Machine Learning and Fund Characteristics Help to Select Mutual Funds with Positive Alpha*

Victor DeMiguel Javier Gil-Bazo Francisco J. Nogales André A. P. Santos

January 19, 2023

Abstract

Nonlinear machine-learning methods select tradable long-only portfolios of mutual funds that earn significant out-of-sample alphas of 2.3% per year net of all costs. In contrast, linear methods deliver insignificant alphas. Machine learning unveils important interactions between fund activeness and past performance—to earn positive alpha, investors should choose more active funds conditional on their having good past performance, but less active funds conditional on poor past performance. Our findings demonstrate that investors can benefit from active management, but only if they have access to the predictions of sophisticated methods that capture complexity in the relation between fund characteristics and performance.

Keywords: Mutual-fund performance; performance predictability; active management; tradable strategies; random forests; gradient boosting; nonlinearities and interactions.

JEL classification: G23; G11; G17.

*Gil-Bazo, Department of Economics and Business, Universitat Pompeu Fabra, Barcelona School of Economics, and UPF Barcelona School of Management, e-mail: javier.gil-bazo@upf.edu; DeMiguel, Management Science and Operations, London Business School, e-mail: avmiguel@london.edu; Nogales, Department of Statistics, Universidad Carlos III de Madrid, fjnm@est-econ.uc3m.es; Santos, CUNEF and University of Edinburgh Business School, andre.santos@cunef.edu. A previous version of this manuscript was circulated under the title "Can Machine Learning Help to Select Portfolios of Mutual Funds?" We are grateful for comments from Eddie Anderson, Fahiz Baba-Yara, Paul Borochin, Andrea Buraschi, Joao Cocco, Pasquale Della Corte, Francisco Gomes, Jin Guo, Martin Haugh, Juan Imbet, Marcin Kacperczyk, Howard Kung, Narayan Naik, Jean Pauphilet, Anna Pavlova, Markus Pelger, Zhen Qi, Alexandre Rubesam, Stephen Schaefer, Henri Servaes, Raman Uppal, Michael Young, and Paolo Zaffaroni, as well as seminar participants at CUNEF, Imperial College Business School, London Business School, Universidad Autónoma de Madrid, Universidad de Zaragoza, University of Bath, University of Bristol and Université Paris Dauphine and conference participants at the 2021 conference of the French Finance Association, 2021 Finance Forum (Spanish Finance Association), 2021 EFMA annual meeting, 2021 conference of the Brazilian Finance Society, 2021 FMA annual meeting, 2021 INFORMS Annual Meeting, 2021 Paris December Finance Meeting, 2022 Global Finance Conference, and 2022 UF Research Conference on Machine Learning in Finance. Javier Gil-Bazo acknowledges financial support from the Spanish Government, Ministry of Science and Innovation's grant PID2020-118541GB-I00, and Spanish Agencia Estatal de Investigación (AEI), through the Severo Ochoa Programme for Centres of Excellence in R&D (Barcelona School of Economics CEX2019-000915-S). Francisco J. Nogales acknowledges the financial support from the Spanish Government through project PID2020-116694GB-I00.

1 Introduction

Mutual-fund research consistently shows that the average active fund earns negative risk-adjusted returns (alpha) after transaction costs, fees, and other expenses (Sharpe, 1966; Jensen, 1968; Gruber, 1996; Ferreira et al., 2013). Moreover, although several studies document the existence of a subset of managers that outperform their benchmarks (Wermers, 2000; Barras et al., 2010; Fama and French, 2010; Kacperczyk et al., 2014; Berk and Van Binsbergen, 2015), it is notoriously difficult to identify the outperforming funds ex ante. We show that machine-learning methods that exploit nonlinearities and interactions in the relation between fund characteristics and performance can help to construct tradable long-only portfolios of mutual funds that earn significant out-of-sample alphas net of all costs. In contrast, linear prediction methods fail to deliver significant alpha. Our results imply that investors can earn economically significant alpha by investing in active mutual funds, but only if they have access to sophisticated prediction methods that capture the complexity in the relation between fund characteristics and performance.

Assets under passive management have recently surpassed those under active management in U.S. domestic equity mutual funds. Many interpret this victory of passive management as a consequence of the persistent inability of the average active manager to outperform cheaper passive alternatives (Gittelsohn, 2019). To determine whether at least some active managers outperform, researchers have investigated if future fund performance can be predicted using past returns. The consensus that emerges from this literature is that positive net alpha does not persist, particularly after accounting for the exposure of mutual-fund returns to the momentum factor (Carhart, 1997).[1]

Lack of persistence in fund net alpha is consistent with the seminal model of Berk and Green (2004), in which investors supply capital with infinite elasticity to those funds they expect to outperform, based on past returns. If there are diseconomies of scale in portfolio management, in equilibrium funds with positive past alpha attract more assets, and thus, earn the same expected net alpha as any other active fund: that of the alternative passive benchmark (zero). However, mutual-fund investors fail to appropriately adjust returns for risk, which suggests that their ability to judge

[1]A notable exception is the study of Bollen and Busse (2005), who find evidence of short-term (quarterly) persistence among top-performing funds.

mutual-fund performance is limited (Berk and Van Binsbergen, 2016; Barber et al., 2016; Evans and Sun, 2021). In addition, frictions may prevent investor flows from driving fund performance to zero (Dumitrescu and Gil-Bazo, 2018; Roussanov et al., 2021). Consequently, whether mutual-fund performance is predictable is ultimately an empirical question that has received considerable attention in the literature. Specifically, several studies have shown that mutual-fund characteristics can be used to predict fund performance; see Jones and Mo (2020) for a review. Typically, these studies rank funds every month or quarter on the basis of some mutual-fund characteristic. They then allocate funds to quintile or decile portfolios and evaluate the performance of the *long-short* portfolios of funds. However, only a small subset of the mutual-fund characteristics considered in the literature can be used to select *long-only* portfolios of funds with positive alpha after transaction costs, fees, and other expenses. This is crucial because open-end funds cannot be easily shorted, and thus, investors can only benefit from active management via long-only portfolios of funds that deliver positive net alpha.

Our goal is to study whether investors can benefit from active management, and thus, we take on the challenge of identifying long-only portfolios of mutual funds with positive future alpha net of all costs. Our approach departs from the existing literature along three dimensions. First, we jointly exploit 17 mutual-fund characteristics to predict fund performance, which allows us to account for the complex nature of the problem. Fund performance is determined by a host of different factors including the manager multifaceted ability, portfolio constraints, manager incentives and agency problems, as well as fund trading costs, fees, and other expenses. Thus, it seems unlikely that using a single variable to predict performance would be as efficient as exploiting a large set of characteristics.

Second, we use three machine-learning methods to forecast fund performance: elastic net, gradient boosting, and random forests. These methods can accommodate irrelevant or highly correlated predictors, and therefore, they allow us to consider multiple characteristics with lower risk of overfitting than Ordinary Least Squares (OLS). In addition, the two decision-tree based methods (gradient boosting and random forests) can exploit nonlinearities and interactions, and thus, they may uncover predictability that would be missed by linear methods such as elastic net or OLS. As a robustness test, in Section 6 we also consider neural networks.

Third, we focus on identifying *tradable* portfolios of funds. In particular, we consider long-only portfolios of mutual funds, we construct the portfolios using exclusively past data, and we evaluate their future (out-of-sample) performance in terms of alpha net of fees, transaction costs, and other expenses. Finally, we employ a dynamic approach—the decision whether to exploit a fund characteristic is taken every time we rebalance the portfolio. By allowing for variation over time in the relation between characteristics and performance, our method can accommodate changes in the determinants of fund performance due to investor learning or shifts in market conditions and manager strategies.

We compare the out-of-sample and net-of-costs performance of the portfolios of funds constructed using the three machine-learning methods, OLS, and two naive strategies (equally weighted and asset-weighted portfolios of all funds). We use monthly data on the returns and 17 characteristics of no-load actively managed US domestic equity mutual funds spanning the 1980 to 2020 period. We consider only no-load funds to ensure that our alphas are net of all costs. We use the first 10 years of data to train the three machine-learning methods and OLS to predict future annual net alpha, estimated using the five-factor model of Fama and French (2015) augmented with momentum. As predictors, we use the value of the 17 fund characteristics in the previous year. We then form a long-only equally weighted portfolio of the funds in the top decile of predicted net alpha, and compute the net return of the portfolio in the following 12 months. For every remaining year, we expand the training sample forward by one year, construct a new top-decile portfolio, and track its net return for the next 12 months. This way, we construct a time series of monthly out-of-sample net returns of the top-decile portfolio spanning the period from 1990 to 2020. Finally, we evaluate the net alpha of the portfolio over the whole out-of-sample period with respect to four models: Carhart (1997) four-factor model; Fama and French (2015) five-factor model (FF5); FF5 augmented with momentum; and FF5 augmented with momentum and the liquidity factor of Pástor and Stambaugh (2003).

We highlight three main findings. First, the two machine-learning methods that exploit nonlinearities and interactions (gradient boosting and random forests) select long-only portfolios of funds that deliver statistically significant alphas net of all costs of 2.26% and 2.54% per year, respectively, relative to the FF5 model augmented with momentum. These alphas are also economically significant—for instance, they double the average expense ratio in our sample (1.13%). In contrast, the portfolios based on the linear methods (elastic net and OLS) deliver annual net alphas of 1.04% and 1.03%, respectively, which are statistically indistinguishable from zero. The equally weighted and asset-weighted portfolios earn negative annual net alphas of −0.26% and −0.48%, respectively, consistent with existing evidence that the average active fund underperforms passive benhmarks after costs. Our findings are similar when we evaluate out-of-sample alpha using other factor models. In summary, while portfolios that exploit predictability in the data help investors to avoid underperforming funds, only the machine-learning methods that exploit nonlinearities and interactions—gradient boosting and random forests—allow them to earn significantly positive net alpha by investing in active funds.

Second, we unveil nonlinearities and interactions in the relation between the most important characteristics and future alpha. In particular, the most important characteristics for the nonlinear methods are measures of fund activeness and past performance. For example, the top predictor is the R^2 of Amihud and Goyenko (2013), obtained by regressing fund returns on the FF5 factors and momentum. Low-R^2 funds track their benchmarks less closely, and thus, R^2 is a measure of activeness. We find that the marginal relation between R^2 and performance is highly nonlinear: consistent with Amihud and Goyenko (2013), there is an inverse relation between R^2 and performance, but only for low and moderate values of R^2, with the relation being roughly flat for higher values of R^2. The nonlinear methods also unveil important interactions between fund-activeness and past-performance measures. For instance, to earn positive alpha investors should choose more active funds (low R^2) conditional on their having good past performance (high value added), but they should choose less active funds (high R^2) conditional on poor past performance (low value added). Thus, more active funds can help investors achieve positive alpha, but only those with good past performance.

