

The Anatomy of Machine Learning-Based Portfolio Performance

Philippe Goulet Coulombe¹, David E. Rapach^{*2}, Erik Christian Montes
Schütte³, and Sander Schwenk-Nebbe³

¹Université du Québec à Montréal

²Federal Reserve Bank of Atlanta

³Aarhus University

February 18, 2025

Abstract

Asset return predictability is routinely assessed by economic value: based on a set of predictors, out-of-sample return forecasts are generated—increasingly via “black box” machine learning models—which serve as inputs for portfolio construction, and performance metrics are computed over an evaluation period. We develop a methodology based on Shapley values—the *Shapley-based portfolio performance contribution* (SPPC)—to directly estimate the contributions of individual or groups of predictors to a performance metric. We illustrate the SPPC in an empirical application measuring the economic value of cross-sectional stock return predictability using a large number of firm characteristics and machine learning.

JEL classifications: C53, C55, C58, G11, G17

Keywords: Asset return predictability, Machine learning, Out-of-sample forecast, Economic value, Shapley value, XGBoost, Firm characteristics

*Corresponding author. Send correspondence to David Rapach, Research Department, Federal Reserve Bank of Atlanta, 1000 Peachtree Street NE, Atlanta, GA 30309; email: dave.rapach@gmail.com. This research was enabled in part by support provided by Calcul Québec and the Digital Research Alliance of Canada. We are grateful to seminar and conference participants at the Federal Reserve Bank of Atlanta, Insightful Minds in Artificial Intelligence Webinar Series, Applied Machine Learning, Economics, and Data Science (AMLEDS) Webinar Series, University of Sheffield Workshop on Financial Econometrics, Louisiana State University, 2024 Nordic Econometric Meeting, 6th Future of Financial Information Conference, 2024 International Symposium on Forecasting, 2024 EEA-ESEM Meetings, and the CEMFI Workshop on Big Data in Asset Management, as well as Andrew Chen, Gustavo Freire, Eric Ghysels, Theis Ingerslev Jensen, Alejandro Lopez-Lira, Francisco Peñarando (CEMFI discussant), Rickard Sandberg, Robert Taylor, Junbo Wang, Michael Weber, and Abalfazl Zareei (FFIC discussant) for insightful comments. The views expressed here are those of the authors and not necessarily those of the Federal Reserve Bank of Atlanta or the Federal Reserve System. Any remaining errors are the authors’ responsibility.

1. Introduction

Asset return predictability is a leading topic in empirical asset pricing. Out-of-sample tests are now routinely employed, as they are viewed as the most rigorous and informative tests of return predictability, particularly in the era of “big data” and machine learning (e.g., Nagel 2021; Martin and Nagel 2022). In terms of analyzing out-of-sample return predictability, in addition to assessing the statistical accuracy of return forecasts,¹ it is now routine to analyze the economic value of return predictability via asset allocation exercises. Specifically, return forecasts based on a set of predictors serve as inputs for constructing a portfolio. Portfolio performance metrics are then computed over a forecast evaluation period (and often compared to those for a benchmark portfolio) to measure the economic value of return predictability from an investment perspective.

A spate of recent studies employs a multitude of firm characteristics and machine learning methods to generate cross-sectional out-of-sample stock return forecasts (e.g., Freyberger, Neuhierl, and Weber 2020; Gu, Kelly, and Xiu 2020; Avramov, Cheng, and Metzker 2023; Han et al. 2024). They construct long-short portfolios by sorting stocks according to their return forecasts for the next month and going long (short) stocks with the highest (lowest) return forecasts. Similarly to studies of aggregate market return predictability,² these studies find that long-short portfolios based on machine learning forecasts provide substantive economic value to investors as reflected by portfolio performance metrics (especially Sharpe ratios), thereby furnishing strong evidence of cross-sectional stock return predictability.

However, the existing literature does not provide a general methodology for estimating the contributions of predictors in machine learning models to portfolio performance that

¹For example, in the context of aggregate equity market return predictability, the popular out-of-sample R^2 statistic (Fama and French 1989; Campbell and Thompson 2008) measures the proportional reduction in mean squared error for a competing forecast based on the information in a set of predictors vis-à-vis a naïve benchmark forecast that ignores the information. Han et al. (2024) develop a modified out-of-sample R^2 statistic for analyzing cross-sectional return predictability.

²See Rapach and Zhou (2022) for a review of the literature on aggregate equity market return predictability, including its economic value.

produces an exact decomposition of a performance metric. In the present paper, we fill this gap by developing a methodology based on Shapley (1953) values to directly estimate the contributions of individual or groups of predictors to a portfolio performance metric. This allows us to exactly decompose portfolio performance in terms of the underlying predictors—in essence, we “anatomize” machine learning-based portfolio performance. Based on the logic of Shapley values, our methodology provides a framework for fairly allocating the contributions of predictors in fitted prediction models with respect to the economic value that they generate. We call our new measure the *Shapley-based portfolio performance contribution*, which we denote for predictor p by SPPC_p .

Our portfolio performance decomposition based on the SPPC_p can be viewed as a machine learning model interpretation tool for finance. With the growing popularity of machine learning models, many of which are “black boxes,” numerous model interpretation tools have been developed in the statistics literature, including variable importance metrics and measures of interactions and nonlinearities (e.g., Dimopoulos, Bourret, and Lek 1995; Friedman 2001; Štrumbelj and Kononenko 2010, 2014; Goldstein et al. 2015; Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017; Greenwell, Boehmke, and McCarthy 2018; Fisher, Rudin, and Dominici 2019; Apley and Zhu 2020).³ Existing model interpretation tools are primarily designed to analyze fitted prediction models based on training sample data. While conventional model interpretation tools are informative for investigating the relevance of predictors in fitted models, in the context of analyzing the economic value of return predictability, they are not designed to directly measure how predictors influence portfolio performance per se, which is the ultimate object of interest in our context. Our new SPPC_p does this for any model in a manner that produces an exact decomposition of portfolio performance, thereby providing a valuable model interpretation tool for deepening our understanding of the roles

³See Molnar (2024) for a textbook treatment of machine learning model interpretation tools.

of individual or groups of predictors in fitted machine learning models when it comes to the economic value of return predictability.⁴

In presenting the $SPPC_p$, we begin by reviewing the use of Shapley values for interpreting fitted prediction models, as developed by Štrumbelj and Kononenko (2010, 2014) and Lundberg and Lee (2017). We use Shapley values as a foundation for our new metric, since they possess a number of properties—including local accuracy, missingness, symmetry, and linearity—that provide an especially attractive model interpretation framework. To fix ideas, we focus on a setting where we forecast individual stock returns based on firm characteristics using a pooled prediction model, as in Freyberger, Neuhierl, and Weber (2020), Gu, Kelly, and Xiu (2020), and Avramov, Cheng, and Metzker (2023). While we focus on this setting, the $SPPC_p$ can be computed for any situation where we use a fitted prediction model (or ensemble of models) to forecast asset returns, with the return forecasts serving as inputs for constructing a portfolio.

We explain how we extend conventional Shapley values to estimate the contributions of predictors to (1) an out-of-sample return forecast, (2) a portfolio return, and (3) a portfolio performance metric, resulting in the $SPPC_p$. The $SPPC_p$ inherits all of the attractive properties of Shapley values. We emphasize that the $SPPC_p$ is very flexible: it is model agnostic (i.e., it applies to any prediction model, including all types of machine learning models), it can be used for any strategy for mapping return forecasts to portfolio weights, and it can

⁴Moehle, Boyd, and Ang (2022) propose tools, some of which are based on Shapley values, that are designed to attribute portfolio performance to “features” such as rebalancing frequency, leverage limits, and ESG constraints. The $SPPC_p$ is very different, as it measures the contributions of the predictors in fitted machine learning models to portfolio performance, thereby providing insight into the underlying sources of the economic value generated by return predictability. Aleti, Bollerslev, and Siggaard (2025) propose a method based on Shapley values to measure the contributions of predictors to portfolio performance, but it only applies to linear models, while our general method can be used for any model, including nonlinear machine learning models.

be computed for any performance metric.⁵ The $SPPC_p$ is estimated via a sampling-based algorithm, and we provide computational details.

We illustrate the use of the $SPPC_p$ in an extensive empirical application investigating the economic value of forecasting individual stock returns using a machine learning model and 207 firm characteristics from Chen and Zimmermann (2022). We generate monthly out-of-sample firm-level stock return forecasts via the **XGBoost** algorithm (Chen and Guestrin 2016), a powerful machine learning technique based on decision trees that performs well in forecasting competitions in a variety of domains. We use the XGBoost forecasts to sort stocks into quintiles and construct a zero-investment portfolio that goes long (short) the fifth (first) quintile, with each leg value-weighted. The long-short portfolio based on the XGBoost forecasts performs impressively, generating annualized Sharpe and Calmar ratios of 1.80 and 1.44, respectively, for the 1973:01 to 2021:12 forecast evaluation period, both of which are well above the corresponding ratios for the aggregate market portfolio (0.47 and 0.14, respectively). The XGBoost portfolio also generates economically large and statistically significant alphas in the context of two leading multifactor models, namely, a six-factor model comprised of the five Fama and French (2015) factors and a momentum factor as well as the Hou et al. (2021) augmented q-factor model. In sum, we find that firm-level stock return forecasts based on a large number of firm characteristics and machine learning produce substantial economic value.

After placing the individual firm characteristics into 20 groups based on economic concepts, we estimate the contributions of the predictor groups to portfolio performance using the $SPPC_p$. The *Risk*, *Earnings*, *Seasonal momentum*, and *Momentum* groups play leading roles in accounting for the substantive Sharpe and Calmar ratios as well as the sizable alphas generated by the XGBoost portfolio. For example, the XGBoost portfolio increases the

⁵The $SPPC_p$ can also be used to measure the contributions of predictors to portfolio performance when machine learning approaches are used to directly estimate optimal portfolio weights (e.g., Kozak, Nagel, and Santosh 2020; Cong et al. 2022; Chen, Pelger, and Zhu 2024; Jensen et al. 2024).

Sharpe ratio by 1.33 vis-à-vis the market portfolio; the four groups together account for 0.86 (65%) of the increase. In contrast, the *Sales* and *Ownership* groups contribute negatively to portfolio performance across the different metrics.

The performance of the XGBoost portfolio tends to diminish after 2002, although it still outperforms the market portfolio, especially during business-cycle recessions. To examine how the contributions of the predictor groups to portfolio performance change over time, we use the $SPPC_p$ to estimate the group contributions for subsamples and rolling windows from the full forecast evaluation period. While the *Risk* and *Momentum* groups typically make substantial positive contributions to portfolio performance through 2002, they often make negative contributions thereafter. Groups making consistently positive and sizable contributions over the full forecast evaluation period include *Earnings*, *Seasonal momentum*, and *Investment*. Overall, the $SPPC_p$ sheds light on how the predictor groups contribute to portfolio performance, thereby improving our understanding of the role of predictors in generating economic value with respect to cross-sectional return predictability in a machine learning model. As such, the $SPPC_p$ is a valuable tool for identifying the key determinants of cross-sectional expected returns in terms of economic value.

Finally, we compare the $SPPC_p$ to an alternative approach that measures the contribution of a predictor group by breaking the connection between the predictor group and the fitted model predictions. This approach is intuitive and less computationally costly than the $SPPC_p$. However, it is not based on the logic of Shapley values and does not share their properties; for example, unlike our $SPPC_p$, it does not provide an exact decomposition of a performance metric. Furthermore, for our application, we find that the estimated contributions of the predictor groups to the portfolio performance metrics often differ substantially across the two approaches. Overall, because the two approaches generally do not provide similar results—and the $SPPC_p$ follows the logic and has the attractive properties of Shapley values—we view the benefits accruing to the $SPPC_p$ as outweighing the increase in its computational cost.

The rest of the paper is organized as follows. Section 2 provides background on the conventional use of Shapley values for model interpretation and explains our extensions, culminating in the $SPPC_p$ to estimate the contributions of predictors in fitted machine learning models to portfolio performance. Section 3 reports results for the empirical application. Section 4 concludes.

2. Methodology

This section presents our methodology for decomposing portfolio performance in terms of the underlying predictors that guide asset allocation using the $SPPC_p$. In line with our application in Section 3, we consider a setting in which we forecast individual stock returns using firm characteristics via a pooled prediction model. It is straightforward to modify the presentation in this section to accommodate other settings.

2.1. Shapley Values

As background, we begin with a description of conventional Shapley values, which are generally viewed as the most informative interpretation tool for fitted prediction models. The intuition for using Shapley values for model interpretation is to exploit the analogy between players in a cooperative game earning a payoff and the predictors in a forecasting model, where the payoff corresponds to the model’s prediction. According to the logic of Shapley values, payoffs are fairly allocated to the players in a game. In the context of prediction, we are interested in fairly allocating the contributions of the predictors to a fitted model’s prediction. This is a nontrivial task, especially for models with interactions between predictors and nonlinearities. Štrumbelj and Kononenko (2010, 2014) and Lundberg and Lee (2017) show how Shapley values can be used to allocate the contributions of the predictors to a pre-

diction made by the model. We adapt their ideas to a panel setting where a model generates predictions for individual stock returns over time based on a set of firm characteristics.⁶

In terms of notation, we index individual predictors by p , with the index set of predictors denoted by $S = \{1, \dots, P\}$. We index cross-sectional units by i and denote the index set of cross-sectional units by $C = \{1, \dots, N\}$.⁷ We denote the P -vector of firm characteristics (i.e., predictors) for stock i in period t by $\mathbf{x}_{i,t} = \begin{bmatrix} x_{1,i,t} & \cdots & x_{P,i,t} \end{bmatrix}'$, while $r_{i,t}$ denotes the return on stock i in period t . The prediction model is given by

$$r_{i,t+1} = f(\mathbf{x}_{i,t}) + \varepsilon_{i,t+1}, \quad (1)$$

where $f(\mathbf{x}_{i,t})$ is the conditional expectation (i.e., prediction) function, and $\varepsilon_{i,t}$ is a zero-mean disturbance term. The fitted model is denoted by \hat{f} . We use $W_j = \{t_{j,\text{start}}, \dots, t_{j,\text{end}} - 1\}$ to represent the window of panel data observations used to train the model. The prediction model can be estimated using an expanding or rolling window along the time dimension; for the former (latter), the cardinality of W_j increases (remains constant). We denote the prediction function evaluated at instance $\mathbf{x}_{i,t}$ and trained using window W_j by $\hat{f}(\mathbf{x}_{i,t}; W_j)$.

The Shapley value measures the marginal contribution of the predictor $x_{p,i,t}$ to the prediction $\hat{f}(\mathbf{x}_{i,t}; W_j)$ given $S \setminus \{p\}$ (i.e., given the presence of all of the other predictors in the model). By relying on insights from coalitional game theory, Shapley values fairly allocate the marginal contributions among the individual predictors. Formally, adapting Štrumbelj and Kononenko (2010, 2014) to our panel data framework, we can express the Shapley value for predictor p and instance $\mathbf{x}_{i,t}$ for a prediction model trained using the panel data

⁶In this section, we focus on regression prediction. The methodology can be straightforwardly applied to classification prediction. We consider both regression and classification in our empirical application in Section 3.

⁷For notational simplicity, we assume that the number of cross-sectional units is the same each period. In our empirical application in Section 3, the number of cross-sectional units (i.e., firms) changes over time. It is straightforward to modify the notation to allow for time variation in the number of cross-sectional units.

observations in window W_j as

$$\phi_p(\mathbf{x}_{i,t}; W_j) = \sum_{Q \subseteq S \setminus \{p\}} \frac{|Q|!(P - |Q| - 1)!}{P!} [\xi_{Q \cup \{p\}}(\mathbf{x}_{i,t}; W_j) - \xi_Q(\mathbf{x}_{i,t}; W_j)] \quad (2)$$

for $p \in S$, $i \in C$, and $t \in W_j$, where Q is a subset of predictors (i.e., a coalition), $Q \subseteq S \setminus \{p\}$ constitutes the set of all possible coalitions of $P - 1$ predictors in S that exclude predictor p , $|Q|$ is the cardinality of Q ,

$$\xi_Q(\mathbf{x}_{i,t}; W_j) = \mathbb{E} \left[\hat{f} \mid X_{k,i,t} = x_{k,i,t} \forall k \in Q; W_j \right] \quad (3)$$

is the value function, and \mathbb{E} is the expectation operator. Equation (3) is the prediction of the fitted model conditional on the predictors in the coalition Q , with the predictors not in Q integrated out. Accordingly, the expression in brackets in Equation (2), $\xi_{Q \cup \{p\}}(\mathbf{x}_{i,t}; W_j) - \xi_Q(\mathbf{x}_{i,t}; W_j)$, is the change in the prediction of the fitted model when we condition on the predictors in the coalition Q and predictor p relative to when we condition on the predictors in Q only. Equation (2) takes a weighted average of the changes in the value function for all possible coalitions of $P - 1$ predictors that exclude p . The change in the value function receives the weight $|Q|!(P - |Q| - 1)!/P!$, where the weights sum to one. In sum, to measure the marginal contribution of p to the prediction corresponding to the instance $\mathbf{x}_{i,t}$, Shapley values rely on coalitions to control for the presence of the other predictors in the model.

The Shapley value in Equation (2) is characterized by local accuracy (or efficiency):

$$\sum_{p \in S} \phi_p(\mathbf{x}_{i,t}; W_j) = \hat{f}(\mathbf{x}_{i,t}; W_j) - \mathbb{E} \left[\hat{f}; W_j \right] \quad (4)$$

for $i \in C$ and $t \in W_j$, where $\mathbb{E} \left[\hat{f}; W_j \right]$ is the baseline prediction corresponding to the unconditional expectation of \hat{f} . This is a natural baseline for a prediction model, as it is the forecast based on the empty coalition set. According to Equation (4), the model prediction

corresponding to instance $\mathbf{x}_{i,t}$ can be exactly decomposed (in terms of the deviation from the baseline prediction) into the sum of the Shapley values for the individual predictors for that instance. Other attractive properties of Shapley values include the following.

