Mosaics of Predictability

Lin William Cong Guanhao Feng Jingyu He Yuanzhi Wang*

First draft: Feb. 2024; this draft: Dec. 2024

Abstract

Existing studies often regard return predictability as an attribute of predictors or models. This paper argues that return predictability is an unobserved yet inherent asset characteristic linked to expected returns, varying across stocks and over time. We propose a novel tree-based clustering method to measure heterogeneous return predictability by grouping asset-return observations with similar levels of predictability. The resulting clusters are characterized using high-dimensional firm characteristics and aggregate predictors. Our empirical analysis reveals significant patterns of heterogeneous return predictability in individual U.S. stocks. First, asset clusters with low trading volumes, high earnings-to-price ratios, and high unexpected earnings exhibit the highest predictability. Second, predictability declines sharply when the dividend yield is low, while it peaks during periods of high dividend yield and low default yield. Furthermore, we identify a new predictability anomaly: highly predictable long-only portfolios generate unexplained alphas of about 1% across various factor models over the past two decades, while a long-short portfolio based on predictability yields even higher alphas.

Key Words: Anomaly, Characteristics, Heterogeneous Predictability, Regime Switches, Goal-oriented Clustering.

*We are grateful to Chunrong Ai, Doron Avramov, Daniele Bianchi, Jian Chen (Discussant), Qihui Chen, Victor DeMiguel, Fuwei Jiang, Hao Ma, Dacheng Xiu, Lei Yang (Discussant), Yi Zhang, and the seminar and conference participants at Cornell Tech, CUHK Shenzhen, 2024 EFMA Annual Meeting, 2024 FMA Asia/Pacific Conference, 2024 FinEML Conference, HKUST Guangzhou, Queen Mary University of London, 4th Big Data and Econometric Conference at Xiamen University, 16th Annual SoFiE Meeting, 8th PKU-NUS Annual International Conference on Quantitative Finance and Economics, Summer Institute of Finance 2024, 2024 Academic Conference on Digital Economy Development and Governance, SIDF 2024, 17th ARMC 2024. Cong (E-mail: will.cong@cornell.edu) is at Cornell University SC Johnson College of Business (Johnson) and NBER; Feng (E-mail: gavin.feng@cityu.edu.hk), He (E-mail: jingyuhe@cityu.edu.hk), and Wang (E-mail: yuanzwang5-c@my.cityu.edu.hk) are at the City University of Hong Kong.

1 Introduction

Forecasting asset returns has long been a central question in asset pricing. Previous research has provided substantial evidence of return predictability across various asset classes[1]. However, literature often regards return predictability as an attribute of the predictors or models themselves. On one hand, researchers have identified several aggregate or macroeconomic predictors for forecasting market returns, such as market dividend yield (e.g., Fama and French, 1988) and default yield (e.g., Keim and Stambaugh, 1986); while firm characteristics, such as size and value (e.g., Fama and French, 1992), can predict cross-sectional variations in returns. On the other hand, numerous methods have been developed to examine return predictability, including predictive regressions (e.g., Stambaugh, 1999), security sorting (e.g., Jensen et al., 1972), cross-sectional models (e.g., Fama and MacBeth, 1973; Han et al., 2024), and machine learning approaches (e.g., Kelly and Pruitt, 2013; Gu et al., 2020; Kelly et al., 2024).

Return predictability is unobservable[2], and it lacks a clear, consistent definition in the literature. Existing studies often evaluate predictability at an aggregate or average level, typically relying on constant-coefficient models that assume homogeneity across observations and uniform predictability. However, it neglects the potential heterogeneity in return predictability, particularly at the level of individual assets.

This paper offers a novel perspective on return predictability by addressing two key questions: which types of assets exhibit higher return predictability, and how does this predictability evolve across different macroeconomic regimes? We propose that, beyond predictors or models, return predictability represents an unobserved yet intrinsic characteristic of assets, potentially linked to the cross section of expected returns. Our findings indicate that return predictability is inherently heterogeneous, varying significantly across stocks and over time.

[1]For example, empirical studies have documented return predictability in the aggregate market (e.g., Welch and Goyal, 2008; Campbell and Thompson, 2008), individual stocks (e.g., Fama and French, 2008; Rapach et al., 2013; Lewellen, 2015), corporate bonds (e.g., Feng et al., 2024), treasury bonds (e.g., Bianchi et al., 2021), and mutual fund alphas (e.g., Kaniel et al., 2023; DeMiguel et al., 2023).

[2]Many equity characteristics are unobservable and require measurements, such as beta and volatility. Concepts related to predictability include anomaly average return, predictor significance, out-of-sample R^2, forecast-implied portfolio, etc.

First, substantial empirical evidence suggests that return predictability is not homogeneous. For example, Avramov et al. (2023) demonstrate that predictability is concentrated in micro-cap stocks, distressed stocks, and during periods of high market volatility. Additionally, Green et al. (2017) document a decline in the significance of characteristics since 2003, indicating a sharp reduction in return predictability. Long-established research also links stock market return predictability to business cycle indicators (e.g., Keim and Stambaugh, 1986; Fama and French, 1989), with further support for the use of advanced time-varying predictive models.[3] Thus, while unobservable, return predictability may vary significantly across stocks and evolve over time.

Second, if heterogeneous return predictability is an inherent characteristic of assets, how can it be measured? Since predictability is closely linked to the signal-to-noise (S2N) ratio or the regression R^2[4], we thus focus on estimating R^2 as the measure of heterogeneous return predictability. Given this measure, an important question arises: what is the potential relationship between predictability and average returns? While Rapach, Strauss, and Zhou (2010) and Kelly, Malamud, and Zhou (2024) argue that there is no direct link between out-of-sample (OOS) predictive R^2 and investment gains, we identify an anomaly through our proposed measure of return predictability, revealing a positive relationship between predictability and average returns.

To the best of our knowledge, this paper is the first to systematically investigate and measure the heterogeneous return predictability of individual stocks. We propose the concept of *mosaics of predictability*, where return predictability varies across time and the cross section, revealing partitioning patterns that resemble mosaics on the panel. Specifically, instead of measuring global-level R^2, we focus on group-level R^2_j, which reflects the predictability of stock-return observations within the same group j. The regression R^2 serves as a summary statistic for the aggregate goodness of fit, but it is impossible to calculate for each stock given the unbalanced short panel of monthly

[3]Studies investigating time-varying return predictability highlight stronger predictability during economic recessions (e.g., Henkel et al., 2011; Dangl and Halling, 2012). Recently, the pockets of predictability have been revisited in Farmer, Schmidt, and Timmermann (2023); Cakici, Fieberg, Neumaier, Poddig, and Zaremba (2024); Farmer, Schmidt, and Timmermann (2024).

[4]In predictive or cross-sectional regressions, a higher predictor coefficient correlates with greater return predictability and a larger long-short portfolio spread, which results in a higher R^2.

data. As demonstrated by Cong et al. (2023), heterogeneous panel data analyses can be framed as a clustering problem, where observations are grouped based on their similar exposure to factors.

The empirical solution involves developing a clustering algorithm that separates predictable observations from less predictable ones. First, we propose measuring return predictability using the R^2 of the cluster-wise predictive model, ensuring that observations within a cluster share similar predictability. Second, we introduce a customized panel tree approach to identify the optimal clustering structure that captures variations in stock return predictability, grouping stocks with similar predictability levels. The decision tree structure then organizes these clusters into mosaics on the panel, described by firm characteristics and aggregate predictors for easier visualization. This approach preserves economic interpretability and highlights key variables associated with different levels of predictability. By leveraging the asymmetric interactions between aggregate predictors and firm characteristics, we measure heterogeneous return predictability through cluster-wise predictive models.

Empirical Highlights. We conduct our empirical analysis using a panel of U.S. monthly individual stock returns from 1973 to 2022, incorporating 58 firm-level characteristics spanning eight major categories, along with eight monthly aggregate predictors.

First, our findings reveal significant heterogeneity in the cross-sectional predictability of stock returns. Specifically, our approach identifies approximately 15 clusters (distinct leaf nodes in a decision tree), as shown in Figure 4. We find that asset clusters with low trading volumes, high earnings-to-price ratios, and high unexpected earnings are the most predictable, while asset clusters with non-low trading volumes, non-high return variance, and non-high unexpected earnings are the least predictable. The predictability gap between highly and less predictable clusters remains consistent out of the sample, validating the robustness of the relationship between unobservable predictability and underlying high-dimensional characteristics.

Second, we demonstrate that our measure of heterogeneous predictability can be linked to a risk anomaly. By ranking clusters based on their return predictability —

effectively sorting assets by predictability, since the uniformity within a cluster — we reveal a significantly positive relationship between predictability and average returns. We find that highly predictable long-only value-weighted portfolios generate significantly unexplained alphas of approximately 1% across various factor models over the two-decade OOS period, while a predictability-based long-short portfolio yields even higher unexplained alphas. These findings demonstrate that predictability-driven strategies yield superior risk-adjusted returns, presenting a contrasting empirical conclusion to Rapach et al. (2010) and Kelly et al. (2024).

Third, building on the work of Cong et al. (2023) and Feng et al. (2024), we extend the clustering analysis to account for regime shifts driven by aggregate macroeconomic predictors or calendar months (structural breaks). This approach learns dynamic patterns in stock return predictability across both cross-sectional and time-series dimensions. We find that the predictability regimes are driven by two macroeconomic variables: dividend yield and default yield, both are key indicators of business cycles (Fama and French, 1988; Keim and Stambaugh, 1986; Fama and French, 1989). Notably, return predictability declines sharply when the dividend yield is low but peaks during periods of high dividend yield and low default yield—conditions typically associated with recessions. These findings are consistent with those of Henkel et al. (2011) and Dangl and Halling (2012). Across three macroeconomic regimes, each decision tree selects different characteristics, with value characteristics consistently emerging as significant predictors. Similar patterns observed through structural breaks by calendar months further validate the robustness of these results.

Finally, while the primary objective of our approach is to cluster asset returns based on predictability, it also produces cluster-wise predictive models as a byproduct. We evaluate the effectiveness of these models using forecast-implied trading strategies and compare their performance to that of homogeneous predictions. Our analysis demonstrates consistent improvements across multiple investment metrics, including average returns, Sharpe ratios, market alphas, and maximum drawdowns.

Methodological Innovations. There is no off-the-shelf clustering method available to solve this finance-driven empirical challenge. We introduce a "divide-and-conquer" goal-oriented clustering approach that sequentially partitions the panel of stock-return observations according to their predictability. The goal is to separate highly predictable observations from less predictable ones while simultaneously providing the group-level predictability measure of R^2. This partitioning is achieved by maximizing the difference in the S2N ratios, or R^2, across clusters using cluster-wise heterogeneous return forecasting models. The panel tree-based clustering offers an interpretable economic framework by describing clusters through firm-specific characteristics for the cross section, or macro aggregate predictors for regime switching. Generalizing the security sorting, our clustering can be viewed as multi-way sorting with multiple characteristics. Our study contributes to the broader literature on asset heterogeneity, as explored by Cong et al. (2023). However, it differs in its economic objective: while Cong et al. (2023) focuses on maximizing the marginal likelihood of heterogeneous factor models, this paper emphasizes distinguishing stock return predictability.

The return prediction literature (e.g., Gu et al., 2020) often fit a homogeneous, time-invariant global model applied uniformly to all asset return observations. However, Feng and He (2022) and Evgeniou et al. (2023) highlight that such global models ignores the inherent heterogeneity of asset returns. Our framework addresses this limitation by fitting cluster-wise predictive models based on an identified clustering structure, grouping asset-return observations with similar predictability as depicted by firm-level characteristics and/or aggregate predictors. The key innovation of our approach is the seamless integration of clustering and cluster-wise model fitting within a unified framework, rather than treating these processes as separate steps. Moreover, the framework is model-agnostic, allowing it to be combined with various machine learning (ML) methods, including Lasso, PCA, and others.

Literature Positions. Our paper contributes to the extensive empirical literature on return predictability. Early studies (e.g., Keim and Stambaugh, 1986; Fama and French, 1988, 1989) identify market-wide predictors (e.g., default yield, dividend yield) for

time-series return predictability over business cycles in aggregate equity and bond markets. Various characteristics, anomalies, or long-short factors (e.g., size, value, momentum) are documented (e.g., Fama and French, 1992, 1993; Jegadeesh and Titman, 1993) in studies on cross-sectional return predictability. However, many empirical findings appear unstable in OOS or post-publication evaluations (Pesaran and Timmermann, 1995; Welch and Goyal, 2008; Harvey et al., 2016; McLean and Pontiff, 2016). In addition to first measuring the stock characteristic of return predictability, our research complements these studies by investigating the heterogeneity in return predictability, which may help explain the inconsistencies of predictability findings.

Our paper contributes to studies on heterogeneous return predictability. Research on time-varying return predictability shows market returns are more predictable during recessions using a regime-switching VAR (Henkel et al., 2011) and a time-varying coefficient model (Dangl and Halling, 2012). In the cross section, Avramov (2002) find small-cap value stocks more predictable than large-cap growth stocks, and Green et al. (2017) show forecast-implied portfolios exploiting characteristics-based predictability are insignificant outside micro-caps since 2003. Avramov et al. (2023) further discover predictability is concentrated in micro-caps, distressed stocks, or during high-volatility periods. Our model systematically analyzes heterogeneity in return predictability, extending traditional time-varying coefficient models by incorporating high-dimensional firm characteristics and aggregate predictors.

This paper, along with Cong et al. (2024) and Cong et al. (2023), is among the first to develop economically guided panel tree-based clustering, part of the emerging AI literature on goal-oriented search — a data-driven approach to optimizing an economic goal in a large, flexible modeling space (e.g., Cong et al., 2020; Feng et al., 2024). The panel trees' "divide-and-conquer" approach mimics how humans solve complex problems by completing constituent tasks. Cong et al. (2023) and Feng et al. (2024) extend the panel-tree framework to fit heterogeneous models by maximizing marginal likelihood, while our paper focuses on separating observations for heterogeneous S2N ratios. Closely related by analyzing endogenous grouped heterogeneity in

financial markets, Ahn et al. (2009) use unsupervised clustering based on return correlations, and Patton and Weller (2022) generalize K-means to group assets by within-group slopes and averages, finding pervasive risk-price heterogeneity. More recently, Evgeniou et al. (2023) apply unsupervised K-means to cluster firms by characteristics and estimate post-cluster heterogeneous predictive models. Due to the finance-driven empirical challenge, only our customized clustering can separate predictable from less predictable observations for predictability "mosaics" in return panels.

The sections are organized as follows: Section 2 introduces the R^2 measure of predictability and the clustering model, while Section 3 describes the data and model evaluation. Section 4 focuses on cross-sectional heterogeneous predictability and presents the anomaly. Section 5 discusses the time-varying predictability and regime switches. Section 6 reports investment performance gains based on cluster-wise models. Finally, Section 7 concludes with appendices detailing algorithms, data, and additional results.

2 Methodology

2.1 Measurement of Return Predictability for Clustering

As discussed in the introduction, return predictability is unobservable and lacks a consensus definition in the literature. We propose that the in-sample R^2 of a predictive model serves as a natural measure of data predictability. While statisticians traditionally interpret R^2 as a property of the model, representing the proportion of return variation explained by the model, we extend this interpretation from a data perspective. Specifically, we view R^2 as a measure of how much variation in a specific dataset can be explained by the best-fitted ML model. Equivalently, it serves as an indication of the difficulty in predicting the dataset.

Intuitively, if the stock universe is already partitioned into multiple clusters, fitting a predictive model (with the necessary turning of parameters) for each cluster allows each model to achieve optimal predictive performance within its respective cluster. However, the resulting R^2 values for each cluster may still vary, reflecting the intrinsic difficulty of predicting returns within that cluster. A cluster with lower R^2

value suggests that its returns contain less signal to be explained by the predictors, making it more challenging to predict out of the sample.

Specifically, we revisit the calculation of R^2 to justify its usage as a measure of return predictability. First, in the literature (e.g., Fama and French, 2008; Lewellen, 2015; Gu et al., 2020), a predictive model is typically represented as:

$$r_{i,t} = \mathrm{E}_{t-1}[r_{i,t}] + \varepsilon_{i,t}, \tag{1}$$

with the assumption $\mathrm{E}[\varepsilon_{i,t}] = 0$ such that the prediction of expected return is unbiased. The information regarding the heterogeneous predictable difficulty can be represented by the S2N ratio in Equation (2):

$$R^2_{i,t} = 1 - \frac{\mathrm{Var}(\varepsilon_{i,t})}{\mathrm{Var}(r_{i,t})} := 1 - \frac{\sigma^2_{\varepsilon,i,t}}{\sigma^2_{i,t}}, \tag{2}$$

where $\sigma^2_{\varepsilon,i,t}$ and $\sigma^2_{i,t}$ denote the variances of $\varepsilon_{i,t}$ and $r_{i,t}$, respectively. Subscripts i and t of this metric emphasize its cross-sectional and time-series variability. Conceptually, when $R^2_{i,t}$ is high for asset i at a specific period t, it is relatively easier for a predictive model to capture the conditional expectation. Conversely, if the noise for asset i is large, even with knowledge of the true $\mathrm{E}_{t-1}[r_{i,t}]$, the resulting R^2 would still be low, let alone when learning the conditional expectation from noisy data. Thus, with appropriate model regularization, the in-sample $R^2_{i,t}$ provides a reasonable measure of the S2N ratio, reflecting the return predictability of various assets across different periods.

Second, estimating $R^2_{i,t}$ for each asset i at time t is challenging due to the lack of data. The literature often addresses this issue by fitting a pooled model and calculating a single R^2 for all assets across all periods (Gu et al., 2020). However, this approach combines all data under one homogeneous predictive model, neglecting the potential heterogeneity of return predictability on the panel. Our approach balances these extremes by dividing asset returns into a few subsets and calculating the R^2 for each group to measure cluster-wise predictability. This goal-oriented clustering method is specifically designed to maximize differences in predictability across clusters. Details

of the splitting process are discussed in the following subsections.

