Machine Learning Mutual Fund Flows

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

This draft: January 13, 2025
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. Using interpretable ML methods, we identify past flows and the Morningstar rating as the most important predictors for net-flows, while other past performance variables are of minor importance. We find that the importance of Morningstar ratings and expenses has increased over time. In addition, the interaction effect of past flows with the Morningstar rating has a substantial impact on future flows. Furthermore, 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 flow prediction, big data, interpretable machine learning

*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

We thank Peter Fiechter, Michael Hasler, Timm Pfeil and Carolina Salva for helpful comments. We benefited from the comments received at presentations at the 30th Annual Meeting of the German Finance Association (DGF), the University of Mannheim and the University of Neuchâtel. The authors acknowledge support by the state of Baden-Württemberg through bwHPC.

1 Introduction

Open-end mutual funds offer liquidity to investors by allowing them to redeem fund shares at their daily net asset value. This feature 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 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 based on mutual fund characteristics and other variables. According to the literature, numerous variables impact flows, and these relationships are often nonlinear (Sirri and Tufano, 1998; Chevalier and Ellison, 1997). This setting creates an ideal scenario for applying 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) in the context of return predictions, we apply a linear OLS regression (baseline model) and ML methods to obtain improved fund flow predictions. 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 predicted flows, thus also contributing to a better understanding of mutual fund investors' decision making. Finally, 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) based on machine learning implied monthly fund flow predictions.

We start our empirical analyses by comparing the predictive power of simple OLS regressions and different ML models for future equity fund flows in the US over the period from January 2000 to December 2023. To this end, we use the out-of-sample R^2 statistic to benchmark our predictions against a fund's simple historical mean flows. 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 use a semi-structured systematic approach to first identify 55 variables that might have an influence on flows based on the prior literature. These variables are mainly linked to past performance and other fund characteristics. In addition, we use nine variables that proxy for aggregate financial market and macroeconomic conditions that might have an influence on aggregate flows into equity funds. To provide a fair comparison between OLS and ML methods the specification of the OLS model is guided by the literature on mutual fund flows and accounts for the well-known nonlinear impact of past performance on fund flows (Sirri and Tufano, 1998; Chevalier and Ellison, 1997). We find that linear models like OLS and elastic net achieve a positive out-of-sample R^2 of about 17% to 18% percent for the full sample using a historical mean flow benchmark. Nonlinear ML-models like random forest and gradient boosting improve this measure by about 4 percentage points, i.e., by more than a fifth, while neural networks also lead to an improvement which is however smaller. These results remain robust across different economic conditions and market environments, indicating that the predictive accuracy of the forecasting models holds steady even during economic downturns, which are often marked by increased outflows from risky asset classes (e.g., Jank, 2012; Pástor and Vorsatz, 2020). Among all statistical models, the forecasts obtained by random forest are the most precise and significantly better (at the 1% level) than the predictions obtained from any other forecasting method based on pairwise forecast accuracy tests using modified Diebold and Mariano (1995) and West (1996) tests.

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 fund 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.[1] Based on the best performing model, random forest, our results

[1] SHAP values break down a model's prediction for a single data point, showing how much each variable

indicate that past flows over the last month are the most influential variable predicting flows, followed by the average flows over the past six and 12 months, the Morningstar rating, and total net assets (TNA). The importance of past flows is consistent with stable fund characteristics being important flow drivers.[2] We also document that various measures of past performance (e.g., alpha, value added, market-adjusted returns) 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 OLS regressions. Interestingly, this is the case for both retail and institutional investor fund share classes. While most other variables are of similar importance for both types of share classes, expense ratios are more important for institutional share class flow predictions.

Looking at the strength of the impact of these predictors over time, we find that flow-related predictors are consistently ranked as the most important predictors over the whole sample period, while the Morningstar rating is not as important in the very early sample years as it later becomes. This pattern is consistent with the increased popularity of these ratings over time. While most other variables do not show a clear time trend, we do find that expense ratios have become more influential in predicting fund flows. Our results also indicate that in the bottom decile of the predicted flow distribution costs seem to play a more important role, i.e., outflows are more cost-sensitive than inflows.

SHAP values also allow us to determine the directional impact of each variable. We find that our most important predictor variable, past monthly flows, usually affects future flow predictions positively, but in some cases also negatively. This finding suggests a nonlinear relationship on fund flow forecasts. In contrast, the directional impact of average flows over the past six and 12 months is typically positively related to predicted flows. The Morningstar rating usually has a positive impact on flow predictions. All of these patterns are quite stable over time. Furthermore, while expense ratios were of marginal relevance at the beginning of the

contributes to the final outcome by comparing it to an average baseline prediction. Essentially, they explain which variables affect the prediction positively or negatively and by how much relative to the baseline forecast, hence providing an interpretable "fair share" of each variable's impact.

[2] In traditional OLS-based mutual fund flow studies stable fund characteristics are often captured by including fund fixed effects in flow regressions. However, in our predictive machine learning framework, it is unclear how to account for such fixed-effects equivalently.

sample period with no clear directional impact, they became clearly negatively related to future flow predictions from 2006 onward. This finding suggests an increased cost-awareness of fund investors, potentially driven by the competitive pressure from low-cost index funds (Cremers et al., 2016).

Motivated by the nonlinear impact of past performance on flows documented in earlier studies (see, e.g., Sirri and Tufano, 1998), we also assess the functional form of the relationship between the Morningstar rating and predicted flows. We document that both, past flows and the Morningstar rating have a convex impact on predicted flows.

SHAP values can also be used to analyze the importance of the impact of interactions between variables.[3] We analyze the importance of all possible interactions between our predictor variables and find the interaction of past flows with the Morningstar rating to have the strongest impact on flow forecasts among all non-entirely flow-related interactions.[4] Analyzing the functional form for the impact of this relationship, we document 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, suggesting that new money (proxied for by recent past flows) is more attentive to performance information contained in ratings.

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: Long-short decile portfolios based on predicted flows from the ML models with the highest forecasting accuracy, random forest and gradient boosting, generate an alpha of 1.92% and 2.52% per year, respectively, which is statistically significant at the 5% and 1% levels. This outperformance is very similar to the results obtained by Kaniel et al. (2023), that are based on computationally more intense feedforward neural networks and a richer information set to directly predict the fund alpha distribution rather than flows. Consistent with our findings, they identify fund flows as one of the most important predictors

[3] SHAP interaction values show how pairs of variables jointly impact a model's prediction. This approach not only reveals how each variable influences the prediction on its own but also how the combination of two variables shifts the outcome differently than if they acted independently.

[4] The strongest interactions are found between lagged monthly flows and lagged 6-month and lagged 12-month flows, respectively.

of future risk-adjusted fund performance. The predictive power of flow predictions for future performance is consistent with a smart money effect (see, e.g., Zheng, 1999).

Overall, our paper contributes to two 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 significantly improve mutual fund flow predictions. Using interpretable ML methods, we are also able to contribute to a better understanding of the intricate impact of the interplay between past ratings and past flows on flow predictions. Second, we also contribute to a growing literature of applying ML techniques in finance.[5] 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) 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 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.

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

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

Many funds have different share classes with different management fees, expense ratios or loads (Khorana et al., 2008). To account for this heterogeneity, our analyses are performed at the share class level. Our sample includes all institutional and retail share classes that focus on the US equity market. Because data coverage on monthly fund TNAs prior to 1991 is scarce and incomplete (see, e.g., Dou et al., 2024) our sample spans from January 1991 to December 2023. We restrict our analyses to share classes of actively managed funds, 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 consider share classes reaching at least 36 monthly observations since inception and exceed a threshold of 15 million in TNA after this incubation period, which is a commonly used cutoff in the literature (see, among others, Elton et al., 2001; Evans, 2010; Doshi et al., 2015; Kaniel et al., 2023).

To extend the set of characteristics we rely on fund and fund manager data provided by Morningstar. According to Massa et al. (2010), Morningstar is a more important source of 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; Ben-David et al., 2022). 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 13,376 unique share classes. Based on this sample Table 1 reports the number of share classes for each year and the corresponding flow distribution described by its mean, standard deviation, lower (Q1) and upper quantile (Q3). It is notable, that the average (median) fund recorded outflows of -0.47% (-0.65%)[6] throughout the sample period. More specifically, every year since 2004 has seen negative fund flows indicating that the market for actively managed equity mutual funds in the U.S. has been shrinking over the last two decades, while it was characterized by strong external growth during the previous years (Sirri and Tufano, 1998).

To be comprehensive in objectively selecting the most important determinants of fund flows, we follow a systematic semi-structured approach. We search for studies on mutual fund flows in the top finance and related journals. We identify potential predictor variables from these studies and focus on variables that can be computed based on readily available information on fund characteristics and historical returns and are thus relatively easy to calculate using data provided by CRSP and Morningstar. We thus exclude holding-based characteristics. Based on this approach, we identify 55 mutual fund characteristics. In addition, we use nine variables that proxy for macroeconomic conditions. In total, we construct a dataset of 64 predictors, shown in Table 2, to be used in forecasting fund flows using machine learning techniques.

2.2 Mutual fund characteristics

Our main variable of interest is monthly fund flows determined as

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

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). As in Cen et al. (2024), we use the natural logarithm of one plus the monthly flows in our computations since flows are the relative growth in the share class TNA

[6]Note that these numbers differ slightly from the descriptive statistics reported in Table 3, which are based on the training sample.

adjusted for net returns.

Using data provided by CRSP, for each share class i in month t we obtain the total net assets ($TNA_{i,t}$), expense ratio ($ER_{i,t}$, Barber et al. (2016)), the turnover ratio (Ivković and Weisbenner, 2009) as well as front-end and back-end loads (Ivković and Weisbenner, 2009; Chen et al., 2010).[7] The return volatility (Franzoni and Schmalz, 2017) for each share class, using returns net of fees, 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 the $Age_{i,t}$ of each share class (Chevalier and Ellison, 1997), which is the rounded number of months between its inception date and the last calendar date of month t. Following Barber et al. (2016), we estimate the market-adjusted return ($MAR_{i,t}$) by subtracting the market return from the fund return.

Several characteristics used in our analyses 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).[8] While there is still a debate about which asset pricing model is used for risk adjustment by mutual fund investors (Berk and van Binsbergen, 2016; Barber et al., 2016; Jegadeesh and Mangipudi, 2021), as Barber et al. (2016), 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, estimated over a 36-month rolling estimation window, for the FF6 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

[7] If no information about front-end or back-end loads is provided by CRSP we assume that no such loads are charged and set the corresponding variable to 0.

[8] Excess returns ($R^e_{i,t}$) are computed by subtracting the one-month T-bill yield which serves as a proxy for the risk-free rate. The monthly risk-free rate data are obtained from Ken French's data library.

momentum return, $\alpha_{i,t}$ is the return left unexplained by the asset pricing model and $\varepsilon_{i,t}$ is the regression residual.[9] 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 (FF3) that adds SMB and HML, the Carhart (1997) four-factor model (FF4) that considers MoM as an additional explanatory variable, the Fama and French (2015) five-factor extension (FF5) that adds RMW and CMA to the three-factor model, and the six-factor model (FF6) that adds MoM to the five-factor model. We compute the monthly realized alpha, $\alpha^r_{i,t}$, of share class i in month t for the six-factor model in Equation 2.2 as

$$\begin{aligned}\alpha^r_{i,t} = R^e_{i,t} - \hat{b}_{i,t}(R_{M,t} - R_{f,t}) - \hat{s}_{i,t}SMB_t - \hat{h}_{i,t}HML_t \\ -\hat{r}_{i,t}RMW_t - \hat{c}_{i,t}CMA_t - \hat{m}_{i,t}MoM_t,\end{aligned} \tag{2.3}$$

where $\hat{b}_{i,t}$, $\hat{s}_{i,t}$, $\hat{h}_{i,t}$, $\hat{r}_{i,t}$, $\hat{c}_{i,t}$ and $\hat{m}_{i,t}$ are the factor loadings of the ith share class excess return with respect to the six-factor model estimated using a rolling-window regression from $t-36$ to $t-1$. Monthly realized alphas for the other factor models are calculated following the same methodology. As in Barber et al. (2016) we account for the impact of past performance on fund flows (flow-performance sensitivity)[10] by including lagged alphas of the five different factor models introduced above. While Barber et al. (2016) use an exponential decay model to weight the sum of the realized alphas over the prior 18 months, we instead use a simple average of the realized alphas over the past three, six, twelve and 18 months. For market-adjusted returns, the same procedure is applied. Our decisions are primarily driven by the predictive framework using machine learning and avoiding any form of forward-looking bias by estimating the decay factor using data not available at the time of the forecast. Furthermore, our approach allows for more flexibility in modeling the impact of performance over different horizons. In addition to the monthly realized factor model alphas we also consider the factor loadings of the FF6 model as predictive variables (Barber et al., 2016).[11] Furthermore, as in DeMiguel et al. (2023)

[9] Note that our computations are based on log-returns and we require at least 30 months of non-missing return observations over the last 36 months.

