3493458

Buczynski, W., Cuzzolin, F., & Sahakian, B. (2021, April). A review of machine learning experiments in equity investment decision-making: Why most published research findings do not live up to their promise in real life. International journal of data science and analytics. https://pmc.ncbi.nlm.nih.gov/articles/PMC8019690/

Chen, M., Hanauer, M. X., & Kalsbach, T. (2024, December 2). *Design choices, machine learning, and the cross-section of Stock returns*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5031755

Cheng, T., & Zhao, A. B. (2022, April 9). *Stock return prediction: Stacking a variety of models*. Journal of Empirical Finance. https://www.sciencedirect.com/science/article/abs/pii/S0927539822000342

Chin, J. T., Lin, H., & Mei, Y. (2022, December 5). *Machine learning and the cross-section of Stock returns*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4282614

Cookson, J. A., Lu, R., Mullins, W., & Niessner, M. (2025, April 10). *Market signals from Social Media*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5187350

Damodaran, A. (2025). *R&D capitalizer*. R&D Capitalizer. https://pages.stern.nyu.edu/~adamodar/pc/R&DConv.xls

Dujava, C. (2024, December 17). *Design choices in ML and the cross-section of stock returns*. QuantPedia. https://quantpedia.com/design-choices-in-ml-and-the-cross-section-of-stock-returns/#:~:text=in%20model%20training%20decreases%20monthly,return%20than%20a%20rolling%20window

Fama, E. F., & French, K. R. (1992). The Cross-Section of Expected Stock Returns. *The Journal of Finance*, *47*(2), 427–465. https://doi.org/10.1111/j.1540-6261.1992.tb04398.x

Fieberg, C., Metko, D., Poddig, T., & Loy, T. (2022, September 28). *Machine learning techniques for cross-sectional equity returns' prediction - or spectrum*. SpringerLink. https://link.springer.com/article/10.1007/s00291-022-00693-w#:~:text=Recently%2C%20Gu%20et%C2%A0al,23%3B%20Lewellen%202015%3B%20Gu

Frankel, R. M., & Lee, C. M. C. (1998, August 25). *Accounting valuation, market expectation, and the book-to-market effect*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6528

Gu, S., Kelly, B. T., & Xiu, D. (2018, April 9). *Empirical asset pricing via machine learning*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3159577

Hanauer, M. X., Soebhag, A., Stam, M., & Hoogteijling, T. (2025, May 6). *Do machine learning models need to be sector experts?*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5224253

Harvey, Liu, & Zhu. (2015). *... and the cross-section of expected returns*. Oxford Academic. https://academic.oup.com/rfs/article-abstract/29/1/5/1843824

Is (systematic) value investing dead?. AQR Capital Management. (2020, May). https://www.aqr.com/Insights/Perspectives/Is-Systematic-Value-Investing-Dead

Jensen, T. I., Kelly, B., & Pedersen, L. H. (2023). Is There a Replication Crisis in Finance? *The Journal of Finance*, *78*(5), 2465–2518. https://doi.org/10.1111/jofi.13249

Jung, M. J., Keeley, J., & Ronen, J. (2017, June 26). *The predictability of analyst forecast revisions*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2991938

Kaczmarski, K. (2023, April 26). *The rise of ETFs and its powerful impact on markets*. Oliver Wyman - Impact-Driven Strategy Advisors. https://www.oliverwyman.com/our-expertise/insights/2023/may/exchange-traded-funds-are-fueling-market-opportunities.html

Kardashian, K. (2024, September). *Where did all the public companies go?*. Tuck School of Business. https://tuck.dartmouth.edu/news/articles/where-did-all-the-public-companies-go

Kavzoglu, T., & Teke, A. (2022). Advanced hyperparameter optimization for improved spatial prediction of shallow landslides using extreme gradient boosting (XGBoost). *Bulletin of Engineering Geology and the Environment*, *81*(5), 201. https://doi.org/10.1007/s10064-022-02708-w

Kelly, B. T., & Xiu, D. (2023, July 13). *Financial machine learning*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4501707

Lekan, Cena, Harry, & Andrewson. (2025b, March). Comparison of neural networks with traditional machine learning models. https://www.researchgate.net/publication/389546882_Comparison_of_Neural_Networks_with_Traditional_Machine_Learning_Models_eg_XGBoost_Random_Forest

Lev, B., & Sougiannis, T. (1999, February 23). *The capitalization, amortization, and value-relevance of R&D*. Journal of Accounting and Economics. https://www.sciencedirect.com/science/article/abs/pii/0165410195004106

Lev, B., & Srivastava, A. (2019, August 28). *Explaining the recent failure of value investing*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3442539

Lo, A. W. (2004, October 15). *The adaptive markets hypothesis: Market efficiency from an evolutionary perspective*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=602222

Mazur, M., Dang, M., & Vega, M. (2021a, January). *Covid-19 and the March 2020 Stock Market Crash. evidence from S&P1500*. Finance research letters. https://pmc.ncbi.nlm.nih.gov/articles/PMC7343658/

Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. *2020 11th International Conference on Information and Communication Systems (ICICS)*, 243–248. https://doi.org/10.1109/ICICS49469.2020.239556

MSCI. (2024, August). *Global Industry Classification Standard (GICS ...* GLOBAL INDUSTRY CLASSIFICATION STANDARD (GICS®) METHODOLOGY . https://www.msci.com/indexes/documents/methodology/1_MSCI_Global_Industry_Classification_Standard_GICS_Methodology_20240801.pdf

Nalepa, J., & Kawulok, M. (2018, January 3). *Selecting training sets for support vector machines: A Review - Artificial Intelligence Review*. SpringerLink. https://link.springer.com/article/10.1007/s10462-017-9611-1?utm_source