Third, readers may ask why we do not use OOS R^2 to guide our clustering approach. The primary reason is that our clustering method relies on this measure to determine the clustering structure of assets. Incorporating OOS data would introduce future information into the clustering process and result in data-snooping bias. Instead, we treat the clustering approach and the cluster-wise predictive model as a whole, both determined using in-sample data. The performance of this system is then independently evaluated on OOS data. Additionally, the expected OOS mean squared error has the well-known bias-variance decomposition:

$$\mathrm{E}_t\left[\left(r_{i,t+1}-\mathrm{E}_t\left[r_{i,t+1}\right]\right)^2\right]=\left(\mathrm{Bias}\{\widehat{\mathrm{E}}_t\left[r_{i,t+1}\right]\}\right)^2+\sigma^2_{\varepsilon,i,t+1}+\mathrm{Var}\{\widehat{\mathrm{E}}_t\left[r_{i,t+1}\right]\}. \tag{3}$$

When predicting stock returns, $\sigma^2_{\varepsilon,i,t+1}$ dominates due to the low S2N ratio. Robust prediction benchmarks, such as zero for individual stocks or the historical average for the market index, may not reduce predictive bias but typically result in almost zero predictive variance. Furthermore, as documented in the literature, the OOS R^2 generally remains below 1%, making it hard to produce robust clustering results based on OOS R^2 in environments with low S2N ratios.

Finally, once our clustering system is established, we evaluate its out-of-sample (OOS) performance by examining the detected clustering patterns and the fitted cluster-wise predictive models. This standard OOS evaluation mitigates concerns about data snooping and further validates the effectiveness of our clustering algorithm.

2.2 Cluster-wise Predictive Modeling

Clustering and predictive modeling are two fundamental tasks in machine learning, typically addressed by different models. However, our approach integrates both tasks within a unified framework, decision trees are customized for clustering, and ML models (e.g., Ridge regression) are used for prediction.

In this section, we assume the clusters are predetermined and focus on introducing the cluster-wise analysis. The subsequent section details our approach for detect-

ing clusters using cluster-wise predictive models.

We denote the data as $\mathcal{D} = (r_{i,t}, \mathbf{z}_{i,t-1}, \mathbf{x}_{t-1}) \mid i = 1, \ldots, N; t = 1, \ldots, T_i$, where $r_{i,t}$ represents the excess return of stock i at time t. Predictors commonly used in the stock return prediction literature include $\mathbf{z}_{i,t-1}$, a C-dimensional vector of firm characteristics, and $\mathbf{x}_{t-1}$, an M-dimensional vector of aggregate predictors.

Ideally, heterogeneous and time-varying expected excess returns, $\mathrm{E}_{t-1}[r_{i,t}]$, would be modeled as $g_{i,t}(\mathbf{z}_{i,t-1}, \mathbf{x}_{t-1})$, a function that varies across assets and time periods. However, due to the limited observations available for individual stock returns, estimating each $g_{i,t}(\cdot)$ separately is infeasible. As a result, many studies (e.g., Gu et al., 2020) adopt a homogeneous predictive function, $g_t(\cdot)$, and update time-varying coefficients using a rolling-window scheme. However, Feng and He (2022) and Evgeniou et al. (2023) emphasize that such homogeneous modeling overlooks the heterogeneity in predictive power across different assets. Moreover, this approach implicitly assumes homogeneous return predictability—i.e., the same R^2 for every asset—which is inconsistent with empirical findings (Hou et al., 2020; Avramov et al., 2023). Additionally, while rolling-window estimation provides robustness, it fails to capture macroeconomic-driven regime shifts in the stock market.

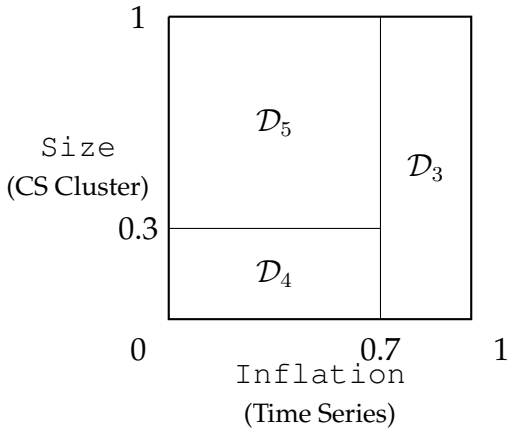
We propose a cluster-wise predictive modeling approach that bridges the gap between individual and pooled models by fitting a cluster-wise predictive model. Unlike the two-step clustering and estimation approach in Evgeniou et al. (2023), we adopt the panel tree framework introduced by Cong et al. (2023), customizing it to cluster assets based on return predictability. This clustering method partitions the stock return panel into multiple clusters (leaf nodes) through characteristics for cross-sectional and aggregate predictors for time-series dimensions. Figure 1 illustrates an example output of our approach, where the panel of stock-return are split to three non-overlapping clusters, described by characteristics or aggregate predictors.[5]

Rather than fitting individual predictive models, $g_{i,t}(\cdot)$, for each asset i in period

[5]A company's cluster membership may evolve over time as its characteristic values change. For example, if a company transitions from a small-cap to a large-cap firm, as reflected by market equity values, its cluster assignment may shift based on the partitioning outcome: $\mathcal{D}_3$: high inflation; $\mathcal{D}_4$: non-high inflation and small-cap; $\mathcal{D}_5$: non-high inflation and non-small cap.

Figure 1: **Clustering Illustration via Partitions**

This figure separates the whole panel of stock returns into three rectangular $\mathcal{D}_3$, $\mathcal{D}_4$, and $\mathcal{D}_5$. The first partition is inflation at 0.7, and the second partition is size at 0.3 when inflation is not high.



t, or a single pooled model, we estimate a small number of cluster-wise homogeneous models that vary across clusters. As a result, the cluster-wise predictive model is:

$$E_{t-1}[r_{i,t}] = g_j(\mathbf{z}_{i,t-1}, \mathbf{x}_{t-1}), \tag{4}$$

where stock-return observations in the j-th cluster follow the same predictive model $g_j(\cdot)$. Our approach simultaneously clusters observations and estimates local predictive models, grouping stock-return observations with similar return predictability into the same cluster. This contrasts with the two-step approach of Evgeniou et al. (2023), which separates based on firm IDs $\{r_i \in j\}$.

Remarkably, our clustering approach allows the user's specific choice of predictive model $g_j(\cdot)$. For simplicity, we illustrate our approach using Ridge regression, which is suggested to be robust under weak signal scenarios by Shen and Xiu (2024). Consequently, the in-sample R_j^2 is calculated with stock returns falling in the specific j-th cluster. Next, we illustrate our tree-based clustering approach step-by-step.

2.3 Clustering: First Split

The objective of our clustering approach is to group asset returns with similar predictability into the same cluster. As discussed in the previous subsection, predictabil-

ity is measured using the in-sample R^2 of a cluster-wise predictive model, such as Ridge regression. Traditional clustering methods in the machine learning literature, such as K-means, are not well-suited for this task. These methods typically minimize within-cluster distances and maximize cross-cluster distances based on characteristics, but they do not account for return predictability. Moreover, their resulting clusters often lack clear interpretability.

We propose a novel model-based clustering approach that partitions the panel of stock returns based on unobservable predictability, distinguishing between highly and less predictable observations. This method optimizes the separation of samples into two subgroups by maximizing the difference in their S2N ratios or R^2 values[6].

Our approach is implemented iteratively to partition the entire panel of stock-return observations, adding one cluster at a time. The results are visualized using a decision tree structure. Unlike the Random Forest algorithm, which constructs multiple trees for prediction only, our method uses lagged (observable) information to iteratively partition the sample, resulting in a deterministic clustering output.

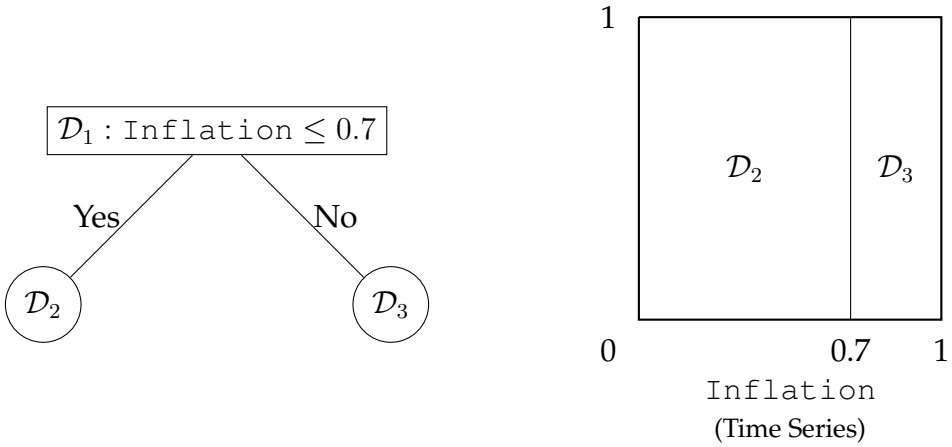
The initial step tries to find the first split point (split predictor and cut-point value) to divide the data into two sets. This process includes fitting cluster-wise predictive models and calculating the absolute R^2 difference for each split candidate, ultimately selecting the one that maximizes the R^2 difference.

Figure 2 illustrates a *candidate* for the first split in a decision tree, where the root node, $\mathcal{D}_1$, containing the entire dataset, is divided into two clusters, or *leaf nodes*—$\mathcal{D}_2$ and $\mathcal{D}_3$. Following terminologies in Cong et al. (2024), leaf nodes refer to the nodes at the bottom of decision trees without subsequent branches. The split is based on the rule "$\text{var}_p(\texttt{Inflation}) \leq c_k(0.7)$," where the p-th variable (firm characteristic, aggregate predictor, or calendar month) is inflation, and the k-th cut-point value is 0.7. To evaluate the quality of this split candidate, we define a goal-oriented split criterion that measures its effectiveness in distinguishing returns with high predictability from those with low predictability.

[6]Alternative clustering objectives could achieve similar goals, but maximizing the R^2 difference is both interpretable and practical.

Figure 2: **Illustration for the First Split**

This figure illustrates one of the first split candidates, such as $\texttt{Inflation} \leq 0.7$. The left figure shows a decision tree that divides the sample, and the right figure shows the corresponding partition plot that only partitions over time.



The first split candidate divides the entire return sample $\mathcal{D}_1$ into two clusters, $\mathcal{D}_2$ and $\mathcal{D}_3$. We fit two cluster-wise predictive models, denoted as $\widehat{g}_2(\cdot)$ and $\widehat{g}_3(\cdot)$, respectively. Generally, for the j-th leaf node, we fit a specific predictive model $g_j(\cdot)$ in Equation (5), and the return predictions will be denoted as $\widehat{r}_{i,t} = \widehat{g}_j(\mathbf{z}_{i,t-1}, \mathbf{x}_{t-1})$, where $\mathbf{z}_{i,t-1}$ and $\mathbf{x}_{t-1}$ are lagged stock characteristics and aggregate macro predictors, respectively.

Notably, Fama and French (2008) have criticized that small-cap stocks with high return variance largely dominate the panel regressions, and Hou et al. (2020) show that many anomalies are replicable due to the dominance of micro-caps (about 60% of all firms) in the cross-sectional regressions. Therefore, we consider using the volatility-weighted Ridge regression for the cluster-wise predictive regressions and further split criterion calculations.

$$
\begin{aligned}
g_j(\cdot) &= \beta_0 + \beta^\intercal \mathbf{s}_{i,t-1} + \varepsilon_{i,t}, \\
\widehat{\beta}_j &= \arg\min_{\beta_0,\beta} \left\{ \frac{1}{N_{\text{leaf}_j}} \sum_{\text{leaf}_j} w_{i,t-1} \left(r_{i,t} - \beta_0 - \beta^\intercal \mathbf{s}_{i,t-1}\right)^2 + \lambda ||\beta||_2^2 \right\}, \\
w_{i,t-1} &= 1/\sigma_{i,t-1}^2
\end{aligned}
\tag{5}
$$

where $\mathbf{s}_{i,t-1} = \{\mathbf{z}_{i,t-1}, \mathbf{x}_{t-1}\}$ include lagged firm characteristics and aggregate predic-

tors, and $w_{i,t-1}$ is the inverse of idiosyncratic return variance. The volatility, $\sigma^2_{i,t-1}$, is estimated on a rolling-window basis, which helps to incorporate both cross-sectional and time-series variation for observation weights within the leaf cluster. The hyper-parameter λ is determined by cross-validation.

Therefore, $\widehat{r}_{i,t} = \widehat{\beta}_{j,0} + \widehat{\beta}_j^{\intercal} \mathbf{s}_{i,t-1}$ is the heterogeneous return forecast for calculating the corresponding S2N ratios, R^2. Within the j-th leaf node:

$$R_j^2 = 1 - \frac{\sum_{\{i,t\} \in \text{leaf}_j} (r_{i,t} - \widehat{r}_{i,t})^2}{\sum_{\{i,t\} \in \text{leaf}_j} r_{i,t}^2}. \tag{6}$$

Since our goal is to separate returns with high predictability from those with lower predictability, it is natural to use the absolute value of the R^2 difference between the left and right clusters as the split criterion:

$$S_{\{\text{leaf}_l, \text{leaf}_r\}} (\text{var}_p, c_k) = \left| R^2_{\text{leaf}_l} - R^2_{\text{leaf}_r} \right|. \tag{7}$$

Intuitively, this criterion evaluates how effectively each split candidate differentiates the R^2 values between the two leaf nodes, regardless of which one is higher. A split candidate achieving a high value for this criterion indicates its success in partitioning stock returns into a highly predictable leaf node and a less predictable one.

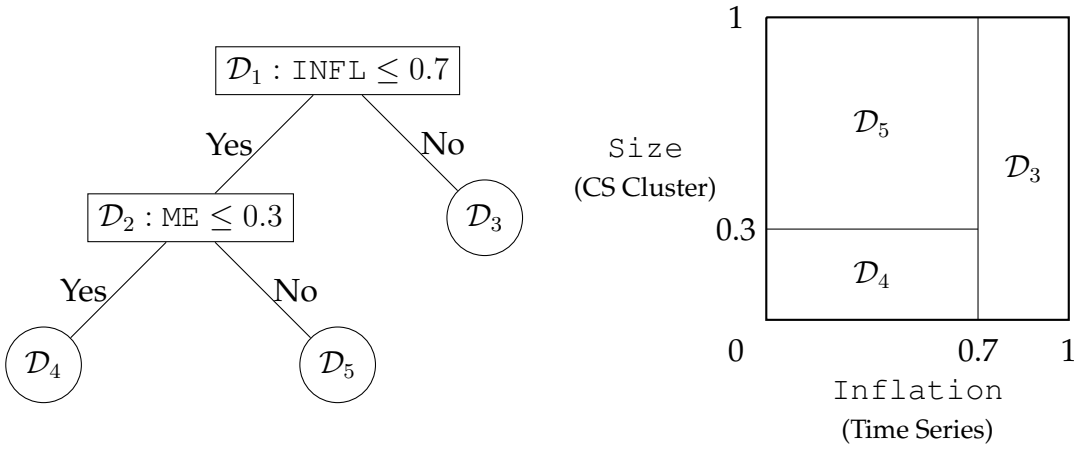
We then evaluate the criterion in Equation (7) across all potential split candidates, considering the full combination of split variables and cut-point values. Different pairs of split candidates, var_p, c_k, result in various partitions of the data, producing non-overlapping sub-samples as leaf nodes, $\mathcal{D}_2$ and $\mathcal{D}_3$, which correspond to cluster-wise predictive models, $\widehat{g}_2(\cdot)$ and $\widehat{g}_3(\cdot)$. These partitions ultimately yield different values for the split criterion in Equation (7), with a successful split candidate maximizing this criterion. Consequently, P candidate variables and K potential cut-point values provide a total of $P \times K$ possible split combinations for the first split, and we evaluate each of them to pick the optimal one as the first split. Then the root node will be divided into two child nodes (clusters).

2.4 Clustering: Subsequent Splits and Stopping Criteria

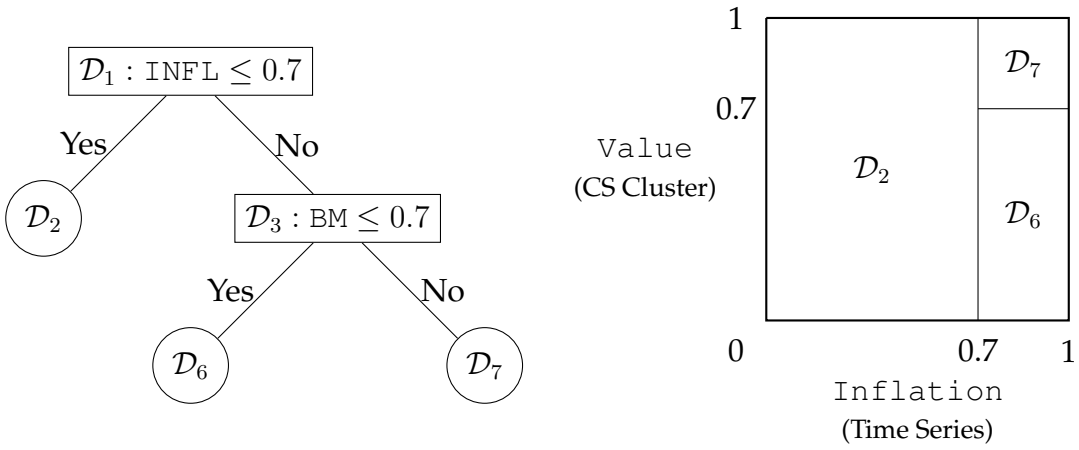
We essentially take a sequential approach to partition the panel into multiple clusters. Once the first split is determined, it creates two leaf nodes, and subsequent splits can occur at either of these nodes to further separate the clusters. Figure 3 illustrates two potential candidates for the second split, which may occur on the left (non-high inflation) or right (high inflation) leaf, depending on different characteristics, highlighting the asymmetric interaction of split predictors.

Figure 3: **Illustration for the Second Split**

This figure illustrates two example candidates for the second split, which can happen on the left or right child node, demonstrating the asymmetric interaction of split predictors.



