Machine Learning Mutual Fund Flows

Jürg Fausch* Moreno Frigg† Stefan Ruenzi‡ Florian Weigert§

This draft: June 3, 2024
First draft: April 30, 2024

Abstract

We present improved out-of-sample predictability of future fund flows using state-of-the-art machine learning methods. Nonlinear machine learning models significantly outperform linear models in terms of out-of-sample R-squared. Past flows and the Morningstar rating are identified as the most important predictors, while characteristics related to past performance are of minor importance. In addition, the interaction effect of past flows with the Morningstar rating has a substantial impact on the forecasts. Our results demonstrate that machine learning-based fund flow predictions can be used to ex-ante differentiate between high and low-performing mutual funds.

JEL Classification: C45, C52, C53, C55, G10, G11, G12, G17, G23

Keywords: Machine learning, fund flows, big data, fund flow predictability, active management

*Jürg Fausch is from the IFZ, Lucerne University of Applied Sciences and Arts, Suurstoffi 1, 6343 Rotkreuz, Switzerland. E-mail: juerg.fausch@hslu.ch

†Moreno Frigg is from the Institute of Financial Analysis, University of Neuchâtel, Rue Abram-Louis-Breguet 2, 2000 Neuchâtel, Switzerland., and the IFZ, Lucerne University of Applied Sciences and Arts, Suurstoffi 1, 6343 Rotkreuz, Switzerland. E-mail: moreno.frigg@unine.ch

‡Stefan Ruenzi is from the Department of Finance, University of Mannheim, L9, 1-2, 68161 Mannheim, Germany. E-mail: stefan.ruenzi@uni-mannheim.de

§Florian Weigert is from the Institute of Financial Analysis, University of Neuchâtel, Rue Abram-Louis-Breguet 2, 2000 Neuchâtel, Switzerland. E-mail: florian.weigert@unine.ch

1 Introduction

Open-end mutual funds offer liquidity to investors by allowing them to redeem fund shares at their daily net asset value. This can create problems for fund managers if flows are unpredictable and volatile, as it makes the liquidity management of the fund difficult. Prior literature has shown that large unexpected flows can negatively influence fund performance (Edelen, 1999) and eventually even lead to distorted asset prices (Coval and Stafford, 2007). Thus, it is of great importance for fund managers to precisely predict future flows.

In this paper, we study the investment decisions of mutual fund investors, reflected in fund flows, and reveal new evidence about the predictability of fund flows linked to mutual fund characteristics. The literature documents that many variables have an impact on flows and that these relationships often are nonlinear (Sirri and Tufano, 1998; Chevalier and Ellison, 1997), imposing an ideal setup for the application of machine learning (ML) methods. These models are not only able to capture the impact of nonlinearities and interactions between a large set of fund characteristics but also mitigate the risk of in-sample model overfitting when meaningfully calibrated. In the same fashion as Gu et al. (2020), DeMiguel et al. (2023) and Bali et al. (2023), we apply a linear OLS regression (baseline model) and machine learning methods to obtain improved fund flow predictions based on various fund characteristics. Our analysis using state-of-the-art ML techniques also allows us to identify nonlinear and hitherto unknown interaction effects that have an impact on flows, thus also contributing to a better understanding of mutual fund investors' decision making. Building on the hypothesis that flow pressure can lead to return effects on the underlying stocks a fund holds and eventually to performance persistence, we aim to differentiate between high and low performing mutual funds (out-of-sample), net of all costs, based on machine learning implied monthly fund flow predictions.

We start our empirical analyses by showing the predictive power of simple OLS regressions and different ML models for future fund flows over the period from January 2000 to December 2020. For this purpose, we apply the out-of-sample R^2 statistic which benchmarks our predictions against the historical mean flows (Campbell and Thompson, 2008). To assess

whether some models deliver significantly better predictions than other models we conduct pairwise comparisons of the predictive accuracy of different forecasting models by utilizing the model-free modified Diebold and Mariano (1995) and West (1996) test statistic.

For our empirical exercise, we first identify 30 variables that might have an influence on flows based on the prior literature. These variables are mainly linked to past performance and fund characteristics. For the simple OLS regression the high number of regressors, some of which are highly correlated, can be problematic. Thus, we first extract principal components out of the potential predictors and then use them as explanatory variables. We find that linear models like OLS and elastic net achieve a positive out-of-sample R^2 of about 14% using a historical mean flow benchmark. Nonlinear ML-models like random forest and gradient boosting improve this measure by about 5 percentage points. These results are robust across the business cycle and the predictive accuracy of the applied forecasting models is maintained even in bad economic times, which are characterized by increased outflows of risky asset classes (e.g., Jank, 2012; Pástor and Vorsatz, 2020). Furthermore, our results are confirmed based on pairwise forecast accuracy tests using modified Diebold and Mariano (1995) and West (1996) tests: Among all statistical models, the forecasts obtained by random forest are significantly better (at the 1% level) than the predictions received by any other forecasting method. This is consistent with DeMiguel et al. (2023) who use a similar dataset to forecast mutual fund performance.

While early ML prediction approaches were often criticized for being a black box with hard to interpret results, over the recent years considerable progress has been made that allows for an analysis of the specific factors that drive flow predictability. We build upon these results to quantify the relative importance of fund characteristics and their interactions for the prediction of fund flows. For that purpose, we estimate model-agnostic SHAP values (Lundberg and Lee, 2017; Lundberg et al., 2020) to determine the contribution of each predictor to the respective fund flow forecasts. Based on the best performing model, random forest, our results indicate that past flows is the most influential characteristic for predicted flows, followed by the Morningstar rating and the volatility of flows. We also document that various measures of past performance (e.g., alpha and value added) are less important in predicting

future fund flows, which is consistent with Ben-David et al. (2022) who show that fund returns and more sophisticated performance measures have little impact on flows once Morningstar ratings are included in the analysis. The importance of past flows is consistent with stable fund characteristics being important flow drivers.

Looking at the strength of the impact of these predictors over time, we find that past flows are consistently ranked as the most important predictor over the whole sample period, while the importance of flow volatility and the Morningstar rating is not as important in the very early sample years. While most other variables do not show a clear time trend, we do find that expense ratios have become more important over time.

SHAP values also allow us to determine the directional impact of each variable and we find that our three most important predictors, i.e., past flows, Morningstar rating and flow volatility, all have a predominantly positive impact on predicted flows over the whole out-of-sample period. However, when considering subperiods before 2008 the relationship between past flows and predicted flows is less conclusive, while the impact of the Morningstar rating and the volatility of flows are non time-varying. Furthermore, while high expense ratios are positively related to future flow prediction in the early out-of-sample period, the opposite is observed from 2006 onwards.

We also assess the functional form of the relationship between past flows and predicted flows as well as the functional impact of the Morningstar rating on the prediction. We document that both, past flows and the Morningstar rating have a convex impact on predicted flows. The latter effect is consistent with a nonlinear impact of past performance on flows documented in earlier studies (e.g., Sirri and Tufano, 1998).

Regarding interaction effects, the interaction of past flows with the Morningstar rating has a considerable impact on flow forecasts. The main insight in this context is that high past inflows and a high Morningstar rating imply a positive impact on predicted flows, while high past inflows and a low Morningstar rating are associated with a negative impact.

In the last step of our analysis, we investigate whether the predictability of fund flows through ML models can help in differentiating high- from low-performing funds out-of-sample

and find this to be the case. A long-short decile portfolio based on predicted flows from the machine learning model with the highest forecasting accuracy, random forest, generates an alpha of 2.76% per year, which is statistically significant at the 1% level. This outperformance is comparable to the results obtained by Kaniel et al. (2023), which use computationally more intense feedforward neural networks and a richer information set to directly predict the alpha distribution. Consistent with our findings, they identify fund flows as one of the most important predictors of future risk-adjusted fund performance.

Overall, our paper contributes to three streams of the literature. First, we contribute to the broad literature on the flow determinants of mutual funds. Since the early work of Chevalier and Ellison (1997) and Sirri and Tufano (1998), many papers have analyzed the impact of various performance measures, fund and fund manager characteristics on flows. We extend this literature by providing evidence that combining information from many individual fund characteristics, allowing for nonlinearities and by uncovering new evidence on the impact of hithertho unnoticed interaction effects, helps to improve mutual fund flow predictions. Second, we also contribute to a growing literature of applying ML techniques in finance.[1] So far, the majority of papers apply ML models to predict stock returns (see, among others, Gu et al., 2020; Chen et al., 2024; Neuhierl et al., 2022; Leippold et al., 2022). Bianchi et al. (2021) and Bali et al. (2022) apply ML models to forecast bond returns, while the predictability of delta-hedged option returns is examined by Bali et al. (2023). Medeiros et al. (2021) and Hauzenberger et al. (2023) use ML techniques to forecast inflation and Goulet Coulombe et al. (2022) provide evidence that nonlinearities in features, exploited by ML models, help to predict macroeconomic variables. Finally, in the mutual fund literature, ML techniques have mainly been applied to predict fund performance (Li and Rossi, 2021; DeMiguel et al., 2023; Kaniel et al., 2023). We complement this literature by showing how ML techniques can be used to improve flow forecasts.

The remainder of the paper is structured as follows. Section 2 describes the data used in the empirical analysis. In Section 3, we provide information about the statistical methods used for prediction and discuss the performance evaluation metrics to evaluate our forecasts. Section 4

[1] For a review on this topic, see Giglio et al. (2022) as well as Kelly and Xiu (2023).

presents the main empirical results of the paper. We look at the out-of-sample predictive ability and discuss which predictors matter most. We also analyze the predictive ability of fund flow forecasts to differentiate ex-ante between high and low performing mutual funds. Finally, we conclude in Section 5.

2 Data

Our primary data sources are the Center for Research in Security Prices (CRSP) Survivor-Bias-Free US Mutual Fund database and MorningstarDirect.

2.1 Sample data

Our sample includes all institutional and retail share classes that focus on the US equity market and spans from January 1980 to December 2020. All share classes that charge front-end or back-end loads are excluded.[2] If there is no information available about these fees we assume that the corresponding share classes do not charge loads. To select the relevant share classes, we apply some filters that are commonly used in the mutual-fund literature. First, we restrict our analyses to share classes of actively managed funds, therefore excluding ETFs and passive funds. More specifically, for all share classes available before 2003 we follow the approach suggested by Gil-Bazo and Ruiz-Verdú (2009) and exclude all share classes that contain specific keywords indicating passively managed funds. Beginning in 2003, we use the corresponding identifier provided by CRSP to select these share classes. Second, we require that each included share class invests more than 70 percent of their total net assets (TNA) into equities. Third, to avoid incubation bias we only include share classes with at least 36 months of age and whose TNA reaches USD five million at least once during their life span (see among others Elton et al., 2001; Evans, 2010).

To extend the set of characteristics we rely on fund manager data provided by Morningstar. In this context, according to Massa et al. (2010), Morningstar is a more important source of

[2]This filter is relevant when building portfolios based on flow predictions and evaluating their performance net of all costs in Subsection 4.3.

information for investors than CRSP and provides us with the Morningstar rating which seems to be an important flow determinant (Del Guercio and Tkac, 2008; Reuter and Zitzewitz, 2021). To match the share classes between these two databases, we use the fund's nine-digit CUSIP (Hillert et al., 2024) as a common unique identifier. To avoid well-known inconsistencies between the two databases (Elton et al., 2001; Berk and van Binsbergen, 2015) we perform various robustness checks. Our final data set contains 9,721 unique share classes corresponding to a matching rate of 93.85 percent. Based on this sample Table 1 reports the number of share classes for each year and the corresponding return distribution described by its mean, standard deviation, lower (Q1) and upper quantile (Q3).

2.2 Mutual fund characteristics

Guided by the mutual fund literature we use 30 characteristics that might have an influence on mutual fund flows. None of these characteristics are holding-based and thus relatively simple to calculate using data provided by CRSP and Morningstar. Our main variable of interest is monthly fund flow determined as

$$flow_{i,t} = \frac{TNA_{i,t} - TNA_{i,t-1}\,(1 + R_{i,t})}{TNA_{i,t-1}}, \tag{2.1}$$

where $R_{i,t}$ denotes the monthly net return of share class i. To account for possible outliers caused by fund mergers and splits, and as common in the literature, monthly flows are cross-sectionally winsorized at the 0.01 and 0.99 percentiles (see, among others, Huang et al., 2022; Hillert et al., 2024).