Nti, I. K., Adekoya, A. F., & Weyori, B. A. (2020, March 11). *A comprehensive evaluation of ensemble learning for stock-market prediction - journal of big data*. SpringerOpen. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00299-5?utm_source=chatgpt.com

Nti, I. K., Adekoya, A. F., & Weyori, B. A. (2020a, March 11). *A comprehensive evaluation of ensemble learning for stock-market prediction - journal of big data*. SpringerOpen. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00299-5

Ortiz-Molina, H., & Phillips, G. M. (2009, June 4). *Real asset illiquidity and the cost of capital*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1413780

Penman, S. H., & Zhang, X.-J. (2000b, January 23). *Accounting conservatism, the quality of earnings, and stock returns*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=201048

perspectives. arXiv.org. https://arxiv.org/abs/1206.5538

Pimco. (n.d.). *Understanding multi-factor strategies: PIMCO*. Pacific Investment Management Company LLC. https://www.pimco.com/us/en/resources/education/understanding-multi-factor-strategies

Putatunda, S., & Rama, K. (2018). A Comparative Analysis of Hyperopt as Against Other Approaches for Hyper-Parameter Optimization of XGBoost. *Proceedings of the 2018 International Conference on Signal Processing and Machine Learning*, 6–10. https://doi.org/10.1145/3297067.3297080

Rasekhschaffe, K., & Jones, R. (2019, March 4). *Machine Learning for Stock Selection*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3330946

Schwartz, M., & Hanauer, M. X. (2024, December 13). *Formula investing*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5043197

Spiliotis, E. (2023, September). *Time Series forecasting with statistical, machine learning, and deep learning methods: Past, present, and future*. SpringerLink. https://link.springer.com/chapter/10.1007/978-3-031-35879-1_3

Swade, A., Hanauer, M. X., Lohre, H., & Blitz, D. (2023a, November 15). *Factor zoo (.zip)*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4605976

8. Appendices

Large language LLM models were utilized during the process of writing this paper. Example prompts include:

"What is the relevant literature on this topic, can you provide me with sources?"

"I don't understand this topic, can you explain it to me clearly?"

"There is a mistake in my code, can you find it and provide details on how to fix it?"

Table 1: Accounting Effects resulting from R&D Capitalization on Key Financial Statements

This table outlines the modifications to key financial statement line items resulting from the capitalization of Research and Development (R&D) Expenditures, as opposed to expensing them directly. Under the default treatment, R&D is reported as a component of Selling, General, and Administrative Expenses (SG&A), fully reducing earnings in the period they are incurred. The modified treatment removes R&D from SG&A and treats it as a capital investment, amortized linearly over its estimated useful life. The adjustments affect profitability metrics (EBITDA, EBIT, Net Income), Balance Sheet items, and the Cash Flow Statement through restated Operating Cash Flow (OCF). All changes are directional and reflect the net effect relative to the default expensing convention.

Line Item	Default	Modified	Δ Default to Modified
Income Statement			
R&D*	Expensed; Part of SG&A	Amortized over useful life	Decreases
D&A*	Standard D&A, unrelated to R&D	Additional D&A for the period	Increases
Net R&D	Standard	(Capitalized) R&D*- D&A*	Increases
SG&A	Includes R&D as an expense	R&D* expense subtracted	Decreases
OPEX	Includes R&D as an expense	R&D expense subtracted	Decreases
EBITDA	R&D included as OPEX	R&D added back	Increases
EBIT	R&D included as OPEX	Additional D&A added back	Increases
Net Income	R&D included as OPEX	Additional D&A added back	Increases
Balance Sheet			
Total Assets	R&D Asset Excluded	Net R&D Asset Included	Increases
Book Equity	Lower Retained Earnings	Net R&D Asset Included	Increases
Cash Flow Statement			
OCF	R&D Expense Deducted	Offset by Net R&D	Increases

All modified line items and the formulas showing how they have been affected by the capitalization of R&D expenditures are available in **Table 4: Variable Definitions and Implementation Methodology**

Table 2: GICS-Specific R&D Amortization Periods

This table reports the amortization periods used to capitalize R&D expenditures by industry in this analysis. Industry classifications follow the Global Industry Classification Standard (GICS) at the six-digit GIND level. Industry names and amortization estimates (in years) are based on data published by Aswath Damodaran (Damodaran, 2025), which were cross-referenced with the official GICS definitions (MSCI, 2024) to ensure alignment. The final column displays the amortization period in quarters (years $\times$ 4), used to adjust R&D-related accounting metrics throughout the analysis. This mapping allows for industry-specific treatment of operational investments.

		Amortization Period	
GIND code	**Industry Name**	**Years**	**Quarters**
101010	Energy Equipment & Services	5	20
101020	Oil, Gas & Consumable Fuels	5	20
151010	Chemicals	10	40
151020	Construction Materials	10	40
151030	Containers & Packaging	5	20
151040	Metals & Mining	5	20
151050	Paper & Forest Products	10	40
201010	Aerospace & Defense	10	40
201020	Building Products	5	20
201030	Construction & Engineering	10	40
201040	Electrical Equipment	10	40
201050	Industrial Conglomerates	5	20
201060	Machinery	10	40
201070	Trading Companies & Distributors	5	20
202010	Commercial Services & Supplies	3	12
202020	Professional Services	3	12
203010	Air Freight & Logistics	10	40
203020	Passenger Airlines (New name)	10	40
203030	Marine Transportation (New Name)	10	40
203040	Ground Transportation (New Name)	5	20
203050	Transportation Infrastructure	5	20
251010	Automobile Components (New Name)	5	20
251020	Automobiles	10	40
252010	Household Durables	5	20
252020	Leisure Products	5	20
252030	Textiles, Apparel & Luxury Goods	3	12
253010	Hotels, Restaurants & Leisure	3	12
253020	Diversified Consumer Services	3	12
254010	Home Improvement Retail	2	8
255010	Distributors	2	8
255020	Internet & Direct Marketing Retail (Discontinued)	3	12
255030	Broadline Retail (New Name)	2	8
255040	Specialty Retail	2	8
301010	Consumer Staples Distribution & Retail	2	8
302010	Beverages	3	12
302020	Food Products	3	12
302030	Tobacco	5	20
303010	Household Products	3	12
303020	Personal Care Products (New Name)	3	12
351010	Health Care Equipment & Supplies	5	20
351020	Health Care Providers & Services	3	12
351030	Health Care Technology	3	12
352010	Biotechnology	10	40