(a) If splitting node $\mathcal{D}_2$ at ME



(b) If splitting node $\mathcal{D}_3$ at BM

Specifically, the second split can occur within the left leaf node $\mathcal{D}_2$, partitioning it into $\mathcal{D}_4$ and $\mathcal{D}_5$, or within the right leaf node $\mathcal{D}_3$, dividing it into $\mathcal{D}_6$ and $\mathcal{D}_7$. Each side still evaluates $P \times K$ split candidates, resulting in a total of $2 \times P \times K$ combinations. When evaluating split candidates for $\mathcal{D}_2$, cluster-wise predictive models $\widehat{g}_4(\cdot)$ and $\widehat{g}_5(\cdot)$ are fitted using the observations in $\mathcal{D}_4$ and $\mathcal{D}_5$, respectively, and the split criterion values are calculated. A similar procedure is applied for candidate splits within $\mathcal{D}_3$. Among all split candidates across $\mathcal{D}_2$ and $\mathcal{D}_3$, the one that maximizes the split criterion is selected as the second split.

This procedure adopts a local-global approach: it evaluates the benefits of splitting each leaf node based on local R^2 differences between sub-samples and globally selects the split that best differentiates return predictability.

All subsequent splits are determined in the same manner. Each time, we examine all existing leaf nodes, search for all possible split candidates, and choose the one with the maximum value as the best partitioning of the specific leaf node. Without prior knowledge of the "correct" clustering pattern, this self-supervised clustering approach partitions stock-return observations into multiple clusters, maximizing the split criterion and predictability heterogeneity between clusters, and then fitting post-cluster heterogeneous predictive models.

Stopping Criteria. Stopping criteria are essential for regularizing in-sample model training, preventing overfitting. The clustering process stops when predetermined conditions are satisfied. We impose a minimum sample size requirement for each leaf cluster, eliminating split candidates whose resulting sub-samples fail to meet this threshold. This ensures the cluster-wise predictive model in each cluster can be fitted with sufficient observations. Additionally, we limit the tree structure's maximum depth and the number of terminal leaves to control its complexity. Finally, a node is no longer split if all split candidates fail to improve predictability -—– specifically, when the R^2 values of both child nodes are smaller than that of their parent node.

Cluster-wise Predictions. Once tree growth terminates, the entire panel of observations is partitioned into multiple non-overlapping clusters based on variable interac-

tions, such as firm characteristics, aggregate predictors, or calendar months. Within each cluster, we refit cluster-specific predictive models using the corresponding observations. As outlined in Section 2.2, we employ Ridge regression for return predictions. The final output of our approach is a decision tree structure that delineates the clustering pattern, with each cluster linked to its own Ridge prediction model. We refer to this as a heterogeneous prediction model, in contrast to a global prediction model that fits a single Ridge regression across all stock-return observations.

Notably, while we demonstrate our clustering approach using Ridge regression in Equation (5) for simplicity, the framework is not restricted to any specific predictive model. Instead, it provides a flexible and generalizable framework for heterogeneous predictions that can incorporate a wide range of ML models.

3 Empirical Data and Evaluations

We apply our approach to U.S. individual stock returns to study their heterogeneous predictability.[7] The monthly sample spans from 1973[8] to 2022, with the first 30 years used for model training and the most recent 20 years for the OOS analyses. The average and median number of stocks in the training sample are 4,840 and 4,772, respectively, while in the test sample, these numbers are 3,911 and 3,696.[9]

Aggregate Predictors. We analyze eight aggregate predictors to define and select macroeconomic regimes characterized by time-varying return predictability. As detailed in Table A.6, these predictors include the 3-month Treasury bill rate, inflation, term spread, default yield, and market-level characteristics such as dividend yield, volatility, net equity issues, and liquidity.

To ensure comparability of these predictors over time, we standardize them to the

[7] We use standard filters (see, e.g., Fama and French, 1992), including: (1) restricting the sample to stocks listed on NYSE, AMEX, or NASDAQ for more than one year; (2) selecting observations for firms with CRSP share codes of 10 and 11; and (3) excluding stocks with negative book equity or negative lagged market equity.

[8] We begin in 1973 as CRSP expanded its data coverage in 1987 to include NASDAQ daily and monthly stock data, with information on domestic common stocks and ADRs traded on the NASDAQ Stock Market starting December 14, 1972.

[9] Note that our algorithm allows the panel data to be unbalanced.

[0, 1] range using their empirical percentile values within a rolling 10-year window.[10] This standardization enables consistent evaluation of each predictor on a unified scale while avoiding look-ahead biases. Consistent with the approach for firm-level characteristics, we select two cut-point values, 0.3 and 0.7, for each aggregate predictor, employing a "top-middle-bottom" sorting method.

Characteristics. As detailed in Table A.7, our dataset comprises 58 firm-level characteristics categorized into eight major groups: size, value, investment, momentum, profitability, liquidity, volatility, and intangibles. Each characteristic is standardized cross-sectionally and scaled uniformly to the range [0, 1] for every month. To mimic the "top-middle-bottom" sorting approach, we define two cut-point values, 0.3 and 0.7, as split-value candidates for each characteristic. These characteristics are utilized for forecasts to construct tree-based clustering and cluster-wise predictive models.

Model Fitting Design. The baseline analyses for cross-sectional partitions use the first 30 years of data for both tree-based clustering and model estimation, with the most recent 20 years serving as testing samples. To enhance the predictive performance, we update the tree-based clustering and cluster-specific predictive models every five years, employing a 30-year rolling window of in-sample data to retrain the decision tree structure. This process is repeated four times over the 20-year OOS data.

For extended analyses involving time-series splits, we conduct a full-sample analysis to account for the long and overlapping nature of business cycles. Additionally, we use cross-validation to optimize the hyperparameters during the post-cluster predictive model training phase.

Performance Evaluation. As mentioned in Section 2.1, we use in-sample R^2 to measure predictability. In addition to the in-sample R^2, we evaluate the OOS R^2 to check whether the predictability gap is consistent. This approach aligns with the standard practice in recent studies. Following Gu et al. (2020), we define the OOS R^2 (R^2_{OOS})

[10] For instance, an inflation value exceeding 0.7 indicates that the current inflation level is higher than 70% of observations in the past ten years.

using zero forecasts as the benchmark,

$$R^2_{OOS,j} = 1 - \frac{\sum_{\{i,t\}\in \text{leaf}_j} (r_{i,t+1} - \widehat{r}_{i,t+1})^2}{\sum_{\{i,t\}\in \text{leaf}_j} r^2_{i,t+1}}, \tag{8}$$

where subscript j represents the specific OOS predictions by the related in-sample cluster-wise predictive model of the j-th cluster.

4 Cross-Sectional Heterogeneous Return Predictability

The baseline model investigates heterogeneous return predictability in the cross section, addressing two key questions: Does heterogeneity exist, and which stock returns are more predictable? This section also demonstrates that our goal-oriented clustering approach effectively can capture this heterogeneity. Unlike methods that assign stocks to fixed clusters across all periods (e.g., Evgeniou et al., 2023), we allow stocks to move between clusters over time. Since clusters are defined by values of characteristics, these transitions occur naturally as stock-specific characteristics evolve.

We refer to the clustering model in this section, which considers only cross-sectional splits, as "CS clusters" to distinguish it from the "TS+CS clusters" model discussed in the Section 5, which incorporates both time-series regime switching and cross-sectional heterogeneity.

4.1 Cluster and Heterogeneous Predictability

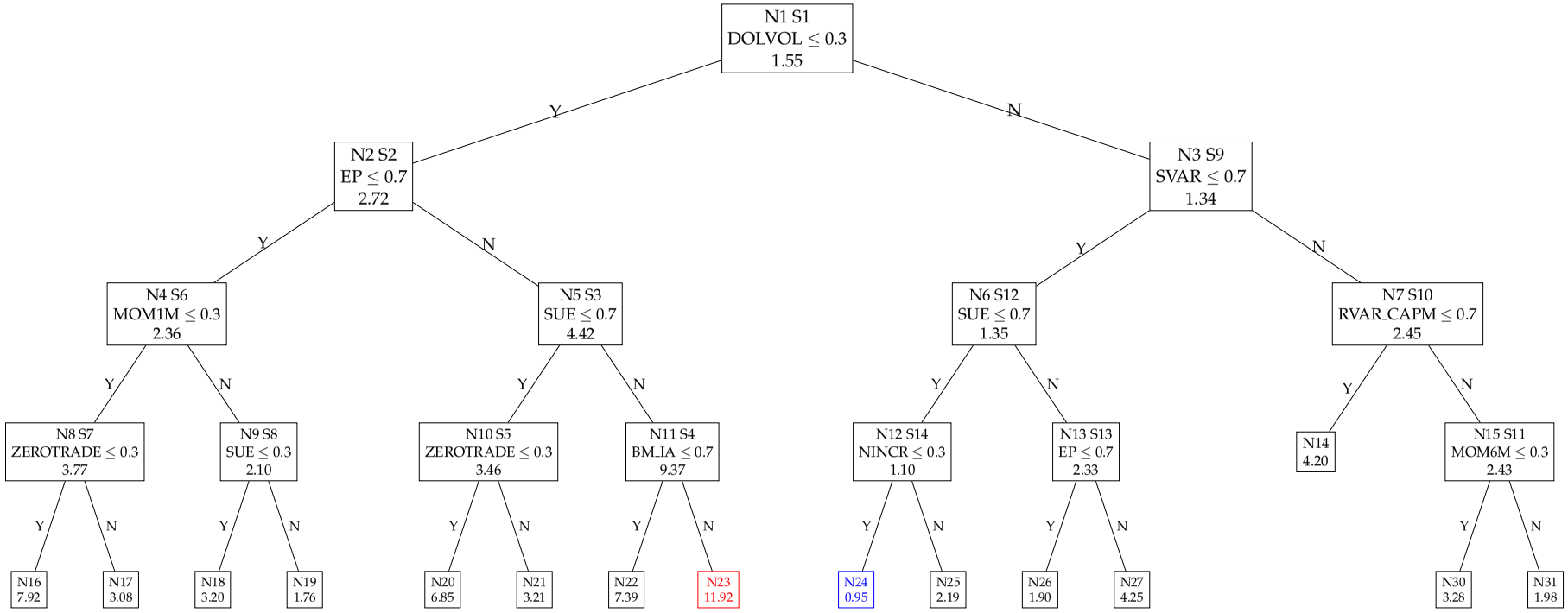
Clustering Pattern. We present our self-supervised cross-sectional clustering results in Figure 4, derived from training the model on data from the first 30-year period (1973–2002). This tree stops growing after reaching 15 final leaves, following our predetermined stopping criteria.[11]

Each leaf node in Figure 4 contains two or three rows of information. The first row identifies the leaf number and split order; for example, the root node represents the initial split and is labeled as leaf node N1 and split S1. The second row displays

[11] A moderately deep decision tree with large leaves (sample size) is both robust and easy to interpret. To achieve this, we set the maximum tree depth to 5 (allowing up to 16 leaves) and specify a minimum leaf size of 10,000 stock-return observations to ensure robust model training. Under these parameters, the algorithm automatically halts after 14 splits.

Figure 4: **Tree-Based Cluster (CS Cluster)**

This figure illustrates the cross-sectional tree-based clustering structure derived from monthly data (1973–2002). The tree partitions stock returns based on the monthly standardization of firm characteristic ranks within the [0,1] range. Terminal leaves represent clusters formed by specific firm characteristic ranges. Each node, including intermediate and terminal leaves, is labeled with an ID ($N\#$) and split order ($S\#$). Cluster-wise R^2 values, indicating return predictability, are shown for each node.



the optimal split rule determined by the algorithm, directing observations meeting the split condition to the left child node, while those not meeting the condition proceed to the right child node. Terminal leaves, which are not subject to further splits, are represented solely by their index number in the first row. The bottom row presents the signal-to-noise (S2N) ratio, R^2, reflecting the cluster-wise return predictability for each node. This design provides an intuitive visualization of the clustering processes and the interactions among different characteristics.

Before any split, the aggregate return predictive ability (R^2) of the homogeneous model is 1.55% (root node). After the first split, the R^2 improves significantly for stock returns with low dollar trading volume ($\texttt{DOLVOL} \leq 0.3$), increasing to 2.72% at N2, while slightly declining for the complement set, reducing to 1.34% at N3. The R^2 difference (2.72 - 1.34)% represents the maximum value identified among all split candidates (58×2) by the algorithm. Stocks with low trading volumes are typically associated with low liquidity and are usually small-cap or distressed stocks, consistent with the findings of Avramov et al. (2023).

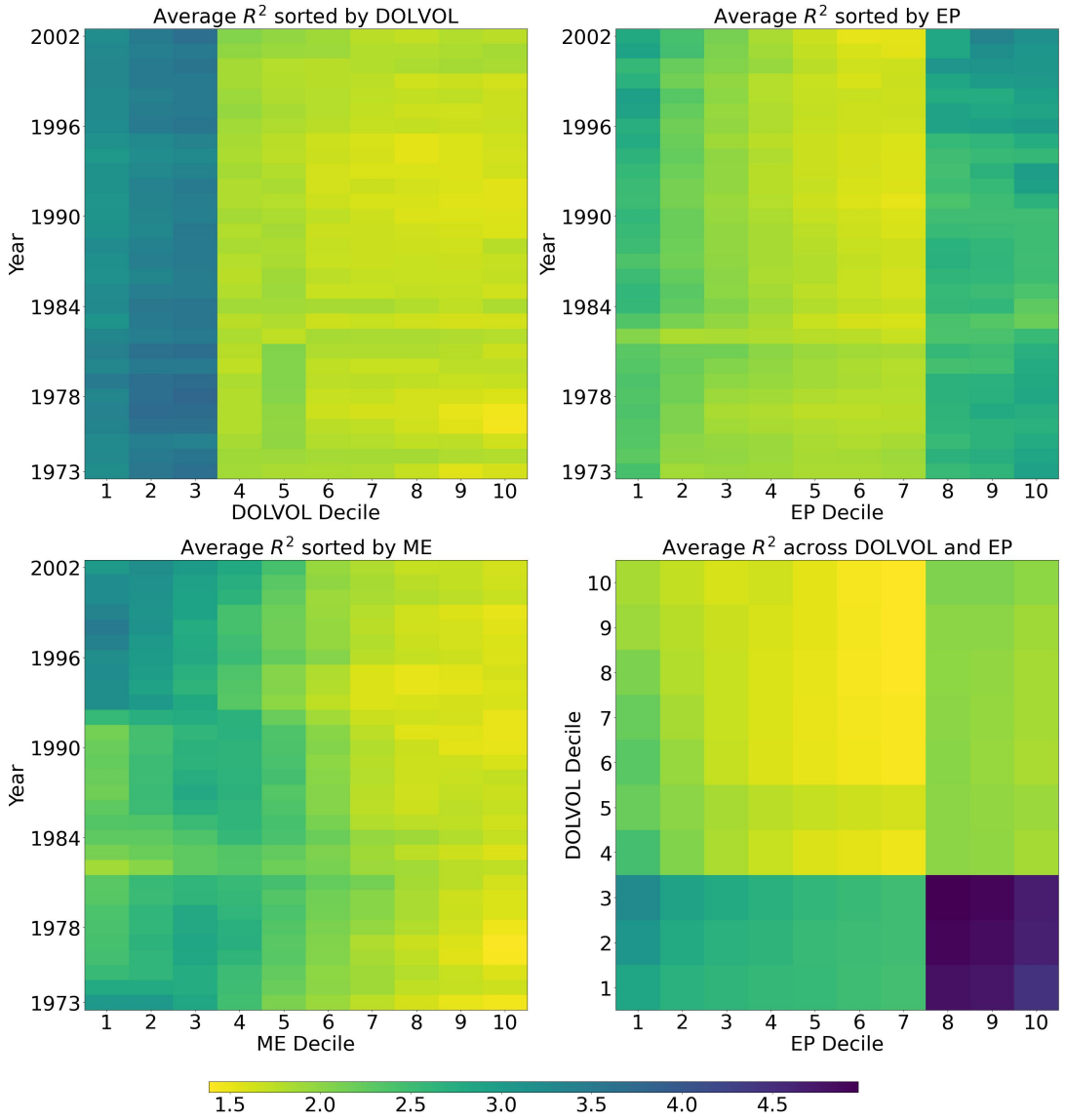
Additionally, Figure 5 highlights marginal information extracted from the decision tree shown in Figure 4. It illustrates the average decile R^2 values for each year and/or decile cluster sorted by various characteristics. A clear pattern emerges, indicating a decline in return predictability for stocks with higher dollar trading volumes, a trend that has persisted over time. The bottom-left sub-figure further corroborates the superior forecast accuracy for small-cap stocks.

The second split selects high earnings-to-price stocks ($\texttt{EP} > 0.7$), resulting in a higher R^2 value of 4.42% at N5 under the low dollar trading volume condition. Value stocks with low trading volumes exhibit greater return predictability, while non-value stocks with small trading volumes show lower forecasting accuracy (2.36% at N4).

In addition, as shown in Figure 5, the EP-sorted decile clusters in the middle exhibit the lowest predictability compared to those at both extremes, with higher value stocks consistently achieving greater accuracy. The bottom-right sub-figure, sorted by DOLVOL and EP, highlights an interaction pattern where stocks with low DOLVOL and

Figure 5: **Mosaics of Predictability by Predictors (CS Cluster)**

These heat maps summarize average return predictability (R^2 values, % in the color bar) for stock returns based on the tree-based clustering in Figure 4. The first three heat maps show average R^2 values for groups sorted by years and deciles of dollar trading volume, earnings-to-price, and market equity value. The fourth heat map displays average R^2 values for 10×10 groups formed by bivariate decile sorting of the top two characteristics.