- Missingness:

$$\forall R \subseteq S \setminus \{p\} : \xi_{R \cup \{p\}}(\mathbf{x}_{i,t}; W_j) = \xi_R(\mathbf{x}_{i,t}; W_j) \Rightarrow \phi_p(\mathbf{x}_{i,t}; W_j) = 0 \quad (5)$$

- Symmetry:

$$\begin{aligned} \forall R \subseteq S \setminus \{p, q\} : \xi_{R \cup \{p\}}(\mathbf{x}_{i,t}; W_j) = \xi_{R \cup \{q\}}(\mathbf{x}_{i,t}; W_j) \Rightarrow \\ \phi_p(\mathbf{x}_{i,t}; W_j) = \phi_q(\mathbf{x}_{i,t}; W_j) \end{aligned} \quad (6)$$

- Linearity: For any real numbers c_1 and c_2 and models $\hat{f}(\mathbf{x}_{i,t}; W_j)$ and $\hat{f}'(\mathbf{x}_{i,t}; W_j)$,

$$\begin{aligned} \phi_p \left(c_1 \left[\hat{f}(\mathbf{x}_{i,t}; W_j) + c_2 \hat{f}'(\mathbf{x}_{i,t}; W_j) \right] \right) = \\ c_1 \phi_p \left(\hat{f}(\mathbf{x}_{i,t}; W_j) \right) + c_1 c_2 \phi_p \left(\hat{f}'(\mathbf{x}_{i,t}; W_j) \right) \end{aligned} \quad (7)$$

Missingness ensures that predictors not used by the fitted model are not attributed any importance. Symmetry and linearity are intuitively appealing properties, and the latter is useful for computing Shapley values for ensembles of prediction models.

In general, it is infeasible to exactly compute the Shapley value in Equation (2) for more than a relatively small number of predictors. The issue is that the prediction function contained in the value function needs to be evaluated for all possible coalitions of predictors with and without p . Štrumbelj and Kononenko (2010, 2014) propose an algorithm using the sampling-based approach of Castro, Gómez, and Tejada (2009). We develop a refined version of their algorithm to estimate the Shapley value in Equation (2). In Section 2.2, we

extend the algorithm to estimate the contributions of the individual predictors to a portfolio performance metric.

We begin by expressing Equation (2) in the following equivalent form:

$$\phi_p(\mathbf{x}_{i,t}; W_j) = \frac{1}{P!} \sum_{\mathcal{O} \in \pi(P)} [\xi_{\text{Pre}_p(\mathcal{O}) \cup \{p\}}(\mathbf{x}_{i,t}; W_j) - \xi_{\text{Pre}_p(\mathcal{O})}(\mathbf{x}_{i,t}; W_j)] \quad (8)$$

for $p \in S$, $i \in C$, and $t \in W_j$, where \mathcal{O} is an ordered permutation for the predictor indices in S , $\pi(P)$ is the set of all ordered permutations for S , and $\text{Pre}_p(\mathcal{O})$ is the set of indices that precede p in \mathcal{O} . To implement the algorithm, we make a random draw m with replacement of an ordered permutation from $\pi(P)$, denoted by \mathcal{O}_m . Using \mathcal{O}_m , we compute the following:

$$\hat{\theta}_{p,m}(\mathbf{x}_{i,t}; W_j) = \frac{1}{|C||W_j|} \sum_{u \in C} \sum_{s \in W_j} \left[\hat{f}(x_{k,i,t} : k \in \text{Pre}_p(\mathcal{O}_m) \cup \{p\}, x_{l,u,s} : l \in \text{Post}_p(\mathcal{O}_m); W_j) - \hat{f}(x_{k,i,t} : k \in \text{Pre}_p(\mathcal{O}_m), x_{l,u,s} : l \in \text{Post}_p(\mathcal{O}_m) \cup \{p\}; W_j) \right] \quad (9)$$

for $p \in S$, $i \in C$, and $t \in W_j$, where $\text{Post}_p(\mathcal{O})$ is the set of indices that follow p in \mathcal{O} .

To integrate out the predictors not in the coalition when computing the conditional expectation in Equation (3), Equation (9) uses “background data” from the training sample (Štrumbelj and Kononenko 2010, 2014; Lundberg and Lee 2017). In effect, Equation (9) samples from the empirical marginal distribution of the training sample for the predictors not in the coalition when integrating them out. This approach represents an “interventional” Shapley value. Alternatively, sampling from the empirical conditional distribution for the predictors not in the coalition constitutes an “observational” Shapley value. In the taxonomy of Chen et al. (2020), the former corresponds with remaining “true to the model” and the latter with remaining “true to the data.” Using insights from Pearl (2009), Janzing, Minorics, and Blöbaum (2020) argue that the interventional Shapely value is more appropriate for fairly allocating the contributions across the predictors. Following the recommendation of Janzing, Minorics, and Blöbaum (2020), and because we are interested in interpreting the

fitted models that generate the out-of-sample forecasts, all of our Shapley-based measures are based on interventional Shapley values and thus remain true to the model.⁸

Using Equation (9), the estimate of the Shapley value $\phi_p(\mathbf{x}_{i,t}; W_j)$ in Equation (8) is given by

$$\hat{\phi}_p(\mathbf{x}_{i,t}; W_j) = \frac{1}{2M} \sum_{m=1}^{2M} \hat{\theta}_{p,m}(\mathbf{x}_{i,t}; W_j) \quad (10)$$

for $p \in S$, $i \in C$, and $t \in W_j$, where M is the number of draws. We increase the computational efficiency of the algorithm in two ways. First, we compute Shapley values for each predictor $p \in S$ for each random draw m (Castro, Gómez, and Tejada 2009). Second, we employ antithetic sampling as a variance-reduction device by computing $\hat{\theta}_{p,m}(\mathbf{x}_{i,t}; W_j)$ in Equation (9) for the original order of a randomly drawn ordered permutation and when the order is reversed (Mitchell et al. 2022). By construction, Equation (10) has all of the properties of Shapley values in Equations (4) to (7), including local accuracy:

$$\sum_{p \in S} \hat{\phi}_p(\mathbf{x}_{i,t}; W_j) = \hat{f}(\mathbf{x}_{i,t}; W_j) - \underbrace{\bar{f}(W_j)}_{\hat{\phi}_\emptyset(W_j)} \quad (11)$$

for $i \in C$ and $t \in W_j$, where

$$\bar{f}(W_j) = \frac{1}{|C||W_j|} \sum_{i \in C} \sum_{t \in W_j} \hat{f}(\mathbf{x}_{i,t}; W_j) \quad (12)$$

is the average in-sample prediction for the model trained using sample W_j . The average in-sample prediction corresponds to the baseline forecast based on the empty coalition set, denoted by $\hat{\phi}_\emptyset(W_j)$ in Equation (11).

⁸Furthermore, Janzing, Minorics, and Blöbaum (2020) and Sundararajan and Najmi (2020) show that, unlike the interventional Shapley value, the observational Shapley value generally does not possess the missingness property in Equation (5), so it can attribute importance to predictors not used by the model, further making interventional Shapley values more appropriate in our context.

To this point, we have followed the convention of computing Shapley values for the in-sample model predictions corresponding to the training sample observations. For developing the SPPC_p in Section 2.2, it is helpful to define the Shapley value corresponding to an out-of-sample observation. Suppose that we train a model using window W_j and generate an out-of-sample return forecast for stock i and period $t_{j,\text{end}} + 1$ based on the fitted model:

$$\hat{r}_{i,t_{j,\text{end}}+1} = \hat{f}(\mathbf{x}_{i,t_{j,\text{end}}}; W_j) \quad (13)$$

for $i \in C$. Modifying Equation (8), we define the Shapley value corresponding to the out-of-sample forecast $\hat{r}_{i,t_{j,\text{end}}+1}$ as

$$\phi_p(\mathbf{x}_{i,t_{j,\text{end}}}; W_j) = \frac{1}{P!} \sum_{\mathcal{O} \in \pi(P)} [\xi_{\text{Pre}_p(\mathcal{O}) \cup \{p\}}(\mathbf{x}_{i,t_{j,\text{end}}}; W_j) - \xi_{\text{Pre}_p(\mathcal{O})}(\mathbf{x}_{i,t_{j,\text{end}}}; W_j)] \quad (14)$$

for $p \in S$ and $i \in C$. We suitably modify the algorithm to estimate Equation (14). After making a random draw m , Equation (9) becomes

$$\hat{\theta}_{p,m}(\mathbf{x}_{i,t_{j,\text{end}}}; W_j) = \hat{r}_{i,t_{j,\text{end}}+1,m,p}(\mathbf{x}_{i,t_{j,\text{end}}}; W_j) - \hat{r}_{i,t_{j,\text{end}}+1,m,\setminus p}(\mathbf{x}_{i,t_{j,\text{end}}}; W_j) \quad (15)$$

for $p \in S$ and $i \in C$, where

$$\begin{aligned} \hat{r}_{i,t_{j,\text{end}}+1,m,p}(\mathbf{x}_{i,t_{j,\text{end}}}; W_j) = \\ \frac{1}{|C||W_j|} \sum_{u \in C} \sum_{s \in W_j} \hat{f}(x_{k,i,t_{j,\text{end}}} : k \in \text{Pre}_p(\mathcal{O}_m) \cup \{p\}, x_{l,u,s} : l \in \text{Post}_p(\mathcal{O}_m); W_j) \end{aligned} \quad (16)$$

and

$$\begin{aligned} \hat{r}_{i,t_{j,\text{end}}+1,m,\setminus p}(\mathbf{x}_{i,t_{j,\text{end}}}; W_j) = \\ \frac{1}{|C||W_j|} \sum_{u \in C} \sum_{s \in W_j} \hat{f}(x_{k,i,t_{j,\text{end}}} : k \in \text{Pre}_p(\mathcal{O}_m), x_{l,u,s} : l \in \text{Post}_p(\mathcal{O}_m) \cup \{p\}; W_j). \end{aligned} \quad (17)$$

The estimate of $\phi_p(\mathbf{x}_{i,t_j,\text{end}}; W_j)$ in Equation (14) is then given by

$$\hat{\phi}_p(\mathbf{x}_{i,t_j,\text{end}}; W_j) = \frac{1}{2M} \sum_{m=1}^{2M} \hat{\theta}_{p,m}(\mathbf{x}_{i,t_j,\text{end}}; W_j) \quad (18)$$

for $p \in S$ and $i \in C$.

The estimate of the Shapley value for the out-of-sample prediction in Equation (18) continues to be characterized by local accuracy (as well as the other properties of Shapley values):

$$\sum_{p \in S} \hat{\phi}_p(\mathbf{x}_{i,t_j,\text{end}}; W_j) = \underbrace{\hat{f}(\mathbf{x}_{i,t_j,\text{end}}; W_j)}_{\hat{r}_{i,t_j,\text{end}+1}} - \hat{\phi}_\emptyset(W_j) \quad (19)$$

for $i \in C$. Equation (19) says that we can exactly decompose the one-step-ahead out-of-sample return forecast for stock i into the contributions of the individual predictors.

2.2. Decomposing Portfolio Performance

Consider an investor who decides on their allocations across the N stocks for period $t_{j,\text{end}} + 1$ based on the set of return forecasts formed using data through period $t_{j,\text{end}}$ ($\hat{r}_{i,t_j,\text{end}+1} = \hat{f}(\mathbf{x}_{i,t_j,\text{end}}; W_j)$ for $i \in C$). The allocation to stock i generally depends on the entire set of return forecasts for $t_{j,\text{end}} + 1$, so we denote the portfolio weight for i by the function

$$w_{i,t_j,\text{end}+1} \left(\left\{ \hat{f}(\mathbf{x}_{i,t_j,\text{end}}; W_j) \right\}_{i \in C} \right) \quad (20)$$

for $i \in C$. The investor could, for example, employ a portfolio optimizer based on the return forecasts or form a long-short portfolio by going long (short) the stocks with the highest (lowest) return forecasts. Our methodology is general, so it applies to any strategy for mapping the return forecasts to the portfolio weights.

The portfolio return for $t_{j,\text{end}} + 1$ is given by

$$r_{t_{j,\text{end}}+1}^{\text{Port}} = \sum_{i \in C} w_{i,t_{j,\text{end}}+1} \left(\left\{ \hat{f}(\mathbf{x}_{i,t_{j,\text{end}}} ; W_j) \right\}_{i \in C} \right) r_{i,t_{j,\text{end}}+1}. \quad (21)$$

Using the logic of Shapley values, we decompose the portfolio return in Equation (21). In terms of the algorithm, we modify Equation (15) as follows:

$$\begin{aligned} \hat{\theta}_{p,m} \left(\left\{ \mathbf{x}_{i,t_{j,\text{end}}} \right\}_{i \in C} ; W_j \right) = & \\ & \sum_{i \in C} \left[w_{i,t_{j,\text{end}}+1} \left(\left\{ \hat{r}_{i,t_{j,\text{end}}+1,m,p}(\mathbf{x}_{i,t_{j,\text{end}}} ; W_j) \right\}_{i \in C} \right) r_{i,t_{j,\text{end}}+1} \right] - \\ & \sum_{i \in C} \left[w_{i,t_{j,\text{end}}+1} \left(\left\{ \hat{r}_{i,t_{j,\text{end}}+1,m,\setminus p}(\mathbf{x}_{i,t_{j,\text{end}}} ; W_j) \right\}_{i \in C} \right) r_{i,t_{j,\text{end}}+1} \right] \end{aligned} \quad (22)$$

for $p \in S$. Equation (22) measures the change in the portfolio return for period $t_{j,\text{end}} + 1$ when we condition on the predictors in the coalition including and excluding the predictor p . To estimate the Shapley-based contribution of predictor p to the portfolio return, Equation (18) becomes

$$\hat{\phi}_p \left(\left\{ \mathbf{x}_{i,t_{j,\text{end}}} \right\}_{i \in C} ; W_j \right) = \frac{1}{2M} \sum_{m=1}^{2M} \hat{\theta}_{p,m} \left(\left\{ \mathbf{x}_{i,t_{j,\text{end}}} \right\}_{i \in C} ; W_j \right) \quad (23)$$

for $p \in S$.

An important issue in implementing the Shapley-based decomposition of the portfolio return is deciding on the return corresponding to the empty coalition set. With regard to the empty coalition set, we have no predictors to determine the portfolio weights in Equation (20), so we need to specify the portfolio return for the empty coalition set, which we denote by $r_{t_{j,\text{end}}+1}^{\text{Base}}$, since the portfolio can be viewed as a “baseline” portfolio. We need $r_{t_{j,\text{end}}+1}^{\text{Base}}$ to compute Equation (22) when p is the first element in the ordered permutation \mathcal{O}_m ; $r_{t_{j,\text{end}}+1}^{\text{Base}}$ also appears in the local accuracy condition in Equation (24) below.

The selection of the baseline portfolio is at the discretion of the researcher. To determine the baseline portfolio, it is sensible to ask, “If I had an empty set of predictors—and so no predictor information—how would I form a portfolio?” A relevant baseline will depend on the context; for example, for a portfolio that broadly invests in equities, the CRSP value-weighted aggregate market portfolio seems a natural choice. In addition to the other properties of Shapley values, the portfolio return decomposition in Equation (23) satisfies local accuracy:

$$\sum_{p \in S} \hat{\phi}_p \left(\{ \mathbf{x}_{i,t_{j,\text{end}}} \}_{i \in C}; W_j \right) = r_{t_{j,\text{end}+1}}^{\text{Port}} - r_{t_{j,\text{end}+1}}^{\text{Base}}. \quad (24)$$

Equation (24) indicates that we can exactly decompose the portfolio return in period $t_{j,\text{end}+1}$ (in terms of the deviation from the baseline portfolio return) into the return contributions made by each of the P predictors.

Finally, we extend our approach to compute the SPPC_p . To do so, we need to take into account the entire series of out-of-sample return forecasts and corresponding portfolio returns over the forecast evaluation period. In terms of the time dimension, we assume that the sample of panel data spans T periods and that data through period T_{in} are used to train the model that generates the first set of out-of-sample return forecasts for period $T_{\text{in}} + 1$. The model is then retrained using panel data through $T_{\text{in}} + 1$ to generate the next set of return forecasts for $T_{\text{in}} + 2$. Continuing in this manner through the end of the available sample, $T - T_{\text{in}} = D$ sets of return forecast are generated, where the final model is trained using panel data through $T - 1$ to compute the last set of return forecasts for T . We define the index set of training windows used to fit the sequence of models as $W = \{1, \dots, D\}$, where $t_{j,\text{end}}$ corresponds to $T_{\text{in}}, T_{\text{in}} + 1, \dots, T - 1$ for $j = 1, 2, \dots, D$, respectively.

The key insight for computing the SPPC_p is to wrap a function corresponding to the performance metric around the portfolio returns. Denoting a performance metric function by $\mathcal{M}(\cdot)$, which depends on the sequence of portfolio returns over the forecast evaluation

period, we modify Equation (22) in the algorithm to

$$\begin{aligned} \hat{\theta}_{p,m} \left(\{ \mathbf{x}_{i,t_j,\text{end}} \}_{i \in C}; W, \mathcal{M} \right) = \\ \mathcal{M} \left(\left\{ \sum_{i \in C} \left[w_{i,t_j,\text{end}+1} \left(\{ \hat{r}_{i,t_j,\text{end}+1,m,p}(\mathbf{x}_{i,t_j,\text{end}}; W_j) \}_{i \in C} \right) r_{i,t_j,\text{end}+1} \right] \right\}_{j \in W} \right) - \\ \mathcal{M} \left(\left\{ \sum_{i \in C} \left[w_{i,t_j,\text{end}+1} \left(\{ \hat{r}_{i,t_j,\text{end}+1,m,\setminus p}(\mathbf{x}_{i,t_j,\text{end}}; W_j) \}_{i \in C} \right) r_{i,t_j,\text{end}+1} \right] \right\}_{j \in W} \right) \end{aligned} \quad (25)$$

for $p \in S$. The Shapley-based estimate of the contribution of predictor p to the portfolio performance metric is then given by

$$\underbrace{\hat{\phi}_p \left(\{ \mathbf{x}_{i,t_j,\text{end}} \}_{i \in C}; W, \mathcal{M} \right)}_{\text{SPPC}_p} = \frac{1}{2M} \sum_{m=1}^{2M} \hat{\theta}_{p,m} \left(\{ \mathbf{x}_{i,t_j,\text{end}} \}_{i \in C}; W, \mathcal{M} \right) \quad (26)$$

for $p \in S$. All of the properties of Shapley values, including local accuracy, apply to the SPPC_p in Equation (26), so the contributions of the predictors to the performance metric sum to the metric for the portfolio in excess of that for the baseline portfolio:

$$\sum_{p \in S} \text{SPPC}_p = \mathcal{M} \left(\left\{ r_{t_j,\text{end}+1}^{\text{Port}} \right\}_{j \in W} \right) - \mathcal{M} \left(\left\{ r_{t_j,\text{end}+1}^{\text{Base}} \right\}_{j \in W} \right). \quad (27)$$

The SPPC_p in Equation (26) allows a researcher to estimate how an individual predictor contributes to portfolio performance, while Equation (27) indicates that the sum of the SPPC_p estimates provides an exact decomposition of portfolio performance relative to the baseline portfolio.