Third, we show that the favorable performance of the gradient-boosting and random-forest portfolios cannot be achieved by exploiting just a few characteristics. For instance, we find that the portfolios that exploit only the three most important characteristics achieve alphas that are smaller than those of the portfolios that exploit all 17 characteristics. However, removing the two most important characteristics from the set of predictors reduces alpha by around 1% per year. These results suggest that although the top characteristics are not sufficient to achieve economically large alphas, they are necessary. Finally, we show that the importance of characteristics varies substantially over time, which highlights the need for a dynamic approach.

We check the robustness of our findings to various methodological choices. First, we show that our results are robust to considering the post-publication decay in predictability documented by McLean and Pontiff (2016). In particular, we show that the performance of our portfolios is similar if we train the machine-learning methods using at each point in time only mutual-fund characteristics and factor models proposed in papers that had already been published. Second, our results continue to hold if we use other performance measures, such as alphas based on the factor models of Cremers et al. (2013), Hou et al. (2015), and Stambaugh and Yuan (2017). Third, the performance of the top-decile portfolio is just as good or even better if we exclude from our sample institutional share classes, which implies that our results are not driven by the presence of share classes targeted to sophisticated investors. Fourth, performance is only slightly weaker if we construct portfolios consisting of funds in the top 5% or 20% of the predicted alpha distribution. Fifth, if we extend the holding period to 24 months instead of 12 months, the performance of the top-decile portfolios selected by gradient boosting and random forests improves substantially. In particular, the annual net alpha for the random-forest portfolio is 4.1%. Sixth, we find that although neural networks can deliver portfolios with positive alphas, they are systematically smaller and less significant than those obtained with gradient boosting and random forests. Seventh, the performance of the machine-learning portfolios is similar if we use a cross-validation method that accounts for time-series properties of the data. Eighth, the performance of the nonlinear machine-learning methods does not decline if we invest in at most one share class per fund.

Finally, Jones and Mo (2020) show that the ability of fund characteristics to predict performance has declined over time due to increased arbitrage activity and mutual-fund competition. Motivated by their work, we study how the alpha of the different portfolios varies from 1991 to 2020. We find that the three prediction-based portfolios (gradient boosting, random forests, and OLS) outperform the two naive portfolios (equally weighted and asset weighted) from 1991 to 2011. Consistent with Jones and Mo (2020), however, the performance of the prediction-based portfolios is similar to that of the naive portfolios from 2012 until 2018. Interestingly, all three prediction-based portfolios outperform the two naive portfolios in the last two years of our sample (2019 and 2020). We also find that the difference in the performance of the nonlinear machine-learning portfolios across different business-cycle and sentiment regimes is not statistically significant.

We emphasize two implications of our work for investment managers and regulators. First, the economically large positive net alphas that we document show that investors can benefit from active management in the mutual-fund industry, but only if they have access to the predictions of sophisticated nonlinear methods. This is consistent with a competition framework à la Berk and Green (2004) in which frictions prevent a substantial fraction of the investor population from accessing the predictions of sophisticated machine-learning methods. Thus, our findings suggest that there is scope for managers of funds of funds, pension-plan administrators, financial advisors, and independent analysts to integrate machine learning with other tools in order to help investors select active mutual funds with positive alpha.

Second, we show that mutual-fund characteristics that do not require fund portfolio holdings are enough to predict positive alpha. This is particularly relevant given the recent debate on the SEC proposal to raise the asset threshold for mandatory portfolio disclosure (Form 13F) from US$ 100 million to US$ 3.5 billion (Aliaj, 2020). While information on portfolio holdings is potentially valuable to investors, it can also reveal portfolio strategies and reduce active managers' incentives to identify mispriced assets, which can be detrimental for market efficiency (Aragon et al., 2013; Shi, 2017). Our results imply that even if no information on portfolio holdings had been available during our sample period, our methods would have identified funds with positive alphas on average. This finding is relevant for the debate on the costs and benefits of mandatory portfolio disclosure.

Our paper contributes to a large literature that documents associations between a single mutual-fund characteristic and fund performance (see Jones and Mo, 2020, for a review). Unfortunately, a strong association between a fund characteristic and performance does not guarantee that long-only portfolios of funds based on that characteristic will earn positive net alphas. For instance, higher expense ratios are negatively associated with net fund alphas in the cross section (in our sample, funds in the bottom decile of the expense-ratio distribution outperform funds in the top decile by 1% per year relative to the FF5 model augmented with momentum), but a portfolio that invests only in the cheapest funds does not outperform passive benchmarks in net terms. In other words, expense ratios help investors to avoid expensive underperforming funds, but not to select outperforming funds with positive net alphas. In fact, only seven of the 27 studies identified by Jones and Mo (2020) report positive and statistically significant in-sample Carhart (1997) alphas after fees and transaction costs for long-only portfolios of mutual funds (Chan et al., 2002; Busse and Irvine, 2006; Mamaysky et al., 2008; Cremers and Petajisto, 2009; Elton et al., 2011; Amihud and Goyenko, 2013; Gupta-Mukherjee, 2014). We contribute to this literature by showing that it is possible to select long-only portfolios of mutual funds with significant positive net alpha by exploiting multiple characteristics and using nonlinear machine-learning methods.

Our paper is related to an emerging literature that uses machine learning to predict fund performance. Wu et al. (2021) predict future *hedge-fund returns* by exploiting characteristics constructed from fund historical returns. Instead, we predict future *mutual-fund alphas* by exploiting both fund historical returns as well as other fund characteristics. Like us, Li and Rossi (2020) use machine learning to select portfolios of mutual funds, but a fundamental difference between the two papers is that they use disjoint sets of predictors: while Li and Rossi (2020) exploit data on *fund holdings* and *stock characteristics*, we exploit data on *fund characteristics*. Our findings complement theirs by showing that investors can select portfolios of mutual funds with positive net alpha by exploiting *solely* the information contained in fund characteristics. Kaniel et al. (2021) use neural networks to predict mutual-fund alpha using a comprehensive set of predictors that includes stock characteristics, fund characteristics, and macroeconomic variables. They not only corroborate our finding that fund characteristics predict performance, but also show that when fund characteristics are included as predictors, stock characteristics no longer help to predict alpha. A key distinguishing feature of our work is the focus on tradable portfolios of mutual funds, which allows us to study whether investors can actually benefit from active management.

In particular, we identify long-only portfolios of mutual funds using exclusively past data, and evaluate their future (out-of-sample) performance net of all costs (including loads). Kaniel et al. (2021) focus on *long-short* portfolios of mutual funds, forecast performance using three-fold cross validation over the entire sample, and do not account for fund loads. Moreover, most of the predictability in *after-fee* alpha documented by Kaniel et al. (2021, Figure 6b) comes from the short leg of their long-short portfolios of funds.

Our paper is also related to studies that use Bayesian methods to construct optimal portfolios of mutual funds (Baks et al., 2001; Pástor and Stambaugh, 2002; Jones and Shanken, 2005; Avramov and Wermers, 2006; Banegas et al., 2013). Unlike these papers, we do not provide recommendations to investors on how they should allocate their wealth across funds given their preferences and priors about managerial skill and predictability. Instead, we try to identify active funds with positive alpha that investors may choose to combine with passive funds and other assets to achieve better risk-return tradeoffs.

Finally, our paper is related to the growing literature that employs machine learning to address empirical problems in Economics and Finance such as predicting global equity-market returns (Rapach et al., 2013); predicting consumer credit-card defaults (Butaru et al., 2016); measuring equity-risk premia (Gu et al., 2020; Chen et al., 2020); detecting predictability in bond risk premia (Bianchi et al., 2021); building test assets that capture nonlinearities and interactions in asset pricing (Feng et al., 2020; Bryzgalova et al., 2019); forecasting inflation (Garcia et al., 2017; Medeiros et al., 2021), and studying the relation between investor characteristics and portfolio allocations (Rossi and Utkus, 2020). Masini et al. (2021) provide a review of applications. In the context of mutual funds, Pattarin et al. (2004), Moreno et al. (2006), and Mehta et al. (2020) employ machine learning to classify mutual funds by investment category, but they do not study fund performance. Chiang et al. (1996) and Indro et al. (1999) use neural networks to predict mutual-fund net asset value and return, respectively. While these authors focus on forecasting accuracy, our goal is to identify funds with superior performance.

2 Data

In this section, we describe the data we use in our analysis. Section 2.1 describes the sample data. Section 2.2 defines the 17 monthly mutual-fund characteristics that we consider. Section 2.3 explains how we transform these monthly characteristics to generate the annual target and predicting variables for the machine-learning methods.

2.1 CRSP sample data

We collect monthly information on US domestic-equity mutual funds from the CRSP Survivor-Bias-Free US Mutual Fund database. To keep our analysis as close as possible to the actual selection problem faced by investors, we perform the analysis at the share-class level.[2] Moreover, we restrict our analysis to share classes that charge no front-end or back-end loads, and thus rebalancing our portfolios of mutual funds does not incur any costs. Our sample includes both institutional and retail share classes and spans from January 1980 to December 2020.

We apply a few filters that are common in the mutual-fund literature. First, we include only share classes of actively managed funds, therefore excluding ETFs and passive mutual funds.[3] Second, we include only share classes of funds with more than 70% of their portfolios invested in equities. Third, to avoid previously documented biases in the CRSP database, we exclude observations of a share class before it reaches 36 months of age and before the first observation with at least US$ 5 million of Total Net Assets (TNA), see Elton et al. (2001) and Evans (2010). Our final sample contains 8,200 unique share classes, of which 7,398 correspond to diversified equity funds (representing 94% of aggregate TNA in the sample) and 802 to sector funds.

2.2 Mutual-fund characteristics

We construct a dataset of 17 share-class characteristics using readily available information on fund characteristics and historical returns. None of our characteristics requires information about

[2]In Section 6 we show that our findings are robust to investing in at most one share class per fund.

[3]We use the index-fund identifier from CRSP, index_fund_flag, to identify funds that aim to replicate an index. When the identifier is missing, we use the fund name to infer whether it is passively managed.

portfolio holdings, and thus, our set of predictors is disjoint from that used by Li and Rossi (2020).[4]

For the ith share class in the mth month, we obtain data on its *return* in excess of the risk-free rate net of expenses and transaction costs ($r_{i,m}$), *total net assets* ($TNA_{i,m}$), *expense ratio* ($ER_{i,m}$), and portfolio *turnover* ratio.[5] In addition, we compute the class *age* as the number of months since its inception date; we estimate the monthly *flows* as the relative growth in the class TNA adjusted for returns net of expenses

$$flow_{i,m} = \frac{TNA_{i,m} - TNA_{i,m-1}\,(1 + r_{i,m})}{TNA_{i,m-1}}; \tag{1}$$

we estimate the *volatility of flows* as the standard deviation of flows in the calendar year; and we compute the *manager tenure* in years.[6] All of these characteristics have been identified in the literature as predictors of mutual-fund performance; see, for instance, Chen et al. (2004); Rakowski (2010); Jones and Mo (2020).