For each share class i in month t we obtain the total net assets ($TNA_{i,t}$), expense ratio ($ER_{i,t}$), the turnover ratio and return net of fees from CRSP. Excess returns ($R^e_{i,t}$) are then computed by subtracting the one-month T-bill yield which serves as a proxy for the risk-free rate.[3] Moreover, several characteristics are based on time-series regressions of share-class excess returns on well-known systematic risk factors (Fama and French, 1993; Carhart, 1997; Fama and French, 2015). While there is still a debate about which asset pricing model is used for

[3]The monthly risk-free rate data are obtained from Ken French's website.

risk adjustment by mutual fund investors (Berk and van Binsbergen, 2016; Barber et al., 2016; Jegadeesh and Mangipudi, 2021), we rely on five different factor models: the capital asset pricing model (CAPM), the Fama and French (1993) three-factor model (FF3), the Carhart (1997) four-factor model (FF4), the Fama and French (2015) five-factor model (FF5) and the Fama and French (2015) five-factor model plus momentum (FF6). The time-series regression for the resulting six-factor model is

$$\begin{aligned} R^e_{i,t} = \alpha_{i,t} + b_{i,t}(R_{M,t} - R_{f,t}) + s_{i,t}SMB_t + h_{i,t}HML_t \\ + r_{i,t}RMW_t + c_{i,t}CMA_t + m_{i,t}MoM_t + \varepsilon_{i,t} \end{aligned} \tag{2.2}$$

where $R^e_{i,t}$ is the excess return on share class i for month t, $R_{f,t}$ is the risk-free rate, $R_{M,t}$ is the market return, SMB_t and HML_t are the size and value-growth returns, MoM_t is the momentum return, $\alpha_{i,t}$ is is the average return left unexplained by the asset pricing model (the α_i estimate) and $\varepsilon_{i,t}$ is the regression residual.[4] The other applied factor models are nested in Equation 2.2 and are the CAPM in which (dropping time subscript) $R_M - R_f$ is the only factor, the Fama and French (1993) three-factor model that adds SMB and HML, the Carhart (1997) four-factor model that considers MoM as an additional explanatory variable, the Fama and French (2015) five-factor extension that adds RMW and CMA to the three-factor model, and the six-factor model that adds MoM to the five-factor model. We estimate the average monthly alpha of share class i for the different factor models implied by the regression in Equation 2.2 based on a rolling-window regression from period $t-36$ to t. In addition to the estimated alphas, the corresponding t-statistics are computed for each forecasting horizon based on the time-series regression in Equation 2.2 and used as additional predictors. The t-statistics are based on heteroskedasticity- and autocorrelation-robust (HAC) standard errors according to Newey and West (1987)[5] and account for the estimation error in alpha and the factor loadings. Following Amihud and Goyenko (2013), for each forecasting horizon, we use the adjusted *adj.* $R^2_{i,t}$ from the six-factor model rolling-window regression as a predictor of future

[4] Note that our computations are based on log-returns.

[5] We set the number of lags to three, implied by the formula $4(T/100)^a$ where a is defined as $2/9$ (Newey and West, 1994).

fund flow. The *adj.* $R^2_{i,t}$ is a measure of activeness and a lower *adj.* $R^2_{i,t}$ means that the share class tracks the benchmark less closely. In addition, we compute the $Age_{i,t}$ of each share class, which is the rounded number of months between its inception date and the last calendar date of month t. The return volatility for each share class is calculated based on a rolling window approach with the requirement of at least 30 months of non-missing return observations over the last 36 months. In addition, we compute volatility of flows (Rakowski, 2010) as the standard deviation of the winsorized flows using the same rolling window procedure. The Sharpe ratio for each share class i is defined as the excess return divided by the return volatility. As an additional predictor we compute the value added as suggested by Berk and van Binsbergen (2015):

$$va_{i,t} = (\alpha_{i,t} + ER_{i,t}/12) \cdot (TNA_{i,t-1}) \tag{2.3}$$

where the realized $\alpha_{i,t}$ for each forcasting horizon is obtained from the six-factor model defined in Equation 2.2 and $ER_{i,t}$ denotes the annual expense ratio of share class i. This version of value added follows DeMiguel et al. (2023) and differs from Berk and van Binsbergen (2015) which estimate before-fee alpha using passively managed index funds as benchmarks. Based on information provided by Morningstar we define a dummy variable, *size of management team*, which is set to 1 if the share class is team-managed and 0 otherwise to consider the impact of the organizational structure on future fund flows (Bär et al., 2011). Finally, we also add the *Morningstar rating* (Del Guercio and Tkac, 2008; Reuter and Zitzewitz, 2021; Evans and Sun, 2021) to the set of fund characteristics used to predict mutual fund flows. The Morningstar rating assigns one to five stars based on a mutual fund's historical risk and load-adjusted returns versus category peers. Table 2 shows the descriptive statistics of the mutual fund characteristics considered in the empirical analysis with the number of observations and the percentage share of missing data (NaN). All variables are measured at the share class level in month t. Figure 1 shows the correlation matrix of the predictor variables. High correlations are observed for characteristics related to the various factor models.

2.3 Data preprocessing

Before we can apply machine learning methods for predicting future mutual fund flows, we need to preprocess the data to generate the target and predictor variables which requires us to make several implementation decisions. The target variable is defined as winsorized monthly mutual fund flows, $flow_{i,t+h}$ for a forecasting horizon of $h = 1$ month. The predictor variables correspond to the mutual fund characteristics outlined in Subsection 2.2. For our analyses, we use non-imputed data and thus avoid relying on a specific imputation method, which implies omitting any observation for which at least one characteristic or the target variable is not available in a given month. Moreover, to determine the predicted flow distribution we build an additional data set, the prediction sample, that is used for the actual prediction and includes share classes with missing target variables to avoid any form of forward-looking bias. More specifically, when utilizing the training dataset to generate predictions at time t, it is important to note that no predictive inference would be obtained for share classes lacking a corresponding target variable at time $t + h$. This is an implication of excluding share classes with missing target variables from the training sample. After all these adjustments, our training sample consists of 603,542 observations (7,072 unique share classes) and the prediction sample contains 607,897 observations and 7,097 unique share classes. Both samples range from January 1989 to December 2020.[6]

To ensure scale invariant predictors, we follow Green et al. (2017) and standardize each variable[7] used in OLS and the applied machine learning algorithms in this paper, except for the neural networks[8], to have a mean of zero and a standard deviation of one. To avoid a look-ahead bias, this standardization is based on the data used in the training set specific to the forecasting cycle. To mitigate potential multicollinearity issues with respect to OLS estimation due to some highly correlated fund characteristics (see Figure 1), we orthogonalize the set of standardized predictors by applying principal component analysis (PCA). As outlined by Gewers et al.

[6]The decrease in the samples from January 1980 to January 1989 is mostly attributable to missing monthly TNA data, which prevents us from reliably computing fund flows, $flow_{i,t}$. The first monthly observations of TNA start in January 1989.

[7]No standardization is required for the variables size of the management team since this predictor is a dummy variable.

[8]In the context of neural networks, we adopt a min-max scaling technique between the range of -1 and 1.

(2021), standardization prior to the application of PCA ensures the equivalence of importance of all characteristics, which therefore excludes an assessment of the most important variables for prediction. The subset of principal components used as predictors in OLS estimation explains more than 95 percent of the variation in the training data and represents a low-dimensional set of features from the larger set of predictive variables.[9] Moreover, note that the number of principal components used for prediction might change in each forecasting cycle. The applied machine learning algorithms do not require such orthogonalization.

3 Machine-Learning algorithms and performance evaluation

Machine learning methods are well-suited for predictive tasks and have recently been applied in the context of empirical asset pricing. We follow this literature (Gu et al., 2020; Li and Rossi, 2021; Bali et al., 2022; DeMiguel et al., 2023; Kaniel et al., 2023; Murray et al., 2024, among others) by comparing and evaluating a variety of machine learning algorithms in their ability to predict future flows of share classes. The applied machine learning models, in an increasing degree of complexity, are elastic net, decision trees, random forest, gradient boosting, and feedforward neural networks (FFNN).[10] We use these methods as they are reported as the most effective machine learning methods in various asset pricing studies (e.g., among others Gu et al., 2020; Bianchi et al., 2021; Li and Rossi, 2021; DeMiguel et al., 2023). The simple linear regression model estimated via OLS serves as a benchmark.

Our data are organized in a panel structure, with months indexed as $t = 1, 2, \ldots, T$ and share classes as $i = 1, 2, \ldots, N$. In this section, we briefly describe these algorithms, the five-fold cross-validation approach used for hyperparameter tuning as well as the performance evaluation methodology.

[9] The main goal of applying PCA in the context of OLS estimation is to orthogonalize highly correlated predictors and not to reduce dimensionality. In contrast to OLS, the elastic net has good properties in handling highly correlated-predictors (Zou and Hastie, 2005; Diebold and Shin, 2019), which makes the application of PCA redundant.

[10] All algorithms are implemented in Python using the *scikit-learn* package, except for neural networks where we rely on *TensorFlow*.

3.1 Linear regression - baseline model

The simple linear predictive regression model serves as a benchmark in our prediction exercise. The model is estimated via ordinary least squares (OLS) by solving the following optimization problem, which yields the pooled OLS estimator:

$$\min_{\theta} \mathcal{L}(\theta) = \min_{\theta} \sum_{t=1}^{T-h} \sum_{i=1}^{N} \left(flow_{i,t+h} - z'_{i,t-1}\theta_{t-1} \right)^2 \tag{3.1}$$

where $flow_{i,t+h}$ are the realized fund flows of the ith share class in month $t+h$, $z_{i,t-1}$ is a P-dimensional vector of standardized characteristics (features) for the ith share class in month $t-1$, and $\theta_{t-1} = (\theta_{0,t-1}, \theta_{1,t-1}, \ldots, \theta_{P,t-1})'$ is the P-dimensional parameter vector. The forecast of fund flow in month $t+h$ is then obtained as:

$$\widehat{flow}_{i,t+h} = z'_{i,t}\hat{\theta}_{t-1}. \tag{3.2}$$

If the number of predictors P is relatively small, the OLS estimate of θ is unbiased and efficient. However, the OLS model tends to perform poorly when the size of the predictors is large (overfitting) and nonlinearities and interactions among predictors are important to the forecast. In our empirical analysis, we consider three machine-learning methods that have been recently used in the empirical asset pricing literature: elastic net, tree-based methods (e.g., decision tree, random forest, gradient boosting) and feed-forward neural networks. All these methods can handle large numbers of predictor variables with less risk of overfitting compared to OLS. In addition, tree-based methods and neural networks are able to exploit nonlinearities and interactions, that are not detected by OLS and linear machine learning models such as elastic net and may result in more accurate forecasts.

3.2 Elastic net

A common approach to mitigate the overfitting issue is achieved by adding a penalty term to the OLS loss function in Equation 3.1:

$$\min_{\theta} \mathcal{L}(\theta; \cdot) = \min_{\theta} [\mathcal{L}(\theta) + \phi(\theta; \cdot)], \tag{3.3}$$

where $\phi(\theta; \cdot)$ is the penalty on θ. We follow Zou and Hastie (2005) and use the elastic net penalty, which consists of two regularization terms and takes the form

$$\phi(\theta; \lambda, \rho) = \lambda (1 - \rho) \sum_{j=1}^{P} |\theta_j| + \frac{1}{2} \lambda \rho \sum_{j=1}^{P} \theta_j^2. \tag{3.4}$$

The elastic net has two positive hyperparameters, $\lambda > 0$ and $\rho > 0$, which are determined separately by cross-validation. The loss function in Equation 3.3 reduces to OLS when $\lambda = 0$. $\rho = 0$ corresponds to the Least Absolute Sum of Squares (LASSO) operator and uses the l_1 norm or absolute value penalties for penalized regression. LASSO imposes sparsity on the model specification in the sense that a subset of the parameters θ is exactly zero and can be used for variable selection. $\rho = 1$ corresponds to a ridge regression model, which relies on l_2-norm parameter penalization, which shrinks all coefficients closer to zero, but does not impose exact zeros anywhere. For values of $\rho \in (0, 1)$, both shrinkage and selection are imposed on the model. Moreover, the elastic net has good properties in handling highly-correlated predictors (Zou and Hastie, 2005; Diebold and Shin, 2019).

3.3 Decision tree

Decision trees are flexible, nonparametric models that split the feature space into K partitions based on one of the predictive variables:

$$g(z_{i,t}; \theta, K, L) = \sum_{k=1}^{K} \theta_{t-1}^{(k)} 1_{\{z_{i,t} \in C_k(L)\}} \tag{3.5}$$

where K denotes the number of terminal nodes, L is the depth, $C_k(L)$ is one of the K partitions of the data, $1_{\{\cdot\}}$ is an indicator function and $\theta_{t-1}^{(k)}$ is defined as the sample average of outcomes within the corresponding partition $k \in K$.

Decision trees are able to jointly incorporate categorical and numerical predictors, are invariant to monotonic transformations, and capture interactions and nonlinearities. Decision trees are grown using the CART algorithm of Breiman et al. (1984), and the decision on which predictor and value to use for a split is determined based on minimizing the l_2 norm. The algorithm recursively splits observations into smaller subsets (decision nodes) until no further split is possible. The predicted $flow$ for each leaf reflects the average of the realized $flow_{t+1}$ of the share classes in the training sample sorted into this leaf. Due to the complexity of the tree, a greedy approach tends to overfit the training data, and thus requires strong regularization to obtain better prediction results. To reduce the complexity of the decision tree, we apply a pre-pruning approach based on a given early stopping criterion. The stopping criterion is based on the maximum depth of the tree L that has, on average, the lowest mean squared forecast error in the data set used for cross-validation.