In sum, the Shapley value ascertains the contribution of an individual predictor p to a prediction by forming a coalition of predictors and measuring the change in the prediction when p is included and excluded in the conditioning set based on the coalition in the fitted prediction model. The change in the value of the prediction corresponding to p is averaged

over many coalitions to arrive at the Shapley value, which fairly allocates the contribution of p to the fitted model’s prediction. We extend the logic of the Shapley value to a performance metric. To accomplish this, we measure the contribution of predictor p to the performance metric by forming a coalition of predictors and computing the change in value of the metric when p is included and excluded in the conditioning set based on the coalition in the fitted prediction model (or sequence of fitted models). We average the change in value corresponding to p over many coalitions, providing a Shapley-based approach for fairly allocating the contribution of p to portfolio performance.

We emphasize that the SPPC_p is very general. It is model agnostic, so it applies to any fitted prediction model (linear or nonlinear, parametric or nonparametric). It also accommodates any rule for mapping the return forecasts to portfolio weights as well as any performance metric.

2.3. Computation

Given the nature of Shapley values, and with more than a few predictors and panel data, estimating the SPPC_p can entail significant computational costs. The [Appendix](#) provides details on implementing the algorithm for calculating the SPPC_p and its computational costs.

An alternative approach for measuring the contribution of a predictor is to “turn off” the predictor so that it does not influence the predictions made by the sequence of fitted models. For example, it is common to scale firm characteristics to lie in the $[-1, 1]$ interval. In this case, we can turn off the predictor by setting it to the midpoint value of zero, so the predictor does not influence the predictions made by the fitted models. (For all of the remaining predictors, we use their original sample values.) We compute the portfolio performance metric with the predictor turned off and compare it to the original performance metric that uses the actual sample values for the predictor (i.e., with the predictor turned on); the difference between the two provides a measure of the predictor’s contribution to the portfolio performance metric. Similarly, we can break the connection between a predictor and the

predictions made by the sequence of fitted models by replacing the predictor value for a given time period and firm with a randomly drawn value from the training sample, in the spirit of permutation feature importance (Breiman 2001). We again compare the performance metric corresponding to the permuted predictor values to the original performance metric that uses the predictor’s actual sample values, and the difference between the two furnishes an estimate of the predictor’s contribution to portfolio performance.⁹

These alternative approaches can be less computationally costly, especially as the number of predictors becomes large. However, unlike the $SPPC_p$, they are not based on the logic of Shapley values and thus do not share their attractive properties. For example, the $SPPC_p$ is characterized by local accuracy, so it provides an exact decomposition of any portfolio performance metric with respect to the set of predictors, while the alternative approaches do not.¹⁰ This may not be a significant concern if an alternative approach and the $SPPC_p$ provide similar measures of the predictor contributions. We examine this issue in the context of our empirical application in Section 3.5.

3. Empirical Application

In this section, we use the $SPPC_p$ from Section 2 to analyze a leading question in empirical asset pricing: Which types of firm characteristics are important for determining cross-sectional expected stock returns? In line with recent studies, we analyze the relevance of firm characteristics using out-of-sample tests (e.g., Lewellen 2015; Green, Hand, and Zhang 2017; Freyberger, Neuhierl, and Weber 2020; Gu, Kelly, and Xiu 2020; Avramov, Cheng, and Metzker 2023; Han et al. 2024). As in these studies, we forecast one-month-ahead cross-sectional returns using a large number of firm characteristics. The cross-sectional return

⁹In the context of developing a framework for constructing an “implementable efficient frontier” via machine learning, Jensen et al. (2024) use this approach to develop an “economic feature importance” measure of how a feature (i.e., predictor) affects realized utility.

¹⁰Cong et al. (2022) develop a convenient “economic distillation” approach based on a polynomial approximation to measure the importance of predictors for portfolio performance. Like the other alternative methods discussed here, the approach does not have the properties of Shapley values, including local accuracy.

forecasts then serve as inputs to form a zero-investment long-short portfolio that is used to measure the economic value of cross-sectional return predictability.

3.1. Data

We use data for a large set of firm characteristics from Chen and Zimmermann (2022), which are available at the [Open Source Asset Pricing](#) website. The data are comprised of 207 firm characteristics from the voluminous literature on cross-sectional expected returns. We use data spanning 1960:01 to 2021:12 (744 months). Following Freyberger, Neuhierl, and Weber (2020) and Gu, Kelly, and Xiu (2020), we transform each characteristic each month by cross-sectionally ranking the characteristics and then mapping the ranks into the $[-1, 1]$ interval. Monthly firm-level stock return data are from CRSP. We consider all firms listed on NYSE, AMEX, and NASDAQ with a market value on CRSP at the end of the previous month and a non-missing value for common equity in the firm’s annual financial statement. We compute the excess return for each stock in a given month using the CRSP risk-free return.¹¹

Table 1 lists the 207 firm characteristics from Chen and Zimmermann (2022) along with their descriptions from [Open Source Asset Pricing](#). In an effort to keep computational costs more manageable, we consolidate the predictors into 20 groups based on economic concepts.¹² We use the 34 categories in Chen and Zimmermann (2022) as a starting point and make various adjustments to arrive at our 20 groups, with a goal of having groups that are reasonably similar in size. As shown in Table 1, the number of characteristics in a group ranges from six (*Volume*) to 14 (*Investment* and *Profitability*), so no groups are inordinately larger than others. Defining the groups quite narrowly enables us to differentiate between

¹¹As in Gu, Kelly, and Xiu (2020), we fill in missing values for a firm characteristic in a given month with the cross-sectional median for the available characteristic observations for that month.

¹²Harvey, Liu, and Zhu (2016), McLean and Pontiff (2016), Freyberger, Neuhierl, and Weber (2020), and Hou, Xue, and Zhang (2020) categorize characteristics into five to six groups based on economic concepts. We specify a larger number of more narrowly defined economic groups, as we use the $SPPC_p$ to estimate the contributions of the groups to various performance metrics.

Table 1. Characteristic groups

The table provides groups for 207 firm characteristics from Chen and Zimmermann (2022) used in the empirical application in Section 3. The characteristics are grouped according to 20 economic categories. The characteristic descriptions are from the [Open Source Asset Pricing](#) website; more information on the data and their sources is provided there.

(1) Description	(2) Description	(3) Description
Panel A: Earnings (9)		
Excluded expenses	Earnings consistency	Earnings streak length
Earnings announcement return	Earnings surprise streak	Analyst earnings per share
Decline in analyst coverage	Earnings surprise	Earnings-to-price ratio
Panel B: Earnings forecast (10)		
Earnings forecast to price	Long-vs-short EPS forecasts	Earnings forecast revisions
EPS forecast revision	Long-term EPS forecast	Up forecast
Δ in forecast and accrual	Predicted analyst forecast error	EPS forecast dispersion
Down forecast EPS		
Panel C: Financing (10)		
Convertible debt indicator	Net equity financing	Leverage component of BM
Δ in current operating liabilities	Net external financing	Market leverage
Δ in financial liabilities	Book leverage (annual)	Net debt to price
Net debt financing		
Panel D: Financing alt (7)		
Composite equity issuance	Initial public offerings	Share issuance (5 year)
Composite debt issuance	Share issuance (1 year)	Spinoffs
Debt issuance		

types of characteristics in terms of a wide variety of economic concepts. All of the results for the $SPPC_p$ in Section 2.2 for individual predictors also hold for predictor groups.

3.2. Portfolio Construction and Prediction Models

Similarly to a number of recent studies (e.g., Freyberger, Neuhierl, and Weber 2020; Gu, Kelly, and Xiu 2020; Avramov, Cheng, and Metzker 2023; Han et al. 2024), we construct a zero-investment long-short portfolio that goes long (short) stocks with the highest (lowest) machine learning return forecasts for the next month. We use 1960:01 to 1972:12 (156 months) as the initial in-sample estimation period and generate firm-level out-of-sample

Table 1 (continued)

(1) Description	(2) Description	(3) Description
Panel E: Investment (14)		
Cash to assets	Growth in book equity	Investment to revenue
Net operating assets	Δ in equity to assets	Δ in PPE and inv/assets
Real estate holdings	Δ in long-term investment	Growth in advertising exp
Tangibility	Δ in net operating assets	Advertising expense
Asset growth	Growth in long term operating assets	
Panel F: Investment alt (12)		
Δ in capital inv (ind adj)	Inventory growth	Deferred revenue
Δ in capex (2 years)	Δ in net noncurrent op assets	Δ in net financial assets
Δ in capex (3 years)	Δ in net working capital	Employment growth
Brand capital investment	Δ in current operating assets	Total accruals
Panel G: Lead lag (9)		
Customer mom	Customers mom	Price delay coeff
Earnings surprise of big firms	Suppliers mom	Price delay SE adjusted
Industry return of big firms	Price delay R square	Conglomerate return
Panel H: Liquidity (11)		
Pastor-Stambaugh liquidity beta	Probability of informed trading	Days with zero trades (v1)
Bid-ask spread	Size	Days with zero trades (v2)
Amihud's illiquidity	Share turnover volatility	Days with zero trades (v3)
Price		
Panel I: Momentum (11)		
Firm age—mom	Mom (12 month)	Mom in high volume stocks
52 week high	Mom (6 month)	Mom based on FF3 resid
Industry mom	Junk stock mom	Trend factor
Intermediate mom	Mom and LT reversal	
Panel J: Ownership (11)		
Sin stock (selection criteria)	Takeover vulnerability	Inst own and idio vol
Active shareholders	Inst own and forecast dispersion	Short interest
Breadth of ownership	Inst own and market to book	Governance index
Inst own among high short interest	Inst own and turnover	

return forecasts and long-short portfolio returns for 1973:01 to 2021:12 (588 months). To keep the fitted prediction model timely when generating the one-month-ahead firm-level return forecasts, we retrain the model each month as additional data become available using a rolling window.

Table 1 (continued)

(1) Description	(2) Description	(3) Description
Panel K: Profitability (14)		
Mohanram G-score	Gross profits / total assets	Return on assets (qtrly)
Piotroski F-score	Inventory growth	Net income / book equity
Cash productivity	Operating profits / book equity	Taxable income to income
Cash-based operating profitability	Operating profitability R&D adjusted	O score
Δ in taxes	Operating leverage	
Panel L: R&D (8)		
Citations to R&D expenses	Organizational capital	IPO and no R&D spending
Patents to R&D expenses	R&D over market cap	Unexpected R&D increase
R&D capital-to-assets	R&D ability	
Panel M: Reversal (7)		
Intangible return using BM	Intangible return using Sale2P	Medium-run reversal
Intangible return using CFtoP	Long-run reversal	Short-term reversal
Intangible return using EP		
Panel N: Risk (12)		
Frazzini-Pedersen beta	Coskewness	Idio risk
CAPM beta	Return skewness	Idio risk (3F model)
Tail risk beta	Idio skewness (3F model)	Idio risk (AHT)
Coskewness using daily returns	Systematic volatility	Max return over month
Panel O: Risk alt (12)		
Real dirty surplus	Industry concentration (equity)	IPO and age
Pension funding status	Volatility smirk near the money	Credit Rating Downgrade
Industry concentration (sales)	Put volatility minus call volatility	Exchange Switch
Industry concentration (assets)	Cash-flow to price variance	Firm age based on CRSP

We compute return forecasts using both classification and regression prediction models. For the classification model, there are five classes, from the bottom 20% to the top 20% of stocks in terms of their returns. We describe the classification and regression prediction models in more detail below, after explaining how we construct the long-short portfolios.

To construct the long-short portfolio for month $t+1$ based on information through month t , we proceed as follows. We generate return forecasts for all available stocks for month $t+1$ using data through month t . Before forming the portfolio, to limit the role of small-cap stocks, we drop stocks with market capitalization below the NYSE 20th percentile at the

Table 1 (continued)

(1) Description	(2) Description	(3) Description
Panel P: Sales (10)		
Δ in asset turnover	Percent operating accruals	Sales growth / overhead growth
Abnormal accruals	Percent total accruals	Revenue growth rank
Accruals	Sales growth / inventory growth	Order backlog
Δ in order backlog		
Panel Q: Seasonal momentum (10)		
Mom without the seasonal part	Off season reversal years 16–20	Return seasonality years 11–15
Off season long-term reversal	Return seasonality years 2–5	Return seasonality years 16–20
Off season reversal years 6–10	Return seasonality years 6–10	Return seasonality last year
Off season reversal years 11–15		
Panel R: Valuation (12)		
Predicted div yield next month	Dividend seasonality	Δ in recommendation
Efficient frontier index	Share repurchases	Consensus recommendation
Dividend initiation	Equity duration	Analyst recs and short-interest
Dividend omission	Analyst optimism	Analyst value
Panel S: Valuation ratio (11)		
Book-to-market and accruals	Cash flow to market	Sales-to-price
Total assets to market	Enterprise component of BM	Net payout yield
Book to market most recent ME	Enterprise multiple	Payout yield
Book to market December ME	Sales-to-price	
Panel T: Volume (6)		
Past trading volume	Option volume to average	Volume to market equity
Option to stock volume	Share volume	Volume trend

end of month t . For the classification model, we take long (short) positions in those stocks predicted to be in the top (bottom) class in month $t + 1$.¹³ To further limit the role of small-cap stocks, the long and short legs are value-weighted. We scale the weights in the long and short legs to sum to 1 and -1 , respectively. For the regression model, we sort stocks in terms of their return forecasts and take long (short) positions in those stocks in the top (bottom) 20% of sorted stocks. We again value weight the long and short legs and scale them to sum to 1 and -1 , respectively.

¹³For the classification model, note that the number of stocks in the long leg does not necessarily equal the number in the short leg.

We generate monthly firm-level out-of-sample return forecasts based on the 207 firm characteristics listed in Table 1 using the XGBoost algorithm (Chen and Guestrin 2016). XGBoost is based on decision trees, which allow for nonlinearities in predictive relations via multiway interactions and higher-order effects of predictors. A decision tree partitions the predictor space into non-overlapping regions and assigns a prediction (or score) for the target in each region. The classification and regression tree (CART) algorithm (Breiman et al. 1984) is typically used to partition the predictor space by applying a sequence of splitting rules. The split at the top of a tree is the “root node,” subsequent splits are “internal nodes,” and the final set of subgroups that define the predictive regions at the bottom of the tree are the “terminal” or “leaf nodes.” Decision trees can be used for both classification and regression problems. For a classification (regression) problem, the prediction is the class with the highest probability (average value of the target observations) in a given leaf node.

XGBoost employs gradient boosting (Breiman 1997; Friedman 2001), which entails constructing an ensemble prediction function additively, where each function in the sequence is a relatively simple model; in the case of a decision tree, each function is a “shallow” tree. Simple models typically have low variance but relatively high bias. Gradient boosting seeks to lower the bias and thus improve out-of-sample performance in light of the bias-variance trade-off by fitting a decision tree to the residuals for the previous tree in the sequence. To help guard against overfitting and further improve out-of-sample performance, stochastic gradient boosting (Friedman 2002) refines conventional gradient boosting by using a randomly drawn subsample of the training data when fitting each decision tree in the sequence. XGBoost is a well known and powerful algorithm that employs stochastic gradient boosting to fit prediction models based on decision trees.¹⁴ It is a leading performer in forecasting competitions and compares favorably to other popular machine learning methods, particu-

¹⁴For the classification (regression) problem, we use log loss (mean squared error) as the objective function when training the prediction model.

larly when the data are “tabular,” as in our application (e.g., Elsayed et al. 2021; Grinsztajn, Oyallon, and Varoquaux 2022).

An important step for improving out-of-sample performance is tuning the hyperparameters for the XGBoost algorithm.¹⁵ We use a “walk-forward” procedure that respects the time-series dimension of the panel data to tune the hyperparameters. When computing the first set of out-of-sample return forecasts for 1973:01, we reserve the last 36 months (1970:01 to 1972:12) of the 1960:01 to 1972:12 initial in-sample estimation sample as a validation period for tuning the hyperparameters. We first train prediction models via XGBoost using data for 1960:01 to 1969:12 and the different combinations of the hyperparameter values. We plug the characteristic values for 1969:12 into the fitted models to generate return forecasts for 1970:01 and compute long-short portfolio weights and the associated portfolio return corresponding to the different combinations of hyperparameter values. Next, we train prediction models via XGBoost using data for 1960:01 to 1970:01 and the different combinations of the hyperparameter values, plug the characteristic values for 1970:01 into the fitted models to generate return forecasts for 1970:02, and compute long-short portfolio weights and the associated portfolio return corresponding to the different combinations of hyperparameter values. We continue in this manner through the end of the validation period and compute Sharpe ratios over the validation period for the long-short portfolio returns corresponding to the different combinations of hyperparameter values. We select the combination of hyperparameter values that produces the highest Sharpe ratio over the validation period. Then, using the selected combination of hyperparameter values and the 1960:01 to 1972:12 sample, we train the prediction model via XGBoost, plug the 1972:12 characteristic values into

¹⁵We tune the following XGBoost hyperparameters: “max_depth,” “reg_alpha,” “reg_lambda,” “subsample,” “colsample_bytree,” “min_child_weight”; we set “n_estimators” to 100. See the [documentation](#) on “XGBoost Parameters” for details on the hyperparameters. Because we tune a large number of hyperparameters and consider a wide grid of values for each, it is very computationally expensive to consider all possible combinations of hyperparameter values. Instead of an exhaustive search, we use [Optuna](#) (Akiba et al. 2019), which employs the Tree-structured Parzen Estimator (TPE) algorithm to conduct a smart search. This substantially reduces computational costs while still making it likely that the selected combination of hyperparameter values is nearly optimal.

the fitted model, and generate the set of return forecasts for the first out-of-sample month (1973:01).