352020	Pharmaceuticals	10	40
352030	Life Sciences Tools & Services	5	20
401010	Banks	2	8
401020	Thrifts & Mortgage Finance (Discontinued)	2	8
402010	Financial Services (New Name)	2	8
402020	Consumer Finance	2	8
402030	Capital Markets	2	8
402040	Mortgage Real Estate Investment Trusts (REITs)	3	12
403010	Insurance	3	12
404010	Life Insurance	3	12
404020	Property and Casualty Insurance	3	12
404030	Multi-Line Insurance	3	12
451010	Technology Services	3	12
451020	IT Services	3	12
451030	Software	3	12
452010	Communications Equipment	10	40
452020	Technology Hardware, Storage & Peripherals	5	20
452030	Electronic Equipment, Instruments & Components	5	20
452040	Technology Distributors	5	20
452050	Electronic Components	5	20
453010	Semiconductors & Semiconductor Equipment	5	20
501010	Diversified Telecommunication Services	5	20
501020	Wireless Telecommunication Services	5	20
502010	Media	3	12
502020	Entertainment	3	12
502030	Interactive Media & Services	2	8
551010	Electric Utilities	10	40
551020	Gas Utilities	10	40
551030	Multi-Utilities	10	40
551040	Water Utilities	10	40
551050	Independent Power and Renewable Electricity Producers	5	20
601010	Diversified REITs (New Name)	3	12
601020	Real Estate Management & Development	3	12
601025	Industrial REITs (New)	3	12
601030	Hotel & Resort REITs (New)	3	12
601040	Office REITs (New)	3	12
601050	Health Care REITs (New)	3	12
601060	Residential REITs (New)	3	12
601070	Retail REITs (New)	3	12
601080	Specialized REITs (New)	3	12
602010	Real Estate Management & Development (New Code)	5	20

Table 3: *Reported* Most Robust Risk Factor Proxies (1971-2021)

This table presents the 30 most robust firm-level feature proxies from Robeco's Factor Zoo (.zip) paper, selected based on their ability to minimize alpha as measured by the Gibbons-Ross-Shanken (GRS) test statistic, spanning the period from 1971 to 2021. The top 15 variables (rows 1–15) correspond to predictors yielding t-statistics above/equal to 3 for GRS alpha reduction. The remaining 15 (rows 16–30) meet a t-stat threshold of at least 2. Each feature is assigned to a high-level factor dimension (e.g., Value, Investment, Quality, Momentum, or Low Risk) and is defined using quarterly or rolling historical data. Variable formulation is based on standard financial theory and constructed using Compustat and CRSP data, as Robeco provides only the variable names and definitions, without any implementation details whatsoever.

Feature	Description	Dimension	Formula
cop_at	cash-based operating profits to book assets	Quality	$\frac{OIBDP_t - \Delta WCAP_{t-t_{-1}}}{Total\ Assets_t}$
noa_gr1a	change in net operating assets	Investment	$\frac{\Delta NOA_{t-t_{-1}}}{Total\ Assets_{t-1}}$
saleq_gr1	quarterly sales growth	Investment	$\frac{\Delta SALESQ_{t-t_{-1}}}{\lvert SALESQ_{t-1}\rvert}$
ival_me[1]	intrinsic value-to-market	Value	$\frac{Intrinsic\ ValueEstimate_t}{Market\ Cap_t}$
resff3_12_1	residual momentum t-12 to t-1	Momentum	$ret_{t-12,t-1} - \hat{\beta}_{mkt}R_{mkt} - \hat{\beta}_{smb}R_{smb} - \hat{\beta}_{sml}R_{sml}$
seas_6_10an	years 6-10 lagged returns, annual	Seasonality	$\prod_{m=61}^{120}(1+r_{t-m})^{\frac{1}{5}} - 1$
debt_me	debt to market cap	Value	$\frac{LT + ST\ Debt_t}{Market\ Cap_t}$
seas_6_10na	years 6-10 lagged returns, nonannual	Low Risk	$\prod_{m=61}^{120}(1+r_{t-m})^{\square} - 1$
zero_trades_252d	number of zero trades (12 months)	Low Risk	$\#Days, Volume_{t=12m} = 0$
cowc_gr1a	change in current operating working capital	Accruals	$\frac{\Delta WCAP_{t-t_{-1}}}{Total\ Assets_{t-1}}$
nncoa_gr1a	change in non-current operating assets	Investment	$\frac{\Delta Non - Current\ Operating\ Assets_{t-t_{-1}}}{Total\ Assets_{t-1}}$
ocf_me	operating cash flows to market cap	Value	$\frac{Operating\ Cash\ Flow_t}{Market\ Cap_t}$
zero_trades_21d	number of zero trades (1 month)	Low Risk	$\#Days, Volume_{t=1m} = 0$
turnover_126d	share turnover	Low Risk	$\frac{\sum_{t=1}^{126} Volume_t}{Shares\ Outstanding_{t=0}}$
rmax5_rvol_21d	highest 5 days of returns scaled by volatility (1 month)	ST Reversal	$\frac{\sum \lvert MaxRet_{t=1m}\rvert}{\sqrt{252}\,.\,\widehat{\sigma_{t=1m}}}$
seas_11_15na	years 11-15 lagged returns, non-annual	Seasonality	$\prod_{m=132}^{180}(1+r_{t-m})^{\square} - 1$
o_score[3]	Ohlson O-Score	Profitability	
niq_at	quarterly return on assets	Quality	$\frac{Quarterly\ Net\ Income}{Total\ Assets}$
seas_16_20an	years 16-20 lagged returns, annual	Seasonality	$\prod_{m=181}^{240}(1+r_{t-m})^{\frac{1}{5}} - 1$
ni_arl	earnings persistence	Debt Issue	$corr.(NI_t, NI_{t-1}, NI_{t-m})$