[10] See, among others, Ippolito (1992); Chevalier and Ellison (1997); Sirri and Tufano (1998); Del Guercio and Tkac (2002); Huang et al. (2007); Frazzini and Lamont (2008); Berk and van Binsbergen (2016); Barber et al. (2016); Goldstein et al. (2017); Roussanov et al. (2020).

[11] Barber et al. (2016) compute the fund's return related to each risk factor using the exponential decay model

we use valued added and the adjusted $adj.\ R^2_{i,t}$ as additional performance-related predictors. More specifically, following Amihud and Goyenko (2013), for each forecasting horizon, we use the adjusted $adj.\ R^2_{i,t}$ from the FF6 factor model rolling-window regression as a predictor of future fund flows. 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. Value added as suggested by Berk and van Binsbergen (2015) is computed as

$$va_{i,t} = (\alpha_{i,t} + ER_{i,t}/12) \cdot (TNA_{i,t-1}), \tag{2.4}$$

where we use the realized $\alpha_{i,t}$ defined in Equation 2.3 and $ER_{i,t}$ denotes the annual expense ratio of share class i.[12]

Based on information provided by Morningstar we define a dummy variable, 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 (Dass et al., 2013). 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 (worst) to five (best) stars, with fixed proportions (10%, 22.5%, 35%, 22.5%, and 10%), based on a mutual fund's historical risk and load-adjusted returns versus category peers.[13] Moreover, we consider variables that account for fund flow persistence (Dou et al., 2024). These predictors include past flows over the previous month, past mean flows over a six and 12 months horizon, respectively, and one year lagged monthly fund flows, i.e., flows in the same calendar month in the previous year to account for fund-specific seasonalities in fund flows (Kamstra et al., 2017). Additionally, we include an integer between 1 and 12 indicating the calendar month as a characteristic to capture aggregate seasonalities. Moreover, we compute the

analogous to the weighting of alphas explained above. In contrast, we directly use the factor loadings of the FF6 factor model based on rolling-window regressions from $t-36$ to t.

[12] 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.

[13] Morningstar launched its mutual fund rating system in 1985. In June 2002, Morningstar revised its rankings significantly. From its inception until the revision, Morningstar ranked all funds against each other based on their returns, while in June 2002 the new rating methodology was that funds were ranked within style categories rather than against the entire fund universe. Ben-David et al. (2022) show that this change in rating methodology did not affect the relationship between flows and ratings and that investors continued to rely on the star rating despite the change in its economic meaning.

volatility of flows as the standard deviation of monthly flows using a rolling window approach with the requirement of at least 30 months of non-missing return observations over the last 36 months. Including lagged flows also captures the impact of non- or slowly varying fund characteristics and is thus to be viewed as analogous to adding fund fixed effects in regression frameworks typically used in fund flow studies (which would be inconsistent with the predictive nature of our ML approach). Finally, we use monthly changes in the Morningstar rating (Del Guercio and Tkac, 2008) as an additional predictor.

2.3 Macroeconomic information

The basic econometric model in fund flow studies (see, e.g., Nanda et al., 2004; Barber et al., 2016; Ben-David et al., 2022) are panel regressions controlling for time-fixed effects. However, how to account for the equivalent of time-fixed effects in a predictive machine learning framework is unclear. Instead, we include various measures of financial markets and macroeconomic uncertainty that may have an impact on aggregate flows in the next period. Ferson and Kim (2012), and in a more recent study, Dou et al. (2024) show that fund flows share a significant degree of common time-series variation with such factors. Guided by these studies we include the following variables: short-term interest rates (3-month T-bill rates), long-term government bond yields, the term spread as the difference between the long term yield on government bonds and the treasury-bill, the credit spread as the difference between long-term corporate bond and long-term government bond yields, stock market returns and stock market volatility based on the CRSP value-weighted index composed of NYSE, Amex and Nasdaq stocks as well as the volatility index VIX (Ben-Rephael et al., 2012). The data are obtained from Amit Goyal's webpage, Ken French's data library and the Chicago Board Options Exchange. To take macroeconomic uncertainty into account, we include the risk aversion and uncertainty (annual volatility percentage) indices of Bekaert et al. (2022) as additional predictors.[14]

In Table 3 we provide the descriptive statistics for our final sample (training sample) used in the empirical analysis. Fund-related variables are measured at the share class level. During

[14]These data are provided by Nancy Xu on https://www.nancyxu.net/risk-aversion-index.

our sample period, the average share class has a negative monthly flow (-0.52%) with a standard deviation of 5.63% and an interquartile range of 2.44%, which indicates that there is considerable cross-sectional variation in fund flows. These flow-related summary statistics are similar to the previous literature (see, e.g., Barber et al., 2016; Ben-David et al., 2022) except that the standard deviation in our sample period (January 1991 to December 2023) is larger. Consistent with the mutual-fund literature the average realized alphas of all factor models in our sample are negative and returns load positively on the market and size factor. The average adjusted R^2 of the FF6 factor model is 0.89, suggesting that it explains a substantial part of the time-series variation in equity mutual fund returns. At the fund level, the average TNA is USD 3.59 billion, while the median TNA is considerably smaller at 695 million. The average annual expense ratio is 1.24% and the average front-end (back-end) loads are 1.05% (0.64%), respectively, with a standard deviation of 2.14% (1.28%). As discussed above, if no information about front-end or back-end loads is provided in CRSP we impute this variable with 0. As a result, the median share class does not charge any loads. The average age is 168 months (14 years). The average Morningstar rating is 3.06, while the first quartile (Q1) corresponds to a Morningstar rating of 2.0 and the third quartile (Q3) is equivalent to a rating of 4.0. The vast majority of changes in the Morningstar rating are upgrades or downgrades of one star (Del Guercio and Tkac, 2008). In our sample, the mean and median change in the Morningstar rating is 0 with a standard deviation of 0.37. Overall, our sample descriptive statistics are consistent with other studies in the mutual-fund literature (e.g. Barber et al., 2016; Ben-David et al., 2022; DeMiguel et al., 2023).

Figure 1 shows the correlation coefficients of predictor variables that exceed a threshold of $|0.6|$ and are thus considered substantially correlated.[15] High correlations (> 0.9) are observed between various factor model alphas as well as between the VIX and the risk aversion index. The target variable, fund flows ($t + 1$), has correlations with the lagged predictors that are all below the 0.6 threshold. The largest correlations with monthly flows (our target variable) are observed with mean flows over the past 6 (0.31) and 12 months (0.31), lagged monthly flows (0.20) and the Morningstar rating (0.17). All other correlations are below 0.15.

[15] We do not show the full correlation matrix due to size and readability restrictions.

2.4 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 and the macroeconomic information in Subsection 2.3. 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.[16] 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 1,183,518 observations (10,557 unique share classes) and the prediction sample contains 1,185,094 observations and 10,562 unique share classes. Both samples range from January 1991 to December 2023.

To ensure scale invariant predictors, we follow Green et al. (2017) and standardize each variable[17] used in OLS and the applied machine learning algorithms in this paper, except for the neural networks[18], 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.

[16] We deviate from this approach in the case of front-end or back-end loads. If no information is available in CRSP, we assume that no such loads are charged and impute the corresponding variable with 0.

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

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

3 Prediction framework and performance evaluation

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 our prediction framework, the five-fold cross-validation approach used for hyperparameter tuning as well as the performance evaluation methodology. Our out-of-sample period spans from January 2000 to December 2023.

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 basic OLS model tends to perform poorly when the number of the predictors is large (overfitting) and when 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.

To provide a conservative estimate of the improvements possible based on modern ML methods, we let the existing literature inform our benchmark OLS model and saturate it with variables capturing the well-known nonlinear impact of past performance on fund flows (Sirri and Tufano, 1998; Chevalier and Ellison, 1997). Specifically, we include quadratic terms of the predictors realized alpha ($\alpha^r_{i,t}$) for various factor models and market-adjusted return ($MAR_{i,t}$) as well as the corresponding means of these variables over the past three, six, twelve and 18 months to account for the convexity of the flow-performance relationship. We also create dummy variables for Morningstar ratings (two-star, three-star, four-star, and five-star, where one-star rated share classes form a reference group) to capture a potential nonlinear impact of MS ratings.

Since our realized alpha estimates are based on five different factor models (see Subsection 2.2), for each forecasting period, we train all factor models on the training set and use 5-fold cross-validation (CV)[19] to eventually determine the factor model used to estimate the realized monthly alpha that minimizes the out-of-sample mean squared forecast error (MSFE) for the respective month. The resulting model, using alphas from possibly different factor models in each forecast period, is then re-trained using all data in the training set and applied to predict fund flows using data from the test set. Such an approach is similar to the ML paradigm of using CV to find the best possible set of hyperparameters and guarantees a fair comparison between OLS and ML algorithms. We refer to this model as OLS mixed model. In addition, as a robustness check, we consider one model specification consisting of all predictor variables (full model) and five additional OLS models where each specification uses alphas from one of the five factor models introduced above (CAPM, FF3, FF4, FF5 and FF6) in combination with the remaining predictor variables.

3.2 Machine learning algorithms

ML methods are well-suited for predictive tasks and have recently been applied in the context of empirical asset pricing. Compared to traditional econometric methods the ML

[19] A more detailed description of k-fold CV for hyperparameter tuning is provided in Subsection 3.3.

literature has focused heavily on out-of-sample performance as the main criterion of interest (Athey and Imbens, 2019). We complement the growing literature on the application of ML in finance (see, among others, Gu et al., 2020; Li and Rossi, 2021; Bali et al., 2022; DeMiguel et al., 2023; Kaniel et al., 2023; Cao et al., 2024; Murray et al., 2024) by comparing and evaluating a variety of machine learning algorithms in their ability to predict future fund flows. The ML models implemented in this paper, in an increasing degree of complexity, are elastic net, decision trees, random forest, gradient boosting, and feedforward neural networks (FFNNs).[20] We primarily rely on these supervised ML methods as they are reported to exhibit strong predictive performance in regression tasks using structured (panel) data (Chen and Guestrin, 2016; Lundberg et al., 2020) and are known as the most effective machine learning algorithms in various financial studies (e.g., among others Gu et al., 2020; Bianchi et al., 2021; Li and Rossi, 2021; Bali et al., 2022; DeMiguel et al., 2023).

We now very shortly describe the ML model approaches that we implement. For a more detailed description, we refer the reader to Appendix A. The elastic net (Zou and Hastie, 2005) is a linear method like OLS but uses regularization and variable selection to reduce overfitting. As the elastic net cannot handle nonlinearities independently, it follows the specification of the full OLS model and is based on all predictor variables including the nonlinear terms for the impact of past performance to allow for a fair comparison. The other machine learning methods are able to identify and exploit nonlinearities and higher order interactions that are undetected by OLS and linear machine learning models, such as elastic net, and thus may result in more accurate forecasts. Moreover, tree ensemble methods (random forest and gradient boosting) are very effective at eliminating irrelevant features while neural networks are more sensitive to those predictors. In addition, neural networks are based on a large number of parameters and therefore require a large number of observations to provide accurate estimates. In an empirical asset pricing setting measuring equity risk premia, Gu et al. (2020) identify tree-based methods and neural networks as the best-performing algorithms. In another application, Kaniel et al. (2023) use feed-forward neural networks (FFNNs) for predicting the performance

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

(alpha) of actively managed mutual funds. In this context, we also evaluate the forecasting performance of two different FFNNs and compare it to the other machine learning algorithms and the baseline OLS models.[21] However, for application on tabular-style datasets (e.g. panel data sets) using individually meaningful predictors, tree-based models consistently outperform standard deep learning models (Chen and Guestrin, 2016; Lundberg et al., 2020).

3.3 Cross-validation, hyperparameter tuning and performance evaluation

Hyperparameter tuning and training: To find the hyperparameters[22] with the best model fit, we use k-fold CV[23] (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 randomly into five folds of approximately equal size. 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 ML 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 possible set of hyperparameters. In this context, we

[21] The first FFNN consists of one to three hidden layers and two to 32 neurons per hidden layer. In contrast, the second FFNN is a more complex network with three to ten hidden layers and 32 to 1,024 neurons per hidden layer.

[22] The specific hyperparameters that undergo tuning for each machine learning (ML) algorithm are outlined in Appendix A.

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

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.[24] 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 1991 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.[25] For predictions in January 2002, new hyperparameters using an extended training sample from January 1991 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 on a share class level, which is defined as

[24] In our application, for all ML algorithms, except random forest and 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. For random forest and neural networks, 100 and 200 trials are used, respectively.