To generate the next set of out-of-sample return forecasts for 1973:02, we use 1960:02 to 1973:01 as the estimation period (so we use a rolling estimation window). We tune the hyperparameters using the walk-forward procedure, with the last 36 months of the estimation period serving as the validation sample. Using the tuned hyperparameters and the 1960:02 to 1973:01 estimation sample, we train the prediction model via XGBoost, plug the 1973:01 characteristic values into the fitted model, and produce the set of return forecasts for 1973:02. Continuing in this fashion, we generate firm-level return forecasts based on the 207 characteristics and XGBoost for each month of the forecast evaluation period. We compute return forecasts using both classification and regression prediction models, which we denote by XGBoost(c) and XGBoost(r), respectively. The firm-level monthly return forecasts serve as inputs for constructing the long-short portfolios, as described above. The return forecasts and long-short portfolio weights are based on information available at the time of forecast formation so that there is no “look-ahead” bias in the long-short portfolio returns. As shown in Section 3.3, the long-short portfolio based on XGBoost(c) performs very well in our empirical application, so the XGBoost algorithm provides an informative machine learning device for analyzing the contributions of cross-sectional stock return predictors to portfolio performance.

3.3. Decomposing Portfolio Performance

Table 2 reports performance metrics for the long-short portfolios based on the XGBoost(c) and XGBoost(r) return forecasts for the 1973:01 to 2021:12 forecast evaluation period. For reference, metrics are also reported for the CRSP value-weighted aggregate market portfolio. The long-short XGBoost(c) portfolio delivers impressive performance. Its annualized mean return is 22.58%, while its annualized volatility is 12.53%, which compare to values of 7.44% and 15.86%, respectively, for the market portfolio excess return. The mean and volatility

for the XGBoost(c) portfolio translate into an annualized Sharpe ratio of 1.80, which is nearly four times larger than that for the market portfolio (0.47). The maximum drawdown for the XGBoost(c) portfolio is only 15.70%, leading to an annualized Calmar ratio of 1.44. Again, the maximum drawdown and Calmar ratio compare favorably to those for the market portfolio (54.36% and 0.14, respectively).

Table 2. Portfolio performance

The table reports performance metrics for a zero-investment long-short portfolio that invests in available stocks in a given month. The long-short portfolio is constructed by sorting stocks according to their excess return forecasts for the available stocks in a given month based on the XGBoost(c) or XGBoost(r) model. The excess return forecasts are based on the 207 firm characteristics listed in Table 1. Before forming the portfolio, stocks with market capitalization below the NYSE 20th percentile are dropped. The XGBoost(c) portfolio goes long (short) stocks predicted to be in the top (bottom) quintile of excess returns; the XGBoost(r) portfolio goes long (short) the 20% of stocks with the highest (lowest) excess return forecasts. The long and short legs are value-weighted. The forecast evaluation period is 1973:01 to 2021:12. “MDD” is the maximum drawdown. “Ann. FF6 alpha” is the annualized alpha for a multifactor model that includes the five Fama and French (2015) factors and a momentum factor; “Ann. Q5 alpha” is the annualized alpha for the Hou et al. (2021) augmented q-factor model; *t*-statistics for the alphas are in brackets; *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. “Market” is the CRSP value-weighted market portfolio.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Portfolio	Ann. mean	Ann. volatility	Ann. Sharpe ratio	MDD	Ann. Calmar ratio	Ann. FF6 alpha	Ann. Q5 alpha
XGBoost(c)	22.58%	12.53%	1.80	15.70%	1.44	19.45% [9.82]***	16.29% [6.93]***
XGBoost(r)	9.33%	13.06%	0.71	36.12%	0.26	5.02% [2.75]***	3.15% [1.11]
Market	7.44%	15.86%	0.47	54.36%	0.14	0%	0%

To examine if the XGBoost(c) portfolio generates significant risk-adjusted returns, we estimate its alphas in the context of two leading multifactor models. The first is a six-factor model comprised of the five Fama and French (2015) factors and a momentum factor (FF6). The second is the Hou et al. (2021) augmented q-factor model (Q5), which adds an

expected growth factor to the four factors from the original q-factor model (Hou, Xue, and Zhang 2015).¹⁶ The XGBoost(c) portfolio generates economically sizable annualized alphas of 19.45% and 16.29% for the FF6 and Q5 multifactor models, respectively. Both alpha estimates are significant at the 1% level. This indicates that exposures to popular systematic risk factors from the literature cannot account for the average return of the XGBoost(c) portfolio. Indeed, the risk-adjusted average returns (i.e., alphas) for the XGBoost(c) portfolio in the last two columns of Table 2 are reasonably close to its unadjusted average return in the second column.

Although the long-short portfolio based on the XGBoost(r) return forecasts outperforms the market portfolio in Table 2, it does not perform as well as the long-short XGBoost(c) portfolio. It is perhaps not surprising that the XGBoost(c) model works better for constructing the long-short portfolio than the XGBoost(r) model. XGBoost(c) is a classification model, and forming the long-short portfolio entails classification, as the portfolio goes long (short) stocks in the top (bottom) quintile in terms of their expected returns. Of course, we need return forecasts based on a regression model in some situations, such as constructing mean-variance optimal portfolios, and the $SPPC_p$ can be used to estimate the contributions of the return predictors to portfolio performance in this context.

Overall, the results in Table 2 show that the XGBoost(c) return forecasts based on the 207 firm characteristics serve as valuable inputs for constructing a long-short portfolio. In what follows, we employ the $SPPC_p$ in Equation (26) to estimate the contributions of each of the 20 predictor groups in Table 1 to the performance metrics for the XGBoost(c) portfolio. While we focus on decomposing the performance metrics for the XGBoost(c) portfolio, we reiterate that the $SPPC_p$ can be used to decompose the performance of any portfolio constructed from return forecasts based on any model and set of predictors. If a portfolio performs

¹⁶The factor data for the FF6 (augmented q-factor) model are from Kenneth French’s [Data Library](#) (Lu Zhang’s [website](#)).

poorly, our method can identify the predictors that are primarily responsible for the subpar performance.

As discussed in Section 2.2, to operationalize our method, we need to select the return for the portfolio corresponding to the empty coalition set, which serves as a baseline. If an investor does not have access to any predictors (the empty coalition set), it seems reasonable to simply hold the market portfolio; thus, we use the excess return for the CRSP value-weighted market portfolio as the baseline. Observe that, because the market excess return appears as a factor in both the FF6 and Q5 multifactor models, the alpha for the baseline portfolio is zero by construction.

Table 3 reports the contributions of the 20 predictor groups to the performance metrics for the XGBoost(c) portfolio. The top four contributions for each metric in terms of improving portfolio performance are in bold. According to the local accuracy property in Equation (27), the baseline contribution and those of the 20 predictor groups sum to the total in the last row of Table 3 (apart from rounding), where the last row equals the corresponding value for the XGBoost(c) portfolio in Table 2.

Beginning with the mean return in Table 3, the baseline value for the market portfolio is 7.44%. Since the mean return for the XGBoost(c) portfolio is 22.58%, the 20 predictor groups together increase the average return by 15.14 percentage points. The top four predictor groups are *Risk*, *Momentum*, *Earnings*, and *Seasonal momentum*, with contributions of 482, 450, 250, and 158 basis points, respectively. These four groups collectively contribute to an increase in the average return of 13.40 percentage points, which is nearly 90% of the increase provided by the XGBoost(c) portfolio. Other groups making contributions above 50 basis points are *Lead lag*, *Investment*, *Profitability*, and *Earnings forecast* (113, 113, 97, and 89 basis points, respectively). A handful of groups make sizably negative contributions to the average return, including *Reversal*, *Ownership*, *Volume*, and *Sales* (−166, −96, −70, and −50 basis points, respectively). With respect to volatility, the XGBoost(c) portfolio lowers it by 3.33 percentage points vis-à-vis the market portfolio. The largest contributors to

Table 3. Portfolio performance contributions

The table reports the contributions of the 20 predictor groups to the performance metrics in Table 2 for the long-short XGBoost(c) portfolio. Table 1 lists the individual firm characteristics in each of the 20 predictor groups. The forecast evaluation period is 1973:01 to 2021:12. The numbers in a column may not add to the value in the “Total” row due to rounding; 0.00 indicates less than 0.005 in absolute value. The top four contributions for each metric in terms of improving portfolio performance are in bold.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Predictor group	Ann. mean	Ann. vol.	Ann. Sharpe ratio	MDD	Ann. Calmar ratio	Ann. FF6 alpha	Ann. Q5 alpha
Baseline	7.44%	15.86%	0.47	54.36%	0.14	0%	0%
Risk	4.82	-0.16	0.35	-3.03	0.24	4.34	4.29
Earnings	2.50	-0.29	0.20	-3.54	0.10	2.72	2.02
Seasonal momentum	1.58	-0.85	0.16	-7.72	0.14	2.38	2.49
Momentum	4.50	2.34	0.15	4.58	0.08	3.25	2.16
Lead lag	1.13	-0.52	0.10	-3.85	0.06	0.95	0.57
Investment	1.13	-0.26	0.10	-6.78	0.11	1.70	0.95
Valuation ratio	0.28	-1.24	0.09	-6.63	0.15	0.38	0.58
Risk alt	0.48	-0.63	0.09	-7.09	0.17	1.27	0.76
Profitability	0.97	-0.35	0.06	-3.38	0.07	0.61	-0.88
Earnings forecast	0.89	0.26	0.05	0.76	0.04	0.85	0.92
Valuation	0.04	-1.06	0.05	-3.73	0.02	0.77	0.48
Financing	0.21	0.02	0.02	-0.12	0.02	0.14	0.29
Financing alt	0.25	-0.09	0.02	-0.69	0.02	0.70	0.29
Volume	-0.70	-1.18	0.02	-3.17	0.04	-0.43	0.17
Liquidity	0.18	1.26	0.02	3.47	0.08	-0.19	0.35
Investment alt	-0.06	0.06	0.00	-1.78	0.02	0.15	0.08
R&D	0.06	0.07	0.00	-0.87	0.00	0.42	-0.03
Reversal	-1.66	-0.38	-0.03	3.11	-0.02	-0.32	1.46
Sales	-0.50	0.00	-0.04	1.01	-0.02	-0.18	-0.65
Ownership	-0.96	-0.32	-0.06	0.79	-0.03	-0.09	-0.01
Total	22.58%	12.53%	1.80	15.70%	1.44	19.45%	16.29%

the reduction in volatility are *Valuation ratio*, *Volume*, *Valuation*, and *Seasonal momentum* (-124, -118, -106, and -85 basis points, respectively).

In terms of the Sharpe ratio in Table 3, the 20 predictor groups collectively increase the Sharpe ratio by a considerable 1.33 for the XGBoost(c) portfolio relative to the market portfolio (0.47 to 1.80). *Risk*, *Earnings*, *Seasonal momentum*, and *Momentum* are the top four predictor groups, with contributions of 0.35, 0.20, 0.16, and 0.15, respectively. Together, these four groups contribute to an increase in the Sharpe ratio of 0.86 or 65% of the total increase. The remaining predictor groups can be divided into three categories. First, there are groups making smaller but still noteworthy contributions in the range of 0.05 to 0.10, led by *Lead lag* and *Investment*, both with contributions of 0.10. Second, a set of groups makes minor contributions of essentially zero to 0.02. Third, three of the groups that contributed negatively to the average return—*Reversal*, *Sales*, and *Ownership*—also make negative contributions to the Sharpe ratio (−0.03, −0.04, and −0.06, respectively).

The XGBoost(c) portfolio reduces the maximum drawdown for the market portfolio from 54.36% to 15.70%, a substantive reduction of 38.66 percentage points. The groups most responsible for the decrease include *Seasonal momentum*, *Risk alt*, *Investment*, and *Valuation ratio*, with contributions of −772, −709, −678, and −663 basis points, respectively. Other groups lowering the maximum drawdown by more than three percentage points are *Lead lag*, *Valuation*, *Earnings*, *Profitability*, *Volume*, and *Risk* (−385, −373, −354, −338, −317, and −303 basis points, respectively).

The results for the Calmar ratio are broadly similar to those for the Sharpe ratio. The XGBoost(c) portfolio increases the Calmar ratio from 0.14 for the market portfolio to 1.44, a substantial increase of 1.30. The *Risk* group makes the largest contribution (0.24) to the increase in the Calmar ratio. In addition to *Risk*, groups making contributions of 0.10 or more include *Risk alt*, *Valuation ratio*, *Seasonal momentum*, *Investment*, and *Earnings* (0.17, 0.15, 0.14, 0.11, and 0.10, respectively). Like the Sharpe ratio, *Reversal*, *Sales*, and *Ownership* make negative contributions (−0.02, −0.02, and −0.03, respectively).

Turning to the alphas for the FF6 and Q5 multifactor models in Table 3, the top four predictor groups in terms of contributions to the Sharpe ratio also comprise the top four

in terms of contributions to the FF6 and Q5 alphas. For the FF6 model, *Risk*, *Momentum*, *Earnings*, and *Seasonal momentum* contribute 434, 325, 272, and 238 basis points, respectively, to alpha; together, these groups account for 12.69 percentage points of the total alpha for the XGBoost(c) portfolio of 19.45%. Other groups contributing more than 50 basis points include *Investment*, *Risk alt*, *Lead lag*, *Earnings forecast*, *Valuation*, *Financing alt*, and *Profitability* (170, 127, 95, 85, 77, 70, and 61 basis points, respectively). *Reversal*, *Sales*, and *Ownership* continue to make negative contributions (−32, −18, and −9 basis points, respectively), along with *Volume* and *Liquidity* (−43 and −19 basis points, respectively).

For the Q5 model, *Risk*, *Seasonal momentum*, *Momentum*, and *Earnings* provide contributions of 429, 249, 216, and 202 basis points, respectively, to alpha. These contributions constitute 10.96 percentage points of the total alpha of 16.29%. Other noteworthy contributions in excess of 50 basis points are made by *Investment*, *Earnings forecast*, *Risk alt*, *Valuation ratio*, and *Lead lag* (95, 92, 76, 58, and 57 basis points, respectively). *Sales* and *Ownership* again make negative contributions (−65 and −1 basis points, respectively), as do *Profitability* and *R&D* (−88 and −3 basis points, respectively).

Overall, the $SPPC_p$ values reported in Table 3 enable us to identify the predictor groups that are primarily responsible for the strong performance of the XGBoost(c) portfolio according to the metrics reported in Table 2. Four groups—*Risk*, *Earnings*, *Seasonal momentum*, and *Momentum*—stand out in Table 3 as leading contributors to the economic value provided by cross-sectional return predictability. These groups make the largest contributions to the Sharpe and Calmar ratios as well as the alphas for the FF6 and Q5 multifactor models. There are also groups that consistently detract from portfolio performance, especially *Sales* and *Ownership*.

Researchers are often interested in how the long and short legs separately affect the performance of long-short portfolios. We can use the $SPPC_p$ to address this issue. As an example, Figure 1 presents waterfall diagrams depicting the contributions of the predictor groups to the FF6 and Q5 multifactor model alphas for the long-short XGBoost(c) portfolio

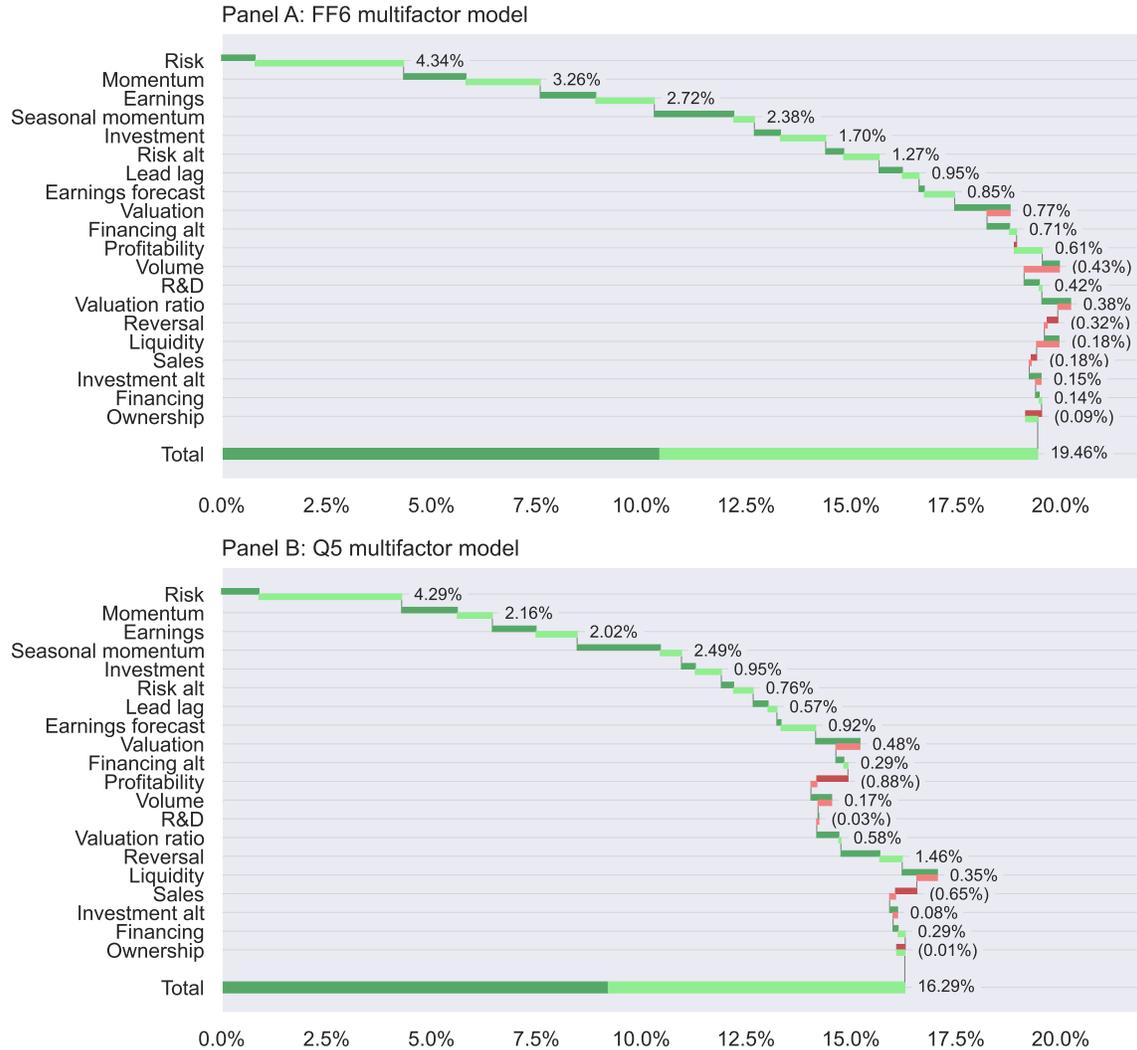


Figure 1. Alpha long- and short-leg contributions. Each panel depicts a waterfall diagram with the contributions of the 20 predictor groups to the annualized alpha for the long-short XG-Boost(c) portfolio in terms of the long and short legs. The dark (light) green segments are the positive contributions of the long (short) leg; the dark (light) red segments are the negative contributions of the long (short) leg. The numbers correspond to the contributions of the long and short legs together reported in the last two columns of Table 3; parentheses indicate a negative number.

in terms of the long and short legs separately. The diagram includes the group contributions to the alphas reported in the last two columns of Table 3 (with negative numbers in parentheses). Panel A (B) of Figure 1 reports results for the FF6 (Q5) model. The dark (light)

green segments are the positive contributions to the risk-adjusted return on the long (short) leg¹⁷; the dark (light) red segments are the negative contributions to the risk-adjusted return on the long (short) leg. The base of the waterfall shows the total contributions of the long and short legs to the alpha. Overall, for both multifactor models, the long and short legs contribute approximately equally to the long-short portfolio alpha.