ivol_ff3_21d	idiosyncratic volatility ff 3-factor model	Low Risk	$R_{i,t} - Rf = \alpha + \hat{\beta}_{mkt} + \hat{\beta}_{smb} + \hat{\beta}_{sml} + \widehat{\varepsilon_t}$ $ivol_{ff3_{21d}} = \widehat{\sigma_{\varepsilon_t}}$
ni_me	earnings to price ratio	Value	$\frac{Net\ Income_t}{Market\ Cap_t}$
dsale_dinv	sales change minus inventory change	Growth	$\frac{\Delta Sales_{t-t_{-1}}}{Sales_{t-1}} - \frac{\Delta Inventory_{t-t_{-1}}}{Inventory_{t-1}}$
ni_be	return on equity	Profitability	$\frac{Net\ Income_t}{Book\ Equity_t}$
noa_at	net operating assets	Debt Issue	$\frac{Operating\ Assets - Operating\ Liabilities}{Total\ Assets}$
age	firm age	Leverage	$Year_t - CRSP\ list\ date$
ret_12_1	price momentum t-12 to t-1	Momentum	$\prod_{m=1}^{12}(1 + r_{t-m})^{\square} - 1$
aliq_mat[2]	liquidity of market assets	Leverage	
nfna_grla	change in net financial assets	Debt Issue	$\frac{Cash_t + ST\ Investments_t - TotalDebt_t}{Total\ Assets_{t-1}}$
at_me	assets to market	Value	$\frac{Total\ Assests_t}{Market\ Cap_t}$

ival_me[1], aliq_mat[2]: are included in this table for completeness, but were ultimately excluded in the final analysis due to ambiguity about their formulation. While these variables may correspond to measures from previous studies (Frankel & Lee, 1998), (Ortiz-Molina & Phillips, 2009), Robeco's "Factor Zoo (.zip)" dataset provides only variable names and brief definitions without implementation details, as mentioned above. Attempts to contact the authors for clarification were unsuccessful during the time of writing.

o_score[3]: $O - score = -1.32 - 0.407 \cdot \log(TA) + 6.03 \cdot \frac{TL}{TA} - 1.43 \cdot \frac{WCAP}{TA} + 0.76 \cdot \frac{CL}{CA} - 1.72 \cdot \text{OENEG(0 or 1)} - 2.37 \cdot \frac{NI}{TA} - 1.83 \cdot \frac{OCF}{TL} + 0.285 \cdot NINEG(0\ or\ 1) - 0.521 \cdot \frac{\Delta NI_t}{|\Delta NI_t| - |\Delta NI_{t-1}|}$

where:

OENEG: 1 if $\text{TA}_t > \text{TL}_t$; 0 otherwise

NINEG: 1 if $\text{NI}_{t-1} < 0\ and\ \text{NI}_t < 0$; 0 otherwise

Table 4: Variable Definitions and Implementation Methodology

This table reports the variable names, line items, and code identifiers from CRSP, which are used for calculating each of the 28 features within the programming environment.

Accounting Variable	Main & Secondary Measures Applicable to all industrial firms	Fin. Firm Specific Measure Sic Codes: 6000-6999	Util. Firm Specific Measures Sic Codes: 4900-4999
Fundamentals: Income Statement Items; Quarterly			
Revenues / "SALES"	"saleq" or "revtq"	"revtq" "saleq" or "finrevq"	"saleq" or "revtq"
Cost of Goods Sold / "COGS"	"cogsq" or ("xoprq" – "xsgaq")	"cogsq" or "xintq" + "xsgaq" or "finxtq" + "xsgaq"	"cogsq" or ("xoprq"- "xsgaq")
Gross Profit / "GP"	Revenues – Cogs	Revenues – Cogs	Revenues – Cogs
Operating Expenses / "OPEX"	"xoprq" or "cogs + xsgaq"	"xoprq", or "cogs + xsgaq", or "finxoprq"	"xoprq" or "cogs + xsgaq"
R&D	"xrdq" if "NA" fill 0	"xrdq" if "NA" fill 0	"xrdq" if "NA" fill 0
EBITDA	"oibdpq" + "xrdq" or Rev – Opex + R&D	"oibdpq" + "xrdq" or Rev – Opex + R&D	"oibdpq" + "xrdq" or "uopiq" + "xrdq" or Rev – Opex + R&D
D&A	"dpq" – general measure	"dpq" – general measure	"dpq" – general measure
R&D Ammortization*	"R&D_ammort" – from capitalization	"R&D_ammort" – from capitalization	"R&D_ammort" – from capitalization
EBIT	"oiadpq" + "R&D*" – "R&D_ammort" or "oidpq" – "dpq- "R&D_ammort" or EBITDA – "dpq" – "R&D_ammort"	"oiadpq" + "R&D*" – "R&D_ammort" or "oidpq" – "dpq- "R&D_ammort" or EBITDA – "dpq" – "R&D_ammort"	"oiadpq" + "R&D*" – "R&D_ammort" or "oidpq" – "dpq- "R&D_ammort" or EBITDA – "dpq" – "R&D_ammort"
Taxes Payable	"txpq"	"txpq"	"txpq"
Net Income / "NI"	"ibq" + "R&D*" – "R&D_ammort" or "niq" + "R&D*" – "R&D_ammort"	"ibq" + "R&D*" – "R&D_ammort" or "niq" + "R&D*" – "R&D_ammort"	"ibq" + "R&D*" – "R&D_ammort" or "niq" + "R&D*" – "R&D_ammort"
Fundamentals: Balance Sheet/Cash Flow Statement/Market Based Items; Quarterly			
Current Assets / "CA"	"actq" or "rectq" + "invtq" + "cheq" + "acoq"	"actq" or "cheq" + "finivstq" + "invtq" + "finreccq" + "finacoq" or "finchq" + finivstq" + "finreccq" + "finacoq"	"actq" or "rectq" + "invtq" + "cheq" + "acoq" or "urectq" + "uinvq" + "uacoq" + "cheq"
Current Liabilities / "CL"	"lctq" or "apq" + "dlcq" + "txpq" + "lcoq"	"lctq" or "finnpq" + "findlcq" + "finlcoq"	"lctq" or "apq" + "dlcq" + "apq" + "txpq" + "lcoq"
Current Operating Assets / "COA"	CA – "cheq"	CA (cash treated as operational for financials)	CA – "cheq"
Current Operating Liabilities / "COL"	CL – max("dlcq",0)	CL (short term debt treated as operational)	CL – max("dlcq",0)
Net Operating Assets / "NOA"	TA – "cheq" – "ivaoq" – TL – "dlttq" – "dlcq"	Not Applicable	TA – "cheq" – "ivaoq" – TL – "dlttq" – "dlcq"