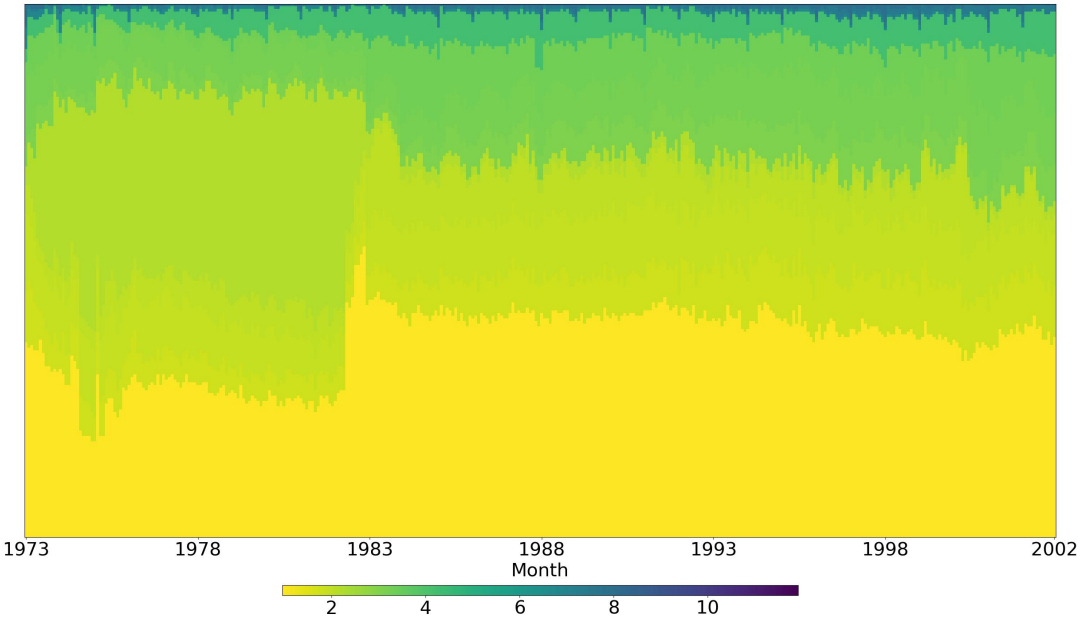


high EP exhibit the highest return predictability, indicated by the darkest shading. These mosaic-like patterns emphasize the differences in return predictability arising from the interactions among various characteristics.

Figure 6 presents the return predictability derived from tree-based clustering in a visually distinct, mosaic-like format. The sample period spans 360 months, corresponding to the tree structure shown in Figure 4, with approximately 4,000 to 5,500

Figure 6: **Mosaics of Predictability by Clusters (CS Cluster)**

This heat map summarizes monthly return predictability (R^2 values, % in the color bar) for stock returns based on the tree-based clustering in Figure 4. The vertical axis shows the proportion of observations per cluster, while the horizontal axis represents months. Colors range from light to dark, indicating increasing levels of return predictability.



observations per month. Notably, since clusters are defined based on characteristic values, the number of stocks within each cluster is not fixed over time. For each month, clusters are arranged in ascending order of predictability, with the height of each segment reflecting the proportion of stocks within that cluster for that month.[12] As R^2 values increase (darker colors), the proportion of observations decreases, indicating that the majority of stock returns are hard to predict. By setting the lower bound of the color bar to 0, it becomes apparent that more than half of the observations shift toward lighter shades of yellow each month. In contrast, only a small fraction of observations exhibit high predictability, represented by the deepest colors, corresponding to R^2 values exceeding 10%. This horizontal, mountain-like cascading pattern, with its variations across cross-sectional, time-series, and color dimensions, provides valuable insights into the heterogeneous nature of stock return predictability.

[12]The denominators for the monthly proportions are based on the total number of observations in each month. The gap around 1982 may be attributed to a surge in IPOs during that period, potentially linked to the launch of the NASDAQ National Market in April 1982, which included larger and more actively traded NASDAQ securities.

Cluster-Wise Performance. Panel A of Table 1 provides summary statistics for each cluster. Ordered by descending predictability, the table reveals substantial heterogeneity across clusters, with R^2 values ranging from 11.92% to 0.95%, reflecting a difference exceeding 10%. These variations in R^2 demonstrate that each cluster exhibits a distinct level of predictability, underscoring the diversity in the cross-sectional dimension. Furthermore, the least predictable cluster, N24, accounts for a large portion of stocks, whereas the most predictable cluster, N23, includes only a small subset. This disparity highlights the limited availability of highly predictable stocks.

Table 1: **Cluster-Wise Performance (CS Cluster)**

This table summarizes cluster-based information from the cross-sectional tree structure in Figure 4. Panel A reports the number of observations per cluster (# obs) and return predictability (R^2 values in %) for each cluster. Panel B shows the monthly average return (Avg in %) and annualized Sharpe ratio (SR) for equal/value-weighted (EW/VW) portfolios based on all observations. Leaf nodes are ordered by descending R^2.

	Panel A: Summary Statistics		Panel B: Profitability			
Leaf	# obs	R^2	Avg_{EW}	SR_{EW}	Avg_{VW}	SR_{VW}
N23	11,627	11.92	4.34	2.17	3.54	1.92
N16	14,695	7.92	3.58	1.08	2.93	0.96
N22	16,341	7.39	3.15	1.73	2.50	1.58
N20	12,116	6.85	1.95	0.89	1.55	0.76
N27	79,752	4.25	1.98	1.26	1.24	0.81
N14	14,402	4.20	0.30	0.10	-0.22	-0.07
N30	115,528	3.28	-0.92	-0.31	-1.55	-0.53
N21	97,131	3.21	0.66	0.39	0.65	0.44
N18	49,275	3.20	-2.20	-1.15	-1.76	-1.15
N17	87,763	3.08	1.52	0.62	0.70	0.36
N25	179,412	2.19	0.94	0.57	0.62	0.40
N31	126,050	1.98	-0.14	-0.05	-0.37	-0.14
N26	178,691	1.90	1.30	0.78	0.68	0.46
N19	133,911	1.76	0.24	0.13	0.42	0.27
N24	625,549	0.95	0.36	0.24	0.32	0.24

We further examine the relationship between return predictability and average returns by constructing equal- and value-weighted portfolios of stocks within each cluster over time. Panel B of Table 1 reports the average returns and Sharpe ratios for these portfolios. Notably, the most predictable cluster (N23, with $R^2 = 11.92\%$) delivers an average monthly return of 3.54% (4.34%) and an annualized Sharpe ratio of 1.92 (2.17) for value- (equal-) weighted strategies, reflecting both strong profitability

and an attractive risk-return trade-off. In contrast, the least predictable cluster (N24, with $R^2 = 0.95\%$) exhibits significantly weaker investment performance.

These results highlight a clear relationship between average returns and return predictability, suggesting the potential for identifying anomalies associated with predictability. A more detailed discussion of this anomaly is postponed to Section 4.2.

Out-of-sample Predictability. Thus far, the easy- and hard-to-predict clusters have been identified using in-sample information. This naturally raises a question: do these patterns persist out of the sample? Importantly, the goal is not to engage in a horse race of OOS R^2 metrics against various ML models. Instead, the primary objective is to evaluate the persistence of the heterogeneous predictability pattern out of the sample, which cannot be detected by other ML models. To evaluate OOS performance, we adopt a five-year rolling clustering setup with a two-fold cross-validation strategy for hyperparameter optimization.[13]

Using the tree structure and cluster-wise performance detailed in Table 1, we categorize all leaves into three groups based on return predictability rankings: high, medium, and low. Observations from clusters such as N23, N16, N22, and N20 are aggregated to form the highly predictable group, while N24 represents the least predictable cluster. The remaining leaves are combined to create the medium predictability group.[14] We use OOS R^2 values, as defined in Equation (8), to validate the cross-sectional patterns of stock return predictability.

Table 2 reports the R^2 statistics for global models (Panel A) and cluster-wise predictive models (Panel B). The "Global" model in Panel A fits all data using a single predictive model without any clustering, and is also evaluated on sub-samples of high, medium, and low predictability clusters identified by our approach. The "Aggregate" model in Panel B represents the aggregation results of predictions from all cluster-wise models. For instance, Panel A global Ridge model achieves 1.62% R^2 out of the sample

[13]The in-sample data is divided into two equally continuous periods. The model is trained on one period and validated on the other. The optimal hyperparameters are selected based on the average MSE, and the model is then retrained using all in-sample data. Finally, the resulting coefficients are applied to predict the next five years of OOS values.

[14]Predictability levels are determined by balancing R^2 values with the proportion of observations, prioritizing clusters with large R^2 gaps while acknowledging that most returns show low predictability.

Table 2: **Evaluations of Predictability (CS Cluster)**

This table presents return predictability (R^2 values, %) for different predictive methods, within-sample and out-of-sample results for cross-sectional splits. Panel A shows the global model, fitted on all data but evaluated on the full sample or clusters identified by our approach. Panel B summarizes the cluster-wise model performance, reporting five samples: Global (no clustering), Aggregate (aggregation of cluster-wise predictions), and High, Medium, and Low, based on the ranking of predictability of clusters.

	In-Sample (1973 - 2002)			Out-of-Sample (2003 - 2022)		
Sample	OLS	Lasso	Ridge	OLS	Lasso	Ridge
		Panel A: Global Forecasts				
Global	1.56	0.57	1.00	0.49	0.35	0.47
High	2.96	1.58	2.02	1.94	1.31	1.62
Medium	1.63	0.55	1.01	0.53	0.32	0.45
Low	1.06	0.41	0.77	0.09	0.27	0.32
		Panel B: Cluster-Wise Forecasts				
Aggregate	2.48	1.71	1.60	0.28	0.55	0.62
High	8.12	7.01	6.50	1.64	1.83	2.05
Medium	2.61	1.82	1.66	0.24	0.56	0.63
Low	0.95	0.29	0.43	0.13	0.26	0.34

for highly predictable clusters, while the cluster-wise model achieves 2.05% in Panel B. The aggregate model demonstrates slight improvements in R^2 values, highlighting the benefits of the cluster-wise predictive model.

When comparing clusters with different levels of predictability, highly predictable clusters consistently outperform their less predictable counterparts, regardless of the model used. For instance, when comparing the values in each row for the highly predictable cluster against the low predictability cluster, every metric—whether in-sample or out-of-sample, OLS or Ridge regression, global or cluster-wise model—shows higher R^2 for the more predictable cluster. These results confirm that predictability is a persistent characteristic of stocks, remaining consistent out of the sample. Furthermore, they validate the effectiveness of our tree-based clustering approach in capturing and identifying this heterogeneity.

4.2 Anomaly of Predictability

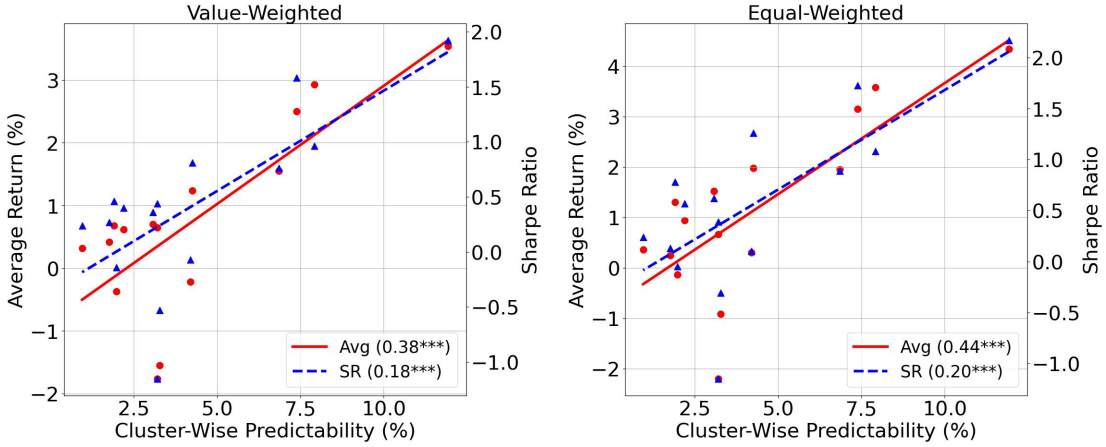
Thus far, we have introduced the novel clustering approach to group stocks based on predictability and demonstrated its variation across the cross section of stocks. It

provides a new measurement of the unobservable return predictability of individual stocks. This naturally raises the question: do these heterogeneous predictability measures capture risks that are orthogonal to established factor models?

To investigate this, we construct value-weighted and equal-weighted long-only portfolios for each cluster and analyze whether their average returns exhibit significant spreads. Panel B of Table 1 highlights the potential return spreads, while Figure 7 visualizes the relationship between cluster predictability and the average returns and Sharpe ratios of these portfolios. The plot reveals a clear, significantly positive relationship: clusters with higher return predictability tend to deliver superior profitability, as reflected in both higher average returns and Sharpe ratios.

Figure 7: **Connections between Predictability and Profitability (CS Cluster)**

This figure shows the relationship between in-sample cluster-wise return predictability (R^2 values, horizontal axis) and the monthly average return (in %, left vertical axis) and annualized Sharpe ratio (right vertical axis) of value/equal-weighted portfolios (see Table 1). The red solid and blue dashed lines represent the fitted scatter plots for the average return (Avg) and Sharpe ratio (SR), respectively.



This result motivates the construction of long-short portfolios to capture anomalies associated with predictability. Our clustering results, organized through a decision tree structure, reveal that each cluster exhibits a distinct level of predictability. This methodology can be interpreted as a generalized sorting approach based on the latent characteristics of heterogeneous predictability.

Specifically, we focus on the value-weighted portfolios depicted in Table 1. At the end of each month, we rank all clusters by their predictability and implement a long-short strategy. This strategy shorts the least predictable cluster portfolio (S1, con-

taining around 35.9% of in-sample observations) and longs the top one to five most predictable cluster portfolios (L1 to L5, which account for 0.7% to 7.7% of in-sample observations). To evaluate the abnormal returns of these anomalies, we run the spanning regressions with the portfolio returns on various widely used factor models.

Table 3: **Testing Anomaly of Predictability**

This table presents summary statistics (Panel A) and abnormal returns (Panel B) for the long-short factor based on cross-sectional return predictability across clusters. "L" and "S" indicate the number of long and short clusters, respectively. Panel A reports the average return (Avg, %), median (Median, %), standard deviation (Std, %), skewness, kurtosis, annualized Sharpe ratio (SR), and maximum drawdown (MDD). Panel B provides abnormal return estimates (alphas, %) and significance (denoted by "*" and t-values in parentheses) from various factor models.

	In-Sample (1973 - 2002)					Out-of-Sample (2003 - 2022)				
	L1-S1	L2-S1	L3-S1	L4-S1	L5-S1	L1-S1	L2-S1	L3-S1	L4-S1	L5-S1
				Panel A: Summary Statistics						
Avg	3.22	2.92	2.67	2.31	2.03	2.78	1.67	1.45	1.32	1.08
Median	2.82	2.23	2.18	1.79	1.67	2.48	1.61	1.45	1.46	1.00
Std	4.94	5.79	4.72	4.58	3.82	4.70	4.36	3.66	3.31	3.30
Skewness	1.09	1.28	0.94	1.04	1.04	1.09	0.24	0.24	0.22	0.45
Kurtosis	3.62	4.16	2.61	3.17	3.06	4.61	1.89	0.52	0.24	1.37
SR	2.26	1.74	1.96	1.75	1.85	2.05	1.33	1.37	1.38	1.13
MDD	9.74	12.84	10.47	10.84	8.66	13.99	13.76	8.68	6.80	8.29
				Panel B: Abnormal Returns						
CAPM	3.26***	2.87***	2.67***	2.31***	2.03***	2.76***	1.71***	1.41***	1.27***	0.92***
	(12.53)	(9.40)	(10.66)	(9.52)	(10.03)	(8.93)	(5.97)	(5.86)	(5.87)	(4.40)
FF3	2.87***	2.53***	2.32***	1.94***	1.68***	2.80***	1.75***	1.45***	1.31***	0.97***
	(13.23)	(11.20)	(13.42)	(12.12)	(12.49)	(9.63)	(6.63)	(7.03)	(7.12)	(5.80)
FF3+IVOL	2.80***	2.69***	2.38***	1.98***	1.69***	2.94***	1.85***	1.52***	1.36***	1.04***
	(12.52)	(11.63)	(13.32)	(11.98)	(12.21)	(10.49)	(7.12)	(7.53)	(7.43)	(6.40)
FF3+MOM	2.94***	2.68***	2.41***	2.04***	1.76***	2.89***	1.80***	1.51***	1.33***	1.01***
	(13.08)	(11.54)	(13.51)	(12.36)	(12.74)	(10.20)	(6.86)	(7.48)	(7.25)	(6.20)
FF5	2.72***	2.56***	2.27***	1.88***	1.61***	2.88***	1.81***	1.52***	1.37***	1.07***
	(12.43)	(11.01)	(12.74)	(11.39)	(11.69)	(9.60)	(6.78)	(7.23)	(7.24)	(6.31)
FF5+MOM+IVOL	2.82***	2.73***	2.39***	2.00***	1.72***	2.98***	1.87***	1.58***	1.40***	1.12***
	(12.57)	(11.52)	(13.10)	(11.91)	(12.24)	(10.48)	(7.10)	(7.73)	(7.42)	(6.84)
Q5	2.70***	2.64***	2.31***	1.93***	1.64***	2.88***	1.74***	1.58***	1.39***	1.09***
	(10.27)	(9.69)	(10.92)	(9.78)	(9.75)	(9.35)	(6.30)	(7.43)	(7.18)	(6.14)
BS6	2.48***	2.60***	2.20***	1.80***	1.49***	2.83***	1.76***	1.51***	1.31***	0.98***
	(10.62)	(10.40)	(11.54)	(10.27)	(10.27)	(9.78)	(6.58)	(7.37)	(6.97)	(5.88)

Panel A of Table 3 summarizes the performance of the anomalies, showing that portfolios formed based on return predictability generate distinct return spreads. Long-short strategies with larger predictability gaps (e.g., "L1-S1") outperform those with smaller gaps (e.g., "L5-S1"), with higher average returns, Sharpe ratios, and profitability across both sample periods. Even the strategy with the largest number of long clus-

ters ("L5-S1") delivers solid performance, with an average monthly return of 1.08% and a Sharpe ratio of 1.13 for the OOS data.

Panel B of Table 3 reports the results of evaluating these anomalies using various factor models, including CAPM, FF3, FF3+IVOL, FF3+MOM, FF5, FF5+MOM+IVOL, Q5, and BS6 (Barillas and Shanken, 2018). All models show statistically significant unexplained alphas, both in-sample and out-of-sample. Notably, the OOS monthly alphas for the strategy with the smallest predictability gap ("L5-S1") are highly significant, ranging from 0.92% to 1.12% with high t-stats. These indicate that the unexplained alphas generated by the strategy remain robust even after adjusting for risk factors explained by traditional asset pricing models.