Moreover, we obtain several characteristics associated with the time-series regression of share-class returns on the five Fama and French (2015) and momentum factors (hereafter, FF5+MOM). In particular, for each share class and month in our sample, we run a "rolling-window" regression of the share-class returns on the FF5+MOM factor returns for the previous 36 months.[7] We then compute *alpha t-stat* (the intercept scaled by its standard error) as well as *beta t-stats*. We use t-stats instead of raw alphas and betas as predictors to account for estimation error (Hunter et al., 2014). In addition, we use the R^2 from the FF5+MOM rolling-window regression as a predictor of fund performance, as proposed by Amihud and Goyenko (2013), who explain that low-R^2 funds track the benchmark less closely, and thus, R^2 is a measure of activeness.[8] We also compute the

[4] Moreover, our dataset allows us to study whether it is possible to select mutual funds even in the absence of information on portfolio holdings, which is important for the current debate on the costs and benefits of mandatory portfolio disclosure (Aliaj, 2020).

[5] We proxy for the risk-free rate using the one-month T-bill rate downloaded from Ken French's website.

[6] We cross-sectionally winsorize flows at the 1st and 99th percentiles; that is, each month we replace extreme observations that are below the 1st percentile or above the 99th percentile with the value of those percentiles. The computation of the standard deviation of flows is based on winsorized flows.

[7] To run each regression, we require at least 30 months of non-missing returns in the 36-month window.

[8] Another popular measure of fund activeness is the active share of Cremers and Petajisto (2009). We do not include this measure in our dataset because we rely only on fund characteristics that do not require information on mutual-fund holdings.

monthly realized alpha for the ith share class in the mth month ($\alpha_{i,m}$) as:

$$\begin{aligned}\alpha_{i,m} = r_{i,m} &- \widehat{\beta}_{MKT,i,m}\, MKT_m - \widehat{\beta}_{SMB,i,m}\, SMB_m - \widehat{\beta}_{HML,i,m}\, HML_m \\ &- \widehat{\beta}_{RMW,i,m}\, RMW_m - \widehat{\beta}_{CMW,i,m}\, CMW_m - \widehat{\beta}_{MOM,i,m}\, MOM_m, \end{aligned} \tag{2}$$

where MKT_m, SMB_m, HML_m, RMW_m, CMW_m, and MOM_m are the returns in month m of the five Fama-French and momentum factors, and $\hat{\beta}_{MKT,i,m}$, $\hat{\beta}_{SMB,i,m}$, $\hat{\beta}_{HML,i,m}$, $\hat{\beta}_{RMW,i,m}$, $\hat{\beta}_{CMW,i,m}$, $\hat{\beta}_{MOM,i,m}$ are the factor loadings of the ith share class excess return with respect to the FF5+MOM factors estimated using the 36-month estimation window ending in month $m-1$.

Finally, we use the realized alpha defined in Equation (2) to compute the *value added* for each class and month, which we define as in Berk and Van Binsbergen (2015):

$$value\ added_{i,m} = (\alpha_{i,m} + ER_{i,m}/12) \times TNA_{i,m-1}. \tag{3}$$

This variable captures the dollar value extracted by the fund's manager from the asset market.[9]

Table 1 lists the 17 share-class characteristics and their definitions, and Table 2 reports the mean, median, standard deviation, and number of class-month observations for each of the characteristics. Consistent with the mutual-fund literature, we observe that the average share class in our sample has negative alpha and loads positively on the market and size factors. The average R^2 is 90.6%, which suggests that the FF5+MOM factors explain most of the time-series variation in equity mutual-fund returns. The total number of class-month observations varies across variables from 582,328 to 679,569.

2.3 Target and predicting variables

We now explain how we transform the 17 mutual-fund characteristics to generate the target and predicting variables for the machine-learning methods. First, we convert our sample from monthly

[9]In their study, Berk and Van Binsbergen (2015) estimate before-fee alpha by regressing fund gross returns on the gross returns of passive mutual funds tracking different indexes. In unreported analyses, we follow their approach and obtain similar results to those based on the FF5+MOM model.

to annual frequency because some of the characteristics are available at the quarterly or even annual frequency, and even those characteristics available at the monthly frequency are very persistent. We compute annual realized alpha by adding the monthly realized alphas in each calendar year. We compute annual flows and value added by averaging their monthly values in each calendar year. Flow volatility is already defined at the annual frequency. For all other characteristics, we use their values in December of each year.

Second, like Green et al. (2017) we standardize each characteristic so that it has a cross-sectional mean of zero and a standard deviation of one. This ensures the estimation process of the machine-learning methods is scale invariant. We set missing characteristic values to zero.

Third, we build our final dataset consisting of the target variable and the pre-processed characteristics that we use as predictors when training the machine-learning methods. Our target variable is the share-class realized alpha in the calendar year. This choice is consistent with our goal to exploit any information in share-class characteristics to generate positive alpha, regardless of the source of alpha. In contrast, Li and Rossi (2020) use fund excess returns as their target variable, which allows them to study whether the returns of mutual funds can be predicted from the characteristics of the stocks they hold. The 17 characteristics we use as predictors are the following one-year-lagged standardized variables: annual realized alpha, alpha t-stat, TNA, expense ratio, age, flows, volatility of flows, manager tenure, value added, R^2, and the t-stats of the market, profitability, investment, size, value, and momentum betas.[10] Figure 1 shows the correlation matrix of the target and predicting variables. The target variable has low correlation with lagged predictors. However, some predictors exhibit substantial positive and negative correlations, with the highest correlation being that between lagged flows and volatility of flows (58%).

[10]We note that both our target variable, annual realized alpha, and some of the predictors are not directly observable and must be estimated from the data. While this may pose a problem for inference, our goal is not to conduct inference but to predict future performance.

3 Machine-learning methods

We use well-known `R` packages to implement the methods—the interested reader can refer to their documentation for a detailed description of the methods.[11] Gu et al. (2020) also provide an extensive description of various machine-learning methods in the context of asset pricing. In the remainder of this section, we briefly describe the methods we consider and the five-fold cross-validation procedure we use to tune their hyper parameters.

We organize our data in panel structure, with years indexed as $t = 1, 2, \ldots, T$ and share classes as $i = 1, 2, \ldots, N_t$. As a benchmark, we use the ordinary least squares (OLS) method:

$$\min_{\theta} \sum_{t=1}^{T-1} \sum_{i=1}^{N_t} (\alpha_{i,t+1} - z'_{i,t}\theta)^2,$$

where $\alpha_{i,t+1}$ is the realized alpha of the ith share class in year $t+1$, $z_{i,t}$ is a K-dimensional vector of standardized characteristics for the ith share class in year t, and θ is the K-dimensional parameter vector. The OLS estimator of realized alpha, $z'_{i,t}\theta$, is a *linear* function of the share-class characteristics. Although OLS provides an unbiased and interpretable prediction, machine-learning methods often outperform OLS for data that exhibit high variance, nonlinearities, and interactions.

We consider three machine-learning methods: elastic net, random forests, and gradient boosting. *Elastic net* is a linear method, like OLS, but uses regularization to alleviate overfitting and provide robust predictions. To capture nonlinearities and interactions, we consider two types of ensembles of decision trees (*random forests* and *gradient boosting*), which often outperform the linear methods on structured (tabular) data like our mutual-fund database; see, for instance, Medeiros et al. (2021).

Another popular machine-learning method is neural networks, which tend to perform well on non-structured data or highly nonlinear structured data. To capture these nonlinearities, neural networks employ a large number of parameters, and hence, they require a large number of observations to deliver accurate estimates. Consequently, neural networks are not as well suited

[11] Specifically, we use `glmnet`, `randomForest`, `xgboost`, and `h2o` packages for implementing elastic net, random forests, gradient boosting, and neural networks, respectively. The documentation for these four packages can be found in Friedman et al. (2010), Liaw and Wiener (2002), Chen et al. (2020), and LeDell et al. (2020), respectively.

to our setting as ensembles of trees. Nonetheless, as a robustness check we evaluate the performance of feed-forward neural networks with up to three hidden layers in Section 6.[12] Below, Sections 3.1, 3.2, and 3.3 describe the three machine-learning methods we consider and Section 3.4 describes how we use five-fold cross validation to tune the hyper parameters of the methods.

3.1 Elastic net

Regularization is often employed to alleviate overfitting in datasets with a large number of predicting variables. The elastic net approach proposed by Zou and Hastie (2005) uses both 1-norm and 2-norm regularization terms to *shrink* the size of the estimated parameters. The general framework for the elastic net, with two regularization terms, is as follows:

$$\min_{\theta} \sum_{t=1}^{T-1} \sum_{i=1}^{N_t} (\alpha_{i,t+1} - z'_{i,t}\theta)^2 + \lambda\rho \, ||\theta||_1 + \lambda(1-\rho) \, ||\theta||_2^2, \tag{4}$$

where $||\theta||_1 = \sum_{k=1}^{K} |\theta_k|$ and $||\theta||_2 = (\sum_{k=1}^{K} \theta_k^2)^{1/2}$ are the 1-norm and 2-norm of the parameter vector θ, and λ and ρ are hyper parameters. The 1-norm term ($\lambda\rho \, ||\theta||_1$) can be used to control the degree of sparsity of the estimated parameter vector θ and the 2-norm term ($\lambda(1-\rho) \, ||\theta||_2^2$) can be used to increase the stability. For the case with $\rho = 0$, the objective function in (4) includes only the 2-norm term, and thus, elastic net is equivalent to ridge regression, which provides a dense estimator of the paremeter vector θ. If, on the other hand, $\rho = 1$, the objective function includes only the 1-norm term, and the Least Absolute Sum of Squares Operator (LASSO) regression is performed, which provides a sparse estimator.[13] We explain in Section 3.4 how we calibrate the two hyper parameters ρ and λ.

3.2 Random forests

Random forests are ensembles of decision trees formed by bootstrap aggregation (Breiman, 2001). Decision trees split a sample recursively into homogeneous and non-overlapping regions shaped

[12] We have not considered other classes of machine-learning methods such as principal-component regression or partial least squares because they are typically outperformed by elastic net; see Elliott et al. (2013).

[13] See Hastie et al. (2009, p. 61–73) for a detailed discussion of LASSO, ridge regression, and elastic net.

like high-dimensional boxes. The procedure to generate these boxes is often represented as a tree, in which the sample is split at each node based on the characteristic that is most relevant at that particular node. The tree grows from the root node to the leaf nodes, and the prediction is the average value of the target variable for the observations in each leaf node. In Section 5 we discuss a specific example of decision tree for two of the characteristics in our dataset.

Decision trees are highly interpretable, but their performance can be poor because of the high variance of their predictions. Random forests reduce the prediction variance by averaging across the predictions of the numerous decision trees in a *forest*. The reduction in prediction variance is inversely related to the correlation between trees, and thus, ideally the trees should be as uncorrelated as possible. To accomplish this, random forests use bootstrap to select the observations for each tree, and consider a random subset of characteristics for each node of a tree.