[25] To check the robustness of the obtained results we also apply a rolling window approach of 12-, 36-, and 60-months and find our results to hold.

$$R^2_{OOS} = 1 - \frac{\sum_{(i,t)\in\tau_{OOS}} \left(flow_{i,t+h} - \widehat{flow}_{i,t+h}\right)^2}{\sum_{(i,t)\in\tau_{OOS}} \left(flow_{i,t+h} - \overline{flow}_{i,t}\right)^2} \tag{3.3}$$

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 (two-sided test) to test $H_0 : R^2_{OOS} = 0$ against $H_1 : R^2_{OOS} \neq 0$.

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

$$\hat{d}_t^{(1,2)} = \left(\hat{e}^{(1)}_{i,t+h}\right)^2 - \left(\hat{e}^{(2)}_{i,t+h}\right)^2 \tag{3.4}$$

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, $\hat{d}_t^{(1,2)} = \mu + \varepsilon_t$:

$$DMW^{(1,2)} = \frac{\hat{\mu}^{(1,2)}}{\hat{\sigma}^{(1,2)}_{\hat{\mu}}} \tag{3.5}$$

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 (3.5). The corrected test statistic is obtained as

$$HLN - DMW^{(1,2)} = \sqrt{\frac{T + 1 - 2h + T^{-1}h(h-1)}{T}} DMW^{(1,2)}. \tag{3.6}$$

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 3.3. In addition, to discriminate between the forecasting performance of two competing models, we apply the DMW test.

Table 4 reports the forecasting performance, measured by the out-of-sample R^2 statistic, for the different baseline OLS specifications for a one-month forecasting horizon for the full sample. Additionally, we also report the forecasting performance for the 10% of the share classes with the largest (top) and lowest (bottom) monthly flows. Predicting extreme flows is of particular interest for fund managers in the context of their liquidity management.

The out-of-sample performance of OLS is fairly robust among the various model specifications. For the full sample, R^2_{OOS} is very similar across model specifications and varies between 17.85% (OLS full model) and 18.21% (OLS-FF4). Interestingly, the mixed model specification does not perform particularly well. All reported out-of-sample R^2 are statistically significant at the 1% level, indicating that each applied model specification performs significantly better relative to the corresponding historical mean flow benchmark. For all models, out-of-sample R^2 is considerably higher for the funds with the 10% predicted largest outflows (ranging from 18.75% and 19.41%) than for funds receiving the 10% of predicted largest inflows, where R^2_{OOS} is only between 1.31% to 1.39%.

Table 5 shows the forecasting performance of all applied machine learning models based on the out-of-sample R^2 for the full sample and the 10% of the share classes with the largest (top) and lowest (bottom) predicted monthly flows. For the full sample, all machine learning algorithms achieve statistically highly significant positive R^2_{OOS} and show predictive ability for one-month future fund flows. In terms of linear statistical methods, the elastic net performs slightly worse than the baseline OLS models, implying that the effect of dimension reduction inherent to the algorithm does not lead to improved out-of-sample forecasts.

Compared to the best-performing OLS model (OLS - FF4, 18.21%), which serves as our OLS benchmark model in the remainder of the paper, the decision tree (18.07%) achieves a slightly lower R^2_{OOS}. In contrast, a substantial increase in predictive accuracy is achieved by the random forest and gradient boosting models, improving the R^2_{OOS} to above 22%. While the two neural networks perform 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 less complex NN I achieves an out-of-sample R^2 which is 1.07 percentage points higher than that of NN II. Therefore, the increased computational effort associated with the more complex neural network, NN II, is not justified given its inferior performance.

Similar to OLS, the predictive accuracy for machine learning algorithms is much higher for the funds in the bottom decile of the predicted flow distribution. However, the two best performing machine learning methods, random forest and gradient boosting, deliver a R^2_{OOS} in the top decile of the predicted flow distribution that is 9.57 and 8.94 percentage points higher compared to the OLS benchmark model. This shows that the best machine learning methods are much better in predicting extreme flows.

Generally, 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.

Overall, 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 returns (Gu et al., 2020; Leippold et al., 2022). While our best-performing forecasting model, random forest, yields an improvement in R^2_{OOS} for the full sample of 4.57 percentage points, the previously mentioned papers on mutual fund performance prediction report an improvement in out-of-sample R^2, based on a comparison of nonlinear ML methods relative to OLS, of between 1.90 (from 0.81 to 2.71, Leippold et al. (2022)) and 3.86 (from -3.46 to 0.40, Gu et al. (2020)) percentage points.

The results are robust when we benchmark our R^2 metric against a forecast value of zero instead of the historical mean. The out-of-sample R^2 of the best-performing ML method, random forest, is now 4.87 percentage points higher relative to the OLS benchmark. Interestingly, the overall level of R^2 for all models drops by about 5 percentage points when we use zero as naive forecast. This suggests that a forecast of zero is a superior naive forecast compared to the historical mean (Table B1).

Instead of applying an expanding window, we repeat the above analysis for the full sample using a 12-, 36-, and 60-month rolling window for training the algorithms (and hyperparameter tuning) to demonstrate the robustness of the forecasting performance documented in Table 5. The results are shown in Table 6 and provide three important insights. First, a longer training window implies higher predictive accuracy, documented by a larger R^2_{OOS} measure. Second, the forecasting ability of machine learning algorithms, with the exception of neural networks, is relatively robust regardless of the corresponding training window, while the baseline OLS model shows considerably inferior performance based on the short 12-month training window. Finally, neural networks employ a larger number of hyperparameters and thus require a large number of observations to provide accurate forecasts. In our setting, neural networks are associated with inferior forecasting performance compared to tree-based ensemble models but outperform linear methods (OLS, elastic net) on larger training windows.

To statistically evaluate the difference in the forecasting abilities of two competing forecasting models we apply modified Diebold and Mariano (1995), West (1996) tests following Equation 3.6. The results of the pairwise tests of predictive accuracy are shown in Table 7. 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 surprisingly delivers inferior 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. Among the linear models, OLS-FF4 significantly outperforms 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, followed by gradient boosting. While gradient boosting is computationally less expensive, random forest takes the most computing time of all applied ML algorithms. In conclusion, these results further corroborate the findings obtained by the out-of-sample R^2 analysis above.

Next, we explore whether the predictive ability of our models depends on the state of the economy and is stable over different economic regimes. 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. Relatedly, Warther (1995) reports positive correlations between aggregate equity market returns and fund flows. In a more recent study, Ben-Rephael et al. (2012) find that aggregate monthly net exchanges to equity funds, as a proxy for shifts between bond and equity funds, are positively contemporaneously correlated with aggregate stock market returns. Furthermore, fund managers are particularly interested in precisely predicting potential outflows during periods of aggregate outflows, where returns and liquidity are typically under increased stress.

Hence, to examine whether our results depend on different market conditions based on the business cycle, monthly aggregate flows, and the return of the aggregate stock market, we compute the out-of-sample R^2 for each algorithm and month, yielding a total of 288 observations. Subsequently, we separate these observations according to the corresponding market conditions (e.g., expansion and recession) and employ Welch (1947) tests to evaluate whether the mean R^2_{OOS} in two states differs significantly. The results are reported in Table 8.[26] Panel A (business cycle) refers to the state of the business cycle and separates the sample in expansions and recessions as defined by the NBER. All forecasting models except gradient boosting and NN (II) perform slightly better during recessions than expansions based on their R^2_{OOS}, but the differences are not statistically significant based on Welch (1947) tests. As in the overall sample, random forest performs best in both states. Panel B splits the sample into periods where monthly aggregate flows are positive (inflows) or negative (outflows). Results are stable across the two regimes. Although some models perform better during periods of aggregate inflows and some worse, the differences in R^2_{OOS} are small and never statistically significant. Again, random forest remains the best-performing model independent of market conditions. In panel C we divide the sample into periods of positive and negative aggregate monthly stock market returns using the CRSP value-weighted index. All statistical models, except random forest and gradient boosting (which are still the best performing models in both regimes), show statistically significant higher accuracy in months with negative stock market returns.

Overall, our findings from Table 8 suggest that the predictive power of our best-performing ML methods in forecasting fund flows, random forest and gradient boosting, is robust and independent of the state of the economy or market regime.

[26] Note that these results differ marginally from the ones obtained in Table 5, as the reported mean R^2_{OOS} is based on monthly means.

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 we identified as the machine learning method with the highest forecasting accuracy in the previous section. For that purpose, we estimate model-agnostic SHAP (SHapley Additive exPlanations) values (Lundberg and Lee, 2017; Lundberg et al., 2020). More specifically, for each forecasting cycle, share class, and characteristic (predictor), a SHAP value is calculated, reflecting the predictor's main effect on the respective prediction along with the averaged interaction effects between the predictor and the remaining 63 characteristics. A positive SHAP value indicates that the characteristic has a positive influence on the respective prediction relative to the average forecast, whereas a negative SHAP value indicates that the characteristic has a negative influence. To evaluate the overall importance of each characteristic, we calculate the mean absolute SHAP values, which represent the average strength of that predictor's influence on the model's predictions.[27] In our analysis, we rely on the treeSHAP algorithm[28] which delivers accurate estimates of SHAP values for random forest (Lundberg et al., 2020).

Table 9 reports the mean of the absolute SHAP values across all observations over the out-of-sample period from January 2000 to December 2023 for the 30 most important characteristics. We find that past monthly flows is the most influential fund characteristic for predicting future fund flows followed by the average flows over the past six and 12 months, respectively. 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). This finding is also consistent with the strong impact of adding fund fixed effects in standard OLS regression models for flows on the models' R^2 observed in previous studies.

The next most important predictors are the Morningstar rating and the size of the share class measured by its TNA. It is noteworthy that various measures of past performance, including realized alpha for various factor models, market-adjusted returns and value added seem to

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

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

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.

In the following, we will analyze the stability of these findings for top-10% and bottom-10%-flow fund share classes as well as for fund share classes targeted towards retail and institutional investors, respectively. Panels A and B of Table 10 show the mean of the absolute SHAP values for the funds in the top (panel A) and bottom (panel B) deciles of the predicted fund flow distribution. For both deciles, flow-related predictors are identified as the most important characteristics, followed by the Morningstar rating and TNA. In the bottom decile of the predicted flow distribution, costs seem to play a more important role: back-end loads and the expense ratio are ranked among the ten most important characteristics for funds with predicted outflows, while they do not appear among the top 15 for funds with the highest predicted inflows.

Panel C and D show the corresponding mean absolute SHAP values for institutional (panel C) and retail (panel D) share classes. The results between these two types of share classes are very similar and flow-related variables and the Morningstar rating are the most relevant features in predicting fund flows. Interestingly, even for institutional investors the Morningstar rating is more important than alphas based on factor models. Furthermore, for institutional investors, the expense ratio is of importance in predicting flows, while for retail funds cost-related characteristics (front-end and back-end loads, expense ratio) are not among the most important predictors of flows. This pattern suggests that our forecasting model assumes that institutional investors are more fee-sensitive than retail investors.

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 9) using the monthly mean absolute SHAP values for random forest. The figure illustrates that from November 2011 onwards, past monthly flows is consistently the most influential fund characteristic for flow predictions over time. However, from January 2000 to October 2011, average flows over the past six months was identified as the most important predictor in the vast majority of months and remained in second place for the rest of the out-of-sample period. These two are followed by the average flows over the past 12 months.

The Morningstar rating is less important during the earlier years of the out-of-sample period, but gained in relevance over time. Starting in the late 2000s, it consistently ranks among the top five variables. This period coincides with the years when Morningstar became a media staple, with mainstream media regularly reporting Morningstar ratings and Morningstar analysts prominently appearing on financial news programs.

While the importance of flow-related measures and the Morningstar rating is generally relatively stable over time, all other predictor variables show a much greater degree of time variation in the out-of-sample period. The most pronounced change in characteristic importance is observed for the expense ratio, which was of marginal relevance at the beginning of the sample and has become much more influential for fund flow predictions recently. To the best of our knowledge, the increased importance of the expense ratio is a new finding and is consistent with investors being more fee-sensitive in more recent years. One possible explanation could be the increased competition from low-cost, passive investment products that have gained in importance over recent years. In this context Cremers et al. (2016) find that actively managed funds charge lower fees when they face more competitive pressure from low-cost index funds.

We also observe that monthly flows lagged by 12 months increased in importance over time, suggesting that fund-specific seasonalities have become more important in more recent years. This effect is likely to be driven by the increased popularity of savings plans, where recurring investments are often automatically made in the same calendar month.