We further examine whether the abnormal return comes from the long-leg, short-leg of the portfolio, or both. Table A.1 reveals that the long-only portfolios with high predictability consistently contribute to abnormal returns, while the short-leg portfolios show mostly non-significant negative alphas. Thus the most predictable stocks (long-leg) contribute to the abnormal returns.

In conclusion, we identify a novel anomaly arising from the cross-sectional heterogeneity of predictability. Portfolios constructed from sorted predictability clusters exhibit significant spreads in average returns, with the corresponding anomaly generating substantial alpha even after accounting for popular factor models in the literature. These findings highlight predictability as a latent stock characteristic that provides information orthogonal to established factor models. Consequently, our results offer new insights into the role of predictability in stock market dynamics. Moreover, predictability-driven strategies deliver superior risk-adjusted returns, presenting an empirical contrast to the findings of Rapach et al. (2010) and Kelly et al. (2024).

5 Heterogeneous Predictability and Regime Switches

The previous section examined the clustering pattern of predictability in the cross section. A natural follow-up question is whether such heterogeneity varies over time. To address this, the current section extends the analysis to the time-series dimension.

Specifically, it is important to note that the split variables are not necessarily to be characteristics (CS Clusters, in Section 4); they can also include macroeconomic variables (TS+CS Clusters). By identifying splits based on macroeconomic variables, we can explain the observed clustering in stock return predictability across different time horizons (Section 5.1). In addition, for robustness checks, we demonstrate that calendar months can also serve as a valid split variable, thereby helping to identify structural breaks in return predictability (Section 5.2).

Notably, unlike the cross-sectional clusters, the time-series heterogeneity analyzed in this section involves splitting the full sample period from 1973 to 2022. The reason is to detect regimes for the full sample.

5.1 Macro-Driven Regime Change

We begin by using aggregate macro predictors to split the time horizons and identify regimes of predictability. This approach divides the entire sample based on macroeconomic variables, providing clear and intuitive interpretations. Notably, because the regimes are defined by macroeconomic values, a single regime may not correspond to a continuous time horizon.

In particular, we restrict the use of aggregate predictors to time-series data for the first two splits,[15] while utilizing firm characteristics for subsequent cross-sectional partitions. This methodology establishes a systematic framework that learns the clustering of predictability across both time-series and cross-sectional dimensions, enabling the implementation of both timing and stock selection strategies.

Clustering Pattern. For the macro-driven regime case, we fit the model using the full sample of data, resulting in a single tree-based clustering structure for the 50-year sample. Figure 8 illustrates the decision tree structure, where the initial splits are based on macroeconomic variables, followed by splits based on firm characteristics.

[15] We also experimented with relaxing these constraints, allowing both aggregate predictors and firm characteristics to be considered as candidates for all splits. However, our results show that in all cases, the trees consistently select aggregate predictors for the first two layers of partitions. This suggests a clear preference for initial separation based on time series data. To mitigate the potential for excessive regime changes, we restrict the first two splits to the time dimension, reserving subsequent splits for cross-sectional characteristics.

This structure allows the clustering pattern to jointly account for both time-series and cross-sectional heterogeneity. This tree ultimately produces 32 leaves[16] corresponding to different levels of predictive ability.

Specifically, the overall in-sample R^2 for the global model across the entire sample is 1.51% before any partitions. Our approach identifies dividend yields and default yields as the two most important macroeconomic predictors for detecting regimes of predictability, and creates the following three regimes:

1. Regime I: $\texttt{X\_DY} \leq 0.7$, when the dividend yield is not very high.

2. Regime II: $\texttt{X\_DY} > 0.7$ and $\texttt{X\_DFY} \leq 0.3$, when the dividend yield is high, and the default yield is low.

3. Regime III: $\texttt{X\_DY} > 0.7$ and $\texttt{X\_DFY} > 0.3$, when the dividend yield is high, and the default yield is not very low.

Notably, these two predictors are highly related to business cycles (Fama and French, 1988; Keim and Stambaugh, 1986; Fama and French, 1989). Thus, we find that the business cycle affects the return predictability of stock returns.

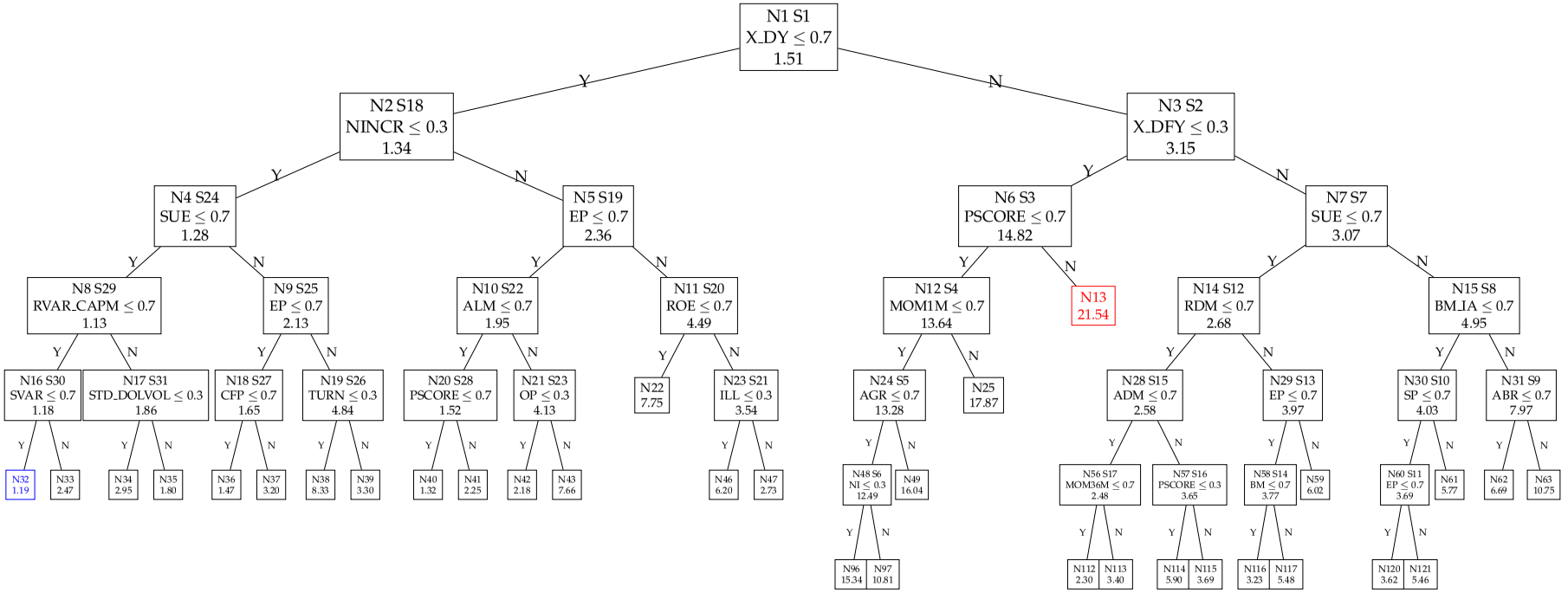
After the first two splits based on macroeconomic variables, this value increases significantly during the period with higher dividend yields and lower default yields (Regime II), reaching 14.82% at node N6. Delving deeper into this subset, the regime with low default yields ($\texttt{X\_DFY} \leq 0.3$) reflects less market uncertainty or corporate default risks, which can lead to greater predictability than the opposite direction (Regime III), with 3.07% at N7. These two periods, conditions typically associated with recessions (consistent with Henkel et al., 2011 and Dangl and Halling, 2012), outperform those when the dividend yield is not-high (Regime I), whose R^2 slightly decreases to 1.34% at node N2.

Furthermore, subsequent partitions based on cross-sectional characteristics further uncover the heterogeneity in stock return predictability within each regime. For

[16] We limit the maximum tree depth within each regime to 5 (resulting in at most 48 leaves) and set the minimum leaf size to 10,000 stock-return observations.

Figure 8: **Tree-Based Cluster (TS+CS Cluster)**

This figure shows the time-series + cross-sectional tree-based clustering structure using monthly data from 1973 to 2022. The tree splits the panel of stock returns based on aggregate predictors by a 10-year rolling percentile standardized to [0,1] for the first two partitions. Then, it splits based on monthly cross-sectional standardization of firm characteristic ranks within [0,1]. The terminal leaves represent distinct clusters identified by the interaction of aggregate predictors and firm characteristics ranges. Each node, including bottom leaves and intermediate nodes, has a unique ID denoted as $N\#$, while the order of the splits is indicated by $S\#$. All nodes are labeled with cluster-wise model R^2 values.



example, leaf node N13 (labeled in red) exhibits the highest R^2 of 21.54%, while the poorest cluster, node N32 (labeled in blue), shows the lowest value at 1.19%. This range is wider than the cross-sectional clustering results presented in Section 4, suggesting that incorporating time-series heterogeneity allows us to identify larger gaps of return predictability.

Figure 9: **Time-Series Regimes (Macro Variables)**

This figure shows time-series regime switches for aggregate predictors from 1973 to 2022 (see Figure 8). Three regimes are represented by different colors: non-high dividend yield (X_DY ≤ 0.7, orange, 426 months), high dividend yield with low default yield (X_DY > 0.7 & X_DFY ≤ 0.3, purple, 16 months), and high dividend yield with non-low default yield (X_DY > 0.7 & X_DFY > 0.3, green, 158 months), reflecting varying predictability based on their position in the color bar. Shaded areas denote NBER recessions, and labeled texts indicate major global events. The vertical axis shows the cross-sectional predictability of clusters, ranging from highly (black) to less predictable (white) under each regime.

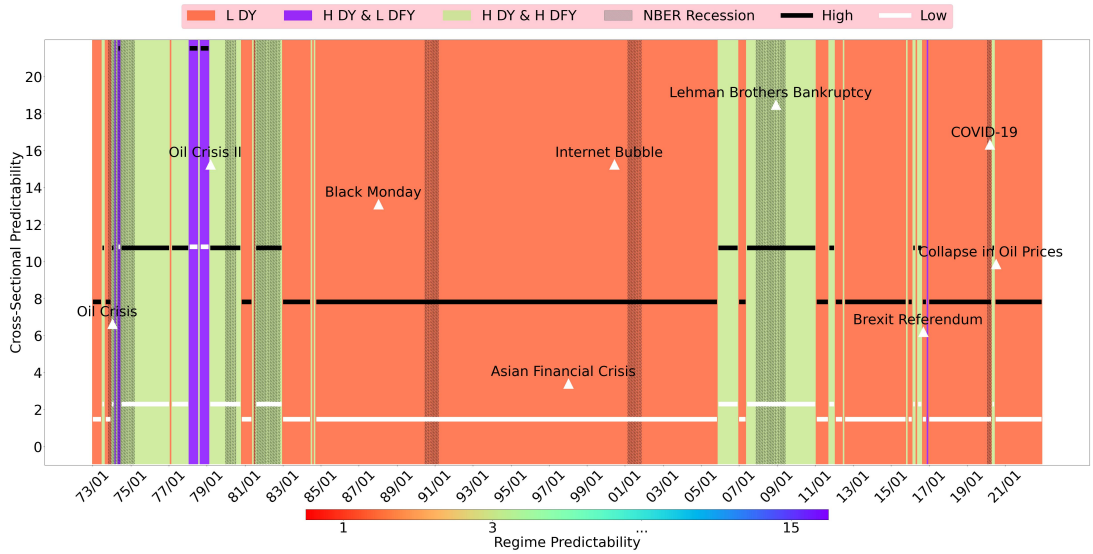
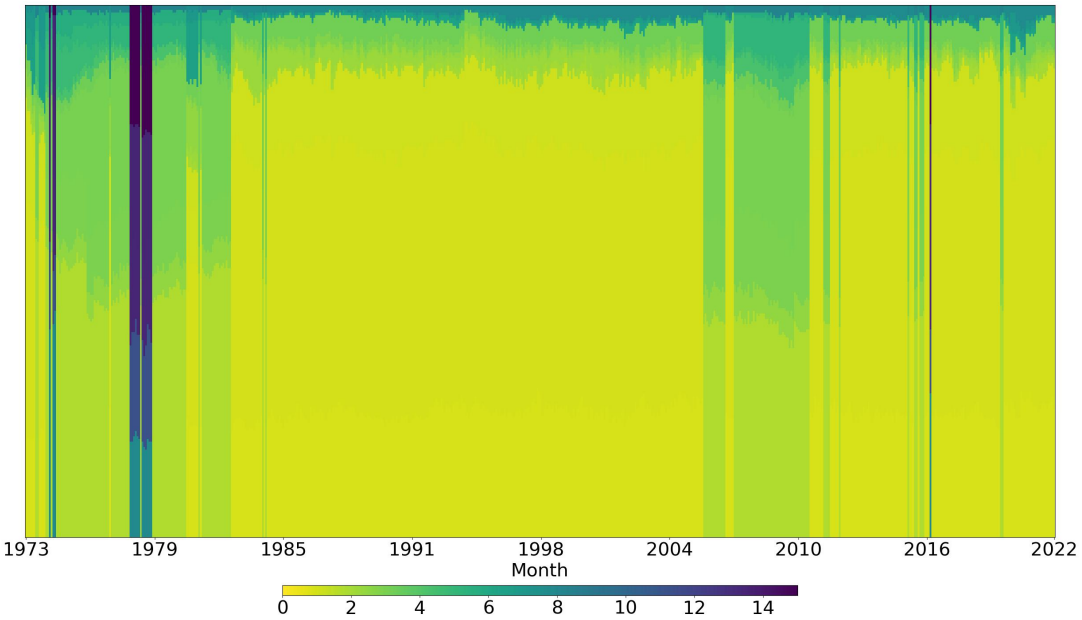


Figure 9 illustrates regime switches from another perspective. The color changes reflect regime shifts detected by the decision tree in Figure 8, with the colors indicating different levels of heterogeneity. The range is broader than in the cross-sectional case (from around 1% in orange to nearly 15% in purple). Cross-sectional splits by firm characteristics further widen the intervals across sub-samples. Highly predictable clusters (black lines) consistently outperform less predictable ones (white lines) and are more robust before additional clustering. Regime transitions are linked to recessions (e.g., 1974, 1981, 1983, 2020) and global events like the Oil Crisis (1973, 1979), Brexit (2016), and the 2020 Oil Price Collapse. Other events, often occurring during

less predictable periods, include the 1987 Black Monday, the 1997-1998 Asian Financial Crisis, the 2000 Internet Bubble, and the 2014 Euro Negative Interest Rate.

Figure 10: **Mosaics of Predictability by Clusters (TS+CS Cluster)**

This heat map summarizes the predictability, R^2 values (% in the color bar), for the panel of stock returns based on the tree-based clustering in Figure 8. The horizontal axis represents months, and colors from light to dark indicate ascending levels of return predictability of each cluster within each month. Vertically, the length of each color bin denotes the proportion of observations for each cluster.



Compared to the CS clusters, the mosaics of stock return predictability for time series in Figure 10 appear more diverse. The horizontal axis represents all 600 months. Sorting the cluster-wise model R^2 values in ascending order from bottom to top and aggregating monthly reveals layered heterogeneity with three hierarchical types. Most clusters show lower predictability in light green, while specific regimes (e.g., 1979 and 2017) show higher predictability for all clusters compared with other regimes. The frequent color shifts across clusters highlight regime variations.

Additionally, we incorporate aggregate predictors with representative characteristics and generate four heat maps using bivariate-sorted deciles (see Figure A.1). The top-left sub-figure highlights the highly predictable periods, while the other three sub-figures show interactions within the cross-sectional dimensions, such as higher earnings-to-price ratios, performance scores, and earnings surprises. These mosaic patterns can be observed before precise interpretations, especially when considering

time-series effects.

Cluster-Wise Performance. Next, we analyze single-leaf statistics to evaluate predictive abilities. Incorporating time-series partitions, Table A.2 translates the tree structure in Figure 8 into descending rankings of predictability within each regime, highlighting heterogeneity.

The gaps between the highest and lowest R^2 values are more pronounced than in the cross-sectional case (Table 1). For instance, the highest R^2 (N13: 21.54%) contrasts sharply with the lowest (N32: 1.19%). Examining the table longitudinally reveals substantial variability across regimes, each exhibiting a wide range of return predictability. Sorting R^2 values in descending order within specific time periods further illustrates the mosaic patterns of predictability across clusters. Investment performance does not uniformly decline with lower predictability, unlike in the cross-sectional analysis, where slight positive correlations are observed between predictability and investment gains in certain time horizons, such as Regimes I and III.

Regime II, spanning only 16 months, represents the most predictable period but shows negative correlations between predictability and returns. This suggests significant opportunities to anticipate return patterns during this challenging financial landscape, warranting alternative strategies like long-short approaches.

Overall, these findings emphasize the importance of analyzing the diversity of return prediction accuracy at the time-series level.

Aggregate Evaluations. Following a similar approach as in the cross-sectional analysis, we perform aggregate evaluations to assess improvements in heterogeneous return predictability after incorporating time-series information. We focus on a large-cap sub-sample analysis. Clusters are aggregated into several sub-samples, with Table 4 presenting the resulting predictability metrics.

When comparing performance without cross-sectional splits, most aggregated clustering results have higher prediction accuracy. Filtering out large-cap stocks further reveals specific predictability trends under varying market conditions.

Among the three regimes, Regime II ($\texttt{X\_DY} > 0.7$ and $\texttt{X\_DFY} \leq 0.3$) consistently

Table 4: **Evaluating Return Predictability (TS+CS Cluster)**

This table reports return predictability (R^2 values, %) for different predictive methods under specific regimes. We present full-sample results from the tree-based cluster model, incorporating cross-sectional and time-series splits. For each regime, we generate two panels: global and cluster-wise predictions. Five samples are presented: Global (homogeneous forecasts), Aggregate (aggregated cluster-wise predictions), and High, Medium, and Low, based on predictive rankings within the tree clusters.