Our random-forest method uses bootstrap with replacement to generate $B = 1,000$ samples from the original data. For each of the bootstrap samples, the method grows a decision tree by selecting a random set of $m < K$ characteristics at each node, and choosing the best out of these m characteristics to split the sample. Section 3.4 discusses how we tune the hyper parameter m. The existing literature shows that random forests achieve good prediction performance, specially when there are many prediction variables and their relation to the target variable is nonlinear and contains interactions; (e.g. Medeiros et al., 2021; Coulombe et al., 2020).

3.3 Gradient boosting

Gradient boosting uses ensembles of decision tress, but instead of aggregating independent decision trees like random forests, gradient boosting aggregates decision trees *sequentially* in order to give more influence to those observations that are poorly predicted by previous trees. As a result, the gradient-boosting method starts from weak decision trees (those with prediction performance only slightly better than random guessing) and converges to strong trees (better performance). In this fashion, boosting achieves improved predictions by reducing not only the prediction variance, but also the prediction bias (Schapire and Freund, 2012).

At each iteration of gradient boosting, a new decision tree is used to fit the *residuals* of the current ensemble of decision trees. Thus, this new decision tree gives more weight to those observations that are poorly predicted by the current ensemble. Then, gradient boosting updates the ensemble using the new decision tree. A key hyper parameter in gradient boosting is the learning rate, which determines the weight the ensemble gives to the most recent decision tree.

Unlike random forests, gradient boosting tends to overfit the data. To avoid overfitting, gradient boosting employs a number of regularization techniques that require tuning additional hyper parameters. For instance, gradient boosting often imposes constraints on the number of decision trees aggregated, the depth and number of nodes of each tree, and the minimum number of observations in a leaf node.

3.4 Cross validation of hyper parameters

For each estimation window, we tune the hyper parameters of the elastic net, random forests, and gradient boosting using five-fold cross-validation; see Hastie et al. (2009, Chapter 7). Specifically, we select a grid of possible values for the hyper parameters. We divide the sample into five equal intervals or "folds." For j from 1 to 5, we remove the jth fold and use the remaining four folds to obtain the predictions corresponding to the different values of the hyper parameters. We then evaluate the prediction error (or cross-validation error) of the prediction associated with each value of the hyper parameters on the jth fold. After completing this process for each of the five folds, we select the value of the hyper parameters that minimizes the average cross-validation error.

An alternative to k-fold cross validation that accounts for the time-series properties of the data is *time-series cross validation*, which reserves a section at the end of the training sample for evaluation. In Section 6, we report the results of a robustness check where we use time-series cross validation. We find that five-fold cross validation performs slightly better, consistent with empirical and theoretical results in Bergmeir et al. (2018) and Coulombe et al. (2020).

4 Performance of machine-learning portfolios

In this section, we first describe our performance-evaluation methodology and then compare the out-of-sample performance of the various portfolios.

4.1 Performance-evaluation methodology

We now describe the procedure we use to select share classes and evaluate the performance of the resulting portfolios. Although the analysis is carried out at the share-class level, for simplicity herein we refer to share classes as funds.

We use the first 10 years of data on one-year ahead realized alphas (from 1981 until 1990) and one-year-lagged fund characteristics (from 1980 until 1989) to train each machine-learning method and OLS. We then use the values of fund characteristics in December of 1990, which are not employed in the training process, to predict fund performance in 1991 using the previously trained method. We form an equally weighted portfolio of the funds in the top decile of the predicted-performance distribution and track its return (net of expenses, fees, loads, and transaction costs) in the 12 months of 1991. If, during that period, a fund that belongs to the portfolio disappears from the sample, the amount invested in that fund is equally distributed among the remaining funds. For every successive year, we expand the training sample forward one year, train the algorithm again on the expanded sample, make new predictions for the following year, construct a new top-decile fund portfolio and track its net return during the next 12 months. This way, we construct a time series of monthly out-of-sample net returns of the top-decile fund portfolio that spans from January 1991 to December 2020 (360 months). The average number of funds selected into the top-decile portfolios is 159 with a minimum of 11 and a maximum of 326.

To evaluate the out-of-sample performance of the top-decile fund portfolio, we run a time-series regression of the 360 out-of-sample monthly portfolio excess returns on contemporaneous risk-factor returns. The portfolio alpha is the intercept of the time-series regression. We consider four risk-factor models to evaluate portfolio performance: the Fama and French (1993) three-factor model augmented with momentum (FF3+MOM) proposed by Carhart (1997); the Fama and

French (2015) five-factor model (FF5); the FF5 model augmented with momentum (FF5+MOM); and the FF5 model augmented with momentum and the aggregate liquidity factor of Pástor and Stambaugh (2003) (FF5+MOM+LIQ). Note however, that in all cases, fund selection is based on performance predicted according to the FF5+MOM model.

4.2 Out-of-sample net performance

Table 3 reports the out-of-sample alpha net of all costs of the top-decile fund portfolios selected by the three machine-learning methods—gradient boosting, random forests, and elastic net—and by OLS. For comparison purposes, we also report the alpha of two naive fund portfolios: an equally weighted and an asset-weighted portfolio of all share classes, both rebalanced annually.

Our main finding is that the two machine-learning methods that exploit nonlinearities and interactions (gradient boosting and random forests) select long-only portfolios of funds that deliver statistically significant net alphas of 18.8 bp and 21.2 bp per month (2.26% and 2.54% per year), respectively, relative to the FF5+MOM model. In contrast, the portfolios based on the linear methods (elastic net and OLS) deliver net alphas of 8.7 bp and 8.6 bp per month (1.04% and 1.03% per year), respectively, which are statistically indistinguishable from zero. The equally weighted and asset-weighted portfolios earn negative net alphas of −2.2 bp and −4 bp per month (−0.26% and −0.48% per year), respectively. Interestingly, the asset-weighted portfolio underperforms the equally weighted portfolio, which implies that the average dollar invested in active funds earns lower risk-adjusted after-cost returns than the average fund. In summary, while portfolios that exploit predictability in the data help investors to avoid underperforming funds, only the machine-learning methods that exploit nonlinearities and interactions (gradient boosting and random forests) allow them to benefit from investing in actively managed funds. Moreover, Table 3 shows that these findings are remarkably stable when we evaluate out-of-sample alpha using the other three factor models we consider.[14]

The positive net alphas achieved by the long-only portfolios of funds selected by gradient-

[14]Furthermore, Section 6 shows that our findings are also robust to evaluating performance with respect to other factor models such as those proposed by Cremers et al. (2013), Hou et al. (2015), and Stambaugh and Yuan (2017).

boosting and random forests are also economically significant. For instance, the median of the *in-sample* alpha *spreads* between the top and bottom quintile portfolios of funds sorted by the predictors considered by Jones and Mo (2020, Table 2) is 21.91 bp per month (2.62% per year). Gradient-boosting and random forests achieve a similar net alpha *for long-only portfolios and out of sample*. Note also that the out-of-sample net alphas achieved by the portfolios of funds selected by gradient boosting and random forests are more than double the average expense ratio in our sample of active funds (1.13%). This means that if the average fund decided to cut down all fees and expenses to zero, it would only boost its net performance by less than half the size of the alpha we find for our best portfolios.

Our best method, random forests, selects a portfolio of mutual funds that earns a net alpha of 20.2 bp per month (2.4% per year) with respect to the FF3+MOM model, which is very similar to that of the best top-decile portfolio of Li and Rossi (2020, Table 4), 2.88% per year. This is somewhat surprising given that the two studies use disjoint sets of predictors: fund characteristics in our case, and stock characteristics combined with fund holdings in Li and Rossi (2020). Thus, our empirical findings complement those of Li and Rossi (2020) by showing that just like manager portfolio holdings, fund traits contain information that can be used to construct portfolios of funds with large positive alpha.[15] Moreover, our findings demonstrate that it is possible to select mutual funds with positive net alpha even in the absence of information on portfolio holdings, which is relevant for the debate on the costs and benefits of mandatory portfolio disclosure (Aliaj, 2020).

Although the alphas of gradient-boosting and random-forest portfolios are significantly different from zero, it is unclear whether they are also significantly different from that of the OLS portfolio. To answer this question, we evaluate the performance of a self-financed portfolio that goes long in the funds included in the gradient-boosting portfolio and short in the funds included in the OLS portfolio. Table 4 shows that the difference in performance between the gradient-boosting and OLS portfolios is positive and significant, ranging from 9.3 bp to 14.6 bp per month (1.1%

[15]Li and Rossi (2020, Sections 5.3 and 6.3) show that a linear combination of fund characteristics cannot improve the information contained in fund holdings and stock characteristics about future fund returns. Nonetheless, we show that using only fund characteristics with machine learning, one can construct portfolios of mutual funds with alphas similar to those obtained by exploiting fund holdings and stock characteristics.

to 1.8% per year) with respect to the four factor models we consider. A similar conclusion holds for the random-forest portfolio, which outperforms the OLS portfolio between 11.9 bp and 17.9 bp per month (1.43% and 2.1% per year), depending on the model. In contrast, the performance of the elastic-net portfolio is statistically indistinguishable from that of the OLS portfolio. Finally, both the equally weighted and asset-weighted portfolios underperform the OLS portfolio, with the difference being statistically significant.

Our main goal is to identify funds with positive alpha net of all costs. Investors may choose to invest only in those funds instead of combining them with benchmark portfolios. Therefore, it is interesting to study how the various portfolios of active funds perform in terms of mean return and risk. To answer this question, Table 5 reports the following measures for each portfolio of funds: mean excess net returns; standard deviation of net returns; Sharpe ratio (mean excess net return divided by standard deviation); Sortino ratio (mean excess net return divided by semi-deviation); maximum drawdown; and value-at-risk (VaR) based on the historical simulation method with 99% confidence. The ranking of mean excess net returns closely mirrors the ranking in alphas. This result is far from obvious because the target variable in our training algorithms is fund alpha, and not fund excess returns, unlike the studies of Wu et al. (2021) and Li and Rossi (2020). Higher mean excess net returns for the prediction-based portfolios are at least partially explained by higher standard deviation. However, our two best methods in terms of alpha, also deliver portfolios with the highest Sharpe ratio: 0.191 and 0.186 for gradient boosting and random forests, respectively, followed closely by the equally weighted portfolio (0.177). Our conclusions do not change if we consider downside risk: gradient boosting and random forests select portfolios of funds with the highest Sortino ratio. In terms of maximum drawdown, the portfolios selected by elastic net and OLS appear to be the riskiest. Finally, the equally weighted and asset-weighted portfolios are the safest in terms of VaR.

