Mosaics of Predictability

Lin William Cong[1] Guanhao Feng[2] Jingyu He[2] Yuanzhi Wang[2]
[1]Cornell University & NBER [2]City University of Hong Kong
2025 AFA Annual Meeting

Highlights

- Propose a novel **tree-based clustering** algorithm to measure **heterogeneous** return predictability.
- Stocks with **low DOLVOL, high EP and SUE** are **more predictable** than others (cross section).
- Predictability peaks during periods of **high dividend yield and low default yield** (time series).
- Identify a new **predictability anomaly**: long-short cluster portfolios yield **high OOS abnormal returns**.

Methodology

Predictability: Signal-to-noise ratio (j^{th} leaf R^2).

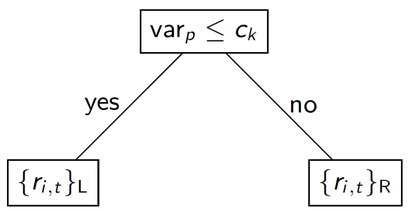
$$R_j^2 = 1 - \frac{\sum_{\{i,t\}\in \text{leaf}_j} (r_{i,t} - \widehat{r}_{i,t})^2}{\sum_{\{i,t\}\in \text{leaf}_j} r_{i,t}^2}$$

$\widehat{r}_{i,t}$: volatility-weighted Ridge regression (avoid dominance of microcaps).

$$\widehat{\beta}_j = \arg\min_{\beta_0,\beta} \left\{ \frac{1}{N_{\text{leaf}_j}} \sum_{\text{leaf}_j} w_{i,t-1} \left(r_{i,t} - \beta_0 - \beta^\intercal \mathbf{s}_{i,t-1}\right)^2 + \lambda ||\beta||_2^2 \right\}$$

$w_{i,t-1} = 1/\sigma_{i,t-1}^2$ (inverse of idiosyncratic return variance)

Goal-oriented Clustering: Decision tree structure.



- Splitting candidates: standardized variables + cutpoints.
- Optimal choice: **maximum R^2 difference**.

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

Feel free to scan for the latest version of paper on SSRN.
Welcome to our regular session, holding on Saturday, Jan. 4, 2025
8:00 – 10:00 AM (PST), Marriott Marquis, Yerba Buena Salon 14 & 15!

Motivation

- Evidence of return predictability from **different asset classes** (aggregate market, individual stock, treasury bond, corporate bond, mutual fund, etc.).
- Predictability: **an attribute of predictors** (macro variables, characteristics, etc.) or **models** (e.g., cross-sectional or panel regression, machine learning).
- Empirical findings based on **homogeneous** (**global**) predictions.

- **Predictability is heterogeneous** for different stocks and varies over time.
- It might be a **characteristics** and an **anomaly** (high predictability → high return)!

Our solution: **Tree-based Clustering** (self-supervised).
Decision tree by **firm characteristics (cross section)** and/or **macro predictors (time series)** to separate panel — **Max predictability (R^2) differences** across groups.

Empirical Results

Figure 1. Cross-sectional Tree-based Clustering

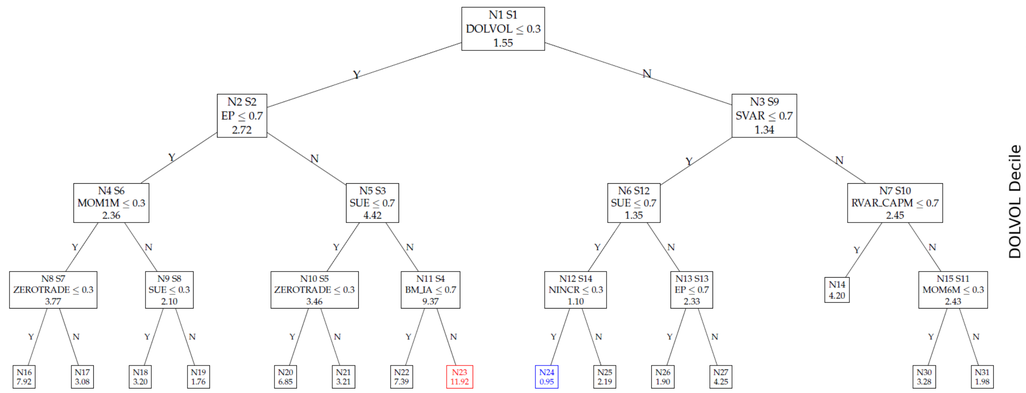


Figure 2. Mosaics of Predictability

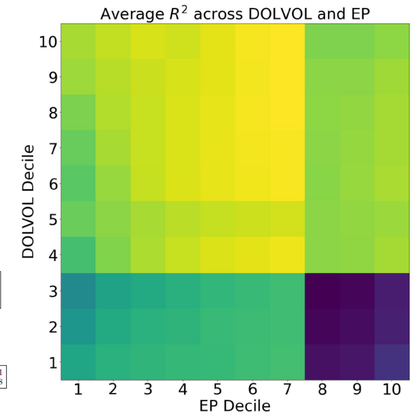


Figure 3. Conn.: Predictability and Profitability

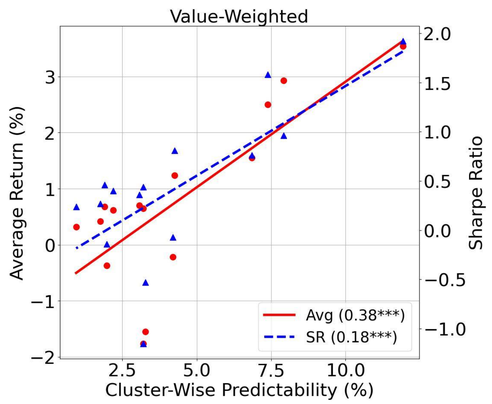


Table 1. Heterogeneous Predictability Anomaly

	1973 - 2002 (in-sample)			2003 - 2022 (out-of-sample)		
	L5	S1	L5-S1	L5	S1	L5-S1
	Panel A: Performance					
Avg (%)	2.35	0.32	2.03	1.81	0.73	1.08
Ann. SR	1.35	0.24	1.85	1.03	0.58	1.13
	Panel B: Unexplained monthly alphas (%)					
CAPM	1.95***	-0.08	2.03***	0.87***	-0.06	0.92***
FF3+MOM	1.67***	-0.09**	1.76***	0.98***	-0.04	1.01***
FF5	1.42***	-0.19***	1.61***	1.00***	-0.08**	1.07***
FF5+MOM+IVOL	1.59***	-0.12***	1.72***	1.05***	-0.07**	1.12***
Q5	1.59***	-0.04	1.64***	1.03***	-0.06	1.09***
BS6	1.35***	-0.14***	1.49***	0.91***	-0.06*	0.98***