There are interesting contrasts in the contributions of some of the predictor groups across the long and short legs. Consider, for example, *Risk*, which makes the largest contributions to the long-short portfolio alphas for both multifactor models. In both cases, the contribution of *Risk* to the alpha is considerably larger for the short leg. Other groups whose short-leg contributions are substantially larger than their long-leg contributions for both models include *Investment* and *Earnings forecast*. In contrast, the long-leg contributions are relatively large for *Seasonal momentum*. For *Momentum* and *Earnings*, two of the leading predictor groups for both models, the groups’ long- and short-leg contributions are fairly similar.

3.4. Predictor Contributions Over Time

To this point, we have computed predictor group contributions for portfolio performance metrics over the full 1973:01 to 2021:12 forecast evaluation period. The $SPPC_p$ can be used to estimate predictor contributions for any subsample of interest. To provide motivation for subsample analysis, Figure 2 shows the cumulative log return for the long-short XGBoost(c) portfolio for the 1973:01 to 2021:12 out-of-sample period. For reference, the cumulative log excess return for the CRSP value-weighted market portfolio is also shown. The XGBoost(c) portfolio performs especially well through approximately 2002, after which its cumulative log return “flattens out” to an extent.¹⁸ The XGBoost(c) portfolio performs considerably bet-

¹⁷We use the negative of the return on the short leg, as this represents a positive contribution to the long-short portfolio alpha.

¹⁸Green, Hand, and Zhang (2017) obtain a similar finding for a long-short portfolio constructed from sorted return forecasts generated using linear Fama and MacBeth (1973) regressions estimated via ordinary least squares based on 94 firm characteristics. A similar pattern is also evident in Han et al. (2024), who construct

ter than the market portfolio during business-cycle recessions throughout the entire 1973:01 to 2021:12 forecast evaluation period. After 2002, the XGboost(c) portfolio continues to perform particularly well during recessions—especially the Great Recession—but its performance is less impressive during expansions *via-à-vis* the pre-2003 period. Investor learning about cross-sectional return predictability from academic studies (McLean and Pontiff 2016) potentially provides a partial explanation for the decrease in performance for the XGBoost(c) portfolio after 2002.

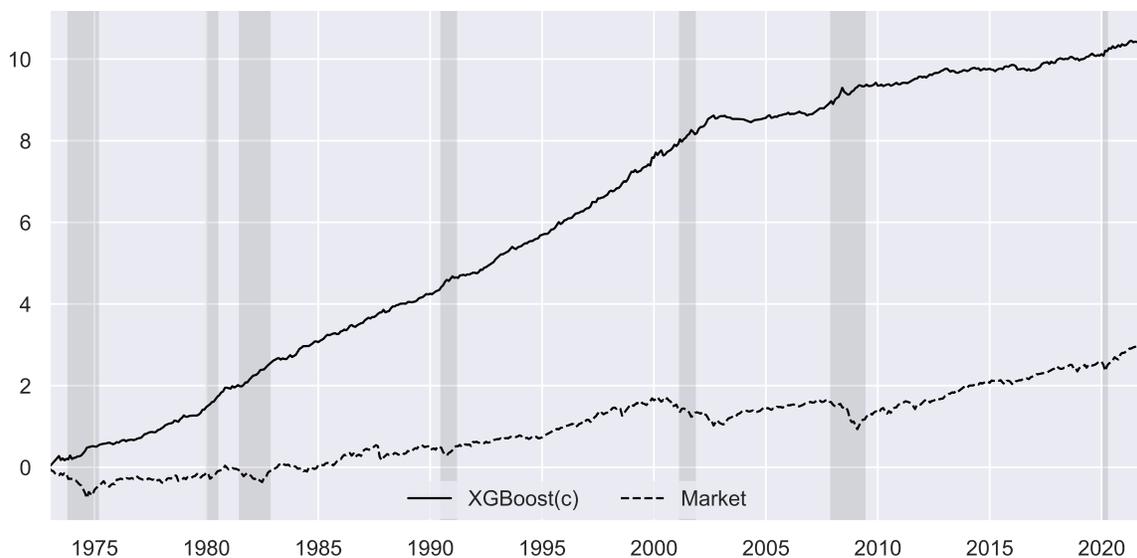


Figure 2. Cumulative log returns. The figure depicts the cumulative log return for the long-short XGBoost(c) portfolio and the cumulative log excess return for the CRSP value-weighted market portfolio. Vertical bars delineate business-cycle recessions as dated by the National Bureau of Economic Research.

Table 4 provides additional information on differences in portfolio performance over time by reporting the performance metrics from Table 2 for the 1973:01 to 2002:12 and 2003:01 to 2021:12 subsamples. Panel A (B) of Table 4 provides results for the long-short XGBoost(c) (market) portfolio. According to the different metrics, the performance of the XGBoost(c)

a long-short portfolio from sorted return forecasts generated using Fama-MacBeth regressions estimated with machine learning methods based on 207 firm characteristics.

portfolio generally declines from the first to the second subsample, in line with Figure 2. The annualized average return for the XGBoost(c) portfolio falls from 29.74% in the first subsample to 11.29% in the second. With limited changes in volatility and the maximum drawdown over the subsamples, the annualized Sharpe (Calmar) ratio declines from 2.38 to 0.95 (2.47 to 0.72). The average return, Sharpe ratio, and Calmar ratio increase for the market portfolio from the first subsample to the second, but they remain below those for the XGBoost(c) portfolio for both subsamples. Stark differences remain in the maximum drawdown between the XGBoost(c) and market portfolios for the two subsamples, with values of 12.06% and 15.70% (51.43% and 51.51%) for the former (latter) for the first and second subsamples, respectively. The annualized alphas for the XGBoost(c) portfolio fall from 24.54% to 9.68% (22.25% to 8.53%) from the first to the second subsample for the FF6 (Q5) model; nevertheless, the alphas remain economically sizable and statistically significant at the 1% level for both models for the second subsample.

Overall, Figure 2 and Table 4 indicate that the performance of the long-short XGBoost(c) portfolio deteriorates to a degree from the 1973:01 to 2002:12 to the 2003:01 to 2021:12 subsample but that it continues to outperform the market portfolio and deliver sizable alphas. Next, we use the $SPPC_p$ to estimate the contributions of the predictor groups to the performance metrics for the two subsamples. Figure 3 provides waterfall diagrams showing the contributions of the predictor groups to the Sharpe ratio, with the market portfolio continuing to serve as the benchmark. For the 1973:01 to 2002:12 subsample in Panel A, 18 of the 20 groups increase the Sharpe ratio for the XGBoost(c) portfolio relative to that for the market portfolio; the two exceptions are *Ownership* and *Sales*. *Risk* makes the largest contribution of 0.60, which is twice as large as the Sharpe ratio for the baseline market portfolio (0.30). Other groups making contributions above 0.20 in the first subsample include *Earnings*, *Momentum*, and *Seasonal momentum* (0.28, 0.27, and 0.21, respectively).

The contributions of the predictor groups often change markedly for the 2003:01 to 2021:12 subsample in Panel B. The baseline Sharpe ratio for the market portfolio increases

Table 4. Portfolio performance for subsamples

Panel A reports performance metrics for the long-short XGBoost(c) portfolio for the 1973:01 to 2002:12 and 2003:01 to 2021:12 subsamples. “MDD” is the maximum drawdown. “Ann. FF6 alpha” is the annualized alpha for a multifactor model that includes the five Fama and French (2015) factors and a momentum factor; “Ann. Q5 alpha” is the annualized alpha for the Hou et al. (2021) augmented q-factor model; t -statistics for the alphas are in parentheses; *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. “Market portfolio” in Panel B is the CRSP value-weighted market portfolio.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Subsample	Ann. mean	Ann. volatility	Ann. Sharpe ratio	MDD	Ann. Calmar ratio	Ann. FF6 alpha	Ann. Q5 alpha
Panel A: XGBoost(c) portfolio							
1973:01–2002:12	29.74%	12.51%	2.38	12.06%	2.47	24.54% [11.51]***	22.25% [7.66]***
2003:01–2021:12	11.29%	11.86%	0.95	15.70%	0.72	9.68% [3.93]***	8.53% [2.73]***
Panel B: Market portfolio							
1973:01–2002:12	5.04%	16.56%	0.30	51.43%	0.10	0%	0%
2003:01–2021:12	11.23%	14.63%	0.77	51.51%	0.22	0%	0%

to 0.77, while the Sharpe ratio for the XGBoost(c) portfolio is 0.95, so the contributions sum to 0.18. Nine (eleven) of the groups contribute positively (negatively) to the Sharpe ratio. The contribution of *Risk* falls from 0.60 in Panel A to 0.04 in Panel B. The contribution of *Momentum* goes from sizably positive (0.27) to negative (−0.08) as we move from Panel A to B. Despite making a positive contribution (0.05) in Panel A, *Reversal* reverses and makes the largest negative contribution (−0.13) in Panel B.

Figure 4 is an analogous version of Figure 3 for the Calmar ratio. The overall story is similar, although many more of the predictor groups (17 out of 20) make a positive contribution in the second subsample in Figure 4. For the first subsample, *Risk*, *Momentum*, *Seasonal momentum*, and *Earnings* make contributions that are twice as large or larger than the Calmar ratio for the baseline market portfolio (0.10). With the exception of *Sales*, all

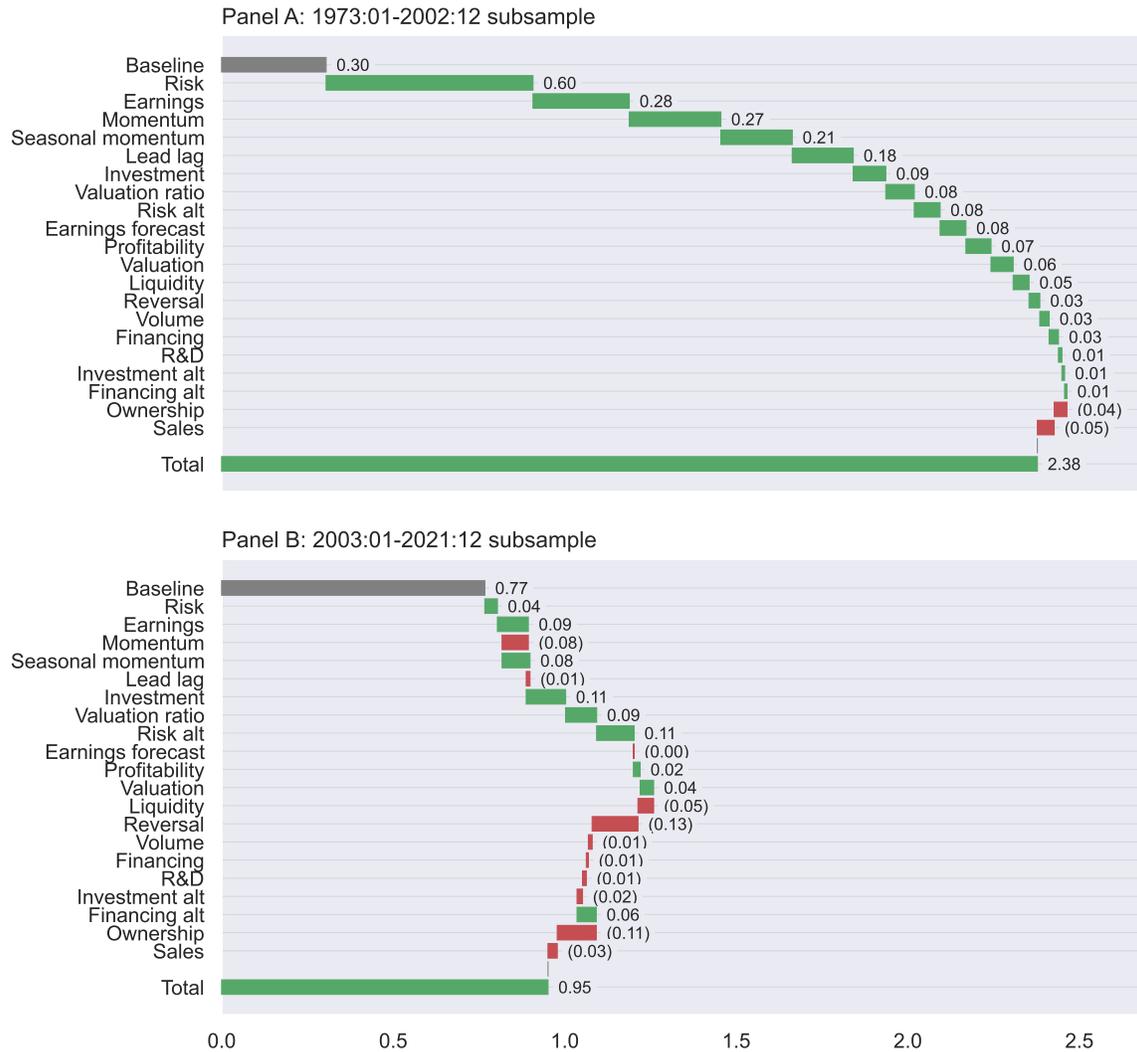


Figure 3. Sharpe ratio contributions for subsamples. Each panel depicts a waterfall diagram with the contributions of the 20 predictor groups to the annualized Sharpe ratio for the long-short XGBoost(c) portfolio. Parentheses indicate a negative number.

of the groups contribute positively to the Calmar ratio in Panel A. The Calmar ratio for the baseline market portfolio increases to 0.22 in the second subsample; given the Calmar ratio of 0.72 for the XGBoost(c) portfolio for the second subsample, the total contribution of the predictor groups is 0.50 (while it is 2.37 for the first subsample). As in Figure 3, *Momentum* goes from making a sizable positive contribution (0.38) in Panel A of Figure 4

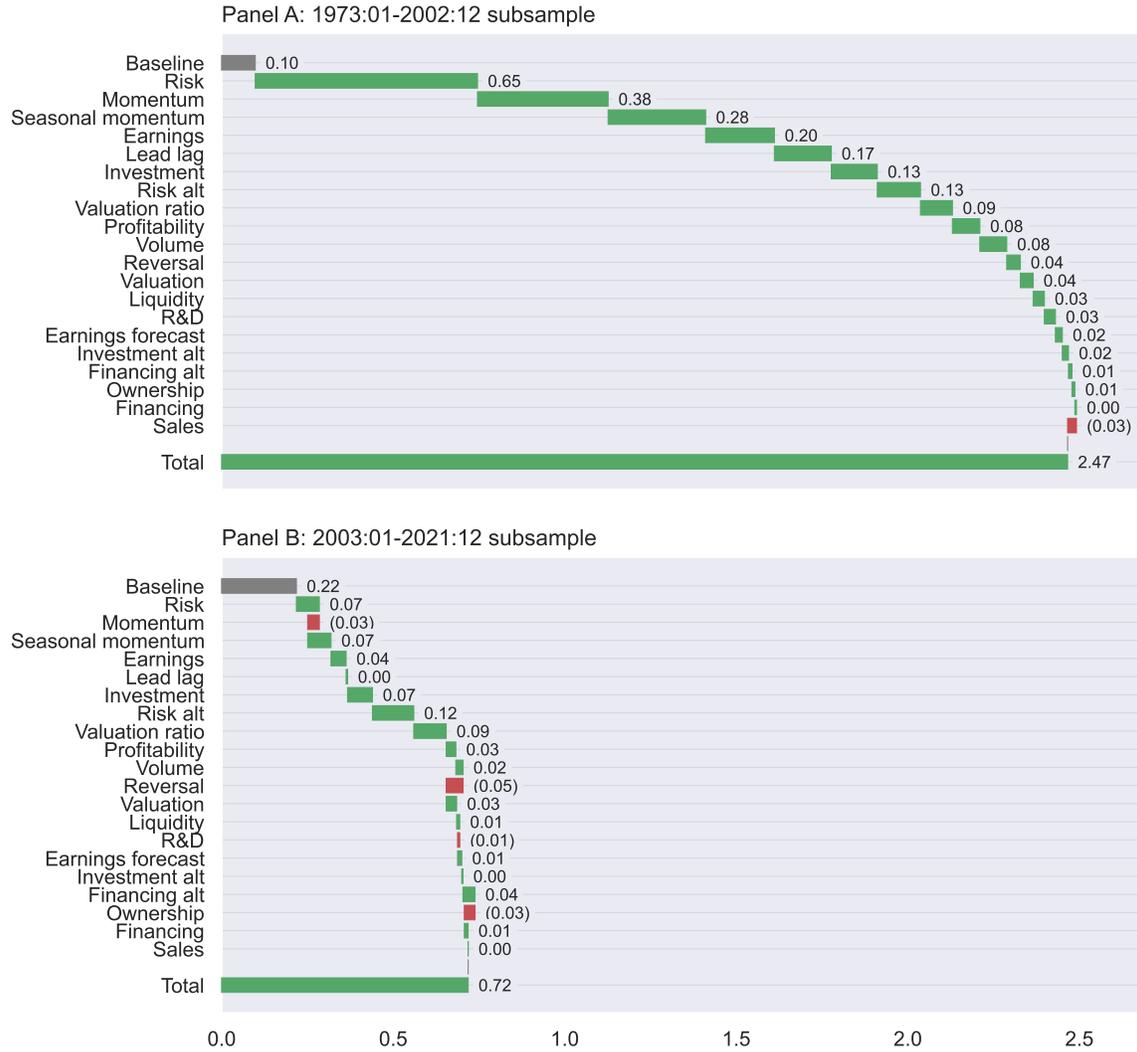


Figure 4. Calmar ratio contributions for subsamples. Each panel depicts a waterfall diagram with the contributions of the 20 predictor groups to the annualized Calmar ratio for the long-short XGBoost(c) portfolio. Parentheses indicate a negative number.

to a negative contribution (-0.03) in Panel B, and *Reversal* goes from making a positive contribution (0.04) in Panel A to the largest negative contribution (-0.05) in Panel B.