While the SHAP values presented hitherto give us some insights into which variables generally matter for flow predictions, they do not tell us anything about the direction in which they influence flows. Thus, we now focus on a better understanding of how various fund characteristics affect the predictions made by random forest in greater detail and analyze the directional impact of these predictors. 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 January 2000 to December 2023 implied by Table 9. Each point represents a SHAP value for one observation from a forecast cycle for a specific predictor. If this point is on the positive domain of the horizontal axis, this means that the respective predictor has a positive impact on predicted flows (as compared to the average prediction). The color of the corresponding points indicates whether the realization of the characteristic that led to the respective prediction was high (violet) or low (yellow). The plots for the various predictors are sorted in descending order with the most important variables at the top. We obtain information on the magnitude (low or high feature value) and the directional impact on the fund flow prediction (SHAP value) for each of these characteristics. More specifically, the beeswarm plot does not only reveal the relative importance of features, but also their actual relationships with the predicted outcome.

With the exception of monthly flows, most variables have a relatively clear relationship. Typically, a positive impact can be seen for features with high values and vice versa. Average flows over longer horizons, Morningstar ratings as well as average realized CAPM alphas over various horizons show a predominantly positive directional impact on predicted flows if the feature value is high. This is consistent with the flow-performance relationship in which flows are positively related to past performance (Chevalier and Ellison, 1997; Sirri and Tufano, 1998; Del Guercio and Tkac, 2002; Huang et al., 2007). A similar finding is observed for market-adjusted returns. Additionally, we find that the relationship of feature value and directional impact of the expense ratio on flow predictions is negative, meaning that expensive funds are predicted to incur outflows. While these findings are consistent with economic rationality of investors, several older studies on mutual fund flows show that funds that charge higher fees typically enjoy higher flows (Sirri and Tufano, 1998; Gil-Bazo and Ruiz-Verdú, 2009; Ivković and Weisbenner, 2009).

However, it is noteworthy that the relationship between the feature value and the directional impact of our most important predictor variable (with the highest mean SHAP value), previous month flows, is ambiguous. In some cases, a positive relationship is observed between high past monthly flows and predicted flows, while in other instances a negative directional impact is found, meaning that high past flows are associated with low predicted flows. This finding can be interpreted as evidence of a more complex, nonlinear underlying true relationship between previous month flows and predicted flows or a significant change in the directional impact of this variable over time. We will discuss this pattern in greater detail below.

To first analyze the temporal stability of the results documented in Figure 3, we divide the out-of-sample period into consecutive two-year intervals. Figure 4 shows the corresponding SHAP values. As in the overall out-of-sample period shown above, the relationship between past monthly flows and their impact on predicted flows is ambiguous for most two-year subperiods. This suggests that the relationship documented in the overall sample is not just driven by some periods with positive and some periods with negative relationships. Rather, these findings are consistent with a nonlinear relationship between past monthly flows and predicted flows.

In contrast, the directional impact of average flows over the past six and 12 months is fairly stable across sub-periods, such that higher average past flows are positively related to predicted flows. Additionally, we observe that a higher Morningstar rating consistently leads to a positive impact on the prediction of fund flows in all sub-periods. Moreover, a high expense ratio is negatively related to predicted flows, and the directional importance has become stronger over the out-sample period.

As discussed above, the relationship between flows in the previous month and predicted flows appears to be nonlinear. Hence, in Figure 5, we further explore the functional form of past monthly flows in more detail by plotting the SHAP value as a function of the corresponding feature value over the whole out-of-sample period. We observe a non-linear relationship between past monthly flows and predicted flows. More specifically, in quadrant I (III) a positive (negative) feature value is associated with a positive (negative) impact on predicted flows, while in quadrant II (IV) a negative (positive) feature value implies a positive (negative) predictive impact.[29] Such a complex non-linear pattern may be influenced by either the main effect or an interaction effect with other predictors. We will explore this possibility in the following.

A major advantage of most modern machine learning methods is their ability to model a large number of potentially relevant interaction terms. Thus, we next analyze interaction

[29] About 93% of the observations are located in quadrant I and III.

effects for all possible combinations of predictor variables.[30] Table 11 shows the strength of the ten most important interactions based on mean absolute SHAP interaction values for gradient boosting.[31] Interactions among the flow-related variables are the three most important interactions, with the strongest interaction found between monthly flows and average flows over the past six months. The table further reveals that monthly flows interacted with the Morningstar rating is the next most important interaction that is not entirely flow related, followed by interactions between mean flows (six months) and the Morningstar rating as well as the interactions of TNA with value added. Interestingly, while value added was not among the top-thirty variables in Table 9, it is important when interacting with TNA in forecasting fund flows. Past performance measures such as the realized alpha of various asset pricing models (CAPM, FF3, FF4) possess some importance as individual predictors as shown in Table 9 but are not important through their interactions with past fund flows or any other combination of predictor variables among the ten most important interactions. However, some fund characteristics like TNA and the expense ratio are not only important as individual predictors but are also important drivers of flows when interacting with each other.

As documented in earlier studies (see, e.g., Sirri and Tufano, 1998), there is evidence of a nonlinear relationship between past performance metrics and subsequent fund flows. Building on these findings, we examine this relationship in our predictive framework for the Morningstar rating which has been identified to be both, the most important performance metric as well as the most important non purely flow-related interaction. For that purpose, we use a violin plot[32] and display for all funds the five Morningstar rating categories and their associated SHAP value over the entire out-of-sample period for random forest, displayed in Figure 6. We observe that the four- and five-star rated funds are predominantly associated with positive flow predictions, as indicated by positive SHAP values, while the lower rated funds are all primarily associated with negative predicted flows. This suggests that the level of the Morningstar rating

[30]This corresponds to $(64 \times (64-1))/2 = 2,016$ interactions.

[31]While random forest is the best performing ML algorithm in our horse race of prediction models, we report in this version of the paper the SHAP interaction values for gradient boosting as preliminary results due to the substantially lower computational burden.

[32]Note that to better visualize the results, the violin plot does not show data (outliers) that are 1.5 $\times$ the interquartile range (IQR) above the third quartile and below the first quartile.

is fundamentally important in predicting fund flows. Overall, the relationship between the Morningstar rating and its associated SHAP value is convex. The predictions for the funds in the two highest rating categories are on average (median) positively impacted, while for all other funds, the Morningstar rating has a negative impact on the forecast, resulting in lower predicted fund flows.

We now further examine the non-linear relationship between past monthly flows and the associated SHAP values from Figure 5 to verify if this pattern could be influenced by the non-linear impact of the most important non-flow related interacted variable, the Morningstar rating. Figure 7 is based on Figure 5, showing 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 2023. This figure provides two 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. This might be due to 'new money' being more sensitive to easily digestible performance information as mirrored in the Morningstar rating. Second, while high past outflows positively impact predictions, low past outflows are mostly associated with negative predicted flows. Overall, this analysis highlights the usefulness of jointly modeling nonlinearities and interaction effects.

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 (see, among others, Gruber, 1996; Zheng, 1999; Keswani and Stolin, 2008), while another strand attributes it to the persistence of fund flows (see, 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 mutual funds with high and low future net returns. 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 3.3 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 (percentile) portfolios over the period from January 2000 to December 2023 (288 monthly observations). Finally, as in Kaniel et al. (2023), we construct a long-short prediction portfolio of each top and bottom decile (percentile) 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 288 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 our baseline OLS model with the highest forecasting accuracy (OLS - FF4), 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.

Panel A of Table 12 reports out-of-sample alphas based on net-returns 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 resulting in significant out-of-sample alphas for the long-short portfolios. The best performing model is gradient boosting which achieves a significant positive alpha relative to all applied factor models. For example relative to the FF5 and FF6 model gradient boosting delivers a highly significant alpha of 21 bp per month (2.52% per year). Interestingly, this number is virtually identical to the outperformance achieved by Kaniel et al. (2023) aimed at directly predicting fund performance (rather than flows).[33] However, it is important to note that linear models (OLS, elastic net) perform reasonably well for long-short portfolios based on the top and bottom decile of the predicted fund flow distribution and achieve higher long-short alphas compared to the decision tree, random forest and the two neural networks. Second, unlike DeMiguel et al. (2023), our machine learning implied fund flow predictions do not allow us to select long-only fund portfolios that provide statistically significant out-of-sample alphas. Thus, the approach of DeMiguel et al. (2023) to directly predict performance (rather than flows) seems to be superior in this context. Third, our approach does allow investors to identify the worst-performing funds implied by the predicted fund flow distribution. For the bottom decile portfolio and the FF5, FF6 and FF7 factor model all of the applied predictive models yield a statistically 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.

Moving to the extremes of the predicted fund flow distribution and analyzing the performance implications of high and low fund flow forecasts for the top and bottom percentiles yields very favorable results for machine learning regarding long-short portfolios. Panel B of Table 12 reports that at the 5% level or higher gradient boosting delivers a significant alpha of about 28 bp per month (3.36% per year) relative to all factor models, while linear models completely fail to do so.

[33] Depending on the factor model analyzed and using only fund-specific predictor variables, they document an outperformance ranging from 1.80% to 2.52%. They further show that adding stock-specific characteristics and sentiment does not result in long-short portfolios with higher alpha.

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 that can be taken into account by modern 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, other machine learning methods such as neural networks also possess a good forecasting ability and still also clearly outperform the baseline OLS model, while less complex machine learning algorithms such as elastic net and decision tree are inferior in their predictive ability compared to OLS.

Our results also show that predictors based on previous flows (past 1-month flows, average flows over six and 12 months), the Morningstar rating, and TNA are the most important variables for predicting future fund flows. In contrast, measures of past performance such as realized alpha from various factor models, value added, or market-adjusted returns are only of minor importance when controlling for the Morningstar rating. 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.

Moreover, our results reveal that over time the expense ratio became much more influential in predicting fund flows and - in contrast to the earlier literature - can be shown to have a strong negative impact on flows. We also identify a convex relationship between past flows and predicted flows as well as a nonlinear impact of the Morningstar 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, suggesting that 'new money' is strongly rating sensitive.

Finally, while not the main focus of our paper, we also show that ML-based fund flow predictions can be used to ex-ante separate high- and low-performing mutual funds.

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 as well how investors can use these predictions to ex-ante separate high from low-performing funds.

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.

Athey, S. and G. W. Imbens (2019): "Machine learning methods that economists should know about," *Annual Review of Economics*, 11, 685–725.

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.*

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.

Bekaert, G., E. C. Engstrom, and N. R. Xu (2022): "The time variation in risk appetite and uncertainty," *Management Science*, 68, 3975–4004.

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.

Ben-Rephael, A., S. Kandel, and A. Wohl (2012): "Measuring investor sentiment with mutual fund flows," *Journal of Financial Economics*, 104, 363–382.

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.

Cao, S., W. Jiang, J. L. Wang, and B. Yang (2024): "From Man vs. Machine to Man + Machine: The art and AI of stock analyses," *Journal of Financial Economics (forthcoming).*

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

Cen, X., W. W. Dou, L. Kogan, and W. Wu (2024): "Fund flows and income risk of fund managers," *Unpublished Working Paper, National Bureau of Economic Research.*

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

Chen, Q., I. Goldstein, and W. Jiang (2010): "Payoff complementarities and financial fragility: Evidence from mutual fund outflows," *Journal of Financial Economics*, 97, 239–262.

Chen, T. and C. Guestrin (2016): "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Association for Computing Machinery, KDD '16, 785–794.

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.

Cremers, M., M. A. Ferreira, P. Matos, and L. Starks (2016): "Indexing and active fund management: International evidence," *Journal of Financial Economics*, 120, 539–560.

Dass, N., V. Nanda, and Q. Wang (2013): "Allocation of decision rights and the investment strategy of mutual funds," *Journal of Financial Economics*, 110, 254–277.

Del Guercio, D. and P. A. Tkac (2002): "The determinants of the flow of funds of managed portfolios: Mutual funds vs. pension funds," *Journal of Financial and Quantitative Analysis*, 37, 523–557.

——— (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.

Doshi, H., R. Elkamhi, and M. Simutin (2015): "Managerial activeness and mutual fund performance," *The Review of Asset Pricing Studies*, 5, 156–184.

Dou, W. W., L. Kogan, and W. Wu (2024): "Common fund flows: Flow hedging and factor pricing," *Journal of Finance (forthcoming)*.

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.

Ferson, W. E. and M. S. Kim (2012): "The factor structure of mutual fund flows," *International Journal of Portfolio Analysis and Management*, 1, 112–143.

Franzoni, F. and M. C. Schmalz (2017): "Fund flows and market states," *Review of Financial Studies*, 30, 2621–2673.

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.

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.

Goldstein, I., H. Jiang, and D. T. Ng (2017): "Investor flows and fragility in corporate bond funds," *Journal of Financial Economics*, 126, 592–613.

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 (2007): "Participation costs and the sensitivity of fund flows to past performance," *Journal of Finance*, 62, 1273–1311.

——— (2022): "Investor learning and mutual fund flows," *Financial Management*, 51, 739–765.

Ippolito, R. A. (1992): "Consumer reaction to measures of poor quality: Evidence from the mutual fund industry," *The Journal of Law and Economics*, 35, 45–70.