3.4 Random forest

Random forests are a modification of bootstrap aggregation or bagging (Breiman, 2001). The basic idea is to randomly draw $b \in B$ bootstrap samples, $\{(z_{i,t}, \alpha_{i,t+h}), (i,t) \in Bootstrap(b)\}$ from the original data set and grow a decision tree

$$g_b(z_{i,t}; \theta_{b,t-1}, K, L) = \sum_{k=1}^{K} \theta_{b,t-1}^{(k)} 1_{\{z_{i,t} \in C_k(L)\}}. \tag{3.6}$$

The final prediction of the random forest is given by the average of the outputs of all B decision trees resulting in an ensemble forecast:

$$g(z_{i,t}; K, L, B) = \frac{1}{B} \sum_{b=1}^{B} g_b(z_{i,t}; \theta_{b,t-1}, K, L). \tag{3.7}$$

To estimate $\theta_{b,t-1}^{(k)}$, we follow the algorithm of Breiman et al. (1984). The hyperparameters to be tuned are the depth of trees L, the number of predictors P considered as split variables at each node, the size of each bootstrapped sample b and the number of bootstrapped samples

B to grow a decision tree.

3.5 Gradient boosting

Gradient boosting is a method that combines multiple oversimplified (shallow) decision trees, known as weak learners, into a single strong learner with potentially greater stability than a single complex tree (Hastie et al., 2009). It is based on an iterative procedure. In the first step, an oversimplified tree with a high forecast error ($flow_{i,t+h} - \widehat{flow}_{i,t+h}$) is computed to predict $\widehat{flow}_{i,t+h}$. As suggested by Hastie et al. (2009) the loss function to be optimized is based on the l_2 norm since the target value is continuous. In the next step, a second shallow tree is used to fit the forecast residuals from the first tree. The predictions from these two trees are then added together to form an ensemble prediction. To prevent the model from overfitting the forecast residuals, the forecast component from the second tree is shrunken by a factor $\nu \in (0, 1)$. This hyperparameter corresponds to the learning rate and determines the weight the ensemble gives to the most recent decision tree. Furthermore, the risk of overfitting is also controlled by the depth of the tree L. At each iteration step, $b \in B$, an additional shallow tree is fitted to the residuals from the preceding ensemble prediction based on the model with $b-1$ trees, and its residual forecast is added to the total ensemble forecast with the pre-defined shrinkage weight of ν, resulting in a stronger model. This iteration continues until there are a total of B trees in the ensemble.

More formally, starting with $\hat{g}_0(z_{i,t}) = 0$, for each boosting iteration b from 1 to B, for each $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$ the negative gradient of the loss function $l(\cdot, \cdot)$ based on the l_2 norm

$$\varepsilon_{i,t+h} \leftarrow -\frac{\partial l\left(flow_{i,t+h}, g\right)}{\partial g} | g = \hat{g}_{b-1}\left(z_{i,t}\right) \tag{3.8}$$

is computed and a shallow regression tree of depth L is fit to the training data $\{(z_{i,t}, \varepsilon_{i,t+h}) : \forall i, \forall t\}$

$$\hat{f}_b(z_{i,t}) \leftarrow g\left(z_{i,t}; \theta, L\right). \tag{3.9}$$

$\hat{g}_b(z_{i,t})$ is updated by adding a shrunken version of the new tree together with an update of the residuals:

$$\hat{g}_b(z_{i,t}) \leftarrow \hat{g}_{b-1}(z_{i,t}) + \nu \hat{f}_b(z_{i,t}), \tag{3.10}$$

$$\varepsilon_{i,t+h} \leftarrow \varepsilon_{i,t+h} - \nu \hat{f}_b(z_{i,t}). \tag{3.11}$$

The output of the boosted model is

$$\hat{g}_B(z_{i,t}; B, \nu, L) = \sum_{b=1}^{B} \nu \hat{f}_b(z_{i,t}). \tag{3.12}$$

The final result is thus an additive model of shallow trees. The three tuning parameters (L, ν, B) are determined by cross-validation. We rely on histogram-based gradient boosting because it is known to reduce computational time while maintaining high accuracy (Ke et al., 2017).

3.6 Feedforward neural network

Feedforward neural networks (FFNN) are very flexible and highly parameterized methods for solving complex machine learning problems due to their ability to entangle many overlapping layers of nonlinear predictor interactions. These algorithms are very often referred to as deep learning. In our empirical analysis, we follow Gu et al. (2020) and Kaniel et al. (2023), and rely on multi-layer feedforward neural networks. These consist of an input layer of raw predictors at least one hidden layer that interacts and nonlinearly transforms the predictors, and an output layer that aggregates hidden layers into a prediction (Svozil et al., 1997).

In a FFNN, each neuron is connected to all neurons in the previous layer, and the connections are one-way from the input layer to the output layer. The number of units in the input layer is equal to the P-dimensional vector of predictor variables z_i defined in Subsection 2.2. The nonlinear transformation applied to the predictor variables is based on an element-wise

activation function. More specifically, each neuron $k \in K$ in the hidden layer j transforms inputs from the previous hidden layer, $x_i^{(j-1)}$, into output according to:

$$x_k^{(j)} = f\left(\theta_{k,0}^{(j)} + \sum_{i=1}^{P} x_i{}^{(j-1)} \theta_{k,i}^{(j)}\right) \tag{3.13}$$

where $f(\cdot)$ represents a nonlinear activation function, θ is a P-dimensional parameter vector that includes an intercept $\theta_{k,0}^{(j)}$ (bias term) and one weight parameter for each predictor ($\theta_{k,i}^{(j)}$). Since multiple options for the nonlinear activation function exist, we aim to identify the optimal function for each hidden layer during the tuning process by considering rectified linear unit (ReLU), sigmoid and tanh. The chosen activation function will be implemented across all neurons within the corresponding hidden layer. A deep neural network that combines multiple layers uses the outputs of one hidden layer as an input to the next hidden layer. In the final hidden layer, the results from each activation are then fed into the output layer, resulting in

$$g(z;\theta) = \theta_0 + \sum_{k=1}^{K} x_k \theta_k, \tag{3.14}$$

a linear regression model in the K activations, providing the ultimate prediction of $\widehat{flow}_{i,t+h}$.

The (weight) parameters in our FFNN are computed iteratively using the back-propagation algorithm *Adam* suggested by Kingma and Ba (2014). The algorithm is initialized by assigning random weights to each connection of neurons. In the learning process, after each epoch[11], the algorithm sequentially updates the weights and biases of the FFNN based on minimizing a mean squared error loss function. To find the best possible architecture of our network, we search for the optimal number of hidden layers, the optimal number of neurons per hidden layer, the optimal activation function per hidden layer, the dropout rate per hidden layer, the number of epochs the model is trained on and the learning rate[12] which determines the iterative update of weights based on minimizing the MSFE. By not deciding ex-ante on the nonlinear activation function, fixing the number of hidden layers or by choosing the number of neurons in

[11] Epochs refer to the number of times the fitting algorithm passes through the full training set.

[12] We start with a learning rate of 0.003 and reduce it with a factor of 0.5 if the mean-squared forecast error (MSFE) does not decrease for two consecutive epochs.

each hidden layer according to the geometric pyramid rule (Masters, 1993) we allow for a lot of flexibility in building the network architecture. In particular, we employ two types of FFNNs that solely differ in the number of hidden layers and number of neurons per hidden layer. Neural network I (NN I) permits a lower set hidden layers (between 1 and 3) and neurons per hidden layer (ranging from 2 to 32) while neural network II (NN II) is a more complex model with 3 to 10 hidden layers and 32 to 1,024 neurons per hidden layer. Due to the stochastic nature and similar to Kaniel et al. (2023) and Chen et al. (2024), all our neural networks are averaged over 10 model fits after finding the optimal hyperparameters.

3.7 Cross-validation, hyperparameter tuning and performance evaluation

Hyperparameter tuning and training: To find the hyperparameters with the best model fit, we use k-fold cross-validation (CV)[13] (Hastie et al., 2009), except for the neural network, where we rely on a randomly selected validation set to minimize computation time. As shown by Bergmeir et al. (2018) and Goulet Coulombe et al. (2022) k-fold CV performs favorably to alternative techniques that explicitly account for the time-series properties of the data. DeMiguel et al. (2023) confirm this finding when predicting mutual fund alpha using machine learning methods. The k-fold CV approach involves dividing the data set into five folds of approximately equal size while maintaining the temporal ordering. Four folds are used as a training sample with a given combination of hyperparameters and the remaining fold is used as a validation set to evaluate the out-of-sample performance of the trained model based on the chosen hyperparameters using the MSFE (cross-validation error). After completing this procedure for each of the five folds, the hyperparameters that minimize the average cross-validation error are selected for the corresponding machine learning algorithms. For all algorithms we rely on the *Optuna* hyper-parameter tuning approach suggested by Akiba et al. (2019). Compared to commonly used greedy grid-search and random-search approaches, *Optuna* provides more sophisticated sampling approaches based on Bayesian optimization techniques to find the best

[13] In our application, we set $k = 5$.

possible set of hyperparameters. In this context, we use the tree-structured Parzen estimator approach (TPE) based on independent sampling suggested by Bergstra et al. (2011). In particular, the TPE sampler models the search space by using one Gaussian mixture model $l(x)$ to the set of parameter values associated with the best objective values, and another Gaussian mixture model $g(x)$ to the remaining parameter values. Finally, the parameter value x that maximizes the ratio $l(x)/g(x)$ is chosen.[14] To optimize the computational time for the hyperparameter tuning of the neural networks, we additionally use the hyperband algorithm proposed by Li et al. (2017) as a pruning mechanism and a callback function that stops the training of the neural network if the value of the MSFE on the validation set does not further decrease after five epochs.

As machine learning algorithms are computationally intensive, we determine the best possible set of hyperparameters of the predictive models every two years instead of every month (similar as in Gu et al., 2020; Leippold et al., 2022). However, after each monthly forecast, we extend the training sample by one month and retrain the algorithms using the previously obtained hyperparameters. In particular, the hyperparameters used in the first forecast of fund flows in January 2000 are based on the training sample using data from January 1989 to November 1999 and the prediction set of December 1999. These hyperparameters remain unchanged for all monthly fund flow forecasts until December 2001 while the training sample is extended every month. For predictions in January 2002, new hyperparameters using an extended training sample from January 1989 to November 2001 are determined. This procedure is repeated until the end of the sample period.

Evaluating out-of-sample predictive ability: To gauge the out-of-sample predictive ability of all applied machine learning models and the benchmark OLS model we compute the out-of-sample R^2 (OOS R^2) statistic (Welch and Goyal, 2008; Campbell and Thompson, 2008) on a share class level, which is defined as

[14] In our application, for all machine learning algorithms, except neural networks, we set the number of trials required for the optimization routine to choose a suitable set of hyperparameter values from a given range to 150 and for the neural networks to 200.

$$R^2_{OOS} = 1 - \frac{\sum_{(i,t)\in\tau_{OOS}} \left(flow_{i,t+h} - \widehat{flow}_{i,t+h}\right)^2}{\sum_{(i,t)\in\tau_{OOS}} \left(flow_{i,t+h} - \overline{flow}_{i,t}\right)^2} \tag{3.15}$$

and measures the reduction in the MSFE compared to a benchmark consisting of the historical mean flow for fund i up to period t, $\overline{flow}_{i,t}$. The forecasting horizon is $h = 1$ month. The predictive power is evaluated on a sample, τ_{OOS}, that is disjoint from the data used in model estimation and hyperparameter tuning. The OOS R^2 statistics pools forecast errors across share classes and over time into a panel-level assessment of each model. We compute the statistic for each monthly forecasting cycle as well as for the entire out-of-sample period. To evaluate the statistical significance of each model we apply the Clark and West (2007) statistic to test $H_0 : R^2_{OOS} \leq 0$ against $H_1 : R^2_{OOS} > 0$.

Forecast comparison: To compare the predictive performance of two forecasting methods, $m = 1, 2$, consisting of machine learning methods and linear models, we apply the Diebold-Mariano-West (DMW) pairwise test (Diebold and Mariano, 1995; West, 1996). For this purpose, we define the loss differential as

$$\hat{d}_t^{(1,2)} = \left(\hat{e}^{(1)}_{i,t+h}\right)^2 - \left(\hat{e}^{(2)}_{i,t+h}\right)^2 \tag{3.16}$$

where $\hat{e}^{(m)}_{i,t+h} = flow_{i,t+h} - \widehat{flow}_{i,t+h}$ refers to the forecast error on share class i at time $t+h$ for method m. The DMW statistic to test the null hypothesis of equal predictive accuracy, $H_0 : E\left(\hat{d}_t^{(1,2)}\right) = 0$, against the two-sided alternative is obtained as the t-statistic of a regression with intercept only, $\hat{d}_t^{(1,2)} = \mu + \varepsilon_t$:

$$DMW^{(1,2)} = \frac{\hat{\mu}^{(1,2)}}{\hat{\sigma}^{(1,2)}_{\hat{\mu}}} \tag{3.17}$$

where $\hat{\mu}^{(1,2)}$ denotes the estimated coefficient (time series average of $\hat{d}_t^{(1,2)}$) and $\hat{\sigma}^{(1,2)}_{\hat{\mu}}$ is the corresponding Newey and West (1987) HAC standard error. It is well known that the DMW test can reject the null hypothesis too often, depending on the sample size and the degree of

serial correlation in the forecast errors. To address this issue and obtain improved small-sample properties we follow Harvey et al. (1997) and make a bias correction to the DMW statistic in (3.17). The corrected test statistic is obtained as

$$HLN - DMW^{(1,2)} = \sqrt{\frac{T + 1 - 2h + T^{-1}h(h-1)}{T}} DMW^{(1,2)}. \tag{3.18}$$

which is compared to the critical values of a Student t-distribution.