	Sample A: All Stocks			Sample B: Large-Cap		
	Regime I: $\mathbb{1}\{\texttt{X\_DY} \leq 0.7\}$					
	OLS	Lasso	Ridge	OLS	Lasso	Ridge
Panel A: Global Forecasts						
Global	1.34	0.65	1.26	1.32	0.89	1.29
High	5.14	3.68	4.96	4.80	3.94	4.78
Medium	1.78	0.99	1.69	1.62	1.11	1.57
Low	1.17	0.51	1.10	1.17	0.77	1.15
Panel B: Cluster-Wise Forecasts						
Aggregate	1.84	1.11	1.79	1.55	1.08	1.53
High	7.79	6.63	7.78	5.20	5.00	5.41
Medium	2.86	2.07	2.84	2.38	1.81	2.37
Low	1.52	0.81	1.47	1.26	0.80	1.24
	Regime II: $\mathbb{1}\{\texttt{X\_DY} > 0.7\}\mathbb{1}\{\texttt{X\_DFY} \leq 0.3\}$					
	OLS	Lasso	Ridge	OLS	Lasso	Ridge
Panel A: Global Forecasts						
Global	15.16	10.86	12.85	20.43	15.56	17.50
High	21.47	14.44	17.42	25.02	18.11	20.55
Medium	14.94	10.79	12.68	18.56	14.33	16.07
Low	10.74	8.18	9.68	18.02	14.76	16.38
Panel B: Cluster-Wise Forecasts						
Aggregate	16.52	12.20	14.23	21.26	16.43	18.48
High	21.97	15.56	18.25	25.73	19.24	21.55
Medium	16.89	12.98	14.73	19.26	15.61	17.11
Low	10.97	7.10	9.48	19.39	14.10	17.19
	Regime III: $\mathbb{1}\{\texttt{X\_DY} > 0.7\}\mathbb{1}\{\texttt{X\_DFY} > 0.3\}$					
	OLS	Lasso	Ridge	OLS	Lasso	Ridge
Panel A: Global Forecasts						
Global	3.08	2.24	2.90	3.22	2.66	3.04
High	7.91	5.48	7.53	6.05	5.02	5.80
Medium	3.40	2.49	3.20	3.51	2.88	3.31
Low	2.35	1.70	2.23	2.61	2.19	2.47
Panel B: Cluster-Wise Forecasts						
Aggregate	3.63	2.78	3.47	3.60	2.81	3.41
High	10.76	10.12	10.49	6.94	6.43	6.72
Medium	4.31	3.18	4.12	4.21	3.18	4.01
Low	2.31	1.83	2.17	2.43	2.05	2.26

outperforms the others. This regime corresponds to periods characterized by high dividend yields and low default yields, typically signaling the start of a business expansion phase. Declining trends are evident in the last three rows within each period.

The highly predictable sub-sample achieves the largest R^2 values, while the least predictable sub-sample underperforms all others, including the "Global" sample without cross-sectional clustering.

Despite differences in magnitude, the patterns of improvement and decline are similar across methods, particularly for return predictions based on cluster-wise models. These results provide insights into the mosaics of return predictability and the heterogeneity of stocks under different characteristics and market conditions.

5.2 Structural Breaks by Calendar Months

Splitting by aggregate predictors generates non-continuous regimes on the time horizon. For robustness check, we also examine results from splitting by calendar months. One advantage of this approach is that calendar months increase over time, leading to single, continuous regimes that resemble structural breaks (e.g., Smith and Timmermann, 2021). This method continues to assess whether dividing continuous periods effectively captures the heterogeneity in stock return predictability.

Figure 11: **Time-Series Regimes (Calendar Months)**

This figure shows time-series regime switches by calendar month from 1973 to 2022. Eight colors represent different regimes, indicating varying predictability based on their position in the color bar. Shaded areas denote NBER recessions, and labeled texts highlight global events affecting the world economy. The vertical axis shows the cross-sectional predictability of clusters, from highly (black) to less predictable (white) under each regime.

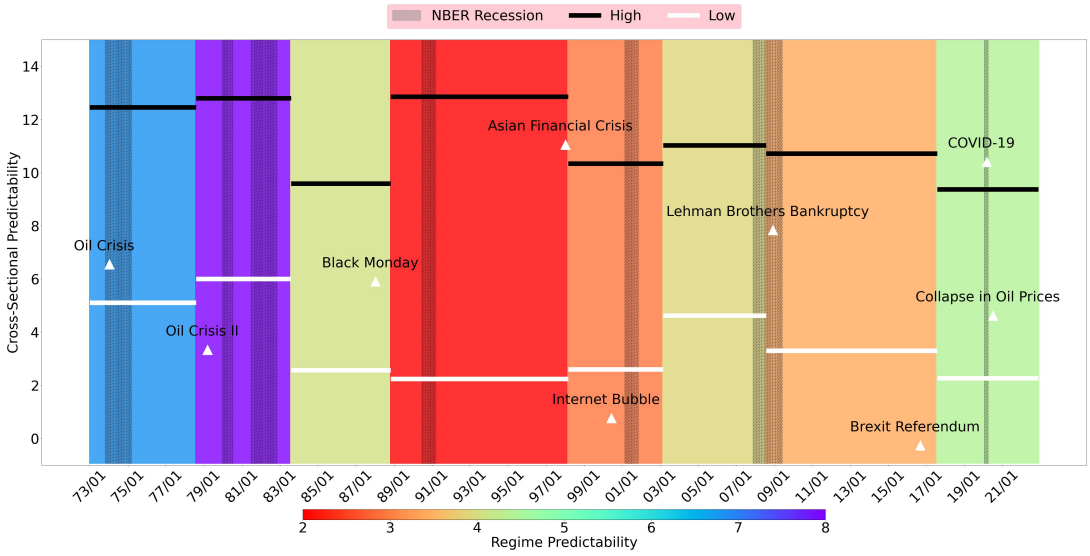


Figure 11 illustrates that stock return predictability varies across time horizons,

although the variability is less than the version using aggregate predictors. For instance, predictability peaks at 7.99% from September 1978 to August 1983 (labeled in purple) but falls to just 2.01% from December 1988 to March 1998 (labeled in red).

Certain events, such as Black Monday (1987), the Asian Financial Crisis (1997–1998), the Euro's official circulation (2002), and the Brexit Referendum (2016), likely triggered regime shifts to different levels of predictability. Conversely, structural breaks may also have contributed to other incidents, such as the Oil Crisis (1979), Lehman Brothers' bankruptcy (2008), and the US-China Trade War (2018).

Splitting further in the cross section within each time-series period would likely increase these heterogeneity gaps. The predictive abilities and investment performance across regimes remain between relatively high (black) and low (white) predictable sub-samples, with trends showing some consistency across periods. These findings confirm the heterogeneous predictability of stock returns in the time-series dimension, highlighting their mosaic-like nature.

6 Investment Gains on Cluster-wise Models

The previous sections primarily focus on illustrating the heterogeneous nature of return predictability, both across the cross section and over time, using our novel clustering approach. In doing so, we uncover an anomaly associated with return predictability. A key byproduct of our clustering analysis is a cluster-wise predictive model. Specifically, we partition assets into multiple clusters, each linked to its cluster-specific predictive model. Together, these cluster-wise models form a heterogeneous predictive model, in contrast to a global predictive model that employs a single, uniform approach to all asset returns.

In this section, we explore the investment gains derived from the cluster-wise predictive model by constructing a new type of forecast-implied portfolio, which evaluates predictive accuracy through portfolio returns. We construct three types of portfolio weights based on model predictions. For equal-weighted and value-weighted portfolios, the stock weights are determined by the absolute value of the respective

weighting schemes, but the sign is determined by the model's prediction. Additionally, we construct forecast-weighted portfolios based on the normalized predictions, controlling for leverage (e.g., Guijarro-Ordonez et al., 2021). In this framework, stocks that are predicted to have high returns will receive a larger long weight, while those with lower predicted negative returns will receive a larger short weight.

Specifically, let weights $w_{i,t}$ be absolute value of original equal-weighted $(1/N_t)$ or value-weighted $(\texttt{size}_{i,t-1}/\sum_i \texttt{size}_{i,t-1})$ portfolios for each cross-sectional dimension at time t. The three portfolio weights are as follows:

$$\begin{aligned} \text{Sign-adjusted Equal/Value-weighted: } \widehat{w}_{i,t} &= \begin{cases} w_{i,t}, & \text{if } \widehat{r}_{i,t} \geq 0 \\ -w_{i,t}, & \text{if } \widehat{r}_{i,t} < 0 \end{cases} \\ \text{Forecast-weighted: } \widehat{w}_{i,t} &= \frac{\widehat{r}_{i,t}}{\sum_i |\widehat{r}_{i,t}|}. \end{aligned} \tag{9}$$

Then, within each leaf cluster j, we define the forecast-implied portfolio as:

$$R_{j,t} = \sum_{\{i,t\} \in \text{leaf}_j} \widehat{w}_{i,t} r_{i,t}. \tag{10}$$

It is important to note that these forecast-implied portfolios differ from cluster portfolios discussed in Section 4.2, where only value-weighted portfolios were created for each cluster to investigate the relationship between predictability and average returns. In contrast, the forecast-implied portfolios in this section are long-short, with the long and short positions determined by the model's predictions, and the absolute weights being either equal- or value-weighted. Intuitively, if the predictive model is more accurate, the resulting portfolio should generate greater investment gains.

Table 5 presents the results of direction-adjusted value- and equal-weighted portfolios, as well as forecast-weighted portfolios, constructed from all observations. These results are derived from the cross-sectional clustering model described in Section 4.

First, our results confirm that highly predictable stocks remain easier to predict out of the sample, and as a result, the corresponding trading strategies generate sig-

Table 5: **Forecast-Implied Investment Performance (CS Cluster)**

This table reports in-sample and out-of-sample investment performance for forecast-implied portfolios (Equation 10). The first two panels show sign-adjusted value- and equal-weighted portfolios, while Panel C reports forecast-weighted portfolios, constructed based on all observations (Equation 9). We present five samples: Global (no clustering), Aggregate (aggregation of clustering results), and High, Medium, and Low, based on predictive rankings within the tree clusters. Each panel includes five columns: monthly average return (Avg, %), standard deviation (Std, %), annualized Sharpe ratio (SR), market alpha (in %), and maximum drawdown (MDD, %).

	In-Sample (1973 - 2002)					Out-of-Sample (2003 - 2022)				
	Panel A: Sign-adjusted Value-Weighted									
	Avg	Std	SR	Alpha	MDD	Avg	Std	SR	Alpha	MDD
Global	0.81	2.87	0.98	0.66^{***}	17.91	0.64	3.53	0.63	0.02	15.52
Aggregate	0.71	3.54	0.70	0.42^{***}	20.02	0.74	3.98	0.64	0.02	16.26
High	2.70	5.92	1.58	2.35^{***}	22.08	1.71	5.31	1.12	0.96^{***}	28.65
Medium	0.88	4.27	0.71	0.52^{***}	20.43	0.86	3.72	0.81	0.29^{**}	15.76
Low	0.61	3.19	0.66	0.41^{***}	19.78	0.65	4.13	0.55	-0.09^{*}	16.78
	Panel B: Sign-adjusted Equal-Weighted									
	Avg	Std	SR	Alpha	MDD	Avg	Std	SR	Alpha	MDD
Global	1.51	2.59	2.02	1.41^{***}	11.85	1.09	3.78	1.00	0.53^{***}	15.91
Aggregate	1.54	2.85	1.87	1.38^{***}	11.55	1.07	3.65	1.02	0.51^{***}	16.74
High	3.41	7.50	1.57	2.98^{***}	25.79	2.70	5.62	1.66	1.90^{***}	16.49
Medium	1.86	3.11	2.07	1.72^{***}	12.34	1.29	3.88	1.16	0.95^{***}	18.75
Low	0.79	2.46	1.11	0.65^{***}	11.62	0.70	4.42	0.55	-0.04	21.56
	Panel C: Forecast-Weighted									
	Avg	Std	SR	Alpha	MDD	Avg	Std	SR	Alpha	MDD
Global	2.28	3.57	2.21	2.16^{***}	17.84	1.44	4.43	1.13	0.75^{***}	19.36
Aggregate	3.09	3.86	2.77	2.95^{***}	12.24	1.83	4.37	1.45	1.22^{***}	19.46
High	4.46	8.72	1.77	3.99^{***}	27.71	3.24	6.87	1.63	2.34^{***}	19.73
Medium	3.10	3.99	2.69	3.02^{***}	23.14	1.95	4.61	1.46	1.54^{***}	19.98
Low	1.18	3.08	1.33	1.02^{***}	13.71	0.86	4.77	0.62	0.06	21.80

nificant investment gains. The sign-adjusted value-weighted portfolio also performs well, indicating that the results are not dominated by small stocks. For example, in Panel A of Table 5, the highly predictable sub-sample shows an OOS Sharpe ratio of 1.12 and an average return of 1.71%. In contrast, clusters with the lowest predictability achieve only about half of these values, with a Sharpe ratio of 0.55 and an average return of 0.65%. Similarly, in Panels B and C, highly predictable stocks generate Sharpe ratios of 1.66 and 1.63, and averages of 2.70% and 3.24%, respectively. These figures

significantly outperform the portfolio of medium- and low-predictability stocks. Similar patterns are observed for market alphas and maximum drawdowns.

Second, the cluster-wise predictive models ("Aggregate") have better performance compared to the homogeneous predictive model ("Global"), which trains on all stocks without accounting for clustering patterns. For example, in Panel C, the OOS average return for the Aggregate model is 1.83%, higher than the 1.44% achieved by the Global model, with Sharpe ratios of 1.45 and 1.13, respectively. The gap is smaller in Panels A and B, primarily because the absolute weights of the two portfolios are identical, with only the sign adjusted by the predictive model. Overall, while both Aggregate and Global models perform similarly in predicting whether a stock price will go up or down in the next month (Panel A and B), the Aggregate model demonstrates superior accuracy in forecasting *how much* the price will go up or down (Panel C).

The time-series and cross-sectional clustering model exhibits similar economic gain for the heterogenous predictive model, for more details, see Appendix II..4.

All evaluations confirm that highly predictable stocks, identified through our cross-sectional tree clusters, significantly dominate the economic profitability of the entire sample. Excluding these highly predictable stocks from the sample markedly reduces potential gains for investors. Extending the findings from the stock return predictability mosaics to the contributions of investment strategies, it is clear that investors can achieve higher economic benefits by allocating funds to stocks with low trading volumes, high earnings surprises, or high valuations (e.g., Basu, 1983; Lakonishok et al., 1994; Piotroski, 2000). Furthermore, the aggregate performance of the cluster-wise predictive model consistently delivers superior forecasts compared to the Global homogeneous model.

7 Conclusion

Existing studies often treat return predictability as a characteristic of the predictors or models employed, evaluating it at an aggregate or average level. This paper offers a novel perspective on the return predictability literature. We argue that return

predictability is an inherent yet unobserved asset characteristic, potentially linked to the cross section of expected returns. Specifically, we demonstrate that return predictability is heterogeneous, varying both across stocks and over time. The central question we address is: Which assets exhibit higher return predictability, and how does this vary across various macroeconomic regimes? Answering this question leads us to uncover what we term the "mosaics of predictability."

To measure heterogeneous return predictability, we introduce a novel tree-based clustering method that groups asset-return observations based on similar predictability levels. Our empirical analysis reveals substantial heterogeneity in return predictability among individual U.S. stocks, advancing our understanding of asset return predictability in several key dimensions. First, we find that asset clusters characterized by low trading volumes, high earnings-to-price ratios, and high unexpected earnings are the most predictable. Second, we observe that return predictability declines sharply when the dividend yield is low, but peaks during periods of high dividend yield and low default yield. Finally, we identify a new predictability anomaly: highly predictable long-only portfolios generate significantly unexplained alphas of approximately 1% over the OOS period of two decades. Moreover, a long-short portfolio based on predictability achieves even greater alphas.

References

Ahn, D.-H., J. Conrad, and R. F. Dittmar (2009). Basis assets. *Review of Financial Studies 22*(12), 5133–5174.

Avramov, D. (2002). Stock return predictability and model uncertainty. *Journal of Financial Economics 64*(3), 423–458.

Avramov, D., S. Cheng, and L. Metzker (2023). Machine learning vs. economic restrictions: Evidence from stock return predictability. *Management Science 69*(5), 2587–2619.

Barillas, F. and J. Shanken (2018). Comparing asset pricing models. *Journal of Finance 73*(2), 715–754.

Basu, S. (1983). The relationship between earnings' yield, market value and return for NYSE common stocks: Further evidence. *Journal of Financial Economics 12*(1), 129–156.

Bianchi, D., M. Büchner, and A. Tamoni (2021). Bond risk premiums with machine learning. *Review of Financial Studies 34*(2), 1046–1089.

Cakici, N., C. Fieberg, T. Neumaier, T. Poddig, and A. Zaremba (2024). Pockets of Predictability: A Replication. *Journal of Finance, forthcoming*.

Campbell, J. Y. and S. B. Thompson (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies 21*(4), 1509–1531.

Cong, L. W., G. Feng, J. He, and X. He (2024). Growing the Efficient Frontier on Panel Trees. *Journal of Financial Economics, Forthcoming*.

Cong, L. W., G. Feng, J. He, and J. Li (2023). Uncommon Factors and Asset Heterogeneity in the Cross Section and Time Series. Technical report, National Bureau of Economic Research.

Cong, L. W., K. Tang, J. Wang, and Y. Zhang (2020). AlphaPortfolio: Direct construction through deep reinforcement learning and interpretable AI. Technical report, Cornell University.

Dangl, T. and M. Halling (2012). Predictive regressions with time-varying coefficients. *Journal of Financial Economics 106*(1), 157–181.

DeMiguel, V., J. Gil-Bazo, F. J. Nogales, and A. A. Santos (2023). Machine learning and fund characteristics help to select mutual funds with positive alpha. *Journal of Financial Economics 150*(3), 103737.

Evgeniou, T., A. Gueciouer, and R. Prieto (2023). Uncovering sparsity and heterogeneity in firm-level return predictability using machine learning. *Journal of Financial and Quantitative Analysis 58*(8), 3384–3419.

Fama, E. F. and K. R. French (1988). Dividend yields and expected stock returns. *Journal of Financial Economics 22*(1), 3–25.