Total Assets / "TA"	"atq" + "R&D_Cap*" or "seqq" + "dlttq" + max(lctq,0) + "R&D_Cap*"	"atq" + "R&D_Cap*"	"atq" + "R&D_Cap*" or "seqq" + "dlttq" + max(lctq,0) + "R&D_Cap*"
Total Liabilities / "TL"	"ltq"	"ltq"	"ltq"
Short-Term Debt / "STDEBT"	"dlcq"	Not Applicable	"dlcq"
Long-Term Debt / "LTDEBT"	"dlttq"	Not Applicable	"dlttq" or "uddq" + "udmbq" + "udoltq" + "udpcoq"
Total Debt / "TOTDEBT"	ST + LT Debt	Not Applicable, debt items are operational liabilities	ST + LT Debt
Adjusted* Book Equity / "BE"	"seqq" + "R&D_Cap*" or "ceqq" + max("pstkq"0) + "R&D_Cap*" or TA + "R&D_Cap*"– TL	"seqq" + "R&D_Cap*" or "ceqq" + max("pstkq"0) + "R&D_Cap*" or TA + "R&D_Cap*"– TL	"seqq" + "R&D_Cap*" or "ceqq" + max("pstkq"0) + "R&D_Cap*" or "uceqq" + max("pstkq"0) + "R&D_Cap*" or TA + "R&D_Cap*"– TL
Net Working Capital / "NWC"	"wcapq" or CA – CL	Not Applicable	"wcapq" or CA - CL
Current Operating Working Capital / "COWC"	COA - COL	COA - COL	COA - COL
Operating Cash Flows / "OCF"	"oancfy_q" + R&D* - R&D_Ammort* or NI + D&A – max("wcapchy_q",0)	"oancfy_q" + R&D* - R&D_Ammort* or NI + D&A – max("wcapchy_q",0)	"oancfy_q" + R&D* - R&D_Ammort* or NI + D&A – max("wcapchy_q",0)
Market Cap / "MC"	"mktvalq"	"mktvalq"	"mktvalq"

Figure 2: ML Rolling Window Training Design

This figure illustrates the rolling window design structure used to train and tune the Amalgam model across time. For each iteration, the training set is anchored, beginning in Q1 1961 and expanding forward to include all available firm-quarter observations up to the start of the validation set. This forms an expanding training window that captures additional firm-quarters as time progresses. A fixed-length validation window of the latest 4 years follows the training period, used for hyperparameter tuning. The test set simulates model deployment (investments in the top decile), generating the out of sample results. This process is repeated on a quarterly basis for the full OOS period (Q1 1981 – Q4 2024). After each quarter's test set prediction, the model is re-fitted by advancing the training and validation windows one quarter at a time.