4 Empirical analysis

In our empirical analysis, we explore the predictive ability of the applied statistical models via out-of-sample R^2 and discuss the importance of fund characteristics and their interactions in predicting future fund flows. Finally, we evaluate whether fund flow prediction can be utilized to consistently differentiate high-performing from low-performing mutual funds. Although our analyses are conducted at the share class level, for simplicity, we refer to share classes as funds in the remainder of this paper.

4.1 Out-of-sample predictability comparison

We begin our analysis by reporting the out-of-sample results of applying OLS regressions and state-of-the-art machine learning algorithms to predict future monthly fund flows. The obtained results are evaluated using the out-of-sample R^2 statistic described in Subsection 3.7. In addition, to discriminate between the forecasting performance of two competing models, we apply the DMW test.

Table 3 shows the forecasting performance of all applied prediction models based on the out-of-sample R^2 measure. Both, OLS and machine learning algorithms, achieve positive out-of-sample R^2 and show predictive ability for one-month future fund flows. All reported out-of-sample R^2 are statistically significant at the 1% level meaning that each applied forecasting model performs significantly better relative to the corresponding benchmark. In terms of linear statistical methods, our baseline model, OLS, performs slightly better than the elastic net,

implying that the effect of dimension reduction inherent to the elastic net does not lead to improved out-of-sample forecasts. Compared to OLS (14.86%), the decision tree raises the R^2_{OOS} by about 1.57 percentage point. A further increase in predictive accuracy is achieved by random forest and gradient boosting, improving the R^2_{OOS} to above 19%. While the two neural networks perform slightly better than the decision tree, they do not beat random forest and gradient boosting. A comparison of the predictive accuracy of NN I and NN II reveals that the more complex NN II achieves a marginally higher out-of-sample R^2 of 0.38 percentage points. Although the NN II is much more computationally intense, this large computational effort is not justified as random forest achieves superior prediction results in much less computational time. The improved predictive ability of machine learning methods like random forest and gradient boosting demonstrate their effectiveness and dominance in capturing nonlinearities and complex interactions between predictors that appear to be relevant in achieving higher prediction accuracy. The magnitude of improvement in out of sample R^2 is comparable to those achieved in studies from the asset pricing literature using machine learning to forecast monthly stock (Gu et al., 2020; Leippold et al., 2022) and bond returns (Bali et al., 2022). These papers report an improvement in out-of-sample R^2 of using nonlinear ML methods relative to OLS in the range of 1.90 to 5.73 percentage points, while our best performing forecasting model, random forest, yields an improvement in R^2_{OOS} of 4.83 percentage points.

Next, we explore whether the predictive ability of our models depends on the state of the economy. As described in Jank (2012), investors react to macroeconomic news by leaving riskier asset classes and entering less risky ones when there is news of an economic downturn. This is empirically confirmed by Pástor and Vorsatz (2020), as during the COVID-19 crisis, actively managed mutual funds experienced rapid outflows during and after the crash. In this context, it is important to note that the models used to forecast fund flows have a high degree of predictive accuracy independent of the state of the economy. Table 4 reports the out-of-sample R^2 computed separately for expansions and recessions. The definition of expansions and recessions follows the NBER convention. All forecasting methods have slightly higher out-of-sample R^2 measures in expansions. To formally test whether the average forecast accuracy

differs significantly in expansions and recessions, we apply a Welch (1947) test. For both OLS and the machine learning algorithms, there is no statistically significant difference in predictive ability concerning the state of the business cycle. This suggests that the predictive power of linear and nonlinear forecasting methods is robust and independent of the state of the economy. However, it is important to note that the monthly average forecast accuracy based on the R^2_{OOS} metric is slightly higher in recessions, resulting in a negative, non-significant t-statistic.

To statistically evaluate the forecasting ability of two competing forecasting models we apply modified Diebold and Mariano (1995), West (1996) tests following Equation 3.18. The results of the pairwise tests of predictive accuracy are shown in Table 5. We find that different classes of linear and nonlinear models differ considerably in their predictive performance. Based on the $HLN - DMW$ statistic suggested by Harvey et al. (1997), machine learning methods that capture nonlinearities and allow for interactions outperform linear models. Regarding neural networks, the more complex network structure of NN II delivers superior predictive performance (1% significance level) compared to the relatively simple architecture of NN I. However, gradient boosting and random forest manage to beat both neural networks while being less computationally expensive. Among the linear models, OLS is able to statistically, significantly outperform the predictions made by the elastic net. In terms of nonlinear models, the forecasts produced by random forest are significantly better than the predictions obtained by any other machine learning algorithm. Overall, random forest is found to be the best performing model with the highest predictive accuracy. In conclusion, these results corroborate the findings obtained by the out-of-sample R^2 analysis above.

4.2 Characteristic importance and interactions

In this subsection, we study the importance of fund characteristics and their interactions for the predictions obtained by random forest, which was identified in Subsection 4.1 as the machine learning method with the highest forecasting accuracy. For that purpose, we estimate model-agnostic SHAP (SHapley Additive exPlanations) values (Lundberg and Lee, 2017; Lundberg et al., 2020), which are based on cooperative game theory to determine the contribution of each

predictor to each individual forecast. Predictors that positively affect the prediction relative to the average forecast have positive SHAP values and vice versa. The magnitude measures the strength of the impact.[15] In our analysis, we rely on the treeSHAP algorithm[16] which delivers fast and accurate estimates of SHAP values for random forest. To quantify the importance of each characteristic, Table 6 reports the mean of the absolute SHAP values across all observations over the out-of-sample period from January 2000 to December 2020.

We find that past flows is the most influential fund characteristic for predicting future fund flows. This finding suggests that relatively stable fund characteristics and/or the relatively stable investment behavior of a fund's clientele explain a large part of the predicted flows (which thus are relatively persistent). The next important predictors are the Morningstar rating and the volatility of flows. It is noteworthy that various measures of past performance, including t-stat alpha and realized alpha for various factor models and value added, seem to be less important in predicting future fund flows. These model-based implications are consistent with the literature showing that mutual fund investors react strongly to simplistic fund rankings (Del Guercio and Tkac, 2008; Reuter and Zitzewitz, 2021; Evans and Sun, 2021) rather than to more sophisticated performance measures based on the CAPM or other commonly used asset pricing models (Ben-David et al., 2022) when allocating capital to mutual funds.

To analyze the time variability in the importance of fund characteristics in predicting mutual fund flows, Figure 2 depicts the importance of the 15 most relevant predictors (according to Table 6) using the monthly mean absolute SHAP values for random forest. The figure illustrates that past flows is consistently the most influential fund characteristic over time. The Morningstar rating and the volatility of flows are less important in the beginning of the out-of-sample period and gain in relevance over time. A similar pattern is observed for predictors such as expense ratio, volatility of flows, age and TNA as they gained in importance at a later stage (around 2005). In contrast, performance-related characteristics (realized alpha and t-stats of various asset pricing models, Sharpe ratio) were more relevant for predicting future fund flows early in the out-of-sample period and lost impact over time.

[15] A general discussion of SHAP can be found in Molnar (2019).

[16] We use the *FastTreeSHAP* package of Yang (2022) implemented in Python for computing the SHAP values.

Next, we aim to understand how various fund characteristics affect the predictions made by random forest in greater detail and thus analyze the directional impact of these predictors. In this context, Figure 3 shows a beeswarm SHAP plot of the 15 most important features identified by their mean absolute SHAP values over the whole out-of-sample period from 2000 to 2020. Predictors are sorted in descending order with the most important variables at the top consistent with the importance ranking established in Table 6. We obtain information about the magnitude (feature value) and the directionality (SHAP value) of each of these characteristics in predicting fund flows. For example, a higher Morningstar rating impacts predicted fund flows positively, while the directional relationship between past flows and predicted flows is less clear. While for most funds, high past flows have a positive impact on predicted flows some of these funds are negatively affected. A similar finding is related to the realized alphas of various factor models (CAPM, FF3, FFC). However, in many cases, a higher realized alpha has a positive directional impact on predicted flows.

To further analyze these directional relationships we divide the out-of-sample period into consecutive two-year intervals. Figure 4 shows the corresponding SHAP values. While the relationship between past flows and their impact on predicted flows is ambiguous for the entire out-of-sample period, a clearer picture emerges when separated into subperiods. For the first four periods, high past flows could negatively or positively affect future flow predictions. Starting in 2008, the relationship became more explicit and high past flows strongly positively impact predicted future flows. Additionally, for all subperiods, we observe that a higher Morningstar rating and higher volatility of flows consistently lead to a positive impact on the prediction of fund flows. In contrast, a higher fund age negatively affects fund flow predictions. Since 2006 a high expense ratio has been negatively related to predicted flows, while earlier in the out-sample period a slightly positive relationship was found. Overall, from 2008 onwards these relationships are stable over time in their directional assessment. However, the magnitude to which changes in individual predictors impact fund flow predictions varies substantially over time.

Next, we compute interaction effects for all possible combinations of predictor variables

for random forest.[17] Table 7 shows the strength of the ten most important interactions based on mean absolute SHAP interaction values. The table reveals that the Morningstar rating, volatility of flows or past performance measures such as t-stat alpha or realized alpha of various asset pricing models (CAPM, FF3, FFC) are not only important as individual predictors as shown in Table 6 but are also important through their interactions with past fund flows. The Morningstar rating with past fund flows is identified as the most important interaction with a mean absolute SHAP value of 0.0194. This interaction is more relevant than the sixth most important individual characteristic, t-stat alpha (FF3)[18], and demonstrates the substantial importance of interactions between predictors in forecasting fund flows.

We further highlight the usefulness of modeling nonlinearities and interaction effects between fund characteristics. Figure 5 and Figure 6 show the convex relationship between past flows and predicted flows (functional form) and adds the interaction effect with the Morningstar rating (colored dots) for the whole out-of-sample period from January 2000 to December 2020 and for two consecutive two-year periods over the same time span. These figures provide three main insights: First, high past inflows and a high Morningstar rating imply a positive impact on predicted flows, while high past inflows and a low Morningstar rating are associated with a negative impact. Second, while high past outflows positively impact predictions, low past outflows are mostly associated with negative SHAP values. Third, the convex relationship between past fund flows and predicted flows and the finding about the interactions of past flows with the Morningstar rating is stable over time.

Finally, in Figure 7, we further explore the nonlinear association between the Morningstar rating and the prediction of fund flows. We observe that a five-star rating is consistently associated with a positive influence on flow predictions, as indicated by positive SHAP values. Conversely, SHAP values for lower-rated funds (one to three stars) are primarily negative, suggesting that the direction of the impact on flow predictions does not vary with the rating level. The influence on predicted flows for funds with a four-star rating, the second-highest Morningstar category, appears more ambiguous.

[17] This corresponds to $(30 \times 29)/2 = 435$ interactions.

[18] Note that the SHAP value of t-stat alpha (FF3) of 0.0175 (see Table 6) also includes interaction values of this predictor with all other characteristics.

4.3 Predicted fund flows and performance

The academic literature documents a significantly positive relationship between past fund flows and future fund performance. One strand of the literature attributes this empirically observed flow-performance relationship to investors' ability to identify managerial skill, known as smart money effect (among others Gruber, 1996; Zheng, 1999; Keswani and Stolin, 2008), while another strand attributes it to the persistence of fund flows (among others Coval and Stafford, 2007; Frazzini and Lamont, 2008; Lou, 2012). Our goal in this section is not to analyze the drivers of the positive flow-performance relationship, but rather to evaluate whether machine learning-based fund flow predictions can be used out-of-sample to consistently differentiate high- from low-performing mutual funds net of all costs. We follow Gu et al. (2020), Bali et al. (2023), DeMiguel et al. (2023) and Kaniel et al. (2023) and form portfolios using machine learning forecasts. Specifically, we first sort individual share classes into equally weighted top and bottom decile portfolios based on their predicted fund flow forecasts for the next month. Second, we compute the average realized return for each portfolio in the corresponding month. For every successive month, we follow the procedure outlined in Subsection 3.7 and expand the training sample by one month, train the algorithm on the expanded sample and construct new portfolios using one-month ahead machine learning implied fund flow predictions. As a result, we obtain a time series of monthly out-of-sample net returns of the top and bottom decile portfolios over the period from January 2000 to December 2020 (252 monthly observations). Finally, as in Kaniel et al. (2023), we construct a long-short prediction portfolio of each top and bottom decile portfolio to evaluate whether spanning based on high and low fund flow forecasts results in positive mutual fund alpha. As in Kaniel et al. (2023) and DeMiguel et al. (2023) the out-of-sample performance of all fund portfolios is evaluated by running a time-series regression of the 252 out-of-sample monthly portfolio excess returns on various risk-factor returns. The portfolio alpha is obtained as the intercept of the time-series regression. We use four risk-factor models to evaluate portfolio performance: the Carhart (1997) four-factor model (FF4), the Fama and French (2015) five-factor model (FF5), the FF5 model augmented with the momentum factor (FF6) and the FF6 model enhanced by the liquidity factor of Pástor

and Stambaugh (2003) (FF7). The fund flow predictions are obtained using the benchmark linear OLS model, a linear machine learning algorithm (elastic net) and five machine learning algorithms that allow for nonlinearities and interactions (decision tree, random forest, gradient boosting, two neural networks). For comparison, the result of a naïve portfolio strategy consisting of an equally weighted portfolio of all share classes is used as a benchmark. Table 8 reports the net of all costs monthly out-of-sample alphas of the top and bottom decile portfolios and the corresponding long-short portfolio implied by the predicted fund flow distribution based on one-month ahead forecasts. Our results reveal three main findings. First, by using fund-flow predictions we are able to ex-ante separate high from low-performing funds, except for the FF4 model, resulting in significant out-of-sample alphas for the long-short portfolios. The best performing model is random forest which delivers a highly significant alpha of 23 bp per month (2.76% per year) relative to the FF7 model. However, it is important to note that linear models (OLS, elastic net) perform reasonably well and achieve higher long-short alphas compared to the decision tree and the simpler of the two neural networks. Second, unlike DeMiguel et al. (2023), which directly forecast fund performance and build portfolios sorts on the predicted alpha distribution, our machine learning implied fund flow predictions do not select long-only fund portfolios that provide statistically significant out-of-sample alphas. Third, however, our approach identifies the worst-performing funds implied by the predicted fund flow distribution. For the bottom decile portfolio and the FF7 model all of the applied predictive models yield a significant negative alpha at the 5% level or higher. This shows that funds associated with high predicted inflows subsequently outperform funds associated with high predicted outflows.