Fama, E. F. and K. R. French (1989). Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics 25*(1), 23–49.

Fama, E. F. and K. R. French (1992). The cross-section of expected stock returns. *Journal of Finance 47*(2), 427–465.

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics 33*(1), 3–56.

Fama, E. F. and K. R. French (2008). Dissecting anomalies. *Journal of Finance 63*(4), 1653–1678.

Fama, E. F. and J. D. MacBeth (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy 81*(3), 607–636.

Farmer, L. E., L. Schmidt, and A. Timmermann (2023). Pockets of predictability. *Journal of Finance 78*(3), 1279–1341.

Farmer, L. E., L. Schmidt, and A. Timmermann (2024). Comment on Cakici, Fieberg, Neumaier, Poddig, and Zaremba: Pockets of Predictability: A Replication. Technical report, University of Viginia.

Feng, G. and J. He (2022). Factor investing: A Bayesian hierarchical approach. *Journal of Econometrics 230*(1), 183–200.

Feng, G., J. He, J. Li, L. Sarno, and Q. Zhang (2024). Currency Return Dynamics: What Is the Role of US Macroeconomic Regimes? Technical report, City University of Hong Kong.

Feng, G., J. He, N. Polson, and J. Xu (2024). Deep Learning of Characteristics-Sorted Factor Models. *Journal of Financial and Quantitative Analysis 59*(7), 3001–3036.

Feng, G., X. He, Y. Wang, and C. Wu (2024). Predicting individual corporate bond returns. *Journal of Banking and Finance, Forthcoming*.

Green, J., J. R. Hand, and X. F. Zhang (2017). The characteristics that provide independent information about average US monthly stock returns. *Review of Financial Studies 30*(12), 4389–4436.

Gu, S., B. Kelly, and D. Xiu (2020). Empirical asset pricing via machine learning. *Review of Financial Studies 33*(5), 2223–2273.

Guijarro-Ordonez, J., M. Pelger, and G. Zanotti (2021). Deep learning statistical arbitrage. Technical report, Stanford University.

Han, Y., A. He, D. E. Rapach, and G. Zhou (2024). Cross-sectional expected returns: New Fama–MacBeth regressions in the era of machine learning. *Review of Finance 28*(6), 1807–1831.

Harvey, C. R., Y. Liu, and H. Zhu (2016). ... and the cross-section of expected returns. *Review of Financial Studies 29*(1), 5–68.

Henkel, S. J., J. S. Martin, and F. Nardari (2011). Time-varying short-horizon predictability. *Journal of Financial Economics 99*(3), 560–580.

Hou, K., C. Xue, and L. Zhang (2020). Replicating anomalies. *Review of Financial Studies 33*(5), 2019–2133.

Jegadeesh, N. and S. Titman (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance 48*(1), 65–91.

Jensen, M. C., F. Black, and M. S. Scholes (1972). The capital asset pricing model: Some empirical tests.

Kaniel, R., Z. Lin, M. Pelger, and S. Van Nieuwerburgh (2023). Machine-learning the skill of mutual fund managers. *Journal of Financial Economics 150*(1), 94–138.

Keim, D. B. and R. F. Stambaugh (1986). Predicting returns in the stock and bond markets. *Journal of Financial Economics 17*(2), 357–390.

Kelly, B., S. Malamud, and K. Zhou (2024). The virtue of complexity in return prediction. *Journal of Finance 79*(1), 459–503.

Kelly, B. and S. Pruitt (2013). Market expectations in the cross-section of present values. *Journal of Finance 68*(5), 1721–1756.

Lakonishok, J., A. Shleifer, and R. W. Vishny (1994). Contrarian investment, extrapolation, and risk. *Journal of Finance 49*(5), 1541–1578.

Lewellen, J. (2015). The Cross-section of Expected Stock Returns. *Critical Finance Review 4*(1), 1–44.

McLean, R. D. and J. Pontiff (2016). Does academic research destroy stock return predictability? *Journal of Finance 71*(1), 5–32.

Patton, A. J. and B. M. Weller (2022). Risk price variation: The missing half of empirical asset pricing. *Review of Financial Studies 35*(11), 5127–5184.

Pesaran, M. H. and A. Timmermann (1995). Predictability of stock returns: Robustness and economic significance. *Journal of Finance 50*(4), 1201–1228.

Piotroski, J. D. (2000). Value investing: The use of historical financial statement information to separate winners from losers. *Journal of Accounting Research*, 1–41.

Rapach, D. E., J. K. Strauss, and G. Zhou (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *Review of Financial Studies 23*(2), 821–862.

Rapach, D. E., J. K. Strauss, and G. Zhou (2013). International stock return predictability: What is the role of the united states? *Journal of Finance 68*(4), 1633–1662.

Shen, Z. and D. Xiu (2024). Can Machines Learn Weak Signals? *University of Chicago, Becker Friedman Institute for Economics Working Paper* (2024-29).

Smith, S. C. and A. Timmermann (2021). Break risk. *Review of Financial Studies 34*(4), 2045–2100.

Stambaugh, R. F. (1999). Predictive regressions. *Journal of Financial Economics 54*(3), 375–421.

Welch, I. and A. Goyal (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies 21*(4), 1455–1508.

Appendix

I. Panel Regression Tree Algorithm

Section 2.3 and 2.4 present the step-by-step tree growing examples, while this section illustrates the complete growing algorithm in pseudo-codes.

Algorithm Panel Regression Tree

procedure PANEL REGRESSION TREE
Input: Asset returns $r_{i,t}$, firm characteristics $z_{i,t-1}$, aggregate predictors x_{t-1}, and tree parameters.
Output: A tree architecture with many split rules.
  for i from 1 to num_iter **do** ▷ Loop over number of iterations
    if current depth $\geq d_{\text{max}}$ **then**
      return.
    else
      Search the tree, find all potential leaf nodes $\mathcal{N}$
      for each leaf node N in $\mathcal{N}$ **do** ▷ Loop over all current leaf nodes
        for each split candidate $\tilde{c}_{p,k,N}$ in $\mathcal{C}_N$ **do**
          Partition data temporally in N according to $\tilde{c}_{p,k,N}$.
          if Left or right child node cannot satisfy minimal leaf size **then**
            continue.
          else
            Obtain cluster-wise return predictions as in (1).
            Calculate the cluster-based R_j^2 by (6).
          end if
        end for
      end for
      Find the best leaf node and split rule that maximizes split criteria for this iteration

$$\tilde{c}_i = \max_{N \in \mathcal{N}, \tilde{c}_{p,k,N} \in \mathcal{C}_N} |R^2_{\text{left}} - R^2_{\text{right}}|$$

      Compare globally for this iteration's split candidates among all leaf nodes.
      Split the node selected at the i-th split rule of the tree $\tilde{c}_i$.
    end if
  end for
  return
end procedure

Note: p, k, N in $\tilde{c}_{p,k,N}$ represent the p-th variable with the k-th value used for leaf node N (Figure 2).

II. Additional Empirical Results

II..1 Cross Section

This section presents additional results related to cross-sectional clustering. Table A.1 reports the performance of two specific long- and short-leg portfolios associated with the "long-short cluster" anomalies detailed in Table 3. Notably, the long-only portfolios comprising highly predictable stocks significantly drive the abnormal returns observed in the long-short strategies.

Table A.1: **Predictability-Implied Anomaly Testing**

This table reports summary statistics (Panel A) and abnormal returns (Panel B) for the long-short factor based on cross-sectional return predictability across clusters. Two representative anomalies, "L1-S1" and "L5-S1," are used to present the long- and short-leg numbers, as shown in Table 3. The numbers after "L" and "S" denote the number of long and short clusters. Panel A provides the average return (Avg, %), median (Median, %), standard deviation (Std, %), skewness, kurtosis, annualized Sharpe ratio (SR), and maximum drawdown (MDD). Panel B reports abnormal return estimates (alphas, %) and their significance (denoted by "*" and t-values in parentheses) from various factor models.

	In-Sample (1973 - 2002)					Out-of-Sample (2003 - 2022)				
	S1	L1	L5	L1-S1	L5-S1	S1	L1	L5	L1-S1	L5-S1
					Panel A: Summary Statistics					
Avg	0.32	3.54	2.35	3.22	2.03	0.73	3.51	1.81	2.78	1.08
Median	0.56	3.50	2.23	2.82	1.67	1.12	3.20	2.29	2.48	1.00
Std	4.57	6.39	6.01	4.94	3.82	4.37	6.43	6.08	4.70	3.30
Skewness	-0.30	0.84	0.46	1.09	1.04	-0.56	0.70	-0.25	1.09	0.45
Kurtosis	2.60	4.34	6.59	3.62	3.06	1.68	4.52	1.91	4.61	1.37
SR	0.24	1.92	1.35	2.26	1.85	0.58	1.89	1.03	2.05	1.13
MDD	22.39	19.02	25.89	9.74	8.66	17.82	19.34	21.39	13.99	8.29
					Panel B: Abnormal Returns					
CAPM	-0.08	3.19***	1.95***	3.26***	2.03***	-0.06	2.70***	0.87***	2.76***	0.92***
	(-1.48)	(12.04)	(9.34)	(12.53)	(10.03)	(-1.37)	(8.86)	(4.21)	(8.93)	(4.40)
FF3	-0.19***	2.68***	1.49***	2.87***	1.68***	-0.05	2.75***	0.91***	2.80***	0.97***
	(-4.14)	(12.53)	(11.09)	(13.23)	(12.49)	(-1.42)	(9.62)	(5.64)	(9.63)	(5.80)
FF3+IVOL	-0.21***	2.59***	1.48***	2.80***	1.69***	-0.06*	2.87***	0.97***	2.94***	1.04***
	(-4.51)	(11.75)	(10.69)	(12.52)	(12.21)	(-1.78)	(10.38)	(6.11)	(10.49)	(6.40)
FF3+MOM	-0.09**	2.85***	1.67***	2.94***	1.76***	-0.04	2.85***	0.98***	2.89***	1.01***
	(-2.11)	(12.98)	(12.46)	(13.08)	(12.74)	(-1.11)	(10.36)	(6.24)	(10.20)	(6.20)
FF5	-0.19***	2.53***	1.42***	2.72***	1.61***	-0.08**	2.80***	1.00***	2.88***	1.07***
	(-4.12)	(11.79)	(10.35)	(12.43)	(11.69)	(-2.04)	(9.50)	(6.00)	(9.60)	(6.31)
FF5+MOM+IVOL	-0.12***	2.70***	1.59***	2.82***	1.72***	-0.07**	2.91***	1.05***	2.98***	1.12***
	(-2.91)	(12.41)	(11.77)	(12.57)	(12.24)	(-2.10)	(10.45)	(6.61)	(10.48)	(6.84)
Q5	-0.04	2.65***	1.59***	2.70***	1.64***	-0.06	2.81***	1.03***	2.88***	1.09***
	(-0.72)	(9.90)	(9.05)	(10.27)	(9.75)	(-1.56)	(9.25)	(5.83)	(9.35	(6.14)
BS6	-0.14***	2.34***	1.35***	2.48***	1.49***	-0.06*	2.77***	0.91***	2.83***	0.98***
	(-3.00)	(10.24)	(9.67)	(10.62)	(10.27)	(-1.76)	(9.84)	(5.73)	(9.78)	(5.88)

II..2 Time Series

Figure A.1 illustrates interactions between macroeconomic predictors and firm characteristics. The discontinuous regime identified by two business cycle predictors—higher dividend yield and lower default yield—exhibits the highest level of predictability. Table A.2 summarizes the cluster-wise information under each regime.

Figure A.1: **Mosaics of Predictability by Predictors (TS+CS Cluster)**

We present four heat maps summarizing average return predictability (R^2 values, % in the color bar) for stock returns based on the tree-based clustering results from Figure 8. Each sub-figure shows the average R^2 values for 10×10 groups, formed by bivariate-sorted deciles of different predictor pairs. Empty grids with $\times$ indicate no observations in those interacted sub-samples.

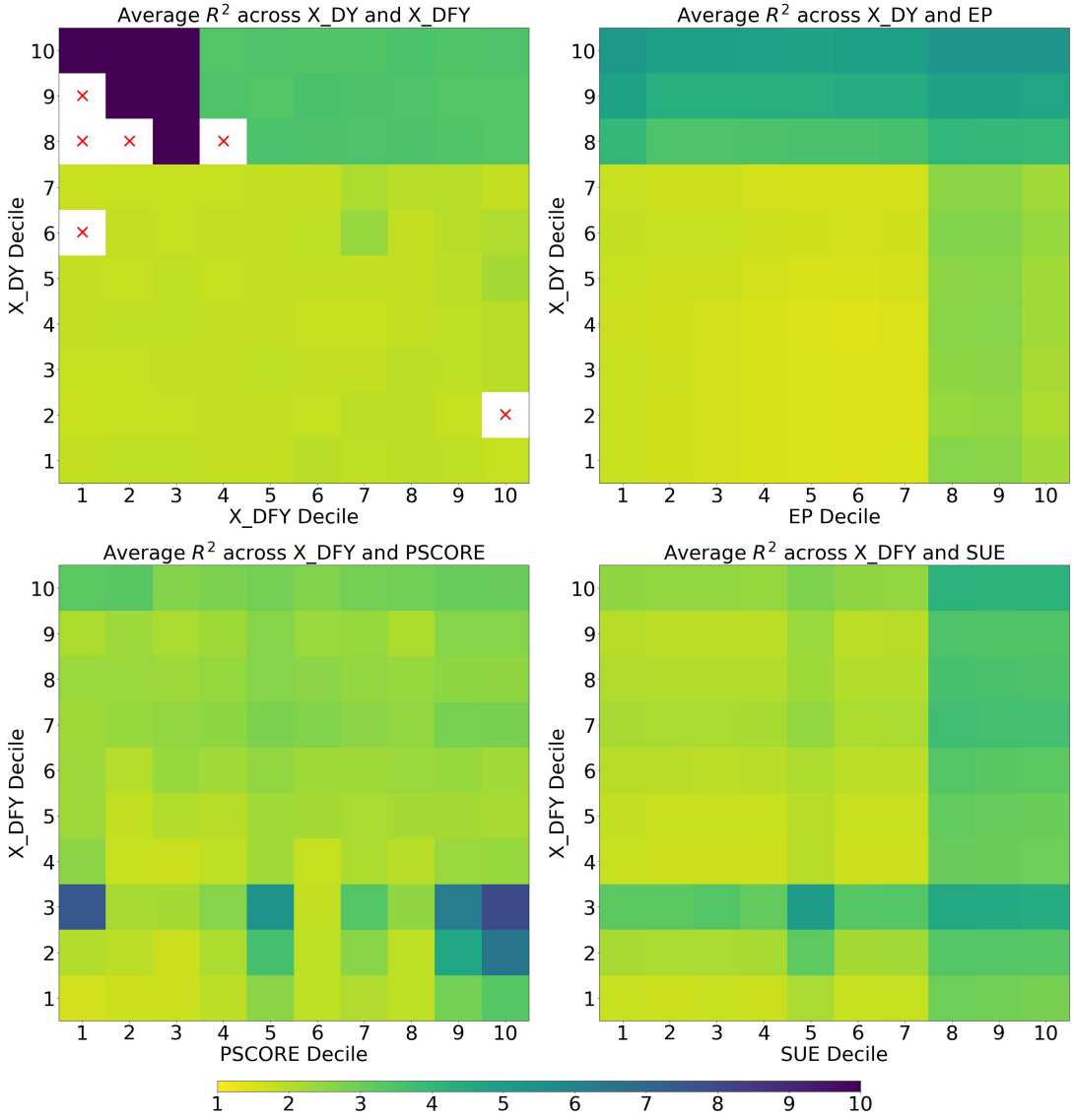


Table A.2: **Cluster-Wise Performance (TS+CS Cluster)**

This table summarizes the cluster-based information under each regime within two panels corresponding to the cross-sectional and time-series tree structure in Figure 8. Panel A counts the number of observations for each cluster ("# obs") and displays the return predictability (R^2 values in %) for each cluster. In Panel B, "Avg" and "SR" denote the monthly average return (in %) and annualized Sharpe ratio for equal/value-weighted (EW/VW) portfolios based on all observations. Each regime of values is arranged downward in the descending order of R^2.

	Panel A: Summary Statistics		Panel B: Profitability			
		Regime I: $\mathbb{1}\{\texttt{X\_DY} \leq 0.7\}$				
Leaf	# obs	R^2	Avg_{EW}	SR_{EW}	Avg_{VW}	SR_{VW}
N38	44,913	8.33	2.52	2.02	1.44	1.14
N22	17,380	7.75	1.47	1.04	1.09	0.64
N43	13,344	7.66	1.08	0.52	1.46	0.66
N46	15,584	6.20	1.51	1.13	1.07	0.74
N39	77,804	3.30	1.74	1.08	0.95	0.64
N37	33,025	3.20	2.30	1.34	1.07	0.71
N34	47,992	2.95	-1.19	-0.36	-1.31	-0.40
N47	19,904	2.73	1.20	0.65	1.24	0.78
N33	14,486	2.47	-0.59	-0.21	-0.67	-0.21
N41	21,982	2.25	1.18	0.64	0.85	0.48
N42	11,929	2.18	0.96	0.37	0.77	0.25
N35	397,816	1.80	-0.85	-0.35	-0.85	-0.35
N36	273,572	1.47	1.15	0.65	0.82	0.60
N40	64,672	1.32	0.60	0.34	0.58	0.37
N32	938,700	1.19	0.39	0.28	0.55	0.45
		Regime II: $\mathbb{1}\{\texttt{X\_DY} > 0.7\}\mathbb{1}\{\texttt{X\_DFY} \leq 0.3\}$				
Leaf	# obs	R^2	Avg_{EW}	SR_{EW}	Avg_{VW}	SR_{VW}
N13	11,785	21.54	-0.03	-0.01	-0.86	-0.55
N25	13,568	17.87	-0.64	-0.31	-0.50	-0.35
N49	10,119	16.04	0.67	0.30	0.01	0.01
N96	10,659	15.34	1.29	0.66	-0.15	-0.09
N97	13,481	10.81	0.69	0.36	-0.13	-0.10
		Regime III: $\mathbb{1}\{\texttt{X\_DY} > 0.7\}\mathbb{1}\{\texttt{X\_DFY} > 0.3\}$				
Leaf	# obs	R^2	Avg_{EW}	SR_{EW}	Avg_{VW}	SR_{VW}
N63	10,928	10.75	4.14	1.65	1.93	0.92
N62	18,667	6.69	3.46	1.47	2.07	1.09
N59	12,040	6.02	1.97	0.84	1.25	0.54
N114	12,910	5.90	0.85	0.29	0.51	0.18
N61	17,589	5.77	2.85	1.27	1.47	0.83
N117	14,648	5.48	2.08	0.76	2.05	0.79
N121	19,622	5.46	2.05	1.12	1.30	0.82
N115	48,112	3.69	1.27	0.56	0.74	0.41
N120	74,177	3.62	1.59	0.85	0.70	0.44
N113	95,937	3.40	0.72	0.39	0.35	0.21
N116	31,856	3.23	1.28	0.51	0.84	0.36
N112	271,598	2.30	0.78	0.41	0.38	0.24

II..3 Structural Break

Table A.3: **Cluster-Wise Performance (+ Structural Break)**

This table shows the performance of the tree-based cluster model using both cross-sectional and time-series splits. In-sample return predictability (R^2, %) is calculated using Equation (6). "# obs" indicates the stock returns count for each cluster. "Avg" and "SR" denote the monthly average return (in %) and annualized Sharpe ratio for cluster-wise equal-weighted and value-weighted forecast-implied portfolios. Regimes are ordered by descending R^2 from left to right.