Although our measures of performance are net of all costs, it is useful to know how much trading the top-decile portfolios require. The last column of Table 5 reports the average annual turnover of the top-decile portfolios. Annual turnover is calculated at the beginning of each calendar year, when the portfolio is rebalanced, as the sum of the absolute values of changes in portfolio weights with respect to the last month of the previous year across all funds in the sample. For instance, a turnover value of one means that 50% of the wealth in the portfolio is reallocated across funds each year. As expected, the naive portfolios have very low turnover. Approximately, only 20% of the portfolio is reallocated from year to year due to changes in the pool of available funds and (for the equally weighted portfolio) also to changes in fund values. In contrast, managing a portfolio based on the performance predictions of elastic net and OLS involves trading roughly 60% of the portfolio value each year, whereas investing based on gradient boosting and random forests requires that 70% of the portfolio value be traded. These findings suggest that to achieve superior performance investing in actively managed funds, portfolio managers must also actively trade their wealth across these funds, and thus, it is important to account for fund loads when we evaluate portfolio performance.

Taken together, the results in this section suggest that it is possible to exploit readily available fund characteristics to select portfolios of mutual funds that significantly outperform (in terms of net alpha) the equally weighted or asset-weighted average mutual fund. This is true even if investors use the worst-performing forecasting methods, elastic net and OLS, to predict performance. In other words, elastic net and OLS help investors avoid underperforming funds. However, neither elastic net nor OLS allow investors to identify funds with significant positive net alpha ex-ante. Only methods that allow for nonlinearities and interactions in the relation between fund characteristics and subsequent performance, namely gradient boosting and random forests, can detect funds with large and significant alphas. Moreover, the resulting portfolios also have the highest Sharpe and Sortino ratios of all the portfolios considered.

5 Understanding the performance of ML portfolios

In this section, we study the drivers of the favorable performance of the machine-learning portfolios that capture nonlinearity and interactions. To do this, we study the importance of characteristics and interactions for gradient boosting and random forest. We then use partial-dependence plots and decision trees to study the nature of the nonlinearities and interactions. Finally, we investigate whether the favorable performance of the nonlinear machine-learning methods survives after exploiting only the most important characteristics or after excluding them.

Figure 2 reports characteristic importance for the four prediction methods: gradient-boosting (GB), random-forest (RF), elastic net (EN), and OLS. We evaluate importance for the last estimation window, which spans the 1980 to 2019 period, and re-scale it from zero to 100.[16] We highlight two findings. First, the three most important characteristics for the two nonlinear methods (gradient boosting and random forest) are R^2, market beta t-stat, and realized alpha in the previous year. The importance of the R^2 characteristic confirms that the activeness measure proposed by Amihud and Goyenko (2013) remains a key predictor even when considered jointly with other characteristics. The second most important characteristic, market beta t-stat can also be interpreted as a measure of activeness. To see this, note that Figure 1 shows that market beta t-stat has a high correlation of 54% with R^2. Also, one would expect less active funds to have highly statistically significant betas on the market. Our second finding is that nonlinear and linear methods differ sharply in characteristic importance. For instance, the nonlinear methods rely heavily on market beta t-stat and realized alpha in the previous year, whereas these two characteristics are much less important for the linear methods. The predictions of elastic net and OLS are, instead, very strongly influenced by the fund three-year precision-adjusted alpha. Finally, while linear models exploit fund expense ratios, the predictive ability of this characteristic is subsumed by other fund characteristics in nonlinear models.

The differences between nonlinear and linear methods in terms of both performance and characteristic importance suggest that there exist nonlinearities and interactions in the relation between characteristics and performance that investors can exploit to select actively managed equity funds. To explore the nature of these nonlinear relations, Figure 3 displays partial-dependence plots for the three most important characteristics (R^2, market beta t-stat, and realized

[16]To quantify the relative importance of each characteristic in the gradient-boosting and random-forest methods, we follow Gu et al. (2020) and compute the mean decrease in impurity (Breiman, 2001), with mean squared error as the impurity measure. For the elastic-net and OLS methods, we compute the importance of each characteristic as the absolute value of its slope coefficient and the absolute value of its slope t-stat, respectively. In unreported results, we have also evaluated characteristic importance for elastic net and OLS using the mean decrease in mean squared error and the findings are very similar.

alpha) associated with gradient boosting and random forest. For each partial-dependence plot, the horizontal axis shows the characteristic cross-sectionally standardized to lie between -1.0 and 1.0 and the vertical axis graphs the average prediction of a given machine-learning method conditional on the characteristic; see Friedman (2001) and Hastie et al. (2009).[17] Reassuringly, Figure 3 shows that the nonlinear patterns identified by the two machine-learning methods are very similar. Importantly, there is a substantial degree of nonlinearity in the relation between the most important fund characteristics and predicted performance. For instance, consistent with Amihud and Goyenko (2013) there is an inverse relation between R^2 and performance for low and moderate values of R^2, but the relation is roughly flat for values of standardized R^2 above 0.3. Nonlinearities are particularly apparent for market beta t-stat, with very low standardized market beta t-stats predicting superior performance and a flat relation for larger betas. Finally, there is a clear positive relation between realized alpha and future performance for values of standardized realized alpha above 0.6, but there is no apparent relation for smaller values of realized alpha.

We now turn our attention to the importance of interactions for the performance of gradient boosting and random forest. We calculate the interaction strength for all pairs of predictors using the approach of Greenwell et al. (2018) and report the strength of the top 20 interactions in Figure 4.[18] The figure reveals that activeness measures such as R^2 and market beta t-stat are not only important as standalone predictors as shown in Figure 2, but are also crucial through their interactions with various measures of *past performance*. To see this, note that one of the activeness measures appears in all but one of the top five interactions for gradient boosting and random forest. Moreover, most of the top interactions include also a measure of past performance.

[17] Specifically, let z_S be the characteristic of interest and z_C a vector containing the rest of the characteristics. Then, the partial-dependence function is estimated as

$$\bar{f}_S(z_S) = \frac{1}{L}\sum_{i=1}^{L} \hat{f}(z_S, z_{iC}),$$

where $\hat{f}(z_S, z_{iC})$ is the prediction of the machine-learning method given the values of z_S and z_{iC} and $\{z_{1C}, z_{2C}, \ldots, z_{nC}\}$ are the L observations of z_C in the sample. That is, the partial dependence function $\bar{f}_S(z_S)$ captures the marginal effect of z_S on the prediction after accounting for the *average* effect of the other predictors z_C.

[18] For every pair of predictors, Greenwell et al. (2018, Section 3.2) propose computing the standard deviation of the conditional partial-dependence function of z_j given z_k as well as that of the conditional partial-dependence function of z_k given z_j. They then average these two standard deviations to obtain a measure of the strength of the interaction between the two predictors.

For instance, three of the five most important interactions for random forest combine a measure of activeness (R^2 or market beta t-stat) with a measure of past performance (alpha t-stat or value added). This suggests that fund activeness is a good predictor of future performance, but only for certain groups of funds defined by their past performance.

To further explore this conjecture, we generate decision trees that exploit a measure of past performance and a measure of activeness to predict future alpha. For instance, Figure 5 depicts a decision tree that exploits value added and R^2 to predict future net alpha. We cross-sectionally standardize the predictors to have zero mean and unit standard deviation as in our main empirical results and consider decision trees with four levels. The decision tree shows that investors should select more active mutual funds (standardized $R^2 \leq -1.1$) conditional on their past performance being good (standardized value added ≥ 0.16). However, investors should choose less active funds (standardized $R^2 \geq -0.48$) conditional on poor past performance (standardized value added < -0.2). Thus, investors may achieve higher net alpha by holding more active funds whose returns are not explained by common risk factors, but only if the funds also have good past performance. Figures 6 and 7 depict the decision trees for other pairwise combinations of measures of activeness and past performance and provide a very similar interpretation. In particular, Figure 6 depicts a decision tree that exploits value added and market beta t-stat and shows that investors should select relatively active funds (standardized market beta t-stat ≤ -1.0), but only provided they have good past performance (standardized value added ≥ 0.16). Likewise, Figure 7 depicts a decision tree that exploits alpha t-stat and R^2 and shows that investors should select relatively active funds (standardized $R^2 \leq -0.3$), but only provided they have good past performance (standardized alpha t-stat ≥ 0.46).

To understand the impact on portfolio composition of the nonlinearities and interactions exploited by machine learning, we compute the fund overlap for the portfolios of the four prediction methods averaged over the out-of-sample period. The fund portfolios selected by the two linear methods (OLS and elastic net) are very similar, with an average 94% fund overlap. The fund portfolios selected by the two nonlinear methods (gradient boosting and random forest) are also quite similar with an average fund overlap of 58%. However, the overlap between the portfolios of the two nonlinear methods and OLS is much smaller, around 40%. This shows that while the shrinkage of elastic net has negligible impact on portfolio composition, the nonlinearities and interactions exploited by gradient boosting and random forest lead to portfolios of funds that differ substantially from the OLS portfolios.

Finally, we study whether the performance of the nonlinear methods is driven by flexibility alone or by flexibility combined with the multivariate approach. In particular, we repeat the out-of-sample evaluation using only the two, three, four, and five most important characteristics for each method and estimation window. Table 6 shows that, when only the two most important characteristics are used to predict performance, the top-decile portfolio of mutual funds selected by gradient boosting earns positive but insignificant alpha according to all factor models considered. Random forests select funds that earn positive alphas that are significant, except for the FF3+MOM model. If we also include the third most important characteristic, performance does not improve for gradient-boosting portfolios but increases marginally for random forests. It takes four regressors for the performance of gradient-boosting portfolios to become statistically significant. In contrast, both the fourth and the fifth regressors have a negligible impact for random forests. Alphas of portfolios obtained with five predictors remain below alphas of portfolios that exploit all characteristics by 3.2 bp per month in the case of gradient boosting and 7 bp per month in the case of random forests, according to the FF5+MOM model. These results suggest that flexibility is not enough to explain the performance of gradient boosting and random forests in selecting portfolios of mutual funds. The methods exploit the predictability contained in many different fund characteristics and their interactions.

Having shown that the most important predictors are not *sufficient* to generate positive alpha, we ask whether they are *necessary*. In Table 7, we repeat the prediction and fund selection exercise for all methods but remove R^2 as a predictor (Panel A) and both R^2 and market beta t-stat (Panel B). Excluding R^2 leads to a small reduction in alphas. However, if we also remove the market beta t-stat, alphas decline substantially (almost 10 bp per month). We conclude that the two most important predictors are necessary for achieving economically large alphas.

To investigate whether the predictive ability of some characteristics changes over time, Figures 8 and 9 graph the importance of each predictor in each year of the out-of-sample period for the gradient-boosting and random-forest portfolios, respectively. The figures show that some of our top characteristics such as R^2 and market beta t-stat remain important over the entire out-of-sample period, but the importance of other characteristics varies over time. Interestingly, Figures 8 and 9 exhibit some remarkable similarities, which suggests that, despite their differences, both methods appear to identify very similar patterns in the data.

These findings suggest that while flexibility enables gradient boosting and random forests to outperform OLS, flexibility alone is not sufficient to achieve the outstanding performance of the the top-decile portfolios selected with these methods. Instead, it is necessary to exploit the information from multiple characteristics. Moreover, the predictive ability of different fund characteristics varies over time. Thus, mutual-fund portfolio selection should be based on multiple fund characteristics and performed dynamically.