5 Conclusion

In this paper, we apply machine learning techniques and a large set of predictors to forecast one-month ahead fund flows. Our results show that nonlinearities and interactions of complex machine learning models prove to be important in predicting future fund flows. Hence, these models significantly outperform standard linear frameworks in terms of out-of-sample R^2. We find that random forest is the best performing machine learning algorithm, followed by gradient

boosting. However, all other machine learning methods such as the decision tree or neural networks also possess a good forecasting ability and outperform the baseline OLS model.

Our results also show that past flows, the Morningstar rating and flow volatility are the most important fund characteristics for predicting future fund flows. In contrast, measures of past performance such as realized alpha from various factor models or value added are only of minor importance. These findings are consistent with the recent literature showing that investors react strongly to simplistic fund rankings rather than to more sophisticated performance measures based on commonly used asset pricing models. We also identify a convex relationship between past flows and predicted flows as well as a nonlinear impact of the Morninstar rating on predicted flows. In addition, the interaction effect of past flows with the Morningstar rating substantially impacts the forecasts. This effect reveals that high past flows and a high Morningstar rating are associated with higher predicted flows, while high past inflows and a low Morningstar rating are associated with a negative impact.

Finally, we show that machine learning-based fund flow predictions can be used for ex-ante separating high- from low-performing mutual funds. Based on long-short decile portfolios the best performing machine learning algorithm, random forest, delivers a significantly positive out-of-sample alpha.

Overall, our findings highlight the importance of the complex interplay between various variables in predicting flows. They show how asset managers could use state-of-the-art ML techniques to better predict flows and optimize their liquidity policy.

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 & Data Mining*, Association for Computing Machinery, KDD '19, 2623–2631.

Amihud, Y. and R. Goyenko (2013): "Mutual fund's R^2 as predictor of performance," *Review of Financial Studies*, 26, 667–694.

Bali, T. G., H. Beckmeyer, M. Moerke, and F. Weigert (2023): "Option return predictability with machine learning and big data," *Review of Financial Studies*, 36, 3548–3602.

Bali, T. G., A. Goyal, D. Huang, F. Jiang, and Q. Wen (2022): "Predicting corporate bond returns: Merton meets machine learning," *Unpublished Working Paper, Georgetown University, University of Lausanne, Singapore Management University, Central University of Finance and Economics*.

Bär, M., A. Kempf, and S. Ruenzi (2011): "Is a team different from the sum of its parts? Evidence from mutual fund managers," *Review of Finance*, 15, 359–396.

Barber, B. M., X. Huang, and T. Odean (2016): "Which factors matter to investors? Evidence from mutual fund flows," *Review of Financial Studies*, 29, 2600–2642.

Ben-David, I., J. Li, A. Rossi, and Y. Song (2022): "What do mutual fund investors really care about?" *Review of Financial Studies*, 35, 1723–1774.

Bergmeir, C., R. J. Hyndman, and B. Koo (2018): "A note on the validity of cross-validation for evaluating autoregressive time series prediction," *Computational Statistics & Data Analysis*, 120, 70–83.

Bergstra, J., R. Bardenet, Y. Bengio, and B. Kégl (2011): "Algorithms for hyper-

parameter optimization," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol. 24.

Berk, J. B. and J. H. van Binsbergen (2015): "Measuring skill in the mutual fund industry," *Journal of Financial Economics*, 118, 1–20.

——— (2016): "Assessing asset pricing models using revealed preference," *Journal of Financial Economics*, 119, 1–23.

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

Breiman, L. (2001): "Random forests," *Machine Learning*, 45, 5–32.

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen (1984): *Classification and regression trees*, Chapman and Hall/CRC.

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, 1509–1531.

Carhart, M. M. (1997): "On persistence in mutual fund performance," *Journal of Finance*, 52, 57–82.

Chen, L., M. Pelger, and J. Zhu (2024): "Deep learning in asset pricing," *Management Science*, 70, 714–750.

Chevalier, J. and G. Ellison (1997): "Risk taking by mutual funds as a response to incentives," *Journal of Political Economy*, 105, 1167–1200.

Clark, T. E. and K. D. West (2007): "Approximately normal tests for equal predictive accuracy in nested models," *Journal of Econometrics*, 138, 291–311.

Coval, J. and E. Stafford (2007): "Asset fire sales (and purchases) in equity markets," *Journal of Financial Economics*, 86, 479–512.

Del Guercio, D. and P. A. Tkac (2008): "Star power: The effect of morningstar ratings on mutual fund flow," *Journal of Financial and Quantitative Analysis*, 43, 907–936.

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

Diebold, F. X. and R. S. Mariano (1995): "Comparing predictive accuracy," *Journal of Business and Economic Statistics*, 13, 253–263.

Diebold, F. X. and M. Shin (2019): "Machine learning for regularized survey forecast combination: Partially-egalitarian lasso and its derivatives," *International Journal of Forecasting*, 35, 1679–1691.

Edelen, R. M. (1999): "Investor flows and the assessed performance of open-end mutual funds," *Journal of Financial Economics*, 53, 439–466.

Elton, E. J., M. J. Gruber, and C. R. Blake (2001): "A first look at the accuracy of the CRSP mutual fund database and a comparison of the CRSP and Morningstar mutual fund databases," *Journal of Finance*, 56, 2415–2430.

Evans, R. B. (2010): "Mutual fund incubation," *Journal of Finance*, 65, 1581–1611.

Evans, R. B. and Y. Sun (2021): "Models or stars: The role of asset pricing models and heuristics in investor risk adjustment," *Review of Financial Studies*, 34, 67–107.

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

——— (2015): "A five-factor asset pricing model," *Journal of Financial Economics*, 116, 1–22.

Frazzini, A. and O. A. Lamont (2008): "Dumb money: Mutual fund flows and the cross-section of stock returns," *Journal of Financial Economics*, 88, 299–322.

Gewers, F. L., G. R. Ferreira, H. F. D. Arruda, F. N. Silva, C. H. Comin, D. R. Amancio, and L. d. F. Costa (2021): "Principal component analysis: A natural approach to data exploration," *ACM Computing Surveys (CSUR)*, 54, 1–34.

Giglio, S., B. Kelly, and D. Xiu (2022): "Factor models, machine learning, and asset pricing," *Annual Review of Financial Economics*, 14, 337–368.

Gil-Bazo, J. and P. Ruiz-Verdú (2009): "The relation between price and performance in the mutual fund industry," *Journal of Finance*, 64, 2153–2183.

Goulet Coulombe, P., M. Leroux, D. Stevanovic, and S. Surprenant (2022): "How is machine learning useful for macroeconomic forecasting?" *Journal of Applied Econometrics*, 37, 920–964.

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

Gruber, M. J. (1996): "Another puzzle: The growth in actively managed mutual funds," *Journal of Finance*, 51, 783–810.

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

Harvey, D., S. Leybourne, and P. Newbold (1997): "Testing the equality of prediction mean squared errors," *International Journal of Forecasting*, 13, 281–291.

Hastie, T., R. Tibshirani, J. H. Friedman, and J. H. Friedman (2009): *The elements of statistical learning: Data mining, inference, and prediction*, vol. 2, Springer.

Hauzenberger, N., F. Huber, and K. Klieber (2023): "Real-time inflation forecasting using non-linear dimension reduction techniques," *International Journal of Forecasting*, 39, 901–921.

Hillert, A., A. Niessen-Ruenzi, and S. Ruenzi (2024): "Mutual fund shareholder letters: Flows, performance, and managerial behavior," *Management Science (forthcoming)*.

Huang, J., K. D. Wei, and H. Yan (2022): "Investor learning and mutual fund flows," *Financial Management*, 51, 739–765.

Jank, S. (2012): "Mutual fund flows, expected returns, and the real economy," *Journal of Banking & Finance*, 36, 3060–3070.

Jegadeesh, N. and C. S. Mangipudi (2021): "What do fund flows reveal about asset pricing models and investor sophistication?" *Review of Financial Studies*, 34, 108–148.

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

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu (2017): "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol. 30.

Kelly, B. T. and D. Xiu (2023): "Financial machine learning," *Unpublished Working Paper, Yale School of Management, University of Chicago*.

Keswani, A. and D. Stolin (2008): "Which money is smart? Mutual fund buys and sells of individual and institutional investors," *Journal of Finance*, 63, 85–118.

Kingma, D. P. and J. Ba (2014): "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*.

Leippold, M., Q. Wang, and W. Zhou (2022): "Machine learning in the Chinese stock market," *Journal of Financial Economics*, 145, 64–82.

Li, B. and A. G. Rossi (2021): "Selecting mutual funds from the stocks they hold: A machine learning approach," *Unpublished Working Paper, Wuhan University, Georgetown University*.

Li, L., K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar (2017): "Hyperband: A novel bandit-based approach to hyperparameter optimization," *Journal of Machine Learning Research*, 18, 6765–6816.

Lou, D. (2012): "A flow-based explanation for return predictability," *Review of Financial Studies*, 25, 3457–3489.

Lundberg, S. M., G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee (2020): "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, 2, 56–67.

Lundberg, S. M. and S.-I. Lee (2017): "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol. 30.

Massa, M., J. Reuter, and E. Zitzewitz (2010): "When should firms share credit with employees? Evidence from anonymously managed mutual funds," *Journal of Financial Economics*, 95, 400–424.

Masters, T. (1993): *Practical neural network recipes in C++*, Academic Press.

Medeiros, M. C., G. F. Vasconcelos, Á. Veiga, and E. Zilberman (2021): "Forecasting inflation in a data-rich environment: The benefits of machine learning methods," *Journal of Business & Economic Statistics*, 39, 98–119.

Molnar, C. (2019): *Interpretable machine learning*, https://christophm.github.io/interpretable-ml-book/, https://christophm.github.io/interpretable-ml-book/.

Murray, S., Y. Xia, and H. Xiao (2024): "Charting by machines," *Journal of Financial Economics*, 153, 103791.

Neuhierl, A., X. Tang, R. T. Varneskov, and G. Zhou (2022): "Option characteristics as cross-sectional predictors," *Unpublished Working Paper, Washington University, University of Texas, Copenhagen Business School, Washington University*.

Newey, W. and K. West (1987): "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix," *Econometrica*, 55, 703–708.

Newey, W. K. and K. D. West (1994): "Automatic lag selection in covariance matrix estimation," *Review of Economic Studies*, 61, 631–653.

Pástor, L. and R. F. Stambaugh (2003): "Liquidity risk and expected stock returns," *Journal of Political Economy*, 111, 642–685.

Pástor, L. and M. B. Vorsatz (2020): "Mutual fund performance and flows during the COVID-19 crisis," *Review of Asset Pricing Studies*, 791–833.

Rakowski, D. (2010): "Fund flow volatility and performance," *Journal of Financial and Quantitative Analysis*, 45, 223–237.

Reuter, J. and E. Zitzewitz (2021): "How much does size erode mutual fund performance? A regression discontinuity approach," *Review of Finance*, 25, 1395–1432.

Sirri, E. R. and P. Tufano (1998): "Costly search and mutual fund flows," *Journal of Finance*, 53, 1589–1622.

Svozil, D., V. Kvasnicka, and J. Pospichal (1997): "Introduction to multi-layer feed-forward neural networks," *Chemometrics and Intelligent Laboratory Systems*, 39, 43–62.

Welch, B. L. (1947): "The generalization of 'STUDENT'S'problem when several different population variances are involved," *Biometrika*, 34, 28–35.

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

West, K. D. (1996): "Asymptotic inference about predictive ability," *Econometrica*, 64, 1067–1084.

Yang, J. (2022): "Fast TreeSHAP: Accelerating SHAP value computation for trees," *Unpublished Working Paper, LinkedIn Corporation*.