Ivković, Z. and S. Weisbenner (2009): "Individual investor mutual fund flows," *Journal of Financial Economics*, 92, 223–237.

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.

Kamstra, M. J., L. A. Kramer, M. D. Levi, and R. Wermers (2017): "Seasonal asset allocation: Evidence from mutual fund flows," *Journal of Financial and Quantitative Analysis*, 52, 71–109.

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.

Khorana, A., H. Servaes, and P. Tufano (2008): "Mutual fund fees around the world," *Review of Financial Studies*, 22, 1279–1310.

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.

Nanda, V., Z. J. Wang, and L. Zheng (2004): "Family values and the star phenomenon: Strategies of mutual fund families," *Review of Financial Studies*, 17, 667–698.

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.

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.

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

Roussanov, N. L., H. Ruan, and Y. Wei (2020): "Mutual fund flows and performance in (imperfectly) rational markets?" *Unpublished Working Paper. Jacobs Levy Equity Management Center for Quantitative Financial Research Paper.*

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.

Warther, V. A. (1995): "Aggregate mutual fund flows and security returns," *Journal of Financial Economics*, 39, 209–235.

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

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 coefficients of predictor variables that exceed a threshold of 0.6 and are thus considered substantially correlated. The sample period spans from January 1991 to December 2023.

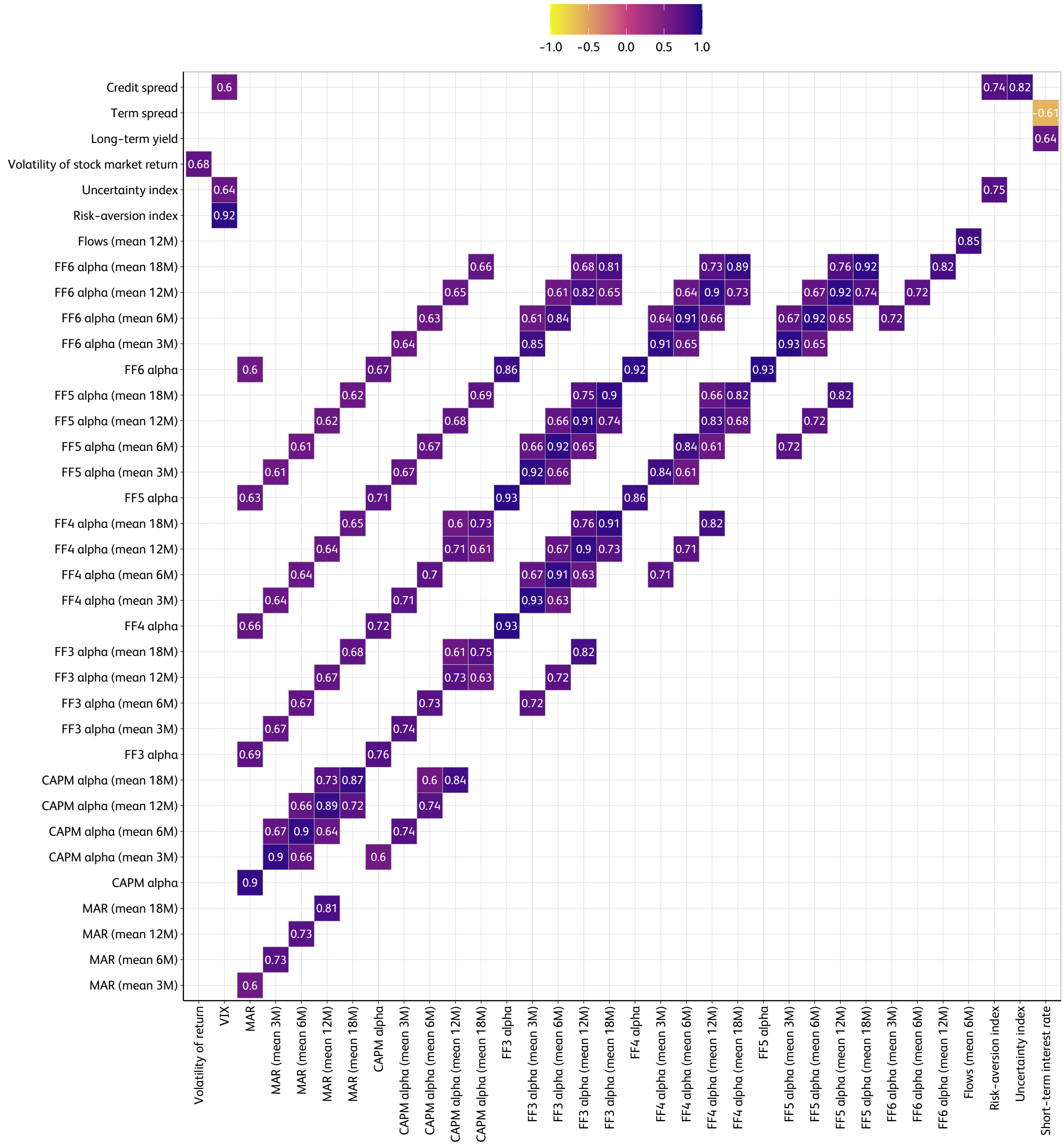

Figure 2: Evolution of characteristic importance for random forest over time

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 2023) 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 64 for the least important predictor.

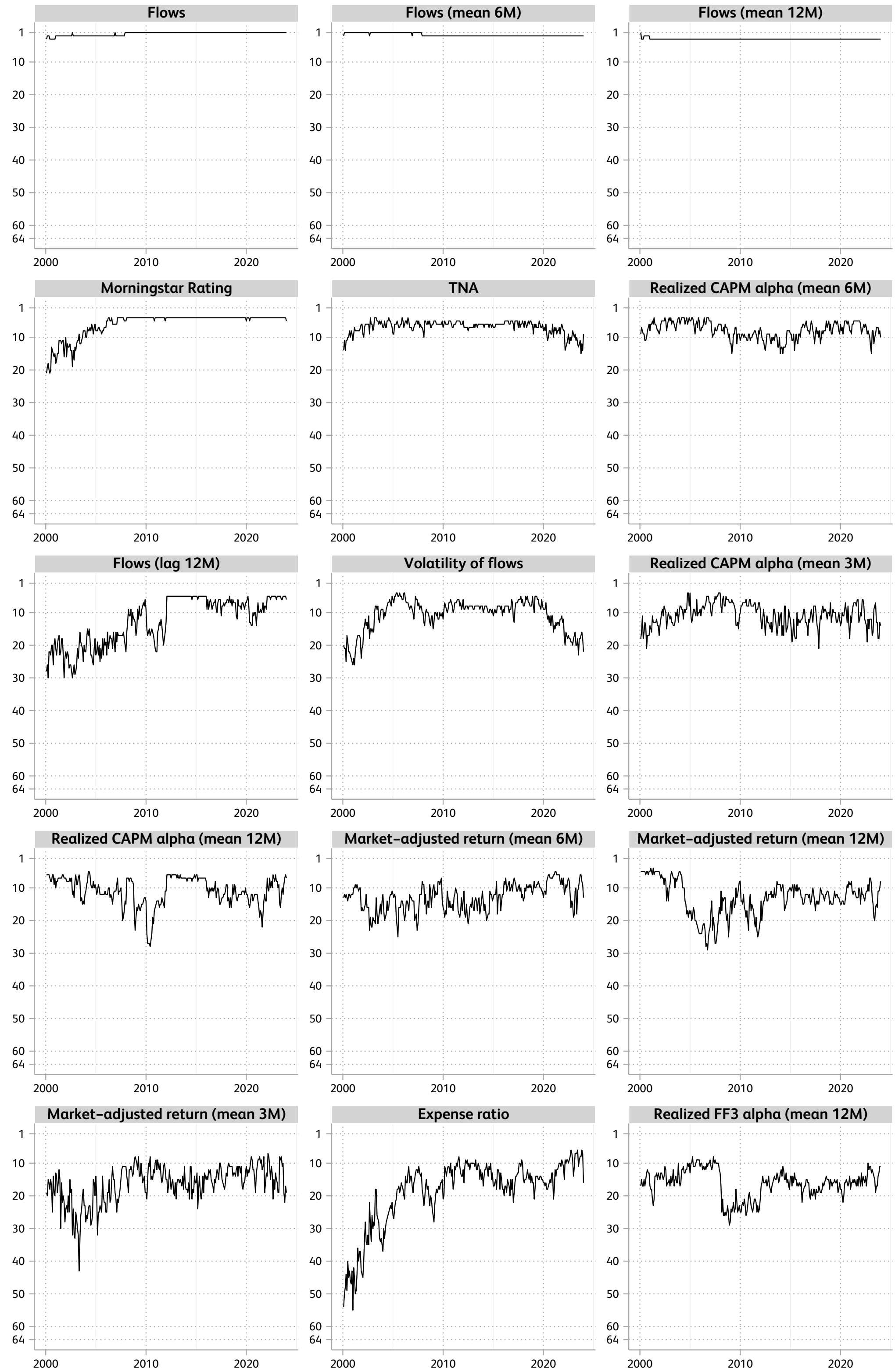

Figure 3: Directional impact of the characteristics for random forest

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 2023) 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).

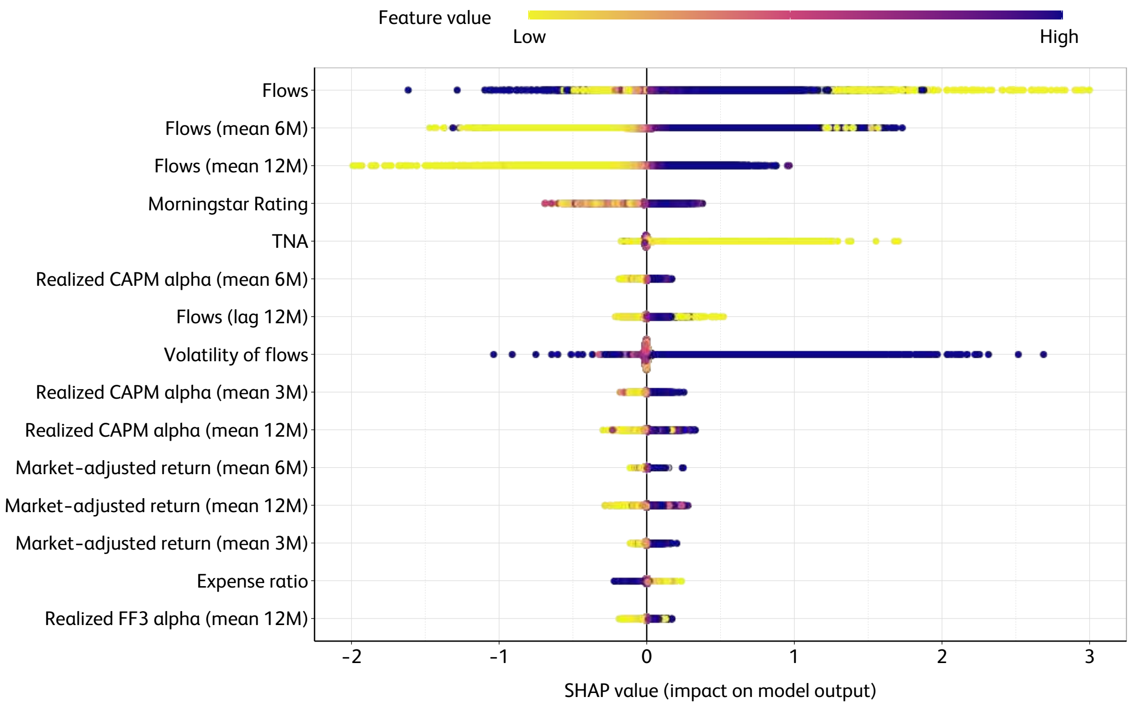

Figure 4: Directional impact of the characteristics for random forest over time

This figure shows a beeswarm SHAP plot indicating 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).