6 Robustness to methodological choices

We now show that our main findings are robust to: (i) considering the post-publication decay in predictability documented by McLean and Pontiff (2016); (ii) using alternative factor models to measure risk-adjusted performance; (iii) building portfolios of only *retail* mutual-fund share classes; (iv) considering alternative cut-off points to select funds; (v) rebalancing the portfolio less frequently; (vi) using neural networks; (vii) using an alternative cross-validation method; and (viii) investing in at most one share class per fund.

First, inspired by the influential work of McLean and Pontiff (2016), we check the robustness of the performance of the machine-learning portfolios to considering post-publication decay in the predictive ability of mutual-fund characteristics. To account for post-publication decay, at each point in time we train the various prediction methods using only factor models and mutual-fund characteristics proposed in papers that have already been published, as listed in Table 8. In particular, at each point in time we compute the target alpha with respect to a published factor model and we only use as predictors mutual-fund characteristics that have already been published. Then, we evaluate the out-of-sample portfolio alphas by regressing the out-of-sample excess monthly portfolio returns net of all costs against the same factor models we consider throughout the manuscript.

Table 9 reports the out-of-sample net alphas of the top-decile portfolios that account for post-publication decay. Comparing the results in Table 9 to those in Table 3, we find that although considering post-publication decay leads to a reduction in the out-of-sample net alpha of the machine-learning portfolios, the gradient-boosting and random-forests portfolios still achieve significant positive out-of-sample net alphas. For instance, accounting for post-publication decay leads to only a 1 bp decline in the monthly alpha of the random-forests portfolio, from 21.2 bp to 20.2 bp. The decline in alpha for funds selected by gradient boosting is slightly larger (about 6 bp), but the portfolio alpha is typically significant at the 10% confidence level. The monthly alpha of funds selected with linear methods (elastic net and OLS) exhibits a larger decline of around 10 bp. These results suggest that the abnormal returns of the nonlinear machine-learning portfolios are robust to considering post-publication decay.

Second, we check if our results are robust to using alternative factor models for evaluating performance. Specifically, in addition to the four models considered in Table 3, we also estimate the risk-adjusted performance of the prediction-based portfolios using the tradable factors of Cremers et al. (2013), the q-factors of Hou et al. (2015) and the mispricing factors of Stambaugh and Yuan (2017). Table 10 shows that the performance results with respect to these alternative factor models are qualitatively similar to those in Table 3. Gradient boosting and random forests yield the best results with the top-decile portfolio earning positive and statistically significant alphas for the three additional models considered. Portfolios based on forecasts by elastic net and and OLS earn positive but insignificant alphas. Equally weighted and asset-weighted portfolios earn the lowest alphas, which tend to be negative.

Third, our sample includes both institutional and retail share classes. It is unclear whether the machine-learning methods considered are simply picking institutional share classes, which

[19] We identify institutional share classes using the institutional indicator in CRSP. When that variable is missing, we use share class names.

usually charge lower costs and are subject to more stringent monitoring by investors (Evans and Fahlenbrach, 2012). To answer this question, we exclude institutional share classes from the sample and repeat the analysis.[19] Table 11 shows that the risk-adjusted performance of the portfolios of retail funds selected by gradient boosting and random forests is in all cases similar or slightly better than that reported in Table 3, where investors can select both institutional and retail share classes. This result suggests that at least part of the value added by active portfolio managers is passed on to retail investors. The fact that the performance of the top-decile portfolio improves if institutional share classes are removed from the sample could be explained by the fact that for these classes the relationship between predictors and performance differs from that for retail classes due to the different nature of competition in this segment of the market. By removing institutional classes, we may improve the accuracy of the function that maps fund characteristics into fund performance. Finally, results for the elastic net, OLS, equally weighted, and asset-weighted portfolios closely parallel those in Table 3.

Fourth, we compute the out-of-sample alpha of the portfolios of funds in the top 5% and 20% of the predicted-performance distribution, instead of the top 10% as in our base case. Table 12 shows that gradient boosting and random forests continue to select portfolios of funds with positive alphas. However, random forests do not yield statistically significant alphas for some of the performance attribution models. Such lack of significance is due to higher standard errors of alphas in the case of the top-5% portfolios and lower average alpha for the top-20% portfolios. In this sense, the 10% cut off seems to be a good compromise. Just like for the top-decile portfolios, elastic net and OLS cannot select portfolios of funds with significant alpha regardless of the threshold employed.

Fifth, we investigate the consequences of decreasing the portfolio rebalancing frequency. Specifically, we repeat the analysis for all prediction-based methods using the same target variable and predictors as in Table 3, but keeping the selected funds in the top-decile portfolio for two and three years. Table 13 displays the results. Biannual portfolio rebalancing improves the performance of portfolios with respect to those obtained with annual rebalancing for all methods and models. In particular, the monthly alpha of the portfolio selected with random forests now ranges between 31.4 bp (3.8% per year) and 40.4 bp (4.8% per year). The performance of the gradient boosting portfolio with biannual rebalancing ranges between 23.3 and 30.4 bp per month (2.8% and 3.6% per year). The performance of the elastic net and OLS portfolio also increases with a holding horizon of 24 months but remains statistically insignificant in all cases but one. However, further increasing the holding period to 36 months, hurts the performance of the resulting portfolios for all methods and models. Only random forests generate portfolios with statistically significant alpha with respect to all models.

Sixth, we investigate the performance of neural networks. Following Gu et al. (2020) and Bianchi et al. (2021), we consider fully connected feed-forward neural networks with up to three hidden layers. Like Gu et al. (2020), we consider neural networks with a single hidden layer of 32 neurons, two hidden layers with 32 and 16 neurons, respectively, and three hidden layers with 32, 16, and eight neurons, respectively.[20] All architectures are fully connected, so each neuron receives an input from all neurons in the layer below. We use the five-fold cross-validation methodology described in Section 3.4 to select the hyper parameters of the neural networks.[21]

Table 14 shows that the neural-network fund portfolios achieve positive alpha for all three architectures we consider, but their alphas are systematically lower than those obtained by the gradient-boosting and random-forest portfolios. Alphas are highest for the one-layer neural network and smallest for the three-layer network. Also, statistical significance is achieved only by portfolios selected by the 1- and 2-layer neural networks, with the exception of FF3+MOM alphas. This suggests that shallow learning is more appropriate than deep learning for the mutual-fund database. Such observation is roughly consistent with Gu et al. (2020), who find that for their stock return database, neural-network performance peaks at just three layers.

[20]Gu et al. (2020) consider feed-forward neural networks with up to five hidden layers, but we do not consider more than three layers because we find that additional layers do not help to improve performance. We have also considered neural networks with a smaller number of neurons. Specifically, we have implemented neural networks with one hidden layer of eight neurons, and two hidden layers of eight and four neurons, but their performance is worse than that of the networks with a higher number of neurons.

[21]Specifically, we employ a 5-fold cross-validation procedure to select the 1-norm and 2-norm weight regularization and the dropout ratios in the input layer and in the hidden layers. In order to avoid overfitting, we also employ early stopping such that the training process is stopped if the mean squared error does not decrease after 10 epochs. We use 50 epochs to train the networks. The activation function is the hyperbolic tangent. The network learning rate is dynamically selected using the method proposed in Zeiler (2012). Finally, we also follow Gu et al. (2020) and we use multiple random seeds to initialize neural network estimation and construct predictions by averaging forecasts from all networks.

Seventh, we study the robustness of our main findings to using *time-series cross-validation* to calibrate the hyper parameters of the machine-learning methods instead of five-fold cross validation as in our base-case analysis. At each estimation window, time-series cross validation uses the first 70% of the data to train the methods and the last 30% of data for pseudo out-of-sample evaluation, and thus, this approach accounts for the time-series properties of the mutual-fund database. Table 15 reports the out-of-sample performance of the fund portfolios obtained with three machine-learning methods (gradient boosting, random forests, and elastic net) when we use times-series cross validation. Comparing Tables 3 and 15, we find that the fund portfolios obtained with time-series cross validation perform slightly worse than those obtained with five-fold cross validation. More importantly, Table 15 shows that our findings are robust to using time-series cross validation: the portfolios obtained with gradient boosting and random forests attain out-of-sample and net-of-all-costs alphas that are positive and statistically significant even when calibrated using time-series cross validation.

Finally, in our base-case results, we allow investors to hold multiple share classes of each fund. Consequently, our equally weighted portfolios of funds could potentially assign a large weight to funds with multiple share classes in the top decile of predicted alpha. Table 16 reports the out-of-sample performance of the top-decile fund portfolios containing only one share class per fund. In particular, when a fund in the top-decile portfolio has more than one share class, we include only the class with highest TNA. The table demonstrates that restricting the portfolio to hold only one share class per fund does not hurt the performance of the nonlinear machine-learning methods.

7 Performance over time and across market conditions

Jones and Mo (2020) show that the ability of fund characteristics to predict performance has declined over time due to increased arbitrage activity and mutual-fund competition. Motivated by their work, we study how the alpha of the different portfolios varies over time in our sample.

[22]We compute monthly alphas as the portfolio excess returns each month minus the product of the factor realization in that month and the portfolio betas estimated over the whole sample period using the FF5 model augmented with momentum.

To do this, we compute the cumulative alpha of the top-decile portfolio for the three prediction methods in each month of the out-of-sample period from 1991 to 2020 as well as the cumulative alpha of the equally weighted and asset-weighted portfolios.[22] Figure 10 shows the time-series of cumulative abnormal returns. The three prediction-based portfolios (gradient boosting, random forests, and OLS) outperform the two naive portfolios (equally weighted and asset weighted) over the whole 30-year period in our sample. In particular, while the gradient-boosting, random-forests, and OLS portfolios achieve cumulative net alphas of 73%, 65%, and 31%, respectively, the equally weighted and asset-weighted portfolios earn negative cumulative net alphas of −8.2% and −14.5%, respectively. Consistent with Jones and Mo (2020), however, the performance of the prediction-based portfolios is similar to that of the naive portfolios from 2012 until 2018. Nevertheless, all three prediction-based portfolios outperform the two naive portfolios in the last two years of our sample (2019 and 2020). In particular, while the gradient-boosting, random-forests, and OLS portfolios achieve cumulative (2019–2020) net alphas of 2.7%, 3.5%, and −0.1%, respectively, the equally weighted and asset-weighted portfolios earn negative cumulative net alphas of −2.9% and −3.8%, respectively.

Li and Rossi (2020) study whether the ability of *mutual-fund holdings and stock characteristics* to predict fund performance varies across market conditions. Inspired by their work, we now investigate whether the ability of *fund characteristics* to select funds with positive alpha changes across market conditions. Like Li and Rossi (2020), we condition estimates of performance on expansions and recessions, as well as on high and low investor sentiment. More specifically, we regress the out-of-sample monthly excess returns of the top decile portfolios selected by gradient boosting and random forests on the Fama and French (2015) five factors and momentum as well as indicator variables for expansions and recessions, and high and low investor sentiment. Expansions and recessions are defined following the NBER convention and are available for our full out-of-sample period. The high (low) investor sentiment indicator equals one if investor sentiment, as defined in Baker and Wurgler (2006, 2007), is above (below) the median of the July 1965 to

December 2018 period.[23] For the regressions with sentiment, we consider a sample period from January 1991 to December 2018 because the sentiment index is not available after December 2018.