Zheng, L. (1999): "Is money smart? A study of mutual fund investors' fund selection ability," *Journal of Finance*, 54, 901–933.

Zou, H. and T. Hastie (2005): "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 301–320.

Figure 1: Correlation matrix between fund characteristics

This figure shows the correlation between the 30 fund characteristics used as predictors. The sample period spans from January 1980 to December 2020.

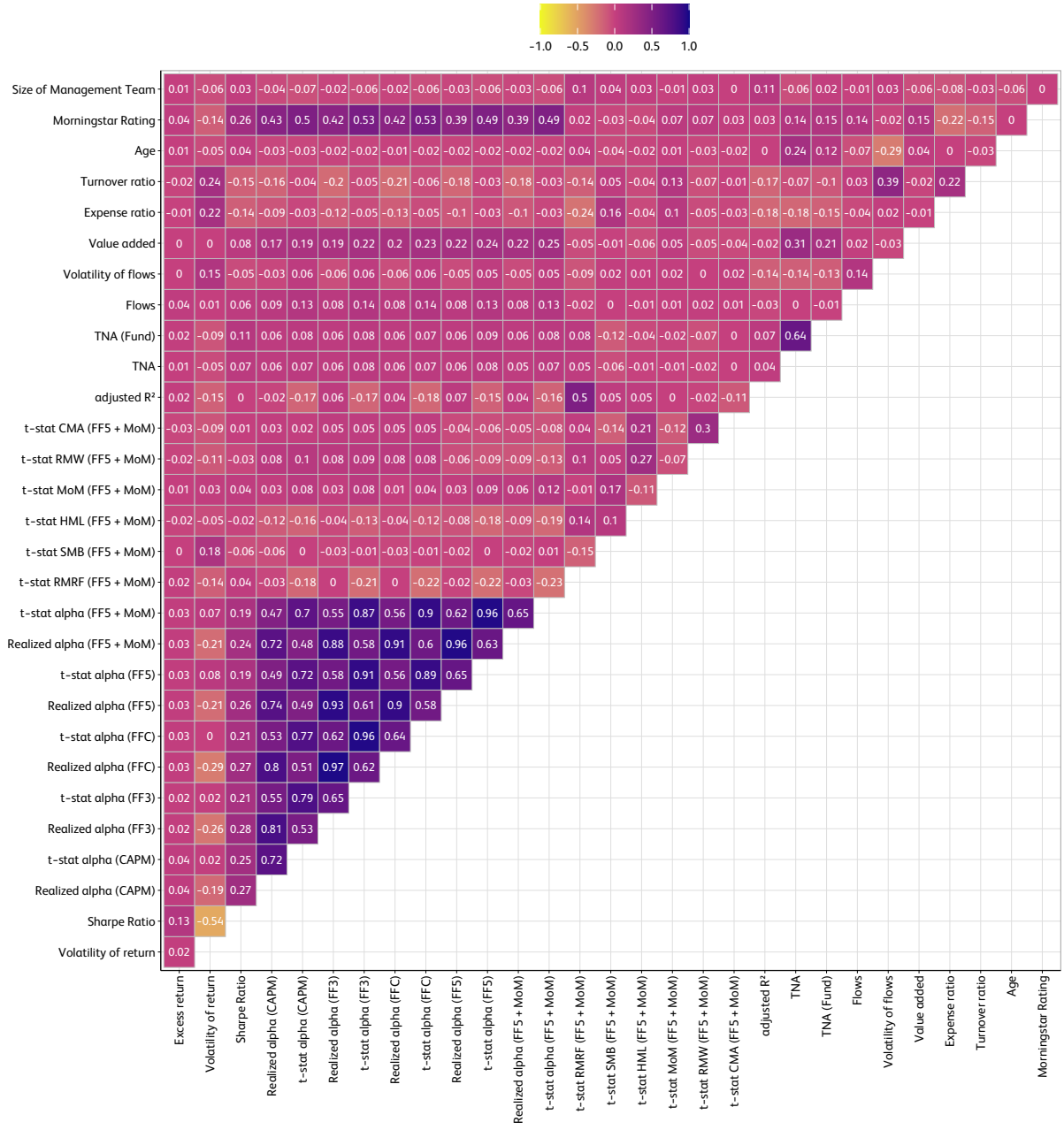


Figure 2: Evolution of characteristic importance for random forest over time

This figure plots the time-variation of characteristic importance of the 15 most important predictors identified by their mean absolute SHAP value over the whole out-of-sample period (January 2000 to December 2020) and a one-month forecasting horizon for random forest. Characteristics are ranked according to their importance and range from 1 for the most important predictor and 30 for the least important predictor.

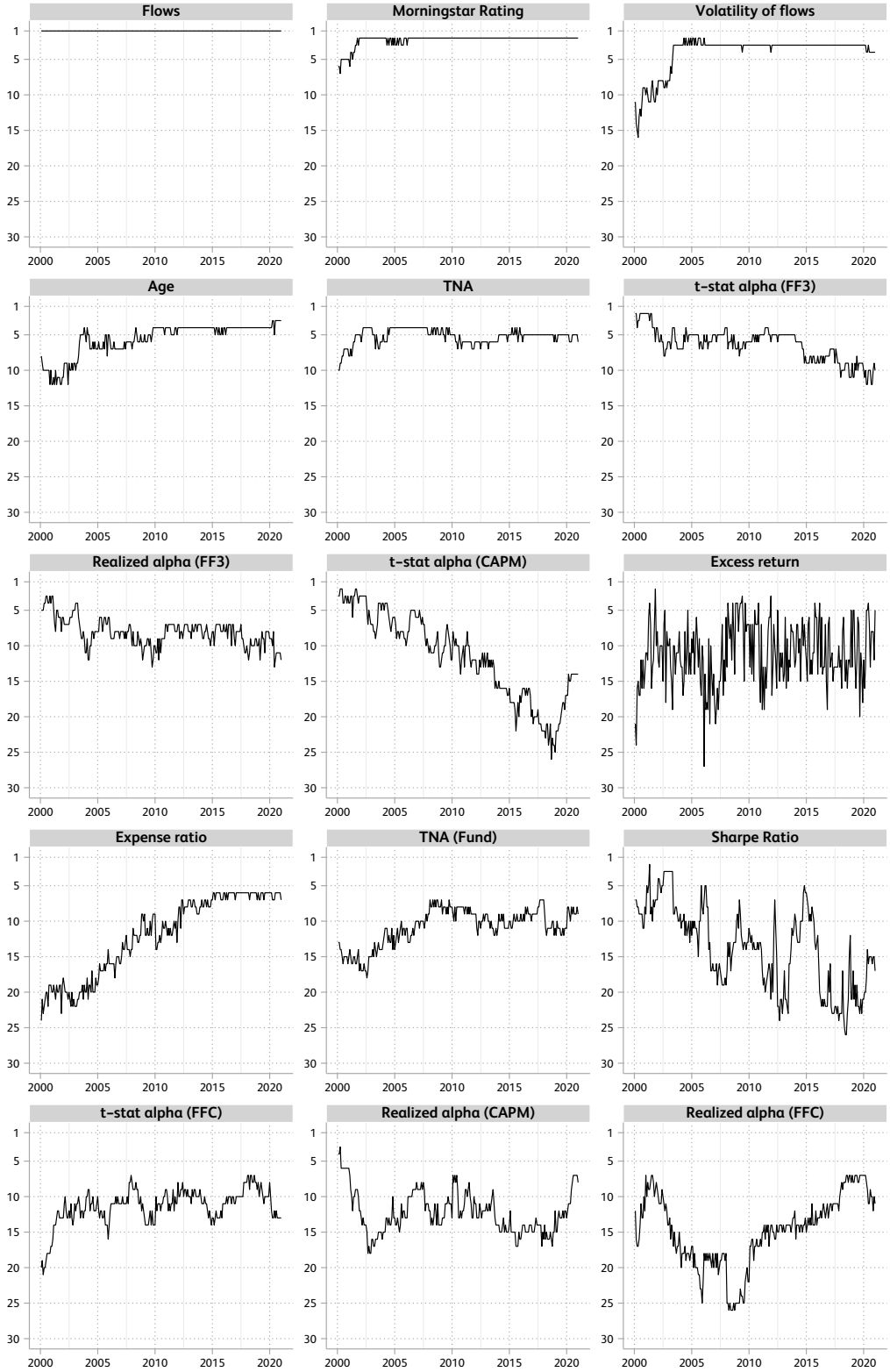


Figure 3: Directional importance of the characteristics for random forest

This figure shows a beeswarm SHAP plot which indicates the directional impact of the 15 most important characteristics identified by their mean absolute SHAP value over the whole out-of-sample period (2000 to 2020) and a one-month forecasting horizon for random forest. Predictors are sorted in descending order with the most important variables at the top. Each dot on the chart corresponds to one SHAP value for a prediction and feature. The magnitude of feature values is color-coded from yellow (low) to purple (high). The points are distributed horizontally according to their SHAP values which measure each feature's directional impact on the predictive model outcome (flow prediction).

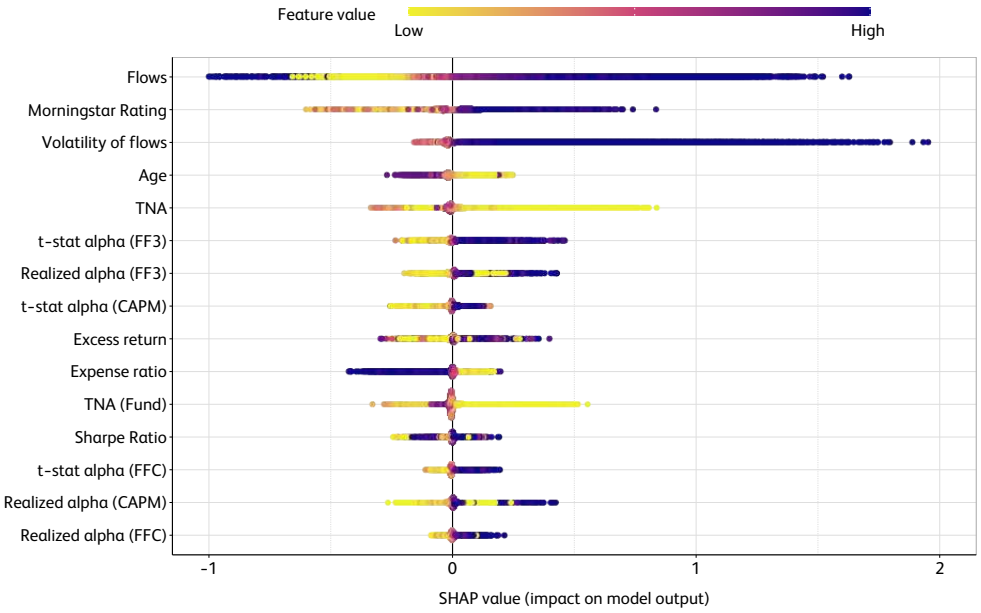


Figure 4: Directional importance of the characteristics for random forest over time

This figure shows a beeswarm SHAP plot which indicates the directional impact of the 15 most important characteristics identified by their mean absolute SHAP value for consecutive two-year periods and a one-month forecasting horizon for random forest. Predictors are sorted in descending order with the most important variables at the top, Each dot on the chart corresponds to one SHAP value for a prediction and feature. The magnitude of feature values is color-coded from yellow (low) to purple (high). The points are distributed horizontally according to their SHAP values which measure each feature's directional impact on the predictive model outcome (flow prediction).