197301-197808	N7	N13	N24	N4	N10	N23	N25	N22						
# obs	12,456	10,468	11,341	19,069	16,687	16,022	11,907	148,017						
R^2	17.80	12.17	11.03	10.78	10.33	9.80	7.85	5.37						
Avg_{EW}	4.49	4.89	5.15	3.50	3.27	2.89	3.68	2.65						
SR_{EW}	1.85	2.42	2.86	2.00	2.33	1.91	2.15	1.76						
Avg_{VW}	3.82	3.97	4.75	2.65	2.61	2.42	3.40	1.67						
SR_{VW}	1.76	2.29	3.02	1.70	2.28	1.89	2.09	1.70						
197809-198308	N15	N14	N19	N10	N11	N16	N18	N6	N17					
# obs	13,683	15,824	11,955	12,017	10,097	21,130	33,132	14,132	107,707					
R^2	16.28	13.25	12.05	11.41	9.76	9.52	9.20	8.34	6.21					
Avg_{EW}	4.40	4.05	3.46	3.86	3.94	2.50	3.53	2.78	2.98					
SR_{EW}	2.74	2.65	2.43	2.55	2.14	2.45	2.38	2.31	2.08					
Avg_{VW}	2.97	2.31	3.45	3.00	2.38	2.06	3.53	1.91	1.86					
SR_{VW}	2.15	1.82	2.30	2.20	1.65	2.20	2.29	1.69	1.83					
198309-198811	N30	N13	N28	N31	N24	N17	N19	N16	N29	N25	N18	N5		
# obs	20,934	11,204	12,120	104,173	11,733	10,950	10,502	10,490	21,443	27,032	46,428	12,370		
R^2	10.08	8.07	7.51	7.11	6.80	6.68	5.69	5.60	4.72	3.17	2.07	-7.77		
Avg_{EW}	2.82	2.03	3.05	2.22	2.98	2.78	2.70	2.46	3.19	1.68	1.86	1.00		
SR_{EW}	1.75	2.06	2.25	1.43	1.76	2.17	2.03	2.25	2.64	1.55	1.83	1.22		
Avg_{VW}	2.04	1.75	2.95	1.61	1.77	2.53	1.99	2.41	3.25	1.13	1.42	0.95		
SR_{VW}	1.33	1.60	2.18	1.12	1.24	2.10	1.42	2.30	2.53	1.05	1.31	1.04		
198812-199803	N14	N31	N12	N20	N18	N30	N26	N21	N19	N23	N17	N27	N22	N16
# obs	14,635	14,915	12,594	14,883	33,550	12,315	107,725	10,176	57,794	18,007	264,478	23,615	14,053	21,101
R^2	12.78	8.11	7.11	6.74	4.87	4.46	3.38	2.75	2.74	2.31	1.94	1.30	0.60	0.39
Avg_{EW}	3.38	3.12	2.10	4.02	1.16	1.75	1.60	2.74	1.59	2.03	2.04	1.54	2.23	0.91
SR_{EW}	3.50	2.66	1.89	2.28	1.77	1.56	2.13	1.50	1.92	1.84	3.02	2.25	2.41	1.97
Avg_{VW}	2.20	2.32	2.13	2.07	1.19	1.52	1.31	2.08	1.23	1.78	1.48	1.10	2.22	0.93
SR_{VW}	2.52	2.36	1.67	1.11	1.35	1.35	1.76	1.07	1.44	1.45	1.69	0.98	2.46	1.32
199804-200303	N6	N15	N10	N19	N14	N18	N11	N16	N17					
# obs	19,659	11,027	10,722	12,155	11,809	54,089	10,912	24,346	197,515					
R^2	12.94	10.97	9.93	9.61	7.50	6.50	5.84	5.28	2.83					
Avg_{EW}	9.61	6.31	7.47	7.02	5.88	5.77	4.49	3.41	1.63					
SR_{EW}	2.53	1.67	2.07	2.12	2.24	3.18	2.37	2.10	1.86					
Avg_{VW}	7.41	6.96	4.94	6.55	4.01	5.79	2.27	3.97	0.63					
SR_{VW}	2.12	1.87	1.38	1.90	1.59	2.69	0.96	2.01	0.64					
200304-200808	N11	N6	N20	N17	N19	N7	N21	N16	N18					
# obs	18,849	20,951	16,527	12,396	52,975	16,343	36,240	30,237	91,152					
R^2	10.75	9.39	7.37	7.18	4.67	4.49	4.29	3.88	2.51					
Avg_{EW}	2.85	3.18	1.75	1.96	1.65	2.85	1.90	2.20	1.27					
SR_{EW}	2.70	1.64	2.76	1.58	1.99	1.88	1.69	2.15	1.81					
Avg_{VW}	1.32	2.02	1.10	1.69	1.46	2.20	0.96	1.01	0.66					
SR_{VW}	1.34	1.02	1.82	1.34	1.91	1.48	1.34	1.21	1.06					
200809-201708	N14	N25	N11	N15	N21	N24	N26	N19	N27	N20	N17	N18	N16	
# obs	10,211	14,442	10,783	20,192	14,600	19,766	13,437	24,431	41,970	47,813	101,001	17,378	64,402	
R^2	11.40	9.41	7.40	7.29	6.84	6.60	6.10	5.37	4.51	4.18	3.44	3.13	1.87	
Avg_{EW}	2.98	2.15	3.94	2.51	3.44	1.53	2.51	1.65	2.29	2.83	1.88	1.48	1.64	
SR_{EW}	2.43	1.78	1.80	1.81	1.84	1.20	2.08	1.34	1.36	1.52	1.45	1.34	1.60	
Avg_{VW}	1.69	2.24	4.00	1.81	1.92	1.28	2.07	1.33	1.97	2.02	1.50	1.39	1.50	
SR_{VW}	1.19	1.44	1.74	1.72	1.33	1.20	1.65	1.20	1.18	1.16	1.34	1.36	1.47	
201709-202212	N13	N22	N30	N12	N31	N23	N18	N10	N14	N19	N17	N16		
# obs	16,828	11,479	14,059	12,671	11,678	13,599	15,872	11,654	12,834	81,762	13,429	11,740		
R^2	13.59	9.39	9.38	9.26	6.79	6.75	6.74	5.00	4.35	3.98	1.54	-6.82		
Avg_{EW}	1.94	4.34	2.72	2.11	2.96	3.77	2.47	3.47	3.00	2.64	1.91	2.11		
SR_{EW}	1.38	1.65	1.60	1.74	1.69	1.75	1.84	2.10	1.34	1.77	1.36	1.36		
Avg_{VW}	2.03	2.73	2.62	1.68	2.80	2.46	1.56	1.44	2.06	1.84	1.67	2.42		
SR_{VW}	1.27	1.15	1.58	1.50	1.57	1.59	1.39	1.34	0.92	1.50	1.26	1.44		

Table A.4: **Evaluating Return Predictability (+ Structural Break)**

This table reports return predictability (R^2, %) for different predictive methods under various regimes. We present full-sample results from the tree-based cluster model, incorporating both structural breaks and cross-sectional splits. For each regime, we show four samples: Aggregate, High, Medium, and Low, based on the predictive rankings within the tree clusters.

	Sample A: All Stocks			Sample B: Large-Cap		
197301-197808	OLS	Lasso	Ridge	OLS	Lasso	Ridge
Aggregate	8.61	7.12	8.21	6.67	5.45	6.34
High	12.94	10.95	12.46	7.37	6.81	7.48
Medium	9.27	7.31	8.95	9.11	7.23	8.83
Low	5.45	4.41	5.11	5.92	4.74	5.51
197809-198308	OLS	Lasso	Ridge	OLS	Lasso	Ridge
Aggregate	8.55	7.11	8.21	11.05	9.46	10.60
High	13.37	11.75	12.80	13.10	11.20	12.44
Medium	9.32	7.83	8.98	9.08	7.39	8.71
Low	6.26	4.90	6.00	9.08	7.99	8.90
198309-198811	OLS	Lasso	Ridge	OLS	Lasso	Ridge
Aggregate	5.28	3.55	4.98	7.78	5.35	7.25
High	10.07	8.22	9.59	13.38	11.08	12.74
Medium	6.88	4.40	6.30	8.49	5.50	7.67
Low	2.51	1.70	2.57	1.41	1.30	2.14
198812-199803	OLS	Lasso	Ridge	OLS	Lasso	Ridge
Aggregate	2.86	2.37	2.79	3.66	2.74	3.52
High	12.94	11.86	12.86	11.25	10.78	11.40
Medium	7.26	5.95	7.19	9.81	6.60	9.36
Low	2.31	1.91	2.24	3.36	2.51	3.22
199804-200303	OLS	Lasso	Ridge	OLS	Lasso	Ridge
Aggregate	7.50	5.78	6.60	6.56	5.04	5.79
High	11.94	10.20	10.34	11.67	10.23	10.58
Medium	6.48	4.62	5.88	6.82	4.66	6.15
Low	3.03	1.60	2.60	2.54	1.22	1.97
200304-200808	OLS	Lasso	Ridge	OLS	Lasso	Ridge
Aggregate	6.71	4.98	6.25	6.27	4.76	5.82
High	11.40	10.32	11.03	9.89	8.95	9.40
Medium	8.58	6.40	8.20	8.40	7.45	8.42
Low	5.11	3.26	4.62	5.55	3.91	5.08
200809-201708	OLS	Lasso	Ridge	OLS	Lasso	Ridge
Aggregate	4.29	2.42	4.27	6.12	3.76	6.12
High	10.88	7.63	10.72	9.59	5.80	9.59
Medium	6.67	4.40	6.61	8.24	5.48	8.23
Low	3.30	1.61	3.30	5.08	2.95	5.08
201709-202212	OLS	Lasso	Ridge	OLS	Lasso	Ridge
Aggregate	3.83	3.18	3.99	5.76	4.95	5.86
High	9.64	8.38	9.37	11.68	9.79	11.19
Medium	5.89	4.27	5.59	6.83	5.38	6.55
Low	1.91	1.64	2.27	3.16	2.98	3.61

II..4 Investment Gains: Forecast-implied Portfolio

This appendix section explores the investment gains of forecast-implied portfolios constructed from TS+CS clusters. The observed patterns are consistent with those of the CS clusters: highly predictable clusters outperform medium- and low-predictability clusters, and the aggregate forecast model delivers superior performance compared to the global model.

Table A.5: **Forecast-Implied Investment Performance (TS+CS Cluster)**

This table reports full-sample and large-cap sub-sample baseline long-short investment performance for forecast-implied portfolios (Equation 10). The first two panels show sign-adjusted value- and equal-weighted portfolios, while Panel C reports forecast-weighted portfolios, all constructed from observations in specific samples (Equation 9). We show five samples: Global (no clustering), Aggregate (aggregation of clustering results), and High, Medium, and Low, based on predictive rankings within the tree clusters. We show results of monthly average return (Avg, %), standard deviation (Std, %), annualized Sharpe ratio (SR), market alpha (in %), and monthly maximum drawdown (MDD, %).

	Sample A: All Stocks					Sample B: Large-Cap				
	Panel A: Sign-adjusted Value-Weighted									
	Avg	Std	SR	Alpha	MDD	Avg	Std	SR	Alpha	MDD
Global	0.92	3.20	0.99	0.81***	15.54	0.89	3.20	0.96	0.77***	16.04
Aggregate	1.10	3.12	1.22	0.95***	13.20	1.06	3.12	1.17	0.91***	13.02
High	1.98	3.82	1.79	1.91**	3.90	1.82	3.71	1.70	1.76*	4.10
Medium	1.39	3.57	1.34	1.23***	13.54	1.32	3.58	1.28	1.16***	13.75
Low	0.85	2.93	1.01	0.69***	13.31	0.83	2.96	0.97	0.66***	13.20
	Panel B: Sign-adjusted Equal-Weighted									
	Avg	Std	SR	Alpha	MDD	Avg	Std	SR	Alpha	MDD
Global	1.62	3.71	1.51	1.59***	19.86	1.10	3.57	1.07	1.00***	18.04
Aggregate	1.83	3.62	1.75	1.77***	17.45	1.27	3.46	1.27	1.13***	14.18
High	3.84	5.21	2.55	3.74***	2.68	2.87	4.54	2.19	2.77***	3.79
Medium	2.31	3.83	2.09	2.17***	17.47	1.63	3.86	1.46	1.47***	16.39
Low	1.46	3.11	1.63	1.50***	17.45	0.95	3.17	1.04	0.84***	13.78
	Panel C: Forecast-Weighted									
	Avg	Std	SR	Alpha	MDD	Avg	Std	SR	Alpha	MDD
Global	2.30	4.37	1.83	2.28***	22.73	1.48	4.17	1.23	1.36***	18.58
Aggregate	2.87	4.62	2.15	2.83***	20.50	1.80	4.13	1.51	1.66***	16.15
High	4.68	5.40	3.00	4.58***	2.17	3.26	4.64	2.43	3.16***	4.35
Medium	3.26	4.77	2.37	3.17***	22.64	2.21	4.84	1.58	2.12***	24.15
Low	2.45	4.14	2.05	2.59***	20.05	1.37	3.72	1.27	1.25***	14.31

III. Predictor Descriptions

Table A.6: **Aggregate Predictors**

No.	Variable Name	Description
1	X_TBL	3-month treasury bill rate
2	X_INFL	Inflation
3	X_TMS	Term spread
4	X_DFY	Default yield
5	X_DY	Dividend yield of S&P 500
6	X_SVAR	Rolling 12-month market excess return volatility
7	X_NI	Net equity issuance of S&P 500
8	X_LIQ	Rolling 12-month Pastor-Stambaugh illiquidity

Table A.7: **Equity Characteristics**

No.	Acronym	Description	Category
1	abr	Cumulative abnormal returns around earnings announcement dates	Momentum
2	acc	Operating Accruals	Investment
3	adm	Advertising Expense-to-market	Intangibles
4	agr	Asset growth	Investment
5	alm	Quarterly Asset Liquidity	Intangibles
6	ato	Asset Turnover	Profitability
7	baspread	Bid-ask spread rolling 3m	Liquidity
8	beta	Beta rolling 3m	Volatility
9	bm	Book-to-market equity	Value
10	bm_ia	Industry-adjusted book to market	Value
11	cash	Cash holdings	Value
12	cashdebt	Cash to debt	Value
13	cfp	Cashflow-to-price	Value
14	chpm	Industry-adjusted change in profit margin	Profitability
15	chtx	Change in tax expense	Momentum
16	cinvest	Corporate investment	Investment
17	depr	Depreciation / PPandE	Momentum
18	dolvol	Dollar trading volume	Liquidity
19	dy	Dividend yield	Value
20	ep	Earnings-to-price	Value
21	gma	Gross profitability	Investment
22	grltnoa	Growth in long-term net operating assets	Investment
23	herf	Industry sales concentration	Intangibles

Table A.7: Equity Characteristics (Continued)

No.	Acronym	Description	Category
24	hire	Employee growth rate	Intangibles
25	ill	Illiquidity rolling 3m	Liquidity
26	lev	Leverage	Value
27	lgr	Growth in long-term debt	Investment
28	maxret	Maximum daily returns rolling 3m	Volatility
29	me	Market equity	Size
30	me_ia	Industry-adjusted size	Size
31	mom12m	Momentum rolling 12m	Momentum
32	mom1m	Momentum	Momentum
33	mom36m	Momentum rolling 36m	Momentum
34	mom60m	Momentum rolling 60m	Momentum
35	mom6m	Momentum rolling 6m	Momentum
36	ni	Net Stock Issues	Investment
37	nincr	Number of earnings increases	Momentum
38	noa	(Changes in) Net Operating Assets	Investment
39	op	Operating profitability	Profitability
40	pctacc	Percent operating accruals	Investment
41	pm	Profit margin	Profitability
42	pscore	Performance Score	Profitability
43	rd_sale	R&D to sales	Intangibles
44	rdm	R&D Expense-to-market	Intangibles
45	rna	Quarterly Return on Net Operating Assets, Quarterly Asset Turnover	Profitability
46	Roa1	Return on Assets	Profitability
47	roe	Return on Equity	Profitability
48	rsup	Revenue surprise	Momentum
49	rvar_capm	Residual variance - CAPM rolling 3m	Volatility
50	svar	Return variance rolling 3m	Volatility
51	seas1a	Seasonality	Intangibles
52	sgr	Sales growth	Value
53	sp	Sales-to-price	Value
54	std_dolvol	Std of dollar trading volume rolling 3m	Volatility
55	std_turn	Std. of Share turnover rolling 3m	Volatility
56	sue	Unexpected quarterly earnings	Momentum
57	turn	Shares turnover	Liquidity
58	zerotrade	Number of zero-trading days rolling 3m	Liquidity