Figures 5 and 6 report analogous results to Figure 1 in terms of the contributions to the multifactor model alphas for the two subsamples. With respect to the alpha for the FF6 model in Figure 5, the results for the first subsample in Panel A are similar to those for

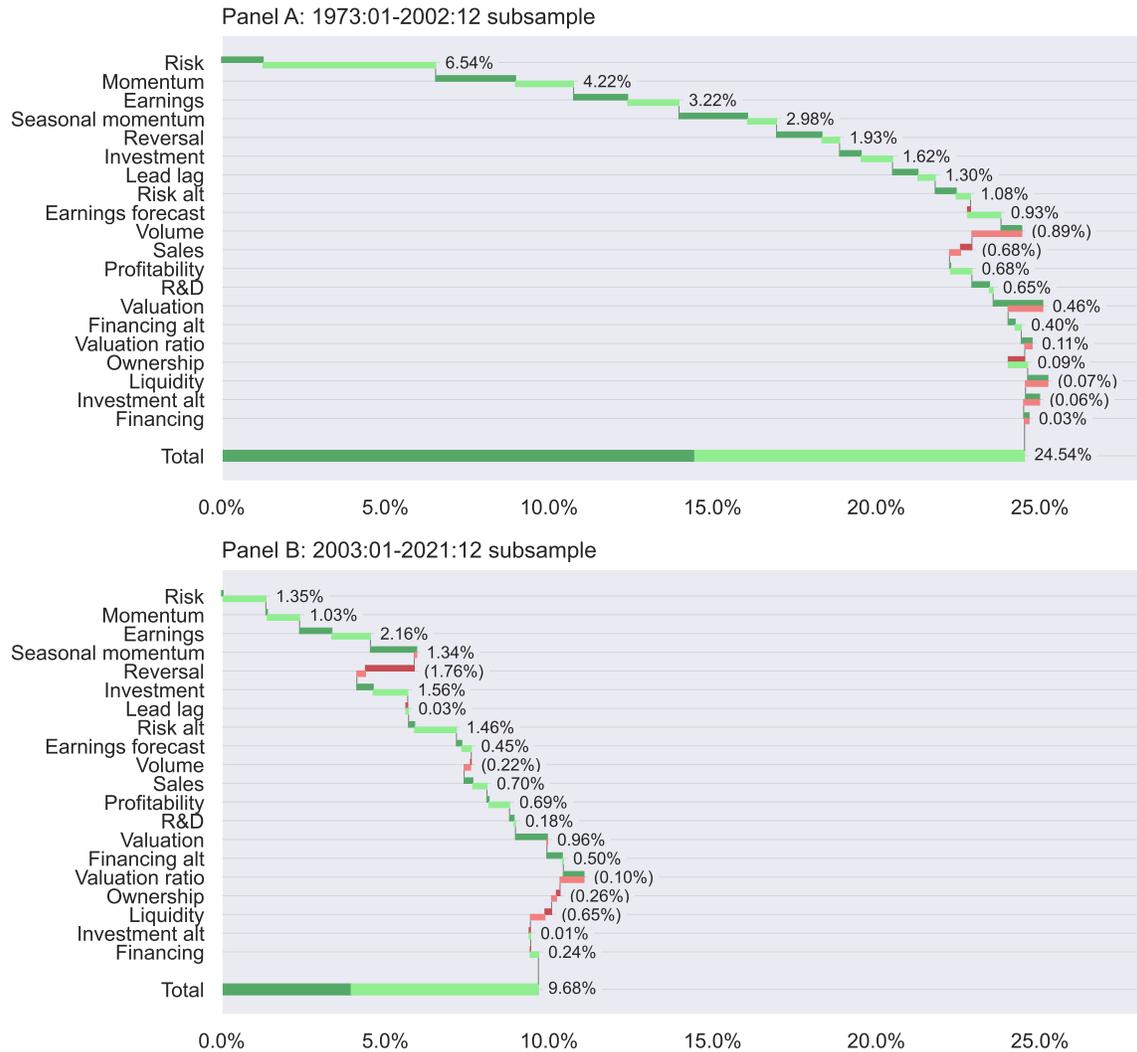


Figure 5. FF6 alpha long- and short-leg contributions for subsamples. Each panel depicts a waterfall diagram with the contributions of the predictor groups to the annualized alpha for the long-short XGBoost(c) portfolio in terms of the long and short legs. Alpha is measured in the context of the FF6 multifactor model. The dark (light) green segments are the positive contributions of the long (short) leg; the dark (light) red segments are the negative contributions of the long (short) leg. Parentheses indicate a negative number.

the full forecast evaluation period in Panel A of Figure 1. *Risk*, *Momentum*, *Earnings*, and *Seasonal momentum* make sizable contributions in the first subsample in Figure 5, with the contribution of *Risk* falling predominantly on the short leg, while the contributions of the

other groups are more evenly distributed across the long and short legs. The alpha falls from 24.54% to 9.68% from the first to the second subsample. *Risk*, *Momentum*, *Earnings*, and *Seasonal momentum* continue to make positive contributions as we move from the first to the second subsample, but their magnitudes are reduced. As with the Sharpe and Calmar ratios over the subsamples, *Reversal* evinces a reversal of its own, making a contribution of 193 basis points in the first subsample (primarily via the long leg), which subsequently falls to -176 basis points in the second subsample (again primarily via the long leg).

The subsample results in Figure 6 for the alpha for the Q5 model are broadly similar to those in Figure 5 for the FF6 model. *Reversal* exhibits an even stronger reversal in Figure 6 as we move from the first to the second subsample. Its contribution is a substantive 385 basis points in the first subsample but falls to -168 basis points in the second (both effects are primarily concentrated in the long leg). *Liquidity* also exhibits a marked turnaround in its contributions across the subsamples, going from a positive contribution of 114 basis points in the first subsample to a negative contribution of -44 basis points in the second.

Table 4 and Figures 3 to 6 investigate changes in portfolio performance and the predictor group contributions by dividing the full forecast evaluation period into non-overlapping subsamples. Another popular strategy for examining how performance changes over time is the use of rolling windows of data over the forecast evaluation period. Again, the $SPPC_p$ can be employed in this context. We compute performance metrics for the long-short XGBoost(c) portfolio based on 60-month rolling windows and then estimate the predictor group contributions via the $SPPC_p$ for the rolling windows. Figure 7 displays the sequences of Sharpe and Calmar ratios as well as FF6 and Q5 alphas computed using the rolling windows. The performance metrics tend to fall when data beyond 2002 are included in the window. In addition, there is a tendency for the metrics to increase when the windows incorporate data from recessions. This is clearly evident for the Sharpe ratio and the alphas around the Great Recession. The performance metrics also markedly increase for windows that include data near the end of the sample, corresponding to the advent of the COVID-19 crisis.

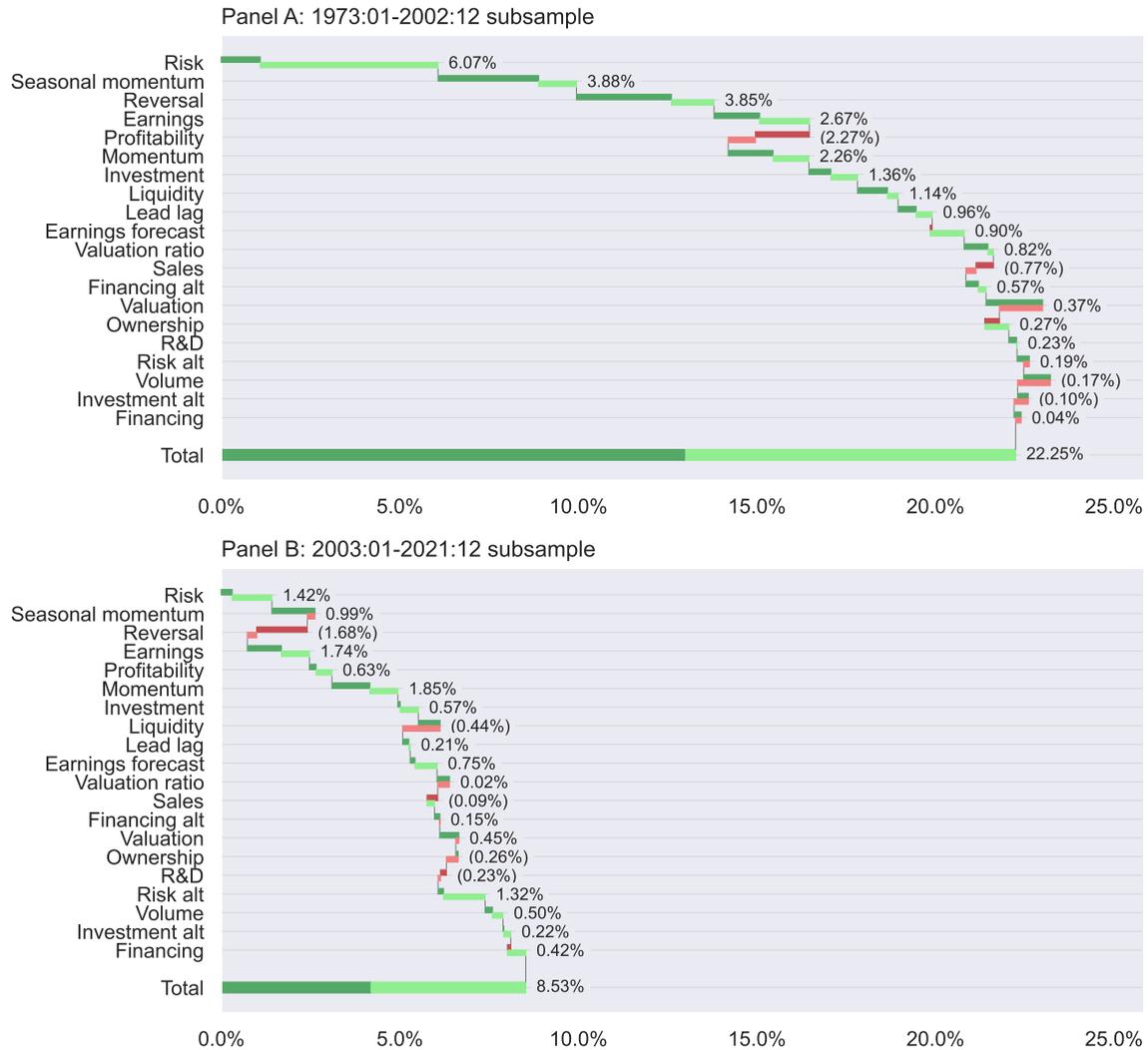


Figure 6. Q5 alpha long- and short-leg contributions for subsamples. Each panel depicts a waterfall diagram with the contributions of the predictor groups to the annualized alpha for the long-short XGBoost(c) portfolio in terms of the long and short legs. Alpha is measured in the context of the Q5 multifactor model. The dark (light) green segments are the positive contributions of the long (short) leg; the dark (light) red segments are the negative contributions of the long (short) leg. Parentheses indicate a negative number.

Figures 8 and 9 depict the contributions of the 20 predictor groups to the Sharpe and Calmar ratios, respectively, computed for the 60-month rolling windows. For a given window,

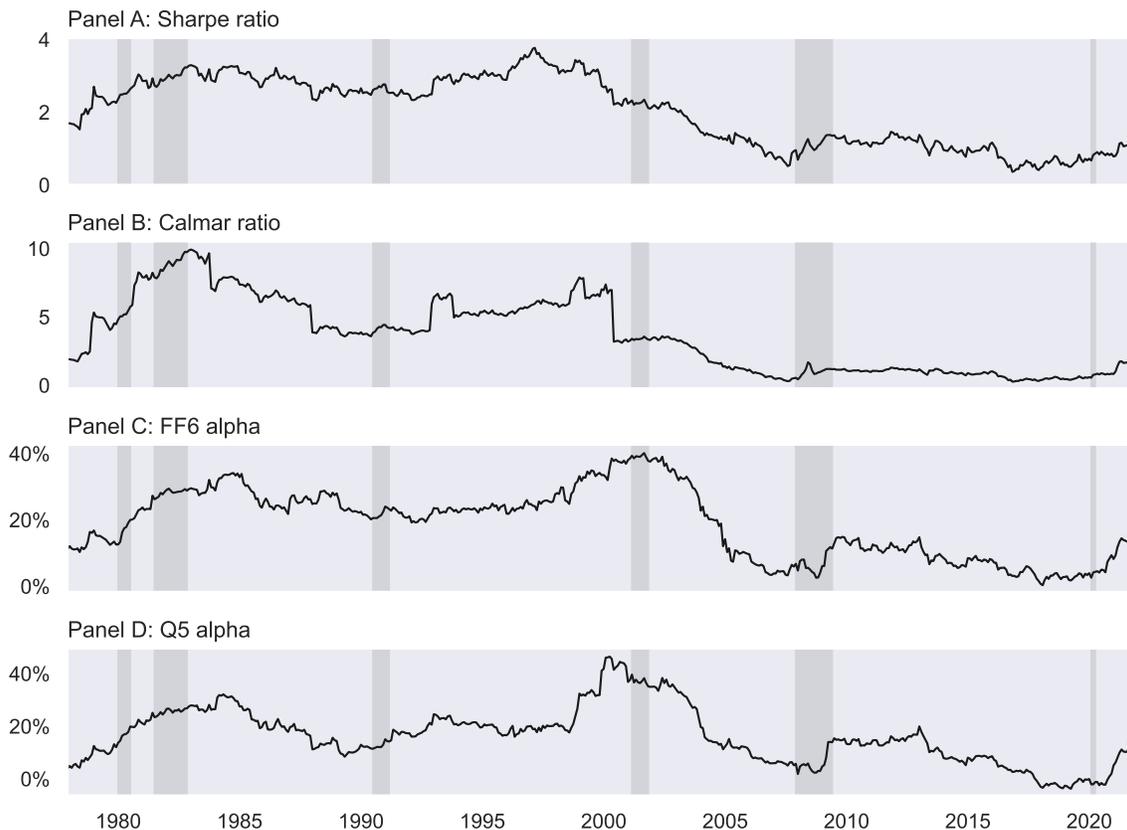


Figure 7. Portfolio performance for 60-month rolling windows. Each panel depicts the annualized performance metric in the panel heading for the long-short XGBoost(c) portfolio. The metrics are computed using 60-month rolling windows over the 1973:01 to 2021:12 forecast evaluation period. The horizontal axis corresponds to the end of the 60-month rolling window. Vertical bars delineate business-cycle recessions as dated by the National Bureau of Economic Research.

we standardize the contributions by the maximum group contribution for that window.¹⁹ The results in Figures 8 and 9 are similar and reveal noteworthy patterns in the group contributions over time. For example, *Risk* and *Momentum* frequently make among the largest contributions for windows ending through the early 2000s, while they subsequently often make sizably negative contributions. Other groups making contributions that substantially

¹⁹Thus, a line in Figure 8 has a maximum value of one when the group’s contribution is equal to the maximum contribution for that window; a standardized contribution of -1 means that the group makes a negative contribution that is equal in magnitude to the maximum contribution for that window.

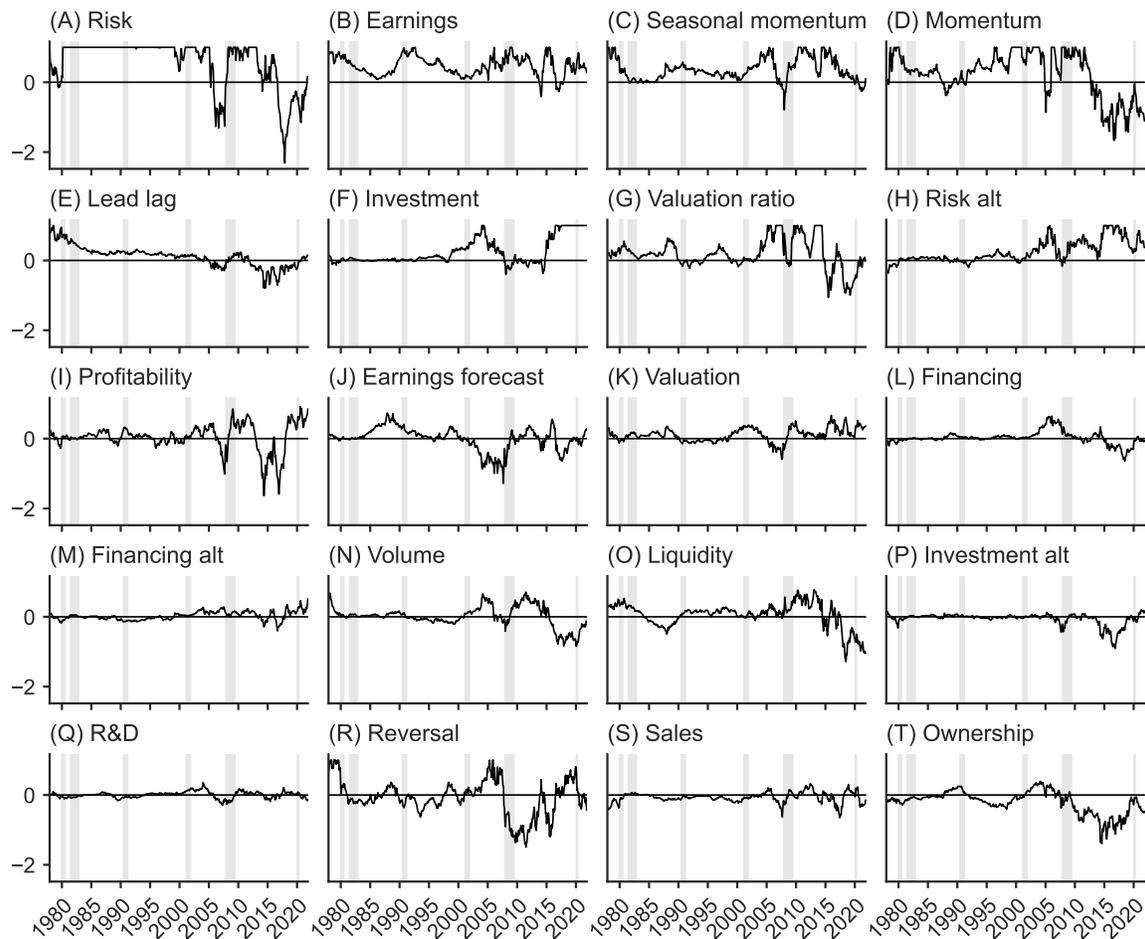


Figure 8. Sharpe ratio contributions for 60-month rolling windows. The figure depicts the contributions of the 20 predictor groups to annualized Sharpe ratios for the long-short XGBoost(c) portfolio. The Sharpe ratios are computed using 60-month rolling windows over the 1973:01 to 2021:12 forecast evaluation period. The horizontal axis corresponds to the end of the 60-month rolling window. The contributions are standardized by the maximum contribution to the Sharpe ratio in a given rolling window. Vertical bars delineate business-cycle recessions as dated by the National Bureau of Economic Research.

vary between positive and negative values include *Profitability*, *Earnings forecast*, and *Reversal*. Groups making more consistent positive contributions over time include *Earnings* and *Seasonal momentum*. In general, the contributions in Figures 8 and 9 appear more stable for windows ending through the early 2000s.