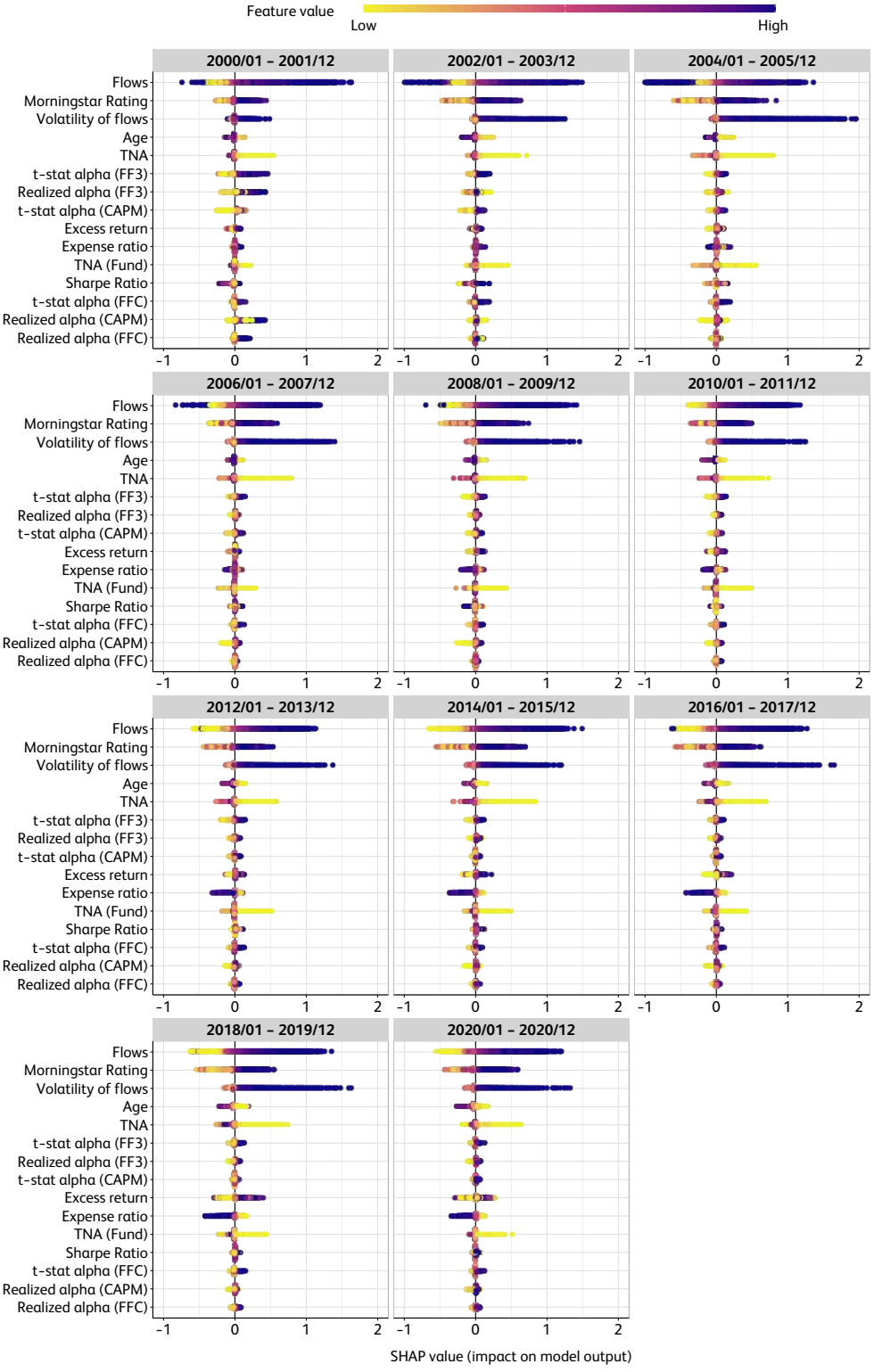


Figure 5: Functional form of past flows and interaction with Morningstar rating

This figure plots the nonlinear (convex) relationship for each observation between past fund flows (predictor) and the associated SHAP value as well as the interaction with the Morningstar rating for the whole out-of-sample period (January 2000 to December 2020) and a one-month forecasting horizon for random forest.

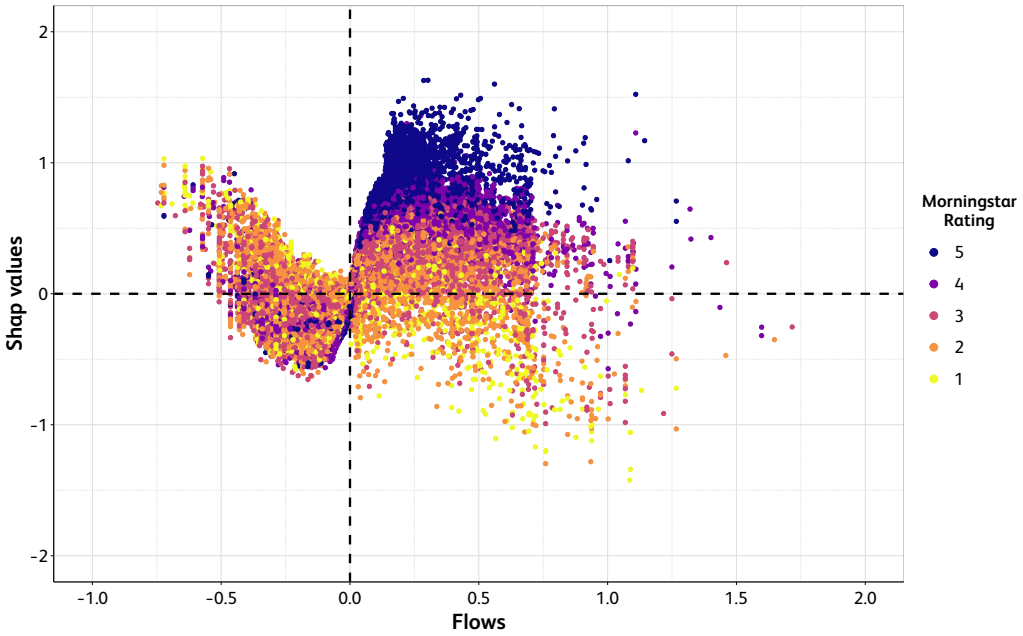


Figure 6: Functional form of past flows and interaction with Morningstar rating over time

This figure plots the nonlinear (convex) relationship for each observation between past fund flows (predictor) and the associated SHAP value as well as the interaction with the Morningstar rating for consecutive two-year periods and a and a one-month forecasting horizon for random forest.

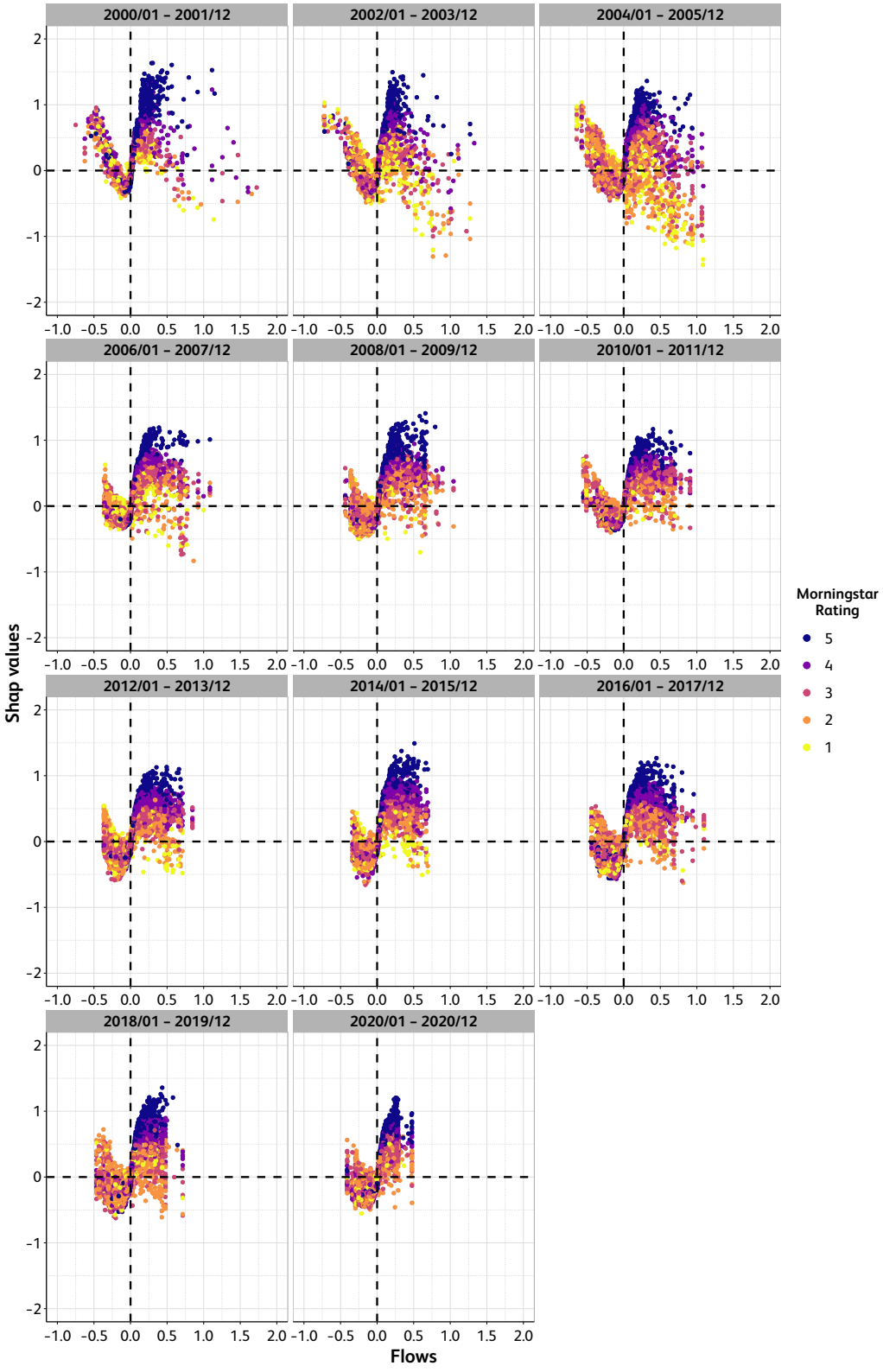


Figure 7: Impact of nonlinearity of the Morningstar rating

This figure plots the five rating categories (stars) of the Morningstar rating and its associated SHAP value over the whole out-of-sample period (January 2000 to December 2020) and a one-month forecasting horizon for random forest. The rating is based on a scale of one to five stars, with one being the worst and five being the best.

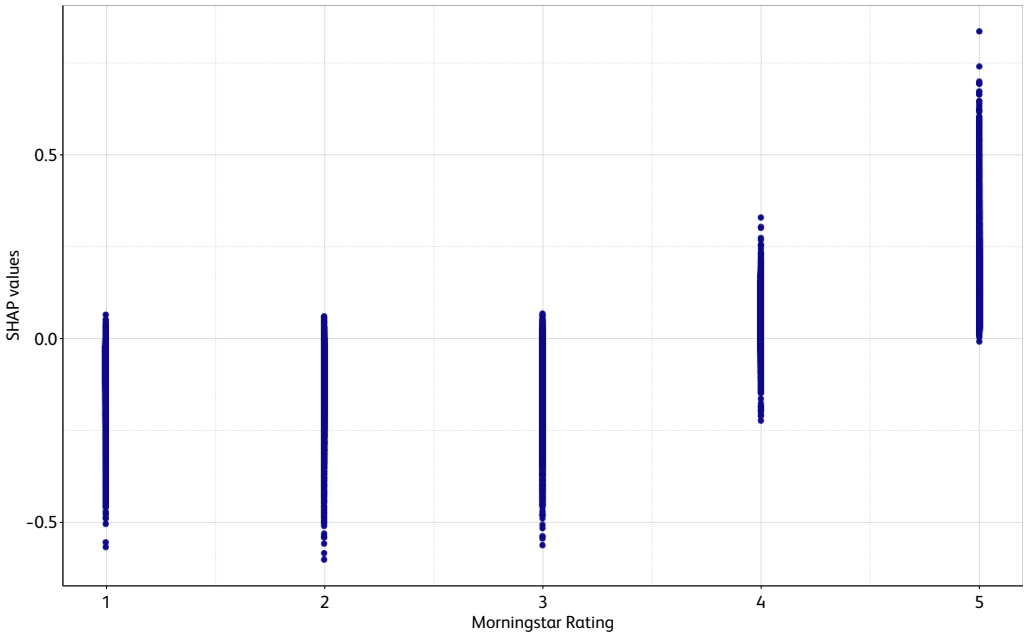


Table 1: Number of share classes per year with monthly average, median and quantile returns

This table reports the number of share classes for each year in our sample and the corresponding return distribution (in percent) described by its mean, lower (Q1) and upper quantile (Q3). The sample period spans from January 1980 to December 2020.

		Return distribution				
Year	Number of share classes	Mean	Std	Q1	Median	Q3
1980	37	2.55	7.33	1.28	3.98	6.47
1981	48	-0.16	5.28	-4.00	-0.25	3.63
1982	42	1.68	5.90	-2.77	0.12	5.77
1983	54	1.66	4.17	-1.23	2.60	4.61
1984	64	-0.34	4.80	-3.46	-0.64	1.82
1985	77	2.04	4.12	-0.84	1.88	4.73
1986	82	1.11	5.38	-1.91	1.92	4.85
1987	82	0.22	10.29	-1.71	1.84	5.29
1988	82	1.00	3.90	-1.65	0.85	3.64
1989	89	2.01	3.60	-0.36	2.01	4.41
1990	96	-0.68	6.17	-4.91	-0.38	3.56
1991	153	2.58	5.00	-0.52	2.82	5.89
1992	172	0.44	3.35	-1.80	0.78	2.70
1993	300	1.09	3.07	-0.74	1.27	2.91
1994	404	-0.06	3.24	-2.49	0.32	2.25
1995	562	2.15	2.67	0.79	2.40	3.70
1996	704	1.50	3.87	-0.26	1.87	3.74
1997	759	1.77	4.85	-1.83	2.34	5.20
1998	1,075	1.05	7.91	-1.64	2.90	6.08
1999	1,068	1.84	5.60	-2.21	1.32	5.15
2000	1,079	-0.04	8.22	-4.68	-0.68	4.76
2001	1,026	-0.92	7.33	-5.08	-0.15	3.90
2002	1,082	-2.01	6.25	-6.51	-1.89	2.25
2003	1,911	2.67	4.92	0.29	2.26	5.37
2004	2,055	1.03	3.65	-0.67	1.44	3.10
2005	1,989	0.58	4.33	-1.80	0.45	3.08
2006	2,256	1.03	2.98	-0.31	1.35	2.69
2007	2,593	0.46	3.36	-1.49	0.88	2.85
2008	4,367	-4.78	8.52	-9.49	-2.81	1.68
2009	4,745	2.36	6.90	-1.11	3.77	6.72
2010	4,881	1.42	5.83	-4.05	2.95	5.96
2011	4,859	-0.23	5.37	-2.50	-0.34	2.49
2012	4,783	1.14	3.55	-0.40	1.84	3.30
2013	4,983	2.37	2.77	0.86	2.74	4.22
2014	5,198	0.65	3.21	-1.70	1.11	2.94
2015	5,338	-0.15	3.92	-2.61	-0.36	1.82
2016	5,439	0.85	3.99	-0.88	0.76	2.82
2017	5,854	1.49	2.49	0.45	1.48	2.55
2018	5,840	-0.70	4.82	-2.83	0.44	2.32
2019	5,825	2.05	4.21	0.35	2.38	4.27
2020	5,698	1.33	8.65	-2.80	2.89	6.40
Total	**9,721**	**0.77**	**5.36**	**-1.58**	**1.36**	**3.70**

Table 2: Summary statistics of share-class characteristics

This table reports monthly descriptive statistics of all fund characteristics considered in the empirical analysis except the expense and turnover ratio which are reported based on annual data. The size of management team is represented as a dummy variable taking a value of zero if the fund is single-managed and one otherwise. All variables are measured at the share-class level for actively managed US domestic equity funds spanning the period from January 1980 to December 2020.