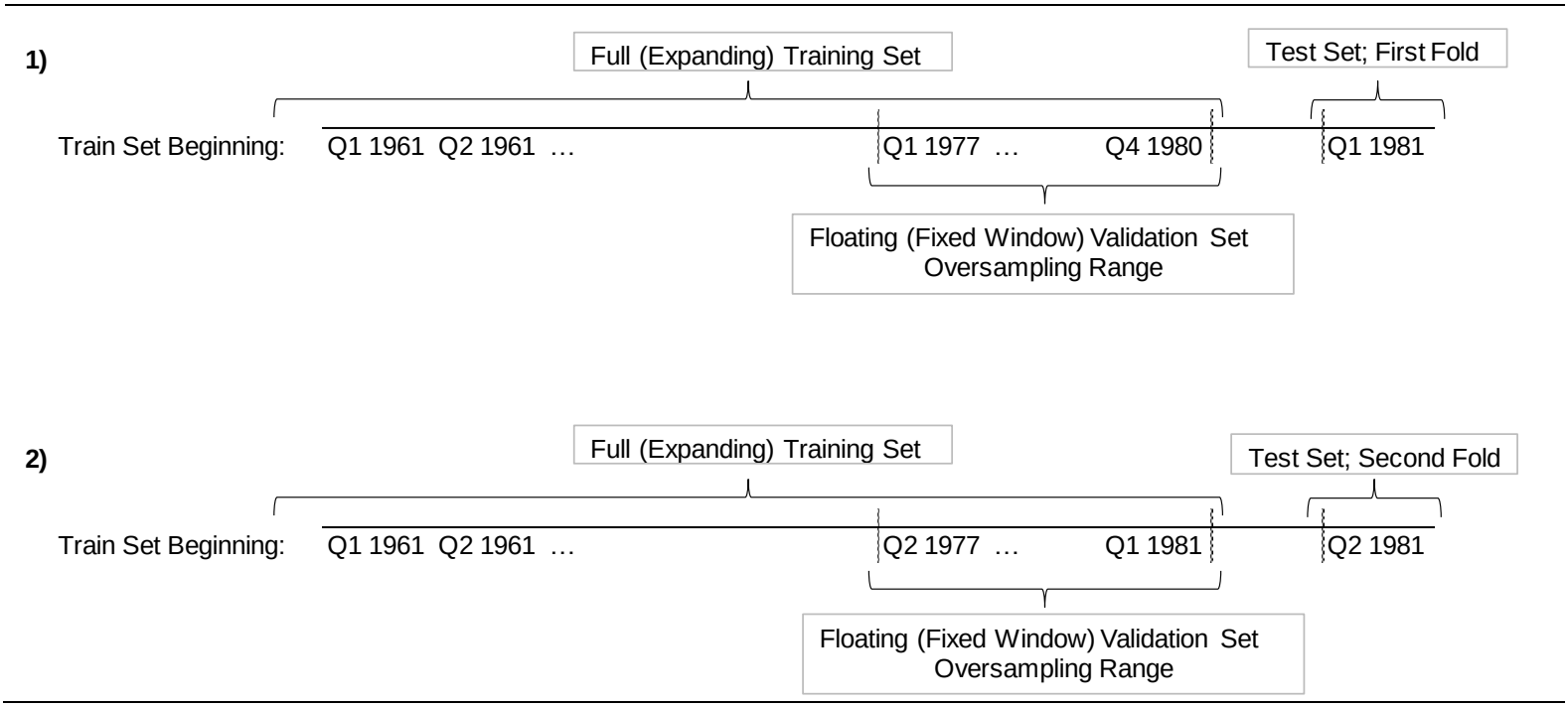


Table 5: Normalization of Feature Proxies via GIND-Level Percentile Ranks

This table illustrates the multi-step transformation process of raw firm-level feature proxies into Global Industry Classification Standard- relative (GICS) percentile ranks, used as model inputs:

The process begins from Panel **A)**, showing a sample of all raw features (**"ni_me", "dsale_dinv", "ni_be", "noa_at"**)[1] for three firms (**PERMNO**) in GICS industry 551020 (Gas Utilities) during 2010Q2.

In Panel B, the raw feature values are transformed into percentile scores within each GICS industry code (GIND) and quarter grouping (**"_pct"** suffix). This procedure captures each firm's relative standing within its industry for the time period (**"y", "qtr"**). Percentile values are bounded between 0 and 1, where a score of 0 corresponds to the lowest observed raw value in that GIND-quarter group and a score of 1 to the highest, with all intermediate values reflecting cross-sectional rankings among industry peers.

Panel **C)** demonstrates that, post-normalization, multiple firms from different industries can share identical percentile scores within a given quarter. To prevent information loss or misinterpretation by the XGBoost Model used in training and prediction, the GIND code is included as a **categorical feature**. This explicitly conditions the model to "recognize" industry context, ensuring that similar percentile values are interpreted *relative* to their originating industry. This normalization design preserves within-industry information while enabling consistent cross-industry identification and learning.

A) Raw Values:

PERMNO	y	qtr	GIND	ni_me	dsale_dinv	ni_be	noa_at
10001	2010	2	551020	0.059327	1.341539007	0.066168	0.757558
12781	2010	2	551020	0.030456	1.028910282	0.043054	0.530067
27756	2010	2	551020	0.009545	0.507949246	0.018874	0.67386

B) Within-Industry Percentile Rankings:

PERMNO	y	qtr	GIND	ni_me_pct	dsale_dinv_pct	ni_be_pct	noa_at_pct
10001	2010	2	551020	1	1	0.8	0.9
12781	2010	2	551020	0.5	0.8	0.5	0.2
27756	2010	2	551020	0.3	0.6	0.4	0.6

C) Holistic Overview of Percentile Rankings Across Different GINDs

PERMNO	y	qtr	GIND	ni_me_pct	dsale_dinv_pct	ni_be_pct	noa_at_pct
12781	2010	2	551020	0.5	0.8	0.5	0.2
75257	2010	2	452010	0.5	0.8	0.5	0.2
89301	2010	2	255040	0.5	0.8	0.5	0.2