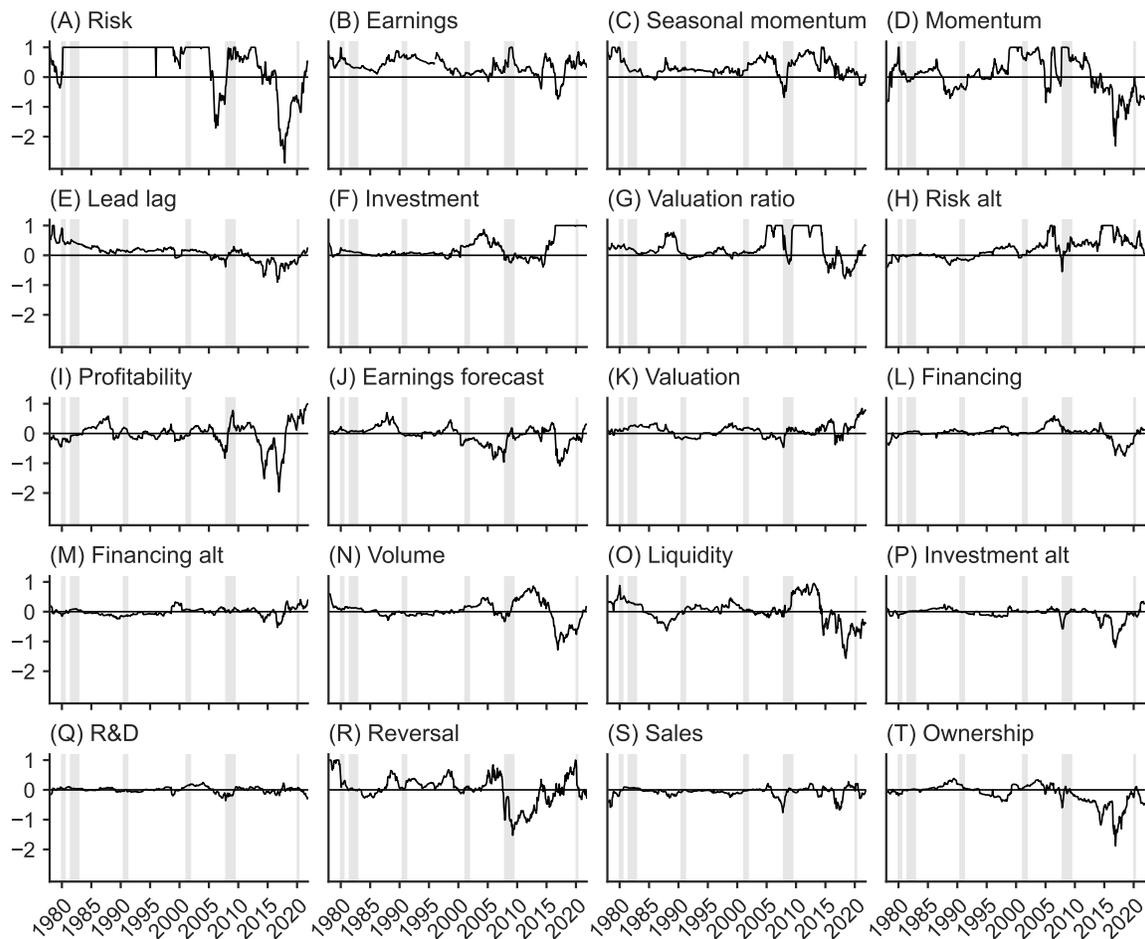


Figure 9. Calmar ratio contributions for 60-month rolling windows. The figure depicts the contributions of the 20 predictor groups to annualized Calmar ratios for the long-short XGBoost(c) portfolio. The Calmar ratios are computed using 60-month rolling windows over the 1973:01 to 2021:12 forecast evaluation period. The horizontal axis corresponds to the end of the 60-month rolling window. The contributions are standardized by the maximum contribution to the Calmar ratio in a given rolling window. Vertical bars delineate business-cycle recessions as dated by the National Bureau of Economic Research.

Figures 10 and 11 show the contributions of the predictor groups to the alphas for the FF6 and Q5 multifactor models, respectively, computed based on the 60-month rolling windows. The two figures tell a similar story with respect to the contributions to the alphas over time. For windows ending prior to the early 2000s, *Risk* nearly always contributes pos-

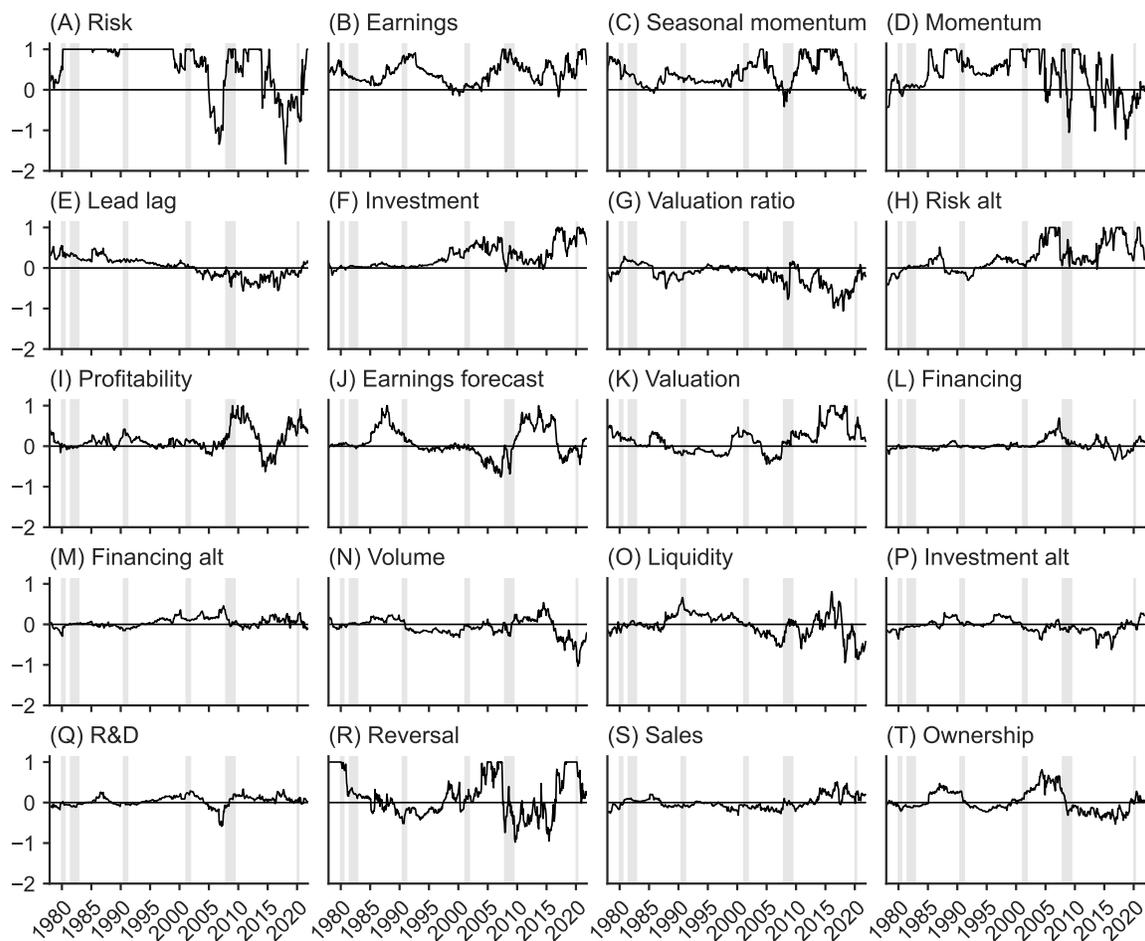


Figure 10. FF6 alpha contributions for 60-month rolling windows. The figure depicts the contributions of the 20 predictor groups to annualized alphas for the long-short XGBoost(c) portfolio. The alphas are estimated in the context of the FF6 multifactor model and computed using 60-month rolling windows over the 1973:01 to 2021:12 forecast evaluation period. The horizontal axis corresponds to the end of the 60-month rolling window. The contributions are standardized by the maximum contribution to the alpha in a given rolling window. Vertical bars delineate business-cycle recessions as dated by the National Bureau of Economic Research.

itively and sizably to the alphas, but it often makes large negative contributions thereafter. *Momentum*, *Valuation ratio*, *Liquidity*, and *Reversal* also often make substantive negative contributions after the early 2000s. *Earnings*, *Seasonal momentum*, and (to a lesser extent) *Investment* make sizably positive contributions on a reasonably consistent basis over time.

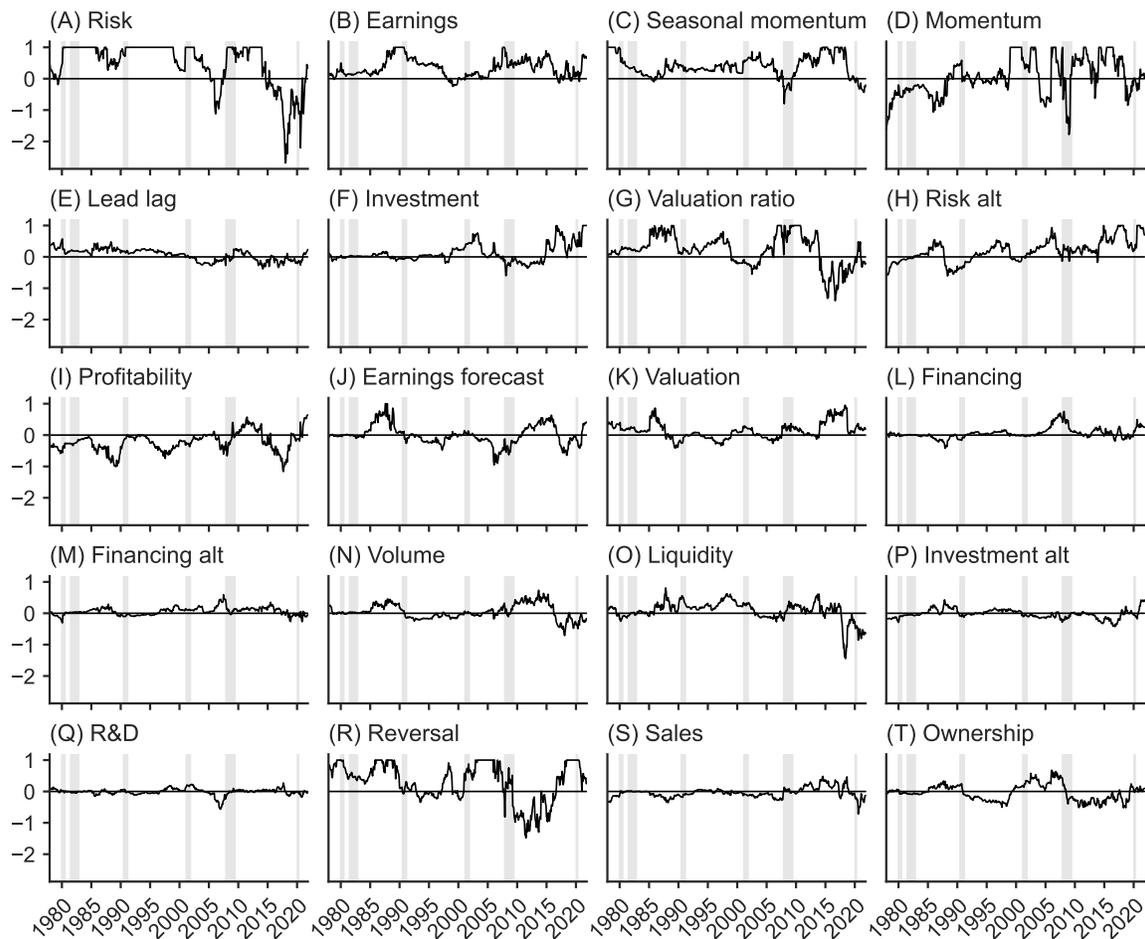


Figure 11. Q5 alpha contributions for 60-month rolling windows. The figure depicts the contributions of the 20 predictor groups to annualized alphas for the long-short XGBoost(c) portfolio. The alphas are estimated in the context of the Q5 multifactor model and computed using 60-month rolling windows over the 1973:01 to 2021:12 forecast evaluation period. The horizontal axis corresponds to the end of the 60-month rolling window. The contributions are standardized by the maximum contribution to the alpha in a given rolling window. Vertical bars delineate business-cycle recessions as dated by the National Bureau of Economic Research.

Overall, similarly to Figures 8 and 9, the contributions tend to exhibit greater fluctuations in Figures 10 and 11 after the early 2000s.

3.5. Alternative Approach

As a final exercise, we decompose the alphas in Table 2 using the alternative method described in Section 2.3. The alternative method turns off each predictor group in turn by setting all of the predictors in the group equal to zero, so the predictor group does not impact the predictions made by the fitted models. We compare the original alpha computed with the predictor group turned on to the alpha computed with the group turned off; the difference measures the contribution of the predictor group to the alpha. This “turn on-off” method is significantly less computationally costly than the $SPPC_p$ for our application. However, as pointed out in Section 2.3, unlike the $SPPC_p$, the turn on-off method does not possess the attractive properties of Shapley values. We focus on the alphas, as the baseline market portfolio has a value of zero in this case. Because the turn on-off method does not readily accommodate a baseline like the market portfolio, we cannot directly compare the $SPPC_p$ and turn on-off decompositions when the performance metric for the baseline is nonzero.²⁰

Table 5 reports decompositions of the FF6 and Q5 alphas. For ease of comparison, the second and fourth columns reproduce the $SPPC_p$ decompositions from the last two columns of Table 3, while the third and fifth columns of Table 5 report decompositions using the turn on-off method. Beginning with the FF6 alpha, we again see from the second column that the $SPPC_p$ provides an exact decomposition of the alpha for the long-short XGBoost(c) portfolio (19.45%). According to the third column, the turn on-off method does not provide an exact decomposition, as the 20 predictor group contributions sum to 27.68%, which is larger than the actual FF6 alpha by 823 basis points. Furthermore, there are a number of substantive discrepancies between the estimated contributions for the two approaches. For example, the signs of the contributions are opposite for five predictor groups (*Volume*, *Liquidity*, *R^{ED}*, *Reversal*, *Sales*), and the magnitudes of the differences are fairly sizable, ranging from 56

²⁰This is an advantage of the Shapley-based $SPPC_p$, as researchers are often interested in the performance of an actively managed portfolio relative to a passively managed benchmark.

Table 5. Alpha contributions using alternative method

The table reports the contributions of the 20 predictor groups to the annualized FF6 and Q5 alphas for the long-short XGBoost(c) portfolio. The contributions in the second and fourth columns are reproduced from the seventh and eighth columns of Table 3; the numbers in the second and fourth columns may not add to the value in the “Total” row due to rounding. The contributions in the third and fifth columns are estimated using the alternative (turn on-off) method. The forecast evaluation period is 1973:01 to 2021:12.

(1) Predictor group	(2) Ann. FF6 alpha		(4) Ann. Q5 alpha	
	(3) SPPC _p	(3) Alternative	(4) SPPC _p	(5) Alternative
Risk	4.34	7.94	4.29	7.35
Earnings	2.72	3.39	2.02	2.88
Seasonal Momentum	2.38	1.73	2.49	1.66
Momentum	3.25	3.86	2.16	2.74
Lead lag	0.95	0.59	0.57	0.28
Investment	1.70	1.03	0.95	-0.23
Valuation Ratio	0.38	0.70	0.58	0.52
Risk alt	1.27	1.16	0.76	0.15
Profitability	0.61	2.28	-0.88	1.22
Earnings Forecast	0.85	0.65	0.92	0.71
Valuation	0.77	1.04	0.48	0.14
Financing	0.14	1.30	0.29	1.45
Financing alt	0.70	0.87	0.29	0.03
Volume	-0.43	0.13	0.17	0.25
Liquidity	-0.19	0.50	0.35	0.88
Investment alt	0.15	0.54	0.08	0.69
R&D	0.42	-0.58	-0.03	-0.78
Reversal	-0.32	0.40	1.46	1.59
Sales	-0.18	0.50	-0.65	-0.63
Ownership	-0.09	-0.35	-0.01	-0.60
Total	19.45%	27.68%	16.29%	20.30%

to 100 basis points (*Volume* and *R&D*, respectively). The contributions of other predictor groups also differ by sizable magnitudes, such as *Risk*, *Profitability*, and *Financing* (360, 167, and 116 basis points, respectively); in these cases, the estimated contributions are higher for

the turn on-off method, helping to explain why the approach overstates the alpha when all of the predictors are taken together.