Characteristic	**Mean**	**Median**	**Std**	**# obs**	**NaN**
Excess return	0.66%	1.20%	5.37%	954,729	0.03%
Volatility of return	4.65%	4.23%	2.01%	836,864	12.37%
Sharpe Ratio	0.17	0.19	0.17	836,864	12.37%
Realized α (CAPM)	-0.13%	-0.12%	0.51%	836,864	12.37%
t-stat α (CAPM)	-0.56	-0.56	1.39	836,864	12.37%
Realized α (FF3)	-0.12%	-0.10%	0.41%	836,864	12.37%
t-stat α (FF3)	-0.67	-0.64	1.42	836,864	12.37%
Realized α (FFC)	-0.12%	-0.11%	0.40%	836,864	12.37%
t-stat α (FFC)	-0.69	-0.67	1.41	836,864	12.37%
Realized α (FF5)	-0.10%	-0.10%	0.41%	836,864	12.37%
t-stat α (FF5)	-0.65	-0.63	1.50	836,864	12.37%
Realized α (FF5 + MoM)	-0.10%	-0.10%	0.39%	836,864	12.37%
t-stat α (FF5 + MoM)	-0.67	-0.65	1.51	836,864	12.37%
t-stat RMRF (FF5 + MoM)	22.62	19.84	14.13	836,864	12.37%
t-stat SMB (FF5 + MoM)	1.83	0.66	4.91	836,864	12.37%
t-stat HML (FF5 + MoM)	-0.00	-0.11	2.99	836,864	12.37%
t-stat MoM (FF5 + MoM)	-0.05	0.00	2.48	836,864	12.37%
t-stat RMW (FF5 + MoM)	-0.25	-0.23	1.94	836,864	12.37%
t-stat CMA (FF5 + MoM)	-0.62	-0.66	2.11	836,864	12.37%
adjusted R^2	0.90	0.94	0.13	836,864	12.37%
TNA	642.66	93.80	2,615.07	933,165	2.29%
TNA (Fund)	1,785.97	365.40	5,682.74	933,165	2.29%
Flows	0.30%	-0.37%	8.02%	926,553	2.98%
Volatility of flows	5.95%	4.01%	6.01%	802,074	16.02%
Value added	0.39	-0.00	7.22	647,843	32.17%
Expense ratio	1.04%	0.99%	0.47%	730,001	23.56%
Turnover ratio	74.63%	52.00%	101.24%	725,922	23.99%
Age	132.25	110.00	104.64	955,048	0.00%
Morningstar rating	3.16	3.00	1.01	840,599	11.98%
Size of Management Team	0.73	1.00	0.45	955,048	0.00%

Table 3: Out-of-sample model comparison based on R^2_{OOS}

This table reports the out-of-sample performances measured by the out-of-sample R^2 statistic, as defined in Equation 3.15, for the applied machine learning methods and the baseline OLS model for a one-month forecasting horizon. The statistical significance of the R^2 measure is evaluated by the Clark and West (2007) statistic (one-sided test). The level of significance is displayed as *, ** and ***, indicating statistical significance at the 10%, 5%, and 1% level, respectively. The out-of-sample period spans from January 2000 to December 2020.

	Out-of-sample R^2
OLS	0.1486***
Elastic net	0.1431***
Decision tree	0.1643***
Random forest	0.1969***
Gradient boosting	0.1911***
Neural network (I)	0.1705***
Neural network (II)	0.1743***
Observations	554,635

Table 4: Out-of-sample model comparison based on R^2_{OOS} in expansions and recessions

This table reports the out-of-sample performances measured by the out-of-sample R^2 statistic, separately for expansions (Exp.) and recessions (Rec.) as defined by the NBER for the applied-machine learning methods and the baseline OLS model for a one-month forecasting horizon. To test whether the average forecasting accuracy differs in expansions and recessions we apply a Welch (1947) test and report the corresponding $t-$statistic in the column Exp.-Rec. The statistical significance of the R^2 measure is evaluated by the Clark and West (2007) statistic (one-sided test). The level of significance is displayed as *, ** and ***, indicating statistical significance at the 10%, 5%, and 1% level, respectively. The out-of-sample period spans from January 2000 to December 2020.

	Out-of-sample R^2		
	Expansion	Recession	Exp. − Rec.
OLS	0.1489***	0.1462***	-1.0534
Elastic Net	0.1434***	0.1410***	-1.1139
Decision Tree	0.1654***	0.1556***	-0.7390
Random Forest	0.1984***	0.1841***	-0.3867
Gradient Boosting	0.1921***	0.1828***	-0.8258
Neural Network (I)	0.1711***	0.1663***	-0.9185
Neural Network (II)	0.1748***	0.1705***	-0.9954
Observations	499,938	54,697	E=221;R=31

Table 5: Forecast comparison based on the modified DMW statistic

This table shows a comparison of the predictive performance of the applied machine learning methods and the baseline OLS model using the modified Diebold and Mariano (1995), West (1996) test statistic defined in Equation 3.18. A positive number indicates that the model in the column outperforms the model in the row. The forecasting horizon corresponds to one month. The level of significance of this outperformance is displayed as *, ** and ***, indicating statistical significance at the 10%, 5%, and 1% level, respectively. The out-of-sample period spans from January 2000 to December 2020.

	Elastic Net	Decision Tree	Random Forest	Gradient Boosting	Neural Network (I)	Neural Network (II)
OLS	-13.87***	10.49***	32.84***	26.02***	14.81***	20.45***
Elastic Net		14.59***	36.91***	29.57***	20.54***	24.27***
Decision Tree			21.99***	16.14***	4.31***	6.51***
Random Forest				-7.04***	-23.52***	-19.21***
Gradient Boosting					-15.69***	-12.19***
Neural Network (I)						3.80***

Table 6: Mean of absolute SHAP-values for random forest

This table reports the mean absolute SHAP values over all forecasting periods (equally weighted) for all fund characteristics for a one-month forecasting horizon for random forest. SHAP values are calculated over the out-of-sample period from January 2000 to December 2020.

Rank	Characteristic	Mean
1	Flows	0.1534
2	Morningstar rating	0.0528
3	Volatility of flows	0.0297
4	Age	0.0204
5	TNA	0.0198
6	t-stat α (FF3)	0.0175
7	Realized α (FF3)	0.0131
8	t-stat α (CAPM)	0.0113
9	Excess return	0.0106
10	Expense ratio	0.0090
11	TNA (Fund)	0.0089
12	Sharpe Ratio	0.0088
13	t-stat α (FFC)	0.0085
14	Realized α (CAPM)	0.0084
15	Realized α (FFC)	0.0069
16	Turnover ratio	0.0063
17	Volatility of return	0.0044
18	Value added	0.0043
19	t-stat α (FF5)	0.0042
20	adjusted R^2	0.0041
21	t-stat SMB (FF5 + MoM)	0.0041
22	t-stat HML (FF5 + MoM)	0.0033
23	Realized α (FF5)	0.0031
24	t-stat CMA (FF5 + MoM)	0.0030
25	t-stat RMW (FF5 + MoM)	0.0028
26	t-stat MoM (FF5 + MoM)	0.0026
27	t-stat alpha (FF5 + MoM)	0.0024
28	t-stat RMRF (FF5 + MoM)	0.0023
29	Realized alpha (FF5 + MoM)	0.0021
30	Size of Management Team	0.0006

Table 7: Mean of absolute SHAP-values (interactions) for random forest

This table reports the mean absolute SHAP interaction values for all monthly forecasting periods (equally weighted) for the ten most important interactions for the random forest. SHAP interaction values are calculated over the out-of-sample period from January 2000 to December 2020.

Rank	Characteristic I	Characteristic II	Mean
1	Flows	Morningstar rating	0.0194
2	Flows	Volatility of flows	0.0095
3	Flows	t-stat α (FF3)	0.0092
4	Flows	Realized α (FF3)	0.0075
5	Flows	t-stat α (CAPM)	0.0071
6	Flows	Age	0.0064
7	Flows	Sharpe Ratio	0.0062
8	Flows	TNA	0.0057
9	Flows	Realized α (CAPM)	0.0045
10	Flows	t-stat α (FFC)	0.0043

Table 8: Monthly out-of-sample alpha (in %) of the top and bottom 10th percentile portfolios

This table reports the monthly out-of-sample alpha of the equally weighted top and bottom 10th percentile portfolios and the corresponding long-short portfolio of the predicted flow distribution based on a one-month ahead forecasts. Additionally, a comparison with a naïve strategy (equally weighted portfolio of all share classes) is provided. The out-of-sample alphas are obtained by regressing the log excess return (long-short log return) of the top and bottom (long-short) portfolios on four different risk-factor models (in log): the Carhart (1997) four-factor model (FF4), the Fama and French (2015) five-factor model (FF5), and the FF5 model augmented with the momentum factor of Carhart (1997) (FF6) and additionally by the liquidity factor of Pástor and Stambaugh (2003) (FF7).The out-of-sample period spans from January 2000 to December 2020 (252 observations). Newey and West (1987) robust standard errors with 12 lags (DeMiguel et al., 2023; Kaniel et al., 2023) are reported in parentheses. The level of significance is displayed as *, ** and ***, indicating statistical significance at the 10%, 5%, and 1% level, respectively.

	Top 10%:				**Bottom 10%:**				**Long-Short:**			
	FF4	FF5	FF6	FF7	FF4	FF5	FF6	FF7	FF4	FF5	FF6	FF7
Equally weighted	-0.05	-0.04	-0.04	-0.05	-0.05	-0.04	-0.04	-0.05				
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)				
OLS	-0.09	0.02	0.02	0.02	-0.12	-0.17**	-0.17**	-0.18**	0.03	0.19*	0.19**	0.19**
	(0.07)	(0.07)	(0.06)	(0.06)	(0.10)	(0.08)	(0.07)	(0.07)	(0.15)	(0.11)	(0.09)	(0.09)
Elastic Net	-0.09	0.03	0.03	0.02	-0.14	-0.18**	-0.18**	-0.19***	0.05	0.20*	0.21**	0.21**
	(0.07)	(0.07)	(0.06)	(0.06)	(0.09)	(0.07)	(0.07)	(0.07)	(0.14)	(0.11)	(0.09)	(0.09)
Decision Tree	-0.05	0.04	0.04	0.04	-0.10	-0.11*	-0.11*	-0.12**	0.05	0.15**	0.16***	0.15***
	(0.05)	(0.07)	(0.06)	(0.06)	(0.07)	(0.06)	(0.06)	(0.06)	(0.08)	(0.06)	(0.05)	(0.05)
Random Forest	-0.04	0.04	0.04	0.05	-0.10	-0.18***	-0.18***	-0.17***	0.07	0.22**	0.22***	0.23***
	(0.05)	(0.07)	(0.06)	(0.06)	(0.10)	(0.07)	(0.07)	(0.07)	(0.11)	(0.09)	(0.08)	(0.08)
Gradient Boosting	-0.04	0.04	0.05	0.04	-0.10	-0.17**	-0.17**	-0.17***	0.06	0.21**	0.21***	0.21***
	(0.05)	(0.07)	(0.07)	(0.06)	(0.10)	(0.07)	(0.07)	(0.07)	(0.11)	(0.09)	(0.08)	(0.08)
Neural Network (I)	-0.10	0.00	0.00	0.00	-0.08	-0.12*	-0.13*	-0.14**	-0.01	0.12	0.13	0.14
	(0.06)	(0.07)	(0.06)	(0.05)	(0.10)	(0.07)	(0.07)	(0.07)	(0.14)	(0.10)	(0.09)	(0.08)
Neural Network (II)	-0.07	0.02	0.02	0.02	-0.11	-0.17**	-0.18**	-0.18***	0.04	0.19*	0.20**	0.20**
	(0.05)	(0.07)	(0.06)	(0.06)	(0.10)	(0.07)	(0.07)	(0.06)	(0.12)	(0.10)	(0.09)	(0.08)