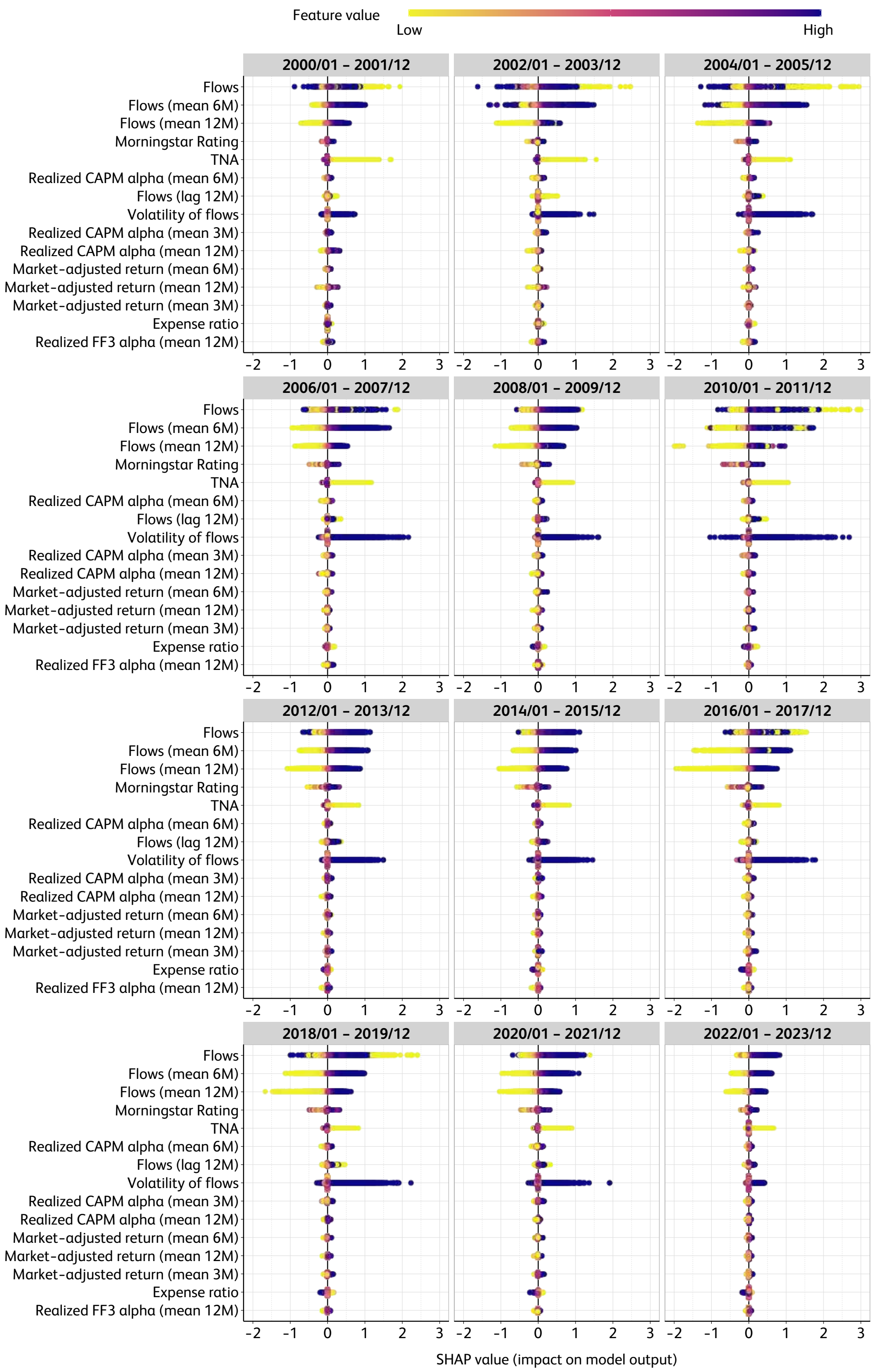

Figure 5: Functional form of past fund flows

This figure plots the nonlinear (convex) relationship for each observation between past monthly fund flows for the whole out-of-sample period (January 2000 to December 2023) and a one-month forecasting horizon for random forest.

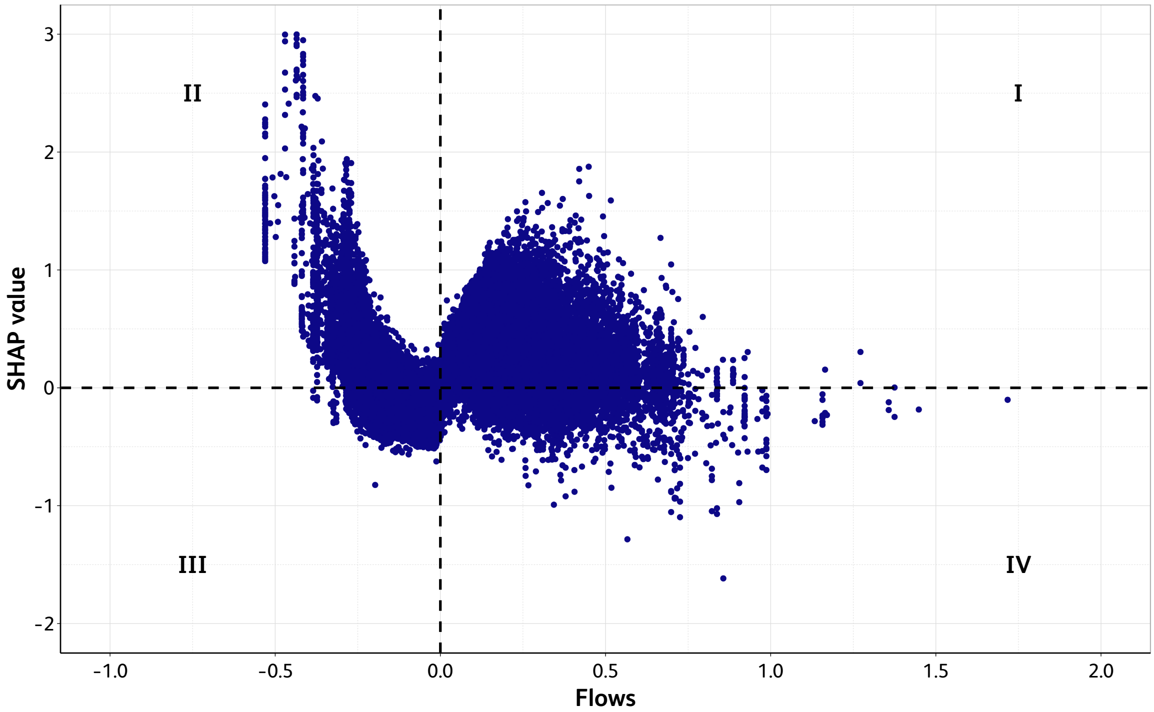

Figure 6: Nonlinear impact of the Morningstar Rating

This figure plots the five rating categories (stars) of the Morningstar rating and their corresponding SHAP values for the whole out-of-sample period (January 2000 to December 2023) and a one-month forecasting horizon for random forest. The Morningstar rating is based on a scale of one to five stars, with one being the worst and five being the best.

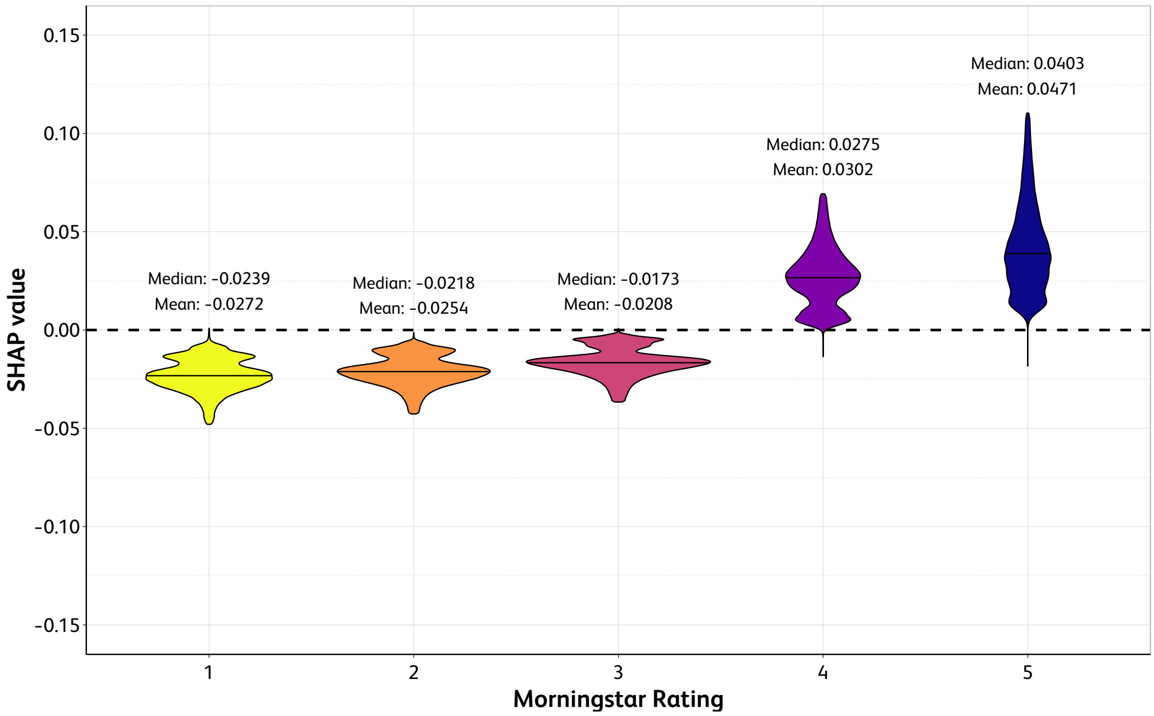

Figure 7: Interaction effect of past fund flows and Morningstar rating

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 2023) and a one-month forecasting horizon for random forest.

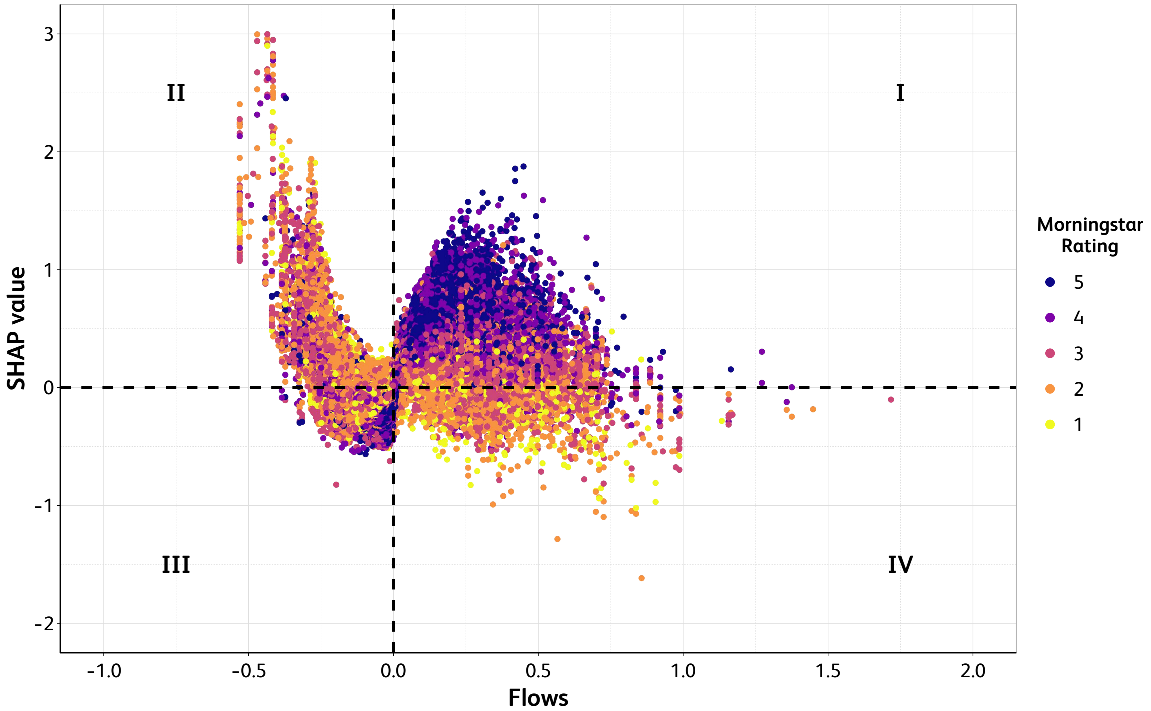

Table 1: Number of share classes per year with monthly average, median and quantile flows

This table reports the number of share classes for each year in our sample and the corresponding flow distribution (in percent) described by its mean, lower (Q1) and upper quantile (Q3). The sample period spans from January 1991 to December 2023.