ni_me: quarterly net income to market equity value; Earnings-to-Price Ratio

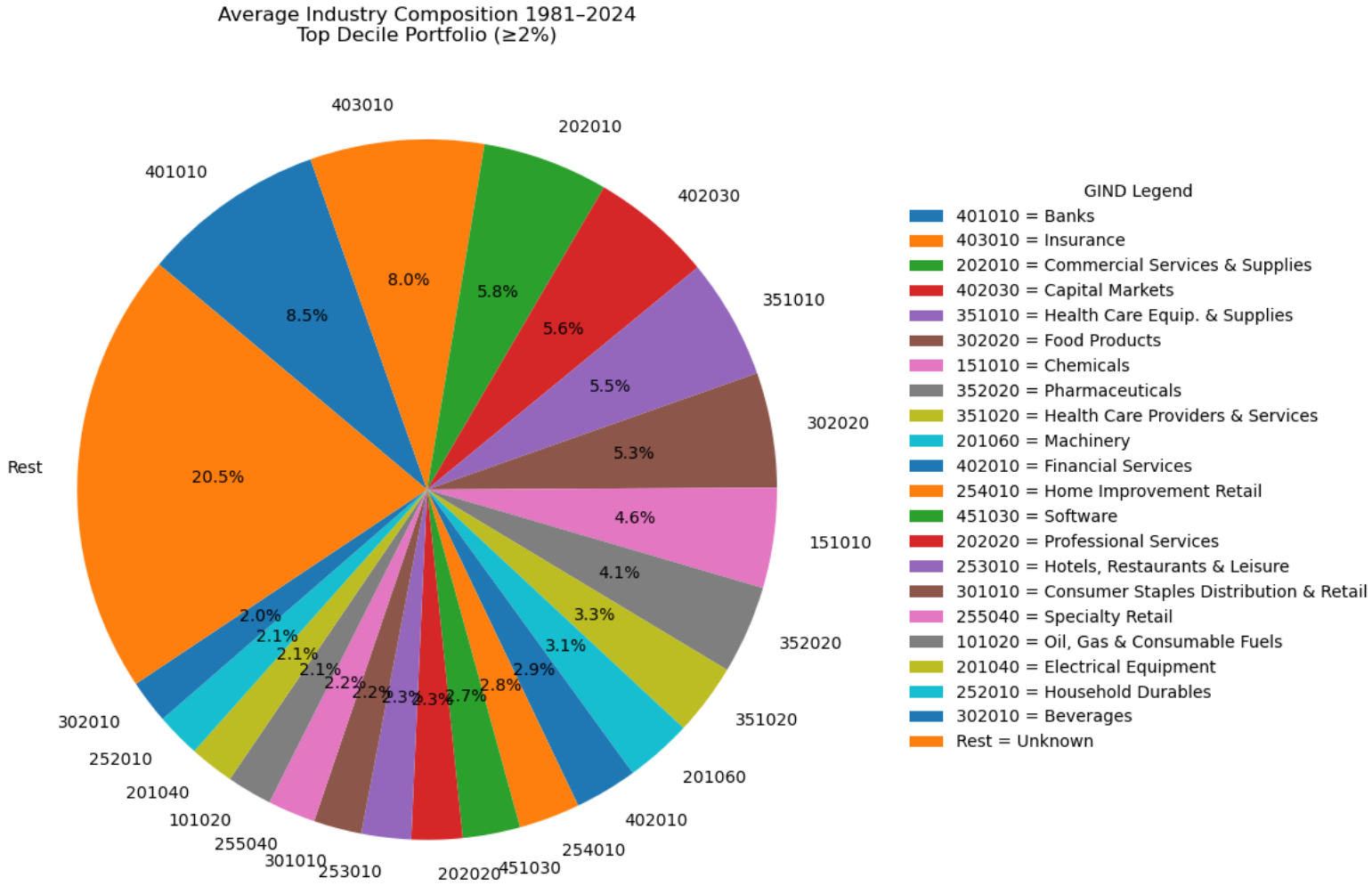
dsale_dinv: quarterly change in sales minus quarterly change in inventory

ni_be: quarterly net income to book equity value; Return on Book Equity

noa_at: ratio of net operating assets (NOA) to total assets

Figure 8: Average Industry Composition of Top and Bottom Decile Portfolios (1981–2024)

This figure showcases the average industry composition across the entire back-test of the model's top and bottom decile portfolios. Industries are categorized using the six-digit GICS Industry Code (GIND), and the charts visualize only those with an average portfolio weight of at least 2%. The remaining firms (>2% of portfolio holdings) are grouped under "Rest".
The top decile portfolio (first graph) reflects the stocks from the most frequently selected industries by the model, with a concentration in Banks, Insurance and various industries under the broad Healthcare Sector.
The bottom decile portfolio (second graph) represents the stocks from the most consistently shorted/avoided industries. It is less fragmented, with the highest weights being in Biotechnology, Financial Services and the leftover, underperforming projections from Banks.
The composition of stocks within different industries and the highlighted differences between the two graphs reveal the systematic "preferences" and "aversions" resulting from the model's selection process over the back-test period.