Table 17 reports estimated alphas for different market conditions and their standard errors with Newey-West adjustment for 12 lags. We also report differences in alphas across market conditions. Our main finding is that the gradient-boosting and random-forest portfolios achieve positive alphas across all market conditions, and although they perform better in recessions and times of high investor sentiment, the differences in alpha across different market conditions are not statistically significant. The random-forest portfolio, in particular, attains positive and *statistically significant* alpha across all market conditions. It performs better in recessions (46.4 bp per month) than in expansions (18.7 bp per month) and in times of high investor sentiment (25.1 bp per month) than in low-sentiment times (19.1 bp per month). However, differences in alpha across market conditions are not statistically significant for this portfolio. The gradient-boosting portfolio also attains positive alpha across market conditions, but its alpha is significant only for periods of expansion and high sentiment. Consistent with the findings of Li and Rossi (2020), just like for the random-forest portfolio, the differences in gradient-boosting alphas across market conditions are not significant.

8 Conclusions

The question of whether mutual-fund investors can earn positive net alpha by investing in active mutual funds has received much attention from academics, practitioners, and regulators. We posit that the pessimistic results that dominate the literature could be a consequence of the methods employed to exploit predictability in fund performance. In particular, we show that machine-learning methods can identify and exploit nonlinearities and interactions in the relation between fund characteristics and performance and help investors to select funds that earn significant and positive alphas net of fees and transaction costs. In contrast, linear forecasting models can

[23]More specifically, we download from Jeffrey Wurgler's website the version of investor sentiment based on the first principal component of five sentiment proxies, where each of the proxies has first been orthogonalized with respect to a set of six macroeconomic indicators.

help investors only to avoid negative alphas. The machine-learning methods reveal that the interactions between measures of fund activeness and past performance are particularly helpful to predict future fund performance. Our results demonstrate that investors can benefit from actively managed mutual funds, but only if they have access to sophisticated predictions that allow flexibility in the relation between fund characteristics and performance. Our work also shows that fund characteristics that do not require information on fund holdings are sufficient to predict fund outperformance, which has implications for the debate on the costs and benefits of portfolio disclosure.

References

Aliaj, O. (2020). Most hedge funds to be allowed to keep equity holdings secret. *Financial Times*, July 11. https://www.ft.com/content/c68ca89c-3f9b-45f9-8205-6dbea70ed859.

Amihud, Y. and R. Goyenko (2013). Mutual fund's R^2 as predictor of performance. *Review of Financial Studies 26*(3), 667–694.

Aragon, G. O., M. Hertzel, and Z. Shi (2013). Why do hedge funds avoid disclosure? Evidence from confidential 13F filings. *Journal of Financial and Quantitative Analysis 48*(5), 1499–1518.

Avramov, D. and R. Wermers (2006). Investing in mutual funds when returns are predictable. *Journal of Financial Economics 81*(2), 339–377.

Baker, M. and J. Wurgler (2006). Investor sentiment and the cross-section of stock returns. *Journal of Finance 61*(4), 1645–1680.

Baker, M. and J. Wurgler (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives 21*(2), 129–152.

Baks, K. P., A. Metrick, and J. Wachter (2001). Should investors avoid all actively managed mutual funds? A study in Bayesian performance evaluation. *Journal of Finance 56*(1), 45–85.

Banegas, A., B. Gillen, A. Timmermann, and R. Wermers (2013). The cross section of conditional mutual fund performance in european stock markets. *Journal of Financial Economics 108*(3), 699–726.

Barber, B. M., X. Huang, and T. Odean (2016). Which factors matter to investors? Evidence from mutual fund flows. *Review of Financial Studies 29*(10), 2600–2642.

Barras, L., O. Scaillet, and R. Wermers (2010). False discoveries in mutual fund performance: Measuring luck in estimated alphas. *Journal of Finance 65*(1), 179–216.

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.

Berk, J. and R. Green (2004). Mutual fund flows and performance in rational markets. *Journal of Political Economy 112*(6), 1269–1295.

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

Berk, J. B. and J. H. Van Binsbergen (2016). Assessing asset pricing models using revealed preference. *Journal of Financial Economics 119*(1), 1–23.

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

Bollen, N. P. and J. A. Busse (2005). Short-term persistence in mutual fund performance. *Review of Financial Studies 18*(2), 569–597.

Breiman, L. (2001). Random forests. *Machine Learning 45*(1), 5–32.

Bryzgalova, S., M. Pelger, and J. Zhu (2019). Forest through the trees: Building cross-sections of stock returns. *Available at SSRN 3493458*.

Busse, J. A. and P. J. Irvine (2006). Bayesian alphas and mutual fund persistence. *Journal of Finance 61*(5), 2251–2288.

Butaru, F., Q. Chen, B. Clark, S. Das, A. W. Lo, and A. Siddique (2016). Risk and risk management in the credit card industry. *Journal of Banking & Finance 72*, 218–239.

Carhart, M. M. (1997). On persistence in mutual fund performance. *Journal of Finance 52*(1), 57–82.

Chan, L. K., H.-L. Chen, and J. Lakonishok (2002). On mutual fund investment styles. *Review of Financial Studies 15*(5), 1407–1437.

Chen, J., H. Hong, M. Huang, and J. D. Kubik (2004). Does fund size erode mutual fund performance? the role of liquidity and organization. *American Economic Review 94*(5), 1276–1302.

Chen, L., M. Pelger, and J. Zhu (2020). Deep learning in asset pricing. *Forthcoming in Managment Science*.

Chen, T., T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, and Y. Li (2020). *xgboost: Extreme Gradient Boosting*. R package version 1.2.0.1.

Chiang, W.-C., T. L. Urban, and G. W. Baldridge (1996). A neural network approach to mutual fund net asset value forecasting. *Omega 24*(2), 205–215.

Coulombe, P. G., M. Leroux, D. Stevanovic, and S. Surprenant (2020). How is machine learning useful for macroeconomic forecasting? *Available at arXiv: https://arxiv.org/abs/2008.12477*.

Cremers, K. M. and A. Petajisto (2009). How active is your fund manager? A new measure that predicts performance. *Review of Financial Studies 22*(9), 3329–3365.

Cremers, M., A. Petajisto, and E. Zitzewitz (2013). Should benchmark indices have alpha? Revisiting performance evaluation. *Critical Finance Review 2*(1), 001–048.

Dumitrescu, A. and J. Gil-Bazo (2018). Market frictions, investor sophistication, and persistence in mutual fund performance. *Journal of Financial Markets 40*, 40–59.

Elliott, G., A. Gargano, and A. Timmermann (2013). Complete subset regressions. *Journal of Econometrics 177*(2), 357–373.

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*(6), 2415–2430.

Elton, E. J., M. J. Gruber, and C. R. Blake (2011). Holdings data, security returns, and the selection of superior mutual funds. *Journal of Financial and Quantitative Analysis 46*(2), 341–367.

Evans, R. B. (2010). Mutual fund incubation. *Journal of Finance 65*(4), 1581–1611.

Evans, R. B. and R. Fahlenbrach (2012). Institutional investors and mutual fund governance: Evidence from retail–institutional fund twins. *Review of Financial Studies 25*(12), 3530–3571.

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*(1), 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*(1), 3–56.

Fama, E. F. and K. R. French (2010). Luck versus skill in the cross-section of mutual fund returns. *Journal of Finance 65*(5), 1915–1947.

Fama, E. F. and K. R. French (2015). A five-factor asset pricing model. *Journal of Financial Economics 116*(1), 1–22.

Feng, G., N. G. Polson, and J. Xu (2020). Deep learning in characteristics-sorted factor models. *Available at SSRN 3243683*.

Ferreira, M. A., A. Keswani, A. F. Miguel, and S. B. Ramos (2013). The determinants of mutual fund performance: A cross-country study. *Review of Finance 17*(2), 483–525.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1–22.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics 29*(5), 1189–1232.

Garcia, M. G., M. C. Medeiros, and G. F. Vasconcelos (2017). Real-time inflation forecasting with high-dimensional models: The case of Brazil. *International Journal of Forecasting 33*(3), 679–693.

Gittelsohn, J. (2019). End of era: Passive equity funds surpass active in epic shift. *Bloomberg*, September 11. https://www.bloomberg.com/news/articles/2019-09-11/passive-u-s-equity-funds-eclipse-active-in-epic-industry-shift.

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

Greenwell, B. M., B. C. Boehmke, and A. J. McCarthy (2018). A simple and effective model-based variable importance measure. *Available at arXiv: https://arxiv.org/abs/1805.04755*.

Gruber, M. J. (1996). Another puzzle: The growth in actively managed mutual funds. *Journal of Finance 51*(3), 783–810.

Gu, S., B. Kelly, and D. Xiu (2020). Empirical asset pricing via machine learning. *Review of*

Financial Studies 33(5), 2223–2273.

Gupta-Mukherjee, S. (2014). Investing in the "new economy": Mutual fund performance and the nature of the firm. *Journal of Financial and Quantitative Analysis 49*(1), 165–191.

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

Hou, K., C. Xue, and L. Zhang (2015). Digesting anomalies: An investment approach. *Review of Financial Studies 28*(3), 650–705.

Hunter, D., E. Kandel, S. Kandel, and R. Wermers (2014). Mutual fund performance evaluation with active peer benchmarks. *Journal of Financial Economics 112*(1), 1–29.

Indro, D. C., C. Jiang, B. Patuwo, and G. Zhang (1999). Predicting mutual fund performance using artificial neural networks. *Omega 27*(3), 373–380.

Jensen, M. C. (1968). The performance of mutual funds in the period 1945–1964. *Journal of Finance 23*(2), 389–416.

Jones, C. S. and H. Mo (2020). Out-of-sample performance of mutual fund predictors. *Review of Financial Studies 34*(1), 149—193.

Jones, C. S. and J. Shanken (2005). Mutual fund performance with learning across funds. *Journal of Financial Economics 78*(3), 507–552.

Kacperczyk, M., S. V. Nieuwerburgh, and L. Veldkamp (2014). Time-varying fund manager skill. *Journal of Finance 69*(4), 1455–1484.

Kaniel, R., Z. Lin, M. Pelger, and S. Van Nieuwerburgh (2021). Machine-learning the skill of mutual fund managers. *Available at SSRN 3977883*.

LeDell, E., N. Gill, S. Aiello, A. Fu, A. Candel, C. Click, T. Kraljevic, T. Nykodym, P. Aboyoun, M. Kurka, and M. Malohlava (2020). *H2O: R Interface for the 'H2O' Scalable Machine Learning Platform*. R package version 3.30.1.3.