The results are broadly similar for the Q5 alpha decompositions in the last two columns of Table 5. In contrast to the exact decomposition provided by the SPPC_p, based on the turn on-off method, the 20 predictor group contributions sum to 20.30%, which is larger than the actual alpha (16.29%) by 401 basis points. Substantive differences in the magnitudes of the estimated contributions are evident for *Risk*, *Profitability*, *Investment*, and *Financing* (306, 210, 118, and 116 basis points, respectively). Overall, we interpret the results in Table 5 as favoring the SPPC_p over the alternative approach for decomposing portfolio performance, even though the latter is less computationally costly.²¹ Because the estimated contributions often differ substantially, it matters which method is used; since the SPPC_p is based on the logic of Shapley values and possesses their attractive properties, it provides a more accurate and informative estimate of a predictor’s contribution, thereby making it the preferred option from a cost-benefit perspective.

4. Conclusion

Asset return predictability is now commonly assessed in terms of economic value. A researcher generates out-of-sample return forecasts for one or more assets, increasingly using a large set of predictors and a machine learning model. The return forecasts then serve as inputs to construct a portfolio over the forecast evaluation period, and performance metrics are used to measure the economic value of return predictability from an investment perspective. While measuring the economic value of return predictability is crucial for assessing the relevance of return predictability, it is also vital to understand the sources of the economic value provided by return predictability.

²¹The results are similar to those in Table 5 if we use the permutation method described in Section 2.3 in lieu of the turn on-off approach.

The information in the underlying predictors in fitted machine learning models is the ultimate source of return predictability and its associated economic value. However, the existing literature does not provide a general procedure for decomposing economic value as measured by a portfolio performance metric into the contributions of the underlying predictors. The present paper fills this gap by developing the $SPPC_p$, a new model interpretation tool for finance founded on Shapley values that directly estimates the contributions of individual or groups of predictors in fitted prediction models to portfolio performance. Based on the logic of Shapley values, the $SPPC_p$ fairly allocates the predictor contributions to the portfolio performance metric. Shapley values have a number of attractive properties inherited by the $SPPC_p$, including local accuracy, so the $SPPC_p$ values for the set of predictors provide an exact decomposition of the performance metric in terms of the underlying predictors. The $SPPC_p$ is very flexible: it can be used for any prediction model, including any machine learning model; any strategy for mapping return forecasts to portfolio weights; and any performance metric. In sum, the $SPPC_p$ provides an anatomy of machine learning-based portfolio performance and constitutes a powerful tool for deepening our understanding of the sources of the economic value generated by return predictability.

We illustrate the use of the $SPPC_p$ in an empirical application investigating firm-level stock return predictability based on 207 firm characteristics from Chen and Zimmermann (2022). We employ the XGBoost algorithm, a powerful machine learning device, to generate monthly out-of-sample return classification forecasts based on the firm characteristics for 1973:01 to 2021:12, where individual stocks are predicted to be in quintiles in terms of their returns for the next month. Based on the forecasts, we construct a zero-investment portfolio that goes long (short) stocks that are predicted to be in the top (bottom) quintile. To minimize the role of small-cap stocks when forming the portfolio, we drop stocks with market capitalization below the NYSE 20th percentile and employ value-weighting in the long and short legs. The long-short portfolio delivers substantial economic value in terms of

Sharpe and Calmar ratios as well as risk-adjusted returns in the context of leading multifactor models.

We categorize the firm characteristics into 20 groups according to economic concepts and estimate the $SPPC_p$ for the predictor groups and portfolio performance metrics. Groups making the largest positive contributions to portfolio performance over the full 1973:01 to 2021:12 forecast evaluation period include *Risk*, *Earnings*, *Seasonal momentum*, and *Momentum*, while *Sales* and *Ownership* make negative contributions. The performance of the long-short portfolio generally declines after 2002, but it still performs relatively well, especially during business-cycle recessions. To shed light on the sources of the change in portfolio performance over time, we estimate the $SPPC_p$ for the predictor groups for subsamples and rolling windows from the full forecast evaluation period. The $SPPC_p$ estimates reveal that the contributions of *Risk* and *Momentum* to the performance metrics typically decrease substantively after 2002, often becoming negative. In contrast, *Earnings*, *Seasonal momentum*, and *Investment* make positive and sizable contributions to portfolio performance on a relatively consistent basis over time, indicating that these characteristic groups are more reliable predictors of cross-sectional stock returns in a machine learning framework when it comes to economic value. Our empirical application demonstrates how the $SPPC_p$ provides insights into the key determinants of cross-sectional expected returns in terms of economic value, including changes in the importance of the determinants over time.

Due to its flexibility, the $SPPC_p$ can be straightforwardly extended to address transaction costs. For example, we can compute a performance metric based on the gross portfolio return and the return net of transaction costs; the difference between the two metrics is the decrease in performance due to transaction costs. Then, we can use the $SPPC_p$ to decompose the performance metric based on the net portfolio return as well as the change in the performance metric due to transaction costs. The latter provides estimates of the contributions of the predictors to the decline in performance due to transaction costs.

Again due to its flexibility, the $SPPC_p$ can also be modified to estimate the predictor contributions to a performance metric when machine learning techniques are used to directly estimate optimal portfolio weights, including those for mean-variance efficient portfolios (e.g., Kozak, Nagel, and Santosh 2020; Cong et al. 2022; Chen, Pelger, and Zhu 2024; Jensen et al. 2024). Since mean-variance efficient portfolio weights relate to the stochastic discount factor, the $SPPC_p$ can provide further insight into the key drivers of the stochastic discount factor.

The voluminous literature on aggregate stock market return predictability considers a lengthy list of predictors, and it would be informative in future research to use the $SPPC_p$ to estimate the contributions of predictors in machine learning models to the economic value generated by aggregate market return predictability. In addition to equities, the $SPPC_p$ can be used to investigate predictor contributions to economic value for portfolios formed from any asset class or combinations of asset classes. In this vein, it would be illuminating to analyze predictor contributions in machine learning models to portfolio performance for a range of asset classes and to explore whether common patterns of predictor importance exist.

Finally, the $SPPC_p$ can potentially be used to improve portfolio performance in real time. For example, at the time of forecast formation, we compute the $SPPC_p$ for each predictor for a holdout out-of-sample period comprised of, say, the last 60 months. Predictors that contribute negatively to portfolio performance over the holdout out-of-sample period are then excluded from the prediction model estimated using data through the current month. Exploring strategies for harnessing the $SPPC_p$ to improve portfolio performance in real time is another fruitful avenue for future research.

Appendix

The computational time for estimating the $SPPC_p$ is dominated by the need to evaluate the fitted prediction function \hat{f} many times. When computing the $SPPC_p$, we need to evaluate the change in the out-of-sample return forecast for firm i when adding predictor p to the coalition of predictors preceding it in the randomly drawn ordered permutation \mathcal{O}_m , which

requires evaluating $\hat{r}_{i,t_j,\text{end}+1,m,p}$ and $\hat{r}_{i,t_j,\text{end}+1,m,\setminus p}$ in Equations (16) and (17), respectively. For $\hat{r}_{i,t_j,\text{end}+1,m,p}(\hat{r}_{i,t_j,\text{end}+1,m,\setminus p})$, we need to integrate out the predictors following p (following and including p) in \mathcal{O}_m . To accomplish this, we effectively average over the panel training sample observations, so we need to evaluate \hat{f} for each observation in the panel training data twice, once with and once without p integrated out. We also need to repeat this process for each predictor $p \in S$. This, however, allows us to eliminate half of the number of evaluations of \hat{f} , so we need to evaluate \hat{f} once for each predictor $p \in S$ with the predictors following p integrated out using each observation in the panel training data.²² To measure the change in the out-of-sample return forecast when adding predictor p to the coalition of predictors preceding it in \mathcal{O}_m for all of the firms, we evaluate $\hat{r}_{i,t_j,\text{end}+1,m,p}$ and $\hat{r}_{i,t_j,\text{end}+1,m,\setminus p}$ for $i \in C$. In effect, we need to evaluate \hat{f} for each predictor and for each observation in the panel training data for each firm.

To this point, we have focused on computations for a single month; we need to perform the computations for each month in the entire out-of-sample period. In total, we need to evaluate \hat{f} for each predictor and each firm for all of the observations in the panel training data and all of the months in the out-of-sample period. Furthermore, we need to do this $2 \times M$ times for the sampling-based approach (taking into account antithetic sampling).

In the empirical application in Section 3, we analyze the contributions of 20 groups of predictors for the 1973:01 to 2021:12 out-of-sample period (588 months). We have an average of approximately 2,000 firms for each month in the out-of-sample period, as well as an average of about 750,000 firm-month observations for the sequence of panel training datasets. There are two dimensions along which to limit computational cost: (1) the number of randomly

²²By way of example, consider three predictors and the ordered permutation $\mathcal{O}_m = \{3, 2, 1\}$, so predictor 3 is added to the coalition of predictors first, then 2, then 1. The effect of adding predictor 3 is measured by computing the forecast conditional on predictor 3 with predictors 2 and 1 integrated out and comparing it to the forecast based on the empty coalition set that integrates out all of the predictors. The effect of adding predictor 2 is measured by computing the forecast conditional on predictors 3 and 2 with predictor 1 integrated out and comparing it to the forecast conditional on predictor 3 with predictors 2 and 1 integrated out. Clearly, we do not need to recompute the latter, as we have already computed it in the first step when predictor 3 is added to the empty predictor coalition set.

drawn ordered permutations M and (2) the proportion of training sample observations to use when integrating out predictors. A decrease in each leads to a proportional reduction in computational time. For our empirical application, we set $M = 50$ (for a total of $2 \times 50 = 100$ ordered permutations with antithetic sampling) and use 10% of the training sample observations when integrating out predictors.²³

Thus, as a first step, we need to evaluate fitted prediction functions approximately $20 \times 588 \times 2,000 \times 0.10 \times 750,000 \times 50 \times 2 = 176,400,000,000,000$ times to compute the required forecasts in Equations (16) and (17) for the out-of-sample period. In this computationally expensive first step, we used 306 and 274 core-months of Intel Xeon Platinum 8260 and Intel Xeon Gold 6148 processors, respectively, with AVX-512 enabled. To substantially reduce computational time after this step, note that Equations (16) and (17) are the lowest unit in any decomposition of a portfolio performance metric. We store the forecasts on disk and cache Equations (16) and (17) in memory when we compute the SPPC_p for a specific performance metric. After the extensive $306 + 274 = 580$ core-months of computations in the first step, we can compute all of the predictor contributions for any performance metric nearly instantly because we no longer need to evaluate \hat{f} and integrate out predictors over the training sample observations. Using the computed forecasts from the first step, we compute the series of portfolio returns inside the curly brackets in Equation (25). To compute the SPPC_p for any performance metric, we only need to evaluate the metric wrapped around the series of portfolio returns about $20 \times 50 \times 2 = 2,000$ times, which incurs no meaningful computational time for the performance metrics considered in the empirical application.

We also estimated the SPPC_p setting $M = 50$ and using 1% of the training sample observations and obtained similar results to those reported in Section 3 based on 10% of the

²³It is standard to set M to a relatively low value and use a subsample of the training sample observations when computing conventional Shapley values; for example, the popular SHAP package in python uses defaults of $M = 10$ and 100 observations from the training sample (corresponding to roughly 0.01% of the training sample observations on average in our application). We use more rigorous settings to improve estimation accuracy.

training sample observations and an independent sample of $M = 50$ ordered permutations. Thus, it appears relatively safe to reduce the computational cost by setting $M = 50$ and using a more limited proportion of training sample observations when integrating out predictors to estimate the SPPC_p .

References

- Akiba, T., S. Sano, T. Yanase, T. Ohta, and M. Koyama (2019). Optuna: A Next-Generation Hyperparameter Optimization Framework. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2623–2631.
- Aleti, S., T. Bollerslev, and M. Siggaard (2025). Intraday Market Return Predictability Culled from the Factor Zoo. *Management Science*, forthcoming.
- Apley, D. W. and J. Zhu (2020). Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 82:4, 1059–1086.
- Avramov, D., S. Cheng, and L. Metzker (2023). Machine Learning Versus Economic Restrictions: Evidence from Stock Return Predictability. *Management Science* 69:5, 2587–2619.
- Breiman, L. (1997). Arcing the Edge. Technical Report 486, Statistics Department, University of California, Berkeley.
- Breiman, L. (2001). Random Forests. *Machine Learning* 45:1, 5–32.
- Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen (1984). *Classification and Regression Trees*. New York: CRC Press.
- 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.
- Castro, J., D. Gómez, and J. Tejada (2009). Polynomial Calculation of the Shapley Value Based on Sampling. *Computer and Operations Research* 36:5, 1726–1730.

- Chen, A. Y. and T. Zimmermann (2022). Open Source Cross-Sectional Asset Pricing. *Critical Finance Review* 27:2, 207–264.
- Chen, H., J. D. Janizek, S. Lundberg, and S.-I. Lee (2020). True to the Model or True to the Data? arXiv Working Paper No. 2006.16234.
- Chen, L., M. Pelger, and J. Zhu (2024). Deep Learning in Asset Pricing. *Management Science* 70:2, 714–750.
- Chen, T. and C. Guestrin (2016). XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.
- Cong, L. W., K. Tang, J. Wang, and Y. Zhang (2022). AlphaPortfolio: Direct Construction Through Deep Reinforcement Learning and Interpretable AI. SSRN Working Paper No. 3554486.
- Dimopoulos, Y., P. Bourret, and S. Lek (1995). Use of Some Sensitivity Criteria for Choosing Networks with Good Generalization Ability. *Neural Processing Letters* 2:1, 1–4.
- Elsayed, S., D. Thyssens, A. Rashed, H. S. Jomaa, and L. Schmidt-Thieme (2021). Do We Really Need Deep Learning Models for Time Series Forecasting? arXiv Working Paper No. 2101.02118.
- 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 (2015). A Five-Factor Asset Pricing Model. *Journal of Financial Economics* 116:1, 1–22.
- Fama, E. F. and J. D. MacBeth (1973). Risk, Return, and Equilibrium: Empirical Tests. *Journal of Political Economy* 81:3, 607–636.
- Fisher, A., C. Rudin, and F. Dominici (2019). All Models Are Wrong, But Many Are Useful: Learning a Variable’s Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research* 20:177, 1–81.

- Freyberger, J., A. Neuhierl, and M. Weber (2020). Dissecting Characteristics Nonparametrically. *Review of Financial Studies* 33:5, 2326–2377.
- Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics* 29:5, 1189–1232.
- Friedman, J. H. (2002). Stochastic Gradient Boosting. *Computational Statistics & Data Analysis* 38:4, 367–378.
- Goldstein, A., A. Kapelner, J. Bleich, and E. Pitkin (2015). Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics* 24:1, 44–65.
- Green, J., J. R. M. Hand, and X. F. Zhang (2017). The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns. *Review of Financial Studies* 30:12, 4389–4436.
- Greenwell, B. M., B. C. Boehmke, and A. J. McCarthy (2018). A Simple and Effective Model-Based Variable Importance Measure. arXiv Working Paper No. 1805.04755.
- Grinsztajn, L., E. Oyallon, and G. Varoquaux (2022). Why Do Tree-Based Models Still Outperform Deep Learning on Typical Tabular Data? In: *Proceedings of the 35th International Conference on Advances in Neural Information Processing Systems*, pp. 507–520.
- Gu, S., B. Kelly, and D. Xiu (2020). Empirical Asset Pricing via Machine Learning. *Review of Financial Studies* 33:5, 2223–2273.
- 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.
- Hou, K., H. Mo, C. Xue, and L. Zhang (2021). An Augmented q-Factor Model with Expected Growth. *Review of Finance* 25:1, 1–41.

- Hou, K., C. Xue, and L. Zhang (2015). Digesting Anomalies: An Investment Approach. *Review of Financial Studies* 28:3, 650–703.
- Hou, K., C. Xue, and L. Zhang (2020). Replicating Anomalies. *Review of Financial Studies* 33:5, 2019–2133.
- Janzing, D., L. Minorics, and P. Blöbaum (2020). Feature Relevance Quantification in Explainable AI: A Causal Problem. In: *23rd International Conference on Artificial Intelligence and Statistics*, pp. 2907–2916.
- Jensen, T. I., B. Kelly, S. Malamud, and L. H. Pedersen (2024). Machine Learning and the Implementable Efficient Frontier. SSRN Working Paper No. 4187217.
- Kozak, S., S. Nagel, and S. Santosh (2020). Shrinking the Cross-Section. *Journal of Financial Economics* 135:2, 271–292.
- Lewellen, J. (2015). The Cross-Section of Expected Stock Returns. *Critical Finance Review* 4:1, 1–44.
- Lundberg, S. M. and S.-I. Lee (2017). A Unified Approach to Interpreting Model Predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4768–4777.
- Martin, I. and S. Nagel (2022). Market Efficiency in the Age of Big Data. *Journal of Financial Economics* 145:1.
- McLean, R. D. and J. Pontiff (2016). Does Academic Research Destroy Return Predictability? *Journal of Finance* 71:1, 5–32.
- Mitchell, R., J. Cooper, E. Frank, and G. Holmes (2022). Sampling Permutations for Shapley Value Estimation. *Journal of Machine Learning Research* 23:1, 2082–2127.
- Moehle, N., S. Boyd, and A. Ang (2022). Portfolio Performance Attribution via Shapley Value. *Journal of Investment Management* 20:3, 33–52.
- Molnar, C. (2024). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

- Nagel, S. (2021). *Machine Learning in Asset Pricing*. Princeton, NJ: Princeton University Press.
- Pearl, J. (2009). *Causality*. Second Edition. Cambridge: Cambridge University Press.
- Rapach, D. E. and G. Zhou (2022). Asset Pricing: Time-Series Predictability. *Oxford Research Encyclopedia of Economics and Finance*. June 20.
- Ribeiro, M. T., S. Singh, and C. Guestrin (2016). “Why Should I Trust You?” Explaining the Predictions of Any Classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144.
- Shapley, L. S. (1953). A Value for n -Person Games. *Contributions to the Theory of Games* 2:28, 307–317.
- Štrumbelj, E. and I. Kononenko (2010). An Efficient Explanation of Individual Classifications Using Game Theory. *Journal of Machine Learning Research* 11:1, 1–18.
- Štrumbelj, E. and I. Kononenko (2014). Explaining Prediction Models and Individual Predictions with Feature Contributions. *Knowledge and Information Systems* 41:1, 647–665.
- Sundararajan, M. and A. Najmi (2020). The Many Shapley Values for Model Explanation. In: *Proceedings of the 37th International Conference on Machine Learning*, pp. 9269–9278.