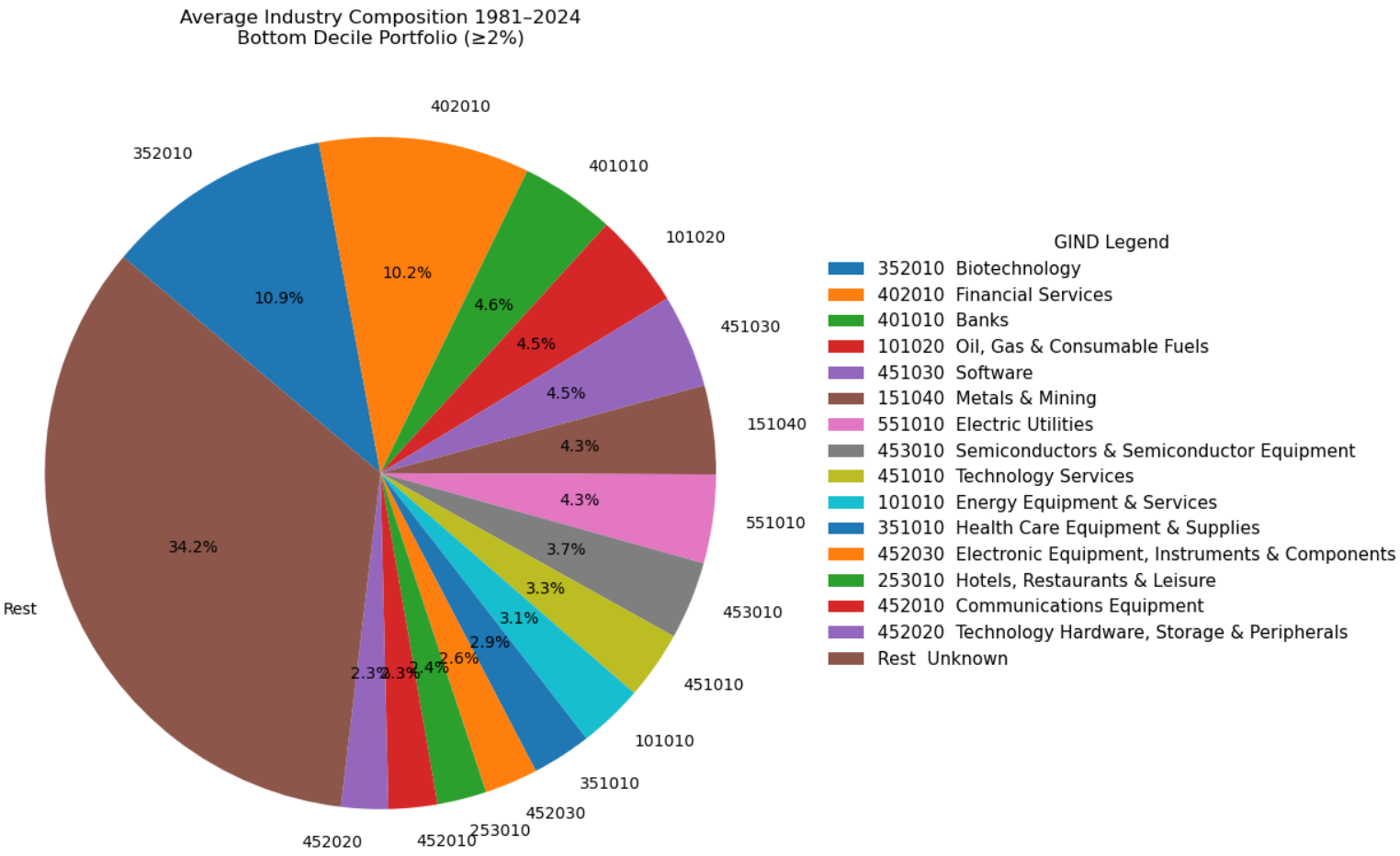

Average Industry Composition 1981–2024
Bottom Decile Portfolio (≥2%)
402010
352010
401010
101020
451030
151040
551010
453010
451010
101010
351010
452030
253010
452010
452020
Rest
10.9%
10.2%
4.6%
4.5%
4.5%
4.3%
4.3%
3.7%
3.3%
3.1%
2.9%
2.6%
2.4%
2.3%
2.3%
34.2%
GIND Legend
352010 Biotechnology
402010 Financial Services
401010 Banks
101020 Oil, Gas & Consumable Fuels
451030 Software
151040 Metals & Mining
551010 Electric Utilities
453010 Semiconductors & Semiconductor Equipment
451010 Technology Services
101010 Energy Equipment & Services
351010 Health Care Equipment & Supplies
452030 Electronic Equipment, Instruments & Components
253010 Hotels, Restaurants & Leisure
452010 Communications Equipment
452020 Technology Hardware, Storage & Peripherals
Rest Unknown

Table 10: Mean Absolute SHAP Values Per Feature

This table shows the mean absolute SHAP % values per feature across 4 decades. Features are grouped within the risk dimension/factor they represent. Each feature's contribution reflects the marginal influence it had on model predictions in a given decade. The sum of all feature contributions per decade (Total Signal Strength) shows the model's overall reliance on signals in that period.

Factors	Features	1980	1990	2000	2010	2020
Value	ocf_me	0.4%	1.4%	2.3%	2.2%	2.7%
	ni_me	7.3%	7.2%	6.2%	6.0%	6.3%
	debt_me	0.3%	0.7%	0.9%	0.9%	1.3%
Industry	GIND	6.5%	6.2%	6.1%	5.4%	6.6%
Low Risk	seas_6_10na	0.1%	0.1%	0.1%	0.0%	0.1%
	turnover_126d	0.8%	0.4%	0.5%	0.5%	0.7%
	zero_trades_21d	0.2%	0.9%	1.1%	0.7%	1.2%
	zero_trades_252d	0.8%	0.5%	0.5%	0.4%	0.6%
	ivol_ff3_21d	1.5%	2.6%	2.4%	2.3%	2.6%
Investment	noa_gr1a	0.8%	0.5%	0.5%	0.3%	0.6%
	saleq_gr1	0.4%	0.9%	1.2%	1.2%	1.9%
	nncoa_gr1a	0.9%	0.8%	0.7%	0.7%	1.0%
Debt Issue	ni_arl	0.2%	0.1%	0.4%	0.3%	0.7%
	noa_at	0.4%	0.8%	1.2%	0.9%	1.2%
	at_me	0.5%	0.7%	0.9%	0.7%	0.8%
Momentum	resff3_12_1	0.1%	0.1%	0.1%	0.0%	0.1%
	ret_12_1	4.3%	3.7%	2.7%	1.2%	1.1%
Profits	age	0.7%	0.4%	0.6%	0.4%	0.7%
	o_score	0.7%	0.2%	0.2%	0.1%	0.3%
	ni_be	1.5%	1.5%	1.3%	0.9%	1.2%
Quality	cop_at	0.1%	0.1%	0.1%	0.1%	0.2%
	niq_at	0.3%	0.6%	0.6%	0.6%	0.9%
Seasonality	seas_11_15na	0.1%	0.1%	0.1%	0.0%	0.2%
	seas_16_20an	0.4%	0.1%	0.0%	0.0%	0.2%
	seas_6_10an	0.5%	0.3%	0.3%	0.1%	0.2%
Accruals	cowc_gr1a	0.3%	0.1%	0.1%	0.0%	0.1%
Growth	dsale_dinv	0.2%	0.2%	0.2%	0.1%	0.2%
Leverage	nfna_grla	1.3%	0.4%	0.2%	0.1%	0.2%
ST Reversal	rmax5_rvol_21d	0.1%	0.1%	0.1%	0.1%	0.2%
	Total Signal Strength	**31.7%**	**31.5%**	**31.6%**	**26.0%**	**34.3%**