		Flow distribution				
Year	Number of share classes	Mean	Std	Q1	Median	Q3
1991	309	0.98	5.10	-0.89	0.15	2.00
1992	339	1.27	4.75	-0.54	0.48	2.24
1993	536	1.33	6.76	-0.78	0.39	2.29
1994	618	0.77	5.73	-1.01	0.12	1.75
1995	787	0.79	6.66	-1.10	0.05	1.74
1996	1,033	0.87	6.28	-1.01	0.21	1.97
1997	1,263	1.00	6.89	-1.03	0.22	2.02
1998	1,611	0.52	6.77	-1.33	-0.03	1.69
1999	1,923	-0.06	6.47	-2.11	-0.44	1.27
2000	2,237	0.50	7.17	-1.67	-0.15	1.76
2001	2,236	0.38	5.77	-1.30	-0.27	1.14
2002	2,374	-0.27	5.60	-1.91	-0.69	0.68
2003	3,753	0.40	6.41	-1.41	-0.24	1.30
2004	4,038	-0.05	5.77	-1.81	-0.54	0.98
2005	4,434	-0.13	7.39	-2.13	-0.75	0.88
2006	4,686	-0.30	5.76	-2.13	-0.73	0.81
2007	4,929	-0.51	5.90	-2.22	-0.79	0.64
2008	6,461	-0.85	5.84	-2.51	-1.00	0.32
2009	6,379	-0.50	5.77	-1.93	-0.70	0.55
2010	6,312	-0.47	5.97	-1.84	-0.73	0.49
2011	6,311	-0.45	6.17	-1.92	-0.75	0.56
2012	6,002	-0.73	5.54	-1.97	-0.85	0.31
2013	6,107	-0.12	5.57	-1.49	-0.44	0.84
2014	6,281	-0.33	5.23	-1.52	-0.51	0.62
2015	6,575	-0.42	5.57	-1.55	-0.54	0.55
2016	6,624	-0.90	5.68	-1.99	-0.87	0.15
2017	6,945	-0.98	6.60	-1.94	-0.82	0.17
2018	6,936	-0.83	5.37	-1.74	-0.73	0.26
2019	6,953	-0.93	5.03	-1.85	-0.80	0.16
2020	6,984	-0.98	5.13	-2.13	-0.89	0.19
2021	7,041	-0.51	4.30	-1.49	-0.53	0.43
2022	7,080	-0.66	3.73	-1.46	-0.57	0.26
2023	7,104	-0.85	3.46	-1.56	-0.66	0.14
Total	**13,376**	**-0.47**	**5.62**	**-1.79**	**-0.65**	**0.55**

Table 2: Predictor variables by category

This table shows all 64 predictor variables sorted into two categories. The first category represents fund-specific characteristics, while the second set of characteristics is related to macroeconomic variables.

	Fund characteristics
(1)	Flows
(2)	Flows (mean past 6 months)
(3)	Flows (mean past 12 months)
(4)	Flows (lagged 12 months)
(5)	Volatility of flows
(6)	Month
(7)	Expense ratio
(8)	Front-end load
(9)	Back-end load
(10)	Turnover ratio
(11)	Age
(12)	Size of management team
(13)	Morningstar rating
(14)	Change in Morningstar rating
(15)	Volatility of return
(16)	TNA
(17)	TNA (Fund)
(18)	MAR
(19)	MAR (mean past 3 months)
(20)	MAR (mean past 6 months)
(21)	MAR (mean past 12 months)
(22)	MAR (mean past 18 months)
(23)	CAPM alpha
(24)	CAPM alpha (mean past 3 months)
(25)	CAPM alpha (mean past 6 months)
(26)	CAPM alpha (mean past 12 months)
(27)	CAPM alpha (mean past 18 months)
(28)	FF3 alpha
(29)	FF3 alpha (mean past 3 months)
(30)	FF3 alpha (mean past 6 months)
(31)	FF3 alpha (mean past 12 months)
(32)	FF3 alpha (mean past 18 months)
(33)	FF4 alpha
(34)	FF4 alpha (mean past 3 months)
(35)	FF4 alpha (mean past 6 months)
(36)	FF4 alpha (mean past 12 months)
(37)	FF4 alpha (mean past 18 months)
(38)	FF5 alpha
(39)	FF5 alpha (mean past 3 months)
(40)	FF5 alpha (mean past 6 months)
(41)	FF5 alpha (mean past 12 months)
(42)	FF5 alpha (mean past 18 months)
(43)	FF6 alpha
(44)	FF6 alpha (mean past 3 months)
(45)	FF6 alpha (mean past 6 months)
(46)	FF6 alpha (mean past 12 months)
(47)	FF6 alpha (mean past 18 months)
(48)	Factor loading RMRF (FF6)
(49)	Factor loading SMB (FF6)
(50)	Factor loading HML (FF6)
(51)	Factor loading RMW (FF6)
(52)	Factor loading CMA (FF6)
(53)	Factor loading MoM (FF6)
(54)	Value added
(55)	Adjusted R^2 (FF6)
	Macroeconomic variables
(56)	Short-term interest rate
(57)	Long-term yield
(58)	Term spread
(59)	Credit spread
(60)	Stock market return
(61)	Volatility of stock market return
(62)	VIX
(63)	Risk aversion index
(64)	Uncertainty index

Table 3: Summary statistics of predictor variables

This table reports monthly descriptive statistics of all fund characteristics (except lags and means of various variables) and macroeconomic variables 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 fund-related variables are measured at the share-class level for actively managed US domestic equity funds and the sample spans the period from January 1991 to December 2023.

Characteristic	**Mean**	**Std**	**Q1**	**Median**	**Q3**
Flows	-0.52%	5.63%	-1.91%	-0.66%	0.53%
Volatility of flows	4.31%	4.13%	1.59%	3.07%	5.63%
Month	6.50	3.44	4.00	7.00	9.00
Expense ratio	1.24%	0.54%	0.88%	1.16%	1.55%
Front-end load	1.05%	2.14%	0.00%	0.00%	0.00%
Back-end load	0.64%	1.28%	0.00%	0.00%	1.00%
Turnover ratio	70.87%	86.05%	28.00%	51.00%	88.00%
Age	168.02	121.47	89.00	137.00	208.00
Size of Management Team	0.71	0.45	0.00	1.00	1.00
Morningstar rating	3.06	1.03	2.00	3.00	4.00
Change in Morningstar rating	0.00	0.37	0.00	0.00	0.00
Volatility of return	5.00%	2.05%	3.51%	4.70%	6.07%
TNA	815.79	3447.95	33.30	116.00	461.00
TNA (Fund)	3,588.93	13,454.07	199.50	695.00	2,124.00
MAR	-0.10%	2.71%	-1.20%	-0.12%	0.98%
CAPM alpha	-0.16%	2.70%	-1.21%	-0.14%	0.93%
FF3 alpha	-0.19%	2.27%	-1.03%	-0.14%	0.70%
FF4 alpha	-0.20%	2.28%	-1.03%	-0.15%	0.69%
FF5 alpha	-0.16%	2.34%	-1.02%	-0.14%	0.73%
FF6 alpha	-0.18%	2.35%	-1.03%	-0.15%	0.71%
Factor loading RMRF (FF6)	0.98	0.19	0.90	0.98	1.06
Factor loading SMB (FF6)	0.19	0.35	-0.07	0.08	0.43
Factor loading HML (FF6)	0.01	0.35	-0.17	0.02	0.19
Factor loading RMW (FF6)	-0.05	0.32	-0.18	-0.02	0.11
Factor loading CMA (FF6)	-0.11	0.41	-0.29	-0.09	0.08
Factor loading MoM (FF6)	0.01	0.18	-0.06	0.00	0.08
Value added	-0.40	47.28	-1.05	-0.02	0.83
Adjusted R^2 (FF6)	0.89	0.14	0.89	0.94	0.96
Short-term interest rate	1.65%	1.83%	0.09%	0.95%	2.74%
Long-term yield	3.56%	1.48%	2.55%	3.44%	4.75%
Term spread	1.91%	1.43%	0.84%	1.85%	3.06%
Credit spread	1.01%	0.40%	0.78%	0.92%	1.13%
Stock market return	0.73%	4.58%	-1.77%	1.33%	3.54%
Volatility of stock market return	4.31%	1.33%	3.05%	4.37%	5.56%
VIX	19.93	8.03	13.95	18.00	23.84
Risk-aversion index	3.07	0.75	2.67	2.86	3.17
Uncertainty index	2.16	0.52	1.86	2.11	2.42

Table 4: Out-of-sample OLS model comparison based on R^2_{OOS}

This table reports the out-of-sample performance measured by the out-of-sample R^2 statistic, as defined in Equation 3.3, for the different baseline OLS models for a one-month forecasting horizon for the full sample including all predictions and the monthly top (bottom) 10% of the share classes with the highest (lowest) flows. The statistical significance of the R^2_{OOS} measure is evaluated by the Clark and West (2007) statistic (two-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 2023.

	Out-of-sample R^2		
	Full sample	**Top 10%**	**Bottom 10%**
OLS - Full Model	0.1785***	0.0138***	0.1875***
OLS - CAPM	0.1800***	0.0139***	0.1886***
OLS - FF3	0.1819***	0.0131***	0.1937***
OLS - FF4	0.1821***	0.0131***	0.1941***
OLS - FF5	0.1820***	0.0134***	0.1939***
OLS - FF6	0.1819***	0.0135***	0.1938***
OLS - Mixed Model	0.1788***	0.0133***	0.1882***
Observations	1,106,802	110,686	110,686

Table 5: Out-of-sample machine learning 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 3.3, for the applied machine learning methods for a one-month forecasting horizon for the full sample including all predictions and the monthly top (bottom) 10% of the share classes with the highest (lowest) flows. The statistical significance of the R^2_{OOS} measure is evaluated by the Clark and West (2007) statistic (two-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 2023.

	Out-of-sample R^2		
	Full sample	**Top 10%**	**Bottom 10%**
Elastic net	0.1721***	-0.0186***	0.1760***
Decision tree	0.1807***	0.0375***	0.1739***
Random forest	0.2278***	0.1088***	0.2010***
Gradient boosting	0.2237***	0.1025***	0.1980***
Neural network (I)	0.1963***	0.0407***	0.2005***
Neural network (II)	0.1856***	0.0204***	0.1937***
Observations	1,106,802	110,686	110,686

Table 6: Out-of-sample machine learning model comparison based on R^2_{OOS} using a 12-, 36-, and 60-month rolling-window on the training data

This table reports the out-of-sample performances measured by the out-of-sample R^2 statistic, as defined in Equation 3.3, for the applied machine learning methods for a one-month forecasting horizon for the full sample including all predictions using a 12-, 36-, and 60-month rolling window for the training data. The statistical significance of the R^2_{OOS} measure is evaluated by the Clark and West (2007) statistic (two-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 2023.

	Out-of-sample R^2		
	12 Months	**36 Months**	**60 Months**
OLS - FF4	-0.0558***	0.1756***	0.1774***
Elastic net	0.1719***	0.1737***	0.1733***
Decision tree	0.1600***	0.1711***	0.1776***
Random forest	0.2158***	0.2228***	0.2254***
Gradient boosting	0.2081***	0.2188***	0.2222***
Neural network (I)	-0.2475***	0.1761***	0.2019***
Neural network (II)	0.1231***	0.1719***	0.1836***
Observations	1,106,802	1,106,802	1,106,802

Table 7: 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 best performing OLS model (OLS-FF4) according to the R^2_{OOS} using the modified Diebold and Mariano (1995), West (1996) test statistic defined in Equation 3.6. 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 2023.

	Elastic Net	Decision Tree	Random Forest	Gradient Boosting	Neural Network (I)	Neural Network (II)
OLS - FF4	-23.78***	-0.64	33.04***	25.99***	14.78***	3.39***
Elastic Net		4.21***	40.87***	32.36***	26.52***	13.98***
Decision Tree			24.47***	19.79***	7.80***	2.40**
Random Forest				-4.22***	-28.56***	-36.49***
Gradient Boosting					-20.37***	-27.30***
Neural Network (I)						-14.00***

Table 8: Out-of-sample model comparison based on R^2_{OOS} in different market conditions

This table reports the out-of-sample performances measured by the out-of-sample R^2 statistic, separately in various market conditions for the applied machine learning methods and the best performing OLS model (OLS-FF4) for a one-month forecasting horizon. Panel A (business cycle) refers to the state of the economy and separates the sample in expansions and recessions as defined by the NBER. Panel B (aggregate flows) splits the sample in periods where aggregate flows are positive (inflows) or negative (outflows) while panel C (market returns) divides the sample into periods of positive and negative aggregate monthly stock market returns using the CRSP value-weighted index. To test whether the average forecasting accuracy differs in each panel we apply a Welch (1947) test and report the corresponding value of the $t-$statistic in the column $t-stat$. 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 2023.

	A. Business cycle			**B. Aggregate flows**			**C. Market returns**		
	Expansion	Recession	$t-stat$	Inflows	Outflows	$t-stat$	Positive	Negative	$t-stat$
OLS - FF4	0.1922	0.1951	-0.1874	0.1949	0.1921	0.1758	0.1830	0.2093	-2.7924***
Elastic Net	0.1828	0.1940	-0.6761	0.1848	0.1838	0.0613	0.1696	0.2091	-4.1563***
Decision Tree	0.1867	0.1916	-0.2918	0.1843	0.1879	-0.1960	0.1801	0.1997	-1.8512*
Random Forest	0.2355	0.2367	-0.0777	0.2473	0.2333	0.8743	0.2314	0.2429	-1.1898
Gradient Boosting	0.2327	0.2245	0.5015	0.2380	0.2306	0.4505	0.2270	0.2402	-1.3586
Neural Network (I)	0.2059	0.2128	-0.4248	0.2008	0.2078	-0.3785	0.1975	0.2227	-2.5293**
Neural Network (I)	0.1955	0.1839	0.6401	0.1890	0.1952	-0.3018	0.1833	0.2132	-2.6596***
Observations	998,720	108,082	E=257	136,111	970,691	I=48	719,601	387,201	P=182
Number of months	257	31	R=31	48	240	O=240	182	106	N =106

Table 9: Mean of absolute SHAP values for random forest

This table reports the mean absolute SHAP values over all forecasting periods for the 30 most important 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 2023.

Rank	Characteristic	Mean
1	Flows	0.1102
2	Flows (mean 6M)	0.1000
3	Flows (mean 12M)	0.0630
4	Morningstar Rating	0.0267
5	TNA	0.0137
6	Realized CAPM alpha (mean 6M)	0.0125
7	Flows (lag 12M)	0.0110
8	Volatility of flows	0.0105
9	Realized CAPM alpha (mean 3M)	0.0103
10	Realized CAPM alpha (mean 12M)	0.0102
11	Market-adjusted return (mean 12M)	0.0091
12	Market-adjusted return (mean 6M)	0.0091
13	Market-adjusted return (mean 3M)	0.0077
14	Expense ratio	0.0072
15	Realized FF3 alpha (mean 12M)	0.0070
16	Realized CAPM alpha	0.0063
17	Realized FF4 alpha (mean 12M)	0.0063
18	Long-term yield	0.0062
19	Month	0.0062
20	TNA (Fund)	0.0060
21	Market-adjusted return	0.0058
22	Back-end load	0.0054
23	Market return	0.0053
24	Market-adjusted return (mean 18M)	0.0052
25	Volatility of market return	0.0047
26	Age	0.0042
27	Realized FF3 alpha (mean 18M)	0.0039
28	Turnover ratio	0.0037
29	Realized CAPM alpha (mean 18M)	0.0036
30	Realized FF4 alpha (mean 18M)	0.0036

Table 10: Mean of absolute SHAP values for random forest - Top and bottom decile of the predicted fund flow distribution and for institutional and retail share classes

This table reports the mean absolute SHAP values over all forecasting periods for the 15 most important 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 2023. Panels A and B report SHAP values for the share classes in the top and bottom decile of the predicted fund flow distribution. Panels C and D show the corresponding SHAP values for institutional and retail share classes.

	A. Top 10%		**B. Bottom 10%**	
Rank	Characteristic	Mean	Characteristic	Mean
1	Flows	0.3006	Flows (mean 6M)	0.1923
2	Flows (mean 6M)	0.2331	Flows (mean 12M)	0.1636
3	Flows (mean 12M)	0.1065	Flows	0.1471
4	Morningstar Rating	0.0583	Morningstar Rating	0.0236
5	TNA	0.0367	TNA	0.0174
6	Volatility of flows	0.0307	Volatility of flows	0.0171
7	Realized CAPM alpha (mean 6M)	0.0197	Flows (lag 12M)	0.0169
8	Flows (lag 12M)	0.0172	Realized CAPM alpha (mean 6M)	0.0142
9	Realized CAPM alpha (mean 3M)	0.0171	Back-end load	0.0127
10	TNA (Fund)	0.0146	Expense ratio	0.0124
11	Realized CAPM alpha	0.0145	Market-adjusted return (mean 6M)	0.0123
12	Realized CAPM alpha (mean 12M)	0.0143	Realized CAPM alpha (mean 12M)	0.0121
13	Turnover ratio	0.0135	Market-adjusted return (mean 12M)	0.0120
14	Market-adjusted return (mean 6M)	0.0128	Realized CAPM alpha (mean 3M)	0.0104
15	Market-adjusted return (mean 12M)	0.0127	Month	0.0097
	C. Institutional		**D. Retail**	
Rank	Characteristic	Mean	Characteristic	Mean
1	Flows	0.1188	Flows	0.1049
2	Flows (mean 6M)	0.1020	Flows (mean 6M)	0.0988
3	Flows (mean 12M)	0.0676	Flows (mean 12M)	0.0602
4	Morningstar Rating	0.0318	Morningstar Rating	0.0236
5	TNA	0.0147	TNA	0.0131
6	Flows (lag 12M)	0.0131	Realized CAPM alpha (mean 6M)	0.0125
7	Realized CAPM alpha (mean 6M)	0.0126	Volatility of flows	0.0116
8	Realized CAPM alpha (mean 3M)	0.0103	Realized CAPM alpha (mean 3M)	0.0104
9	Realized CAPM alpha (mean 12M)	0.0102	Realized CAPM alpha (mean 12M)	0.0102
10	Market-adjusted return (mean 6M)	0.0097	Flows (lag 12M)	0.0098
11	Market-adjusted return (mean 12M)	0.0092	Market-adjusted return (mean 12M)	0.0090
12	Volatility of flows	0.0088	Market-adjusted return (mean 6M)	0.0088
13	Expense ratio	0.0085	Market-adjusted return (mean 3M)	0.0074
14	Market-adjusted return (mean 3M)	0.0080	Realized FF3 alpha (mean 12M)	0.0068
15	Realized FF3 alpha (mean 12M)	0.0073	Realized CAPM alpha	0.0067

Table 11: Mean absolute SHAP interaction values

This table reports the mean absolute SHAP interaction values for all monthly forecasting periods (equally weighted) for the ten most important interactions for gradient boosting. SHAP interaction values are calculated over the out-of-sample period from January 2000 to December 2023.

Rank	Characteristic I	Characteristic II	Mean
1	Flows	Flows (mean 6M)	0.0256
2	Flows	Flows (mean 12M)	0.0088
3	Flows (mean 6M)	Flows (mean 12M)	0.0081
4	Flows	Morningstar Rating	0.0079
5	Flows (mean 6M)	Morningstar Rating	0.0061
6	TNA	Value added	0.0057
7	Flows	TNA	0.0051
8	Expense ratio	TNA	0.0041
9	TNA	Volatility of Flows	0.0039
10	Morningstar Rating	TNA	0.0037

Table 12: Monthly out-of-sample alpha (in %) of the top and bottom decile (1th percentile) portfolios

Panel A (Panel B) reports the monthly out-of-sample alpha of the equally weighted top and bottom decile (1th percentile) portfolios and the corresponding long-short portfolio of the predicted flow distribution based on a one-month ahead forecasts. Additionally, in both panels, 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 2023 (288 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.

Panel A - top and bottom decile portfolios

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

Table 12: Monthly out-of-sample alpha (in %) of the top and bottom decile (1th percentile) portfolios (continued)

Panel B - top and bottom 1th percentile portfolios

	Top 1%:				**Bottom 1%:**				**Long-Short:**			
	FF4	FF5	FF6	FF7	FF4	FF5	FF6	FF7	FF4	FF5	FF6	FF7
Equally weighted	-0.08*	-0.08*	-0.08*	-0.08*	-0.08*	-0.08*	-0.08*	-0.08*				
	(0.05)	(0.04)	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.04)				
OLS	-0.06	-0.05	-0.05	-0.05	-0.28*	-0.26**	-0.26**	-0.27**	0.08	0.08	0.08	0.09
	(0.09)	(0.11)	(0.09)	(0.09)	(0.15)	(0.12)	(0.11)	(0.11)	(0.17)	(0.18)	(0.14)	(0.14)
Elastic Net	-0.08	-0.07	-0.07	-0.06	-0.28	-0.27	-0.27*	-0.27*	0.01	0.03	0.03	0.03
	(0.1)	(0.13)	(0.11)	(0.11)	(0.21)	(0.17)	(0.16)	(0.16)	(0.23)	(0.24)	(0.19)	(0.20)
Decision Tree	0.15	0.13	0.13	0.12	-0.18	-0.18*	-0.18*	-0.18*	0.27*	0.26**	0.26**	0.26**
	(0.11)	(0.10)	(0.10)	(0.10)	(0.12)	(0.10)	(0.10)	(0.10)	(0.14)	(0.13)	(0.12)	(0.12)
Random Forest	0.02	0.02	0.02	0.02	-0.23**	-0.25***	-0.25***	-0.25***	0.20	0.22	0.22*	0.22*
	(0.08)	(0.09)	(0.07)	(0.07)	(0.10)	(0.08)	(0.08)	(0.08)	(0.13)	(0.13)	(0.11)	(0.11)
Gradient Boosting	0.06	0.07	0.07	0.07	-0.26**	-0.25***	-0.25***	-0.25***	0.27**	0.28**	0.28***	0.28***
	(0.07)	(0.08)	(0.07)	(0.07)	(0.11)	(0.09)	(0.09)	(0.08)	(0.12)	(0.11)	(0.10)	(0.10)
Neural Network (I)	0.03	0.06	0.06	0.05	-0.24*	-0.22**	-0.22**	-0.22**	0.21	0.22	0.22*	0.21*
	(0.08)	(0.08)	(0.07)	(0.07)	(0.12)	(0.10)	(0.10)	(0.10)	(0.14)	(0.13)	(0.12)	(0.12)

A Machine Learning Methods

A.1 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:

$$\min_{\theta} \mathcal{L}(\theta;\cdot) = \min_{\theta}[\mathcal{L}(\theta) + \phi(\theta;\cdot)], \tag{A.1}$$

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

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

The elastic net has two positive hyperparameters, $\lambda > 0$ and $\rho > 0$, which are determined separately by cross-validation. The loss function in Equation A.1 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).

A.2 Decision tree

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

$$g\left(z_{i,t};\theta,K,L\right)=\sum_{k=1}^{K}\theta_{t-1}^{(k)}1_{\{z_{i,t}\in C_k(L)\}} \tag{A.3}$$

where K denotes the number of terminal nodes, L is the depth, $C_k\left(L\right)$ is one of the K partitions of the data, $1_{\{.\}}$ 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.

A.3 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, $\{(z_{i,t},\alpha_{i,t+h}),(i,t)\in Bootstrap\left(b\right)\}$ from the original data set and grow a decision tree

$$g_b\left(z_{i,t};\theta_{b,t-1},K,L\right)=\sum_{k=1}^{K}\theta_{b,t-1}^{(k)}1_{\{z_{i,t}\in C_k(L)\}}. \tag{A.4}$$

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:

$$g(z_{i,t}; K, L, B) = \frac{1}{B}\sum_{b=1}^{B} g_b(z_{i,t}; \theta_{b,t-1}, K, L). \tag{A.5}$$

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.

A.4 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

$$\varepsilon_{i,t+h} \leftarrow -\frac{\partial l\left(flow_{i,t+h}, g\right)}{\partial g} |g = \hat{g}_{b-1}\left(z_{i,t}\right) \tag{A.6}$$

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

$$\hat{f}_b(z_{i,t}) \leftarrow g\left(z_{i,t}; \theta, L\right). \tag{A.7}$$

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

$$\hat{g}_b\left(z_{i,t}\right) \leftarrow \hat{g}_{b-1}\left(z_{i,t}\right) + \nu \hat{f}_b(z_{i,t}), \tag{A.8}$$

$$\varepsilon_{i,t+h} \leftarrow \varepsilon_{i,t+h} - \nu \hat{f}_b(z_{i,t}). \tag{A.9}$$

The output of the boosted model is

$$\hat{g}_B\left(z_{i,t}; B, \nu, L\right) = \sum_{b=1}^{B} \nu \hat{f}_b(z_{i,t}). \tag{A.10}$$

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).

A.5 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:

$$x_k^{(j)} = f\left(\theta_{k,0}^{(j)} + \sum_{i=1}^{P} x_i{}^{(j-1)} \theta_{k,i}^{(j)}\right) \tag{A.11}$$

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

$$g(z;\theta) = \theta_0 + \sum_{k=1}^{K} x_k \theta_k, \tag{A.12}$$

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[34], 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

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

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[35] 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 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.

[35] 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.

B Additional Robustness Check

Table B1: Out-of-sample machine learning model comparison based on $R^2_{OOS,Null}$

This table reports the out-of-sample performances measured by the out-of-sample R^2 statistic, based on a null forecast, for the applied machine learning methods for a one-month forecasting horizon for the full sample including all predictions and the monthly top (bottom) 10% of the share classes with the highest (lowest) flows. The statistical significance of the $R^2_{OOS,Null}$ measure is evaluated by the Clark and West (2007) statistic (two-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 2023.

	Out-of-sample R^2		
	Full sample	**Top 10%**	**Bottom 10%**
OLS - FF4	0.1285***	0.1232***	0.1389***
Elastic net	0.1179***	0.0951***	0.1195***
Decision tree	0.1271***	0.1449***	0.1172***
Random forest	0.1772***	0.2082***	0.1463***
Gradient boosting	0.1728***	0.2026***	0.1430***
Neural network (I)	0.1436***	0.1477***	0.1457***
Neural network (II)	0.1322***	0.1297***	0.1384***
Observations	1,106,802	110,686	110,686