Li, B. and A. G. Rossi (2020). Selecting mutual funds from the stocks they hold: A machine learning approach. *Available at SSRN 3737667*.

Liaw, A. and M. Wiener (2002). Classification and regression by randomforest. *R News 2*(3), 18–22.

Mamaysky, H., M. Spiegel, and H. Zhang (2008). Estimating the dynamics of mutual fund alphas and betas. *Review of Financial Studies 21*(1), 233–264.

Masini, R. P., M. C. Medeiros, and E. F. Mendes (2021). Machine learning advances for time series forecasting. *Available at ArXiv: https://arxiv.org/abs/2012.12802*.

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

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*(1), 1–22.

Mehta, D., D. Desai, and J. Pradeep (2020). Machine learning fund categorizations. *Available at ArXiv: https://arxiv.org/abs/2006.00123*.

Moreno, D., P. Marco, and I. Olmeda (2006). Self-organizing maps could improve the classification of spanish mutual funds. *European Journal of Operational Research 174*(2), 1039–1054.

Pástor, L. and R. F. Stambaugh (2002). Investing in equity mutual funds. *Journal of Financial Economics 63*(3), 351–380.

Pástor, L. and R. F. Stambaugh (2003). Liquidity risk and expected stock returns. *Journal of Political Economy 111*(3), 642–685.

Pattarin, F., S. Paterlini, and T. Minerva (2004). Clustering financial time series: An application to mutual funds style analysis. *Computational Statistics & Data Analysis 47*(2), 353–372.

Rakowski, D. (2010). Fund flow volatility and performance. *Journal of Financial and Quantitative Analysis 45*(1), 223–237.

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

Rossi, A. G. and S. P. Utkus (2020). Who benefits from robo-advising? Evidence from machine learning. *Available at SSRN 3552671*.

Roussanov, N., H. Ruan, and Y. Wei (2021). Marketing mutual funds. *The Review of Financial Studies 34*(6), 3045–3094.

Schapire, R. E. and Y. Freund (2012). *Boosting: Foundations and Algorithms*. MIT Press.

Sharpe, W. F. (1966). Mutual fund performance. *Journal of Business 39*(1), 119–138.

Shi, Z. (2017). The impact of portfolio disclosure on hedge fund performance. *Journal of Financial Economics 126*(1), 36–53.

Stambaugh, R. F. and Y. Yuan (2017). Mispricing factors. *Review of Financial Studies 30*(4), 1270–1315.

Wermers, R. (2000). Mutual fund performance: An empirical decomposition into stock-picking talent, style, transactions costs, and expenses. *Journal of Finance 55*(4), 1655–1695.

Wu, W., J. Chen, Z. Yang, and M. L. Tindall (2021). A cross-sectional machine learning approach for hedge fund return prediction and selection. *Management Science 67*(7), 4577–4601.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *Available at ArXiv: htpps://arxiv.org/abs/1212.5701*.

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*(2), 301–320.

Table 1: **Share-class characteristics: Definitions**

This table lists the 17 monthly mutual-fund share-class characteristics that we consider. The first column gives the name of each characteristic and the second column provides its definition.

Variable	**Definition**
realized alpha	Monthly realized alpha calculated using Equation (2)
flows	Monthly flows calculated using Equation (1)
value added	Dollar value extracted by the fund's manager from asset market calculated using Equation (3)
volatility of flows	Standard deviation of monthly flows in calendar year
total net assets (TNA)	Total assets minus total liabilities at end of month
expense ratio	Annual expenses as percentage of assets under management
age (months)	Number of months since share-class's inception date
manager tenure (years)	Number of years since beginning of manager's mandate
turnover ratio	Minimum of annual aggregate sales and annual aggregate purchases divided by total net assets
alpha t-stat	Alpha t-stat from rolling-window regression on FF5+MOM factors for previous 36 months
market beta t-stat	Market beta t-stat from rolling-window regression on FF5+MOM factors for previous 36 months
profitability beta t-stat	Profitability beta t-stat from rolling-window regression on FF5+MOM factors for previous 36 months
investment beta t-stat	Investment beta t-stat from rolling-window regression on FF5+MOM factors for previous 36 months
size beta t-stat	Size beta t-stat from rolling-window regression on FF5+MOM factors for previous 36 months
value beta t-stat	Value beta t-stat from rolling-window regression on FF5+MOM factors for previous 36 months
momentum beta t-stat	Momentum beta t-stat from rolling-window regression on FF5+MOM factors for previous 36 months
R^2	R-squared from rolling-window regression on FF5+MOM factors for previous 36 months

Table 2: **Share-class characteristics: Descriptive statistics**

This table reports monthly descriptive statistics (mean, median, standard deviation, and number of class-month observations) for the mutual-fund share-class characteristics we consider. All variables are measured at the fund share-class level and correspond to US domestic equity funds in the 1980-2020 period.

	Mean	Median	Standard deviation	Class-month observations
monthly return	0.86%	1.25%	5.27%	679,025
monthly realized alpha	-0.14%	-0.14%	2.25%	637,614
alpha t-stat	-0.455	-0.453	1.218	637,960
TNA (USD mill.)	647.3	85.3	2587.6	679,569
expense ratio	1.13%	1.05%	0.65%	672,787
age (months)	145.9	116.0	112.1	679,569
flows	0.004	-0.004	0.042	676,175
manager tenure (years)	8.179	6.921	5.370	625,426
turnover ratio	0.827	0.570	1.235	671,892
volatility of flows	0.055	0.028	0.078	676,175
value added	-0.312	-0.014	11.449	582,328
market beta t-stat	16.492	14.861	10.656	637,960
profitability beta t-stat	-0.120	-0.119	1.466	637,960
investment beta t-stat	-0.449	-0.492	1.535	637,960
size beta t-stat	1.556	0.687	3.835	637,960
value beta t-stat	0.027	-0.075	2.201	637,960
momentum beta t-stat	0.030	0.045	1.888	637,960
R^2	0.906	0.943	0.122	637,960

Table 3: **Out-of-sample alpha of fund portfolios**

This table reports the monthly out-of-sample alphas (in %) net of all costs of the top-decile fund portfolios obtained with three machine-learning methods (gradient boosting, random forests, and elastic net), with Ordinary Least Squares (OLS), and with two naive strategies (equally weighted and asset-weighted portfolios of all available funds). Alphas are computed by regressing the out-of-sample excess monthly portfolio returns net of all costs against the Fama and French (1993) three-factor model augmented with momentum (FF3+MOM), the Fama and French (2015) five factors (FF5), and the FF5 model augmented with momentum (FF5+MOM) and with the liquidity risk factor of Pástor and Stambaugh (2003) (FF5+MOM+LIQ). The out-of-sample period spans from January 1991 to December 2020. We report standard errors with Newey-West adjustment for 12 lags in parentheses. One, two, and three asterisks indicate that the alpha is significant at the 10%, 5%, and 1% level, respectively.

	FF3+MOM	FF5	FF5+MOM	FF5+MOM + LIQ
Gradient boosting	0.174**	0.216***	0.188***	0.187**
	(0.068)	(0.076)	(0.071)	(0.073)
Random forests	0.202**	0.248***	0.212**	0.213**
	(0.082)	(0.092)	(0.084)	(0.085)
Elastic net	0.041	0.069	0.087	0.095
	(0.064)	(0.066)	(0.068)	(0.068)
OLS	0.039	0.069	0.086	0.095
	(0.063)	(0.065)	(0.067)	(0.067)
Equally weighted	-0.023	-0.011	-0.022	-0.021
	(0.045)	(0.045)	(0.044)	(0.045)
Asset weighted	-0.046	-0.035	-0.040	-0.039
	(0.037)	(0.036)	(0.036)	(0.036)

Table 4: **Out-of-sample alpha with respect to OLS**

This table reports the monthly out-of-sample alphas (in %) net of all costs of the portfolio that goes long in the funds selected by one of the methods we consider (gradient boosting, random forests, elastic net, equally weighted, asset weighted) and short in the funds selected by OLS. For instance, "gradient boosting minus OLS" refers to a long-short portfolio that is long on the prediction-based top-decile portfolio obtained with the gradient-boosting method and short on the top-decile portfolio obtained with the OLS method. Alphas are computed by regressing the out-of-sample excess monthly long-short portfolio returns net of all costs against the Fama and French (1993) three-factor model augmented with momentum (FF3+MOM), the Fama and French (2015) five factors (FF5), and the FF5 model augmented with the momentum factor (FF5+MOM) and with the liquidity risk factor of Pástor and Stambaugh (2003) (FF5+MOM+LIQ). The out-of-sample period spans from January 1991 to December 2020. We report standard errors with Newey-West adjustment for 12 lags in parentheses. One, two, and three asterisks indicate that the alpha is significant at the 10%, 5%, and 1% level, respectively.

	FF3+MOM	FF5	FF5+MOM	FF5+MOM + LIQ
Gradient boosting minus OLS	0.135***	0.146***	0.102**	0.093**
	(0.042)	(0.050)	(0.041)	(0.042)
Random forests minus OLS	0.162***	0.179***	0.126**	0.119**
	(0.052)	(0.063)	(0.049)	(0.050)
Elastic net minus OLS	0.001	-0.001	0.001	0.000
	(0.011)	(0.011)	(0.011)	(0.011)
Equally weighted minus OLS	-0.062	-0.080	-0.108**	-0.115**
	(0.047)	(0.051)	(0.048)	(0.048)
Asset weighted minus OLS	-0.086*	-0.105**	-0.126**	-0.133***
	(0.047)	(0.052)	(0.050)	(0.049)

Table 5: **Out-of-sample mean excess return and risk**

For each fund portfolio, this table reports the following monthly out-of-sample performance metrics: mean excess returns net of all costs; standard deviation; Sharpe ratio (mean excess return divided by the standard deviation); Sortino ratio (mean excess return divided by the semi-deviation); maximum drawdown; and value-at-risk (VaR) based on the historical simulation method with 99% confidence. The last column reports the average annual portfolio turnover.

	Mean	Standard deviation	Sharpe ratio	Sortino ratio	Maximum drawdown	VaR 99%	Turnover
Gradient boosting	0.90%	4.69%	0.191	0.292	50.2%	12.6%	1.483
Random forests	0.92%	4.94%	0.186	0.285	55.7%	13.5%	1.435
Elastic net	0.80%	4.82%	0.167	0.247	58.5%	12.2%	1.201
OLS	0.81%	4.82%	0.167	0.249	58.3%	12.2%	1.227
Equally weighted	0.78%	4.40%	0.177	0.262	51.7%	10.2%	0.404
Asset weighted	0.73%	4.44%	0.165	0.242	53.1%	10.7%	0.368

Table 6: **Out-of-sample alpha of fund portfolios using only the most important characteristics**

This table reports the monthly out-of-sample alphas (in %) net of all costs of the top-decile fund portfolios obtained with the gradient-boosting and random forest methods exploiting only the top-2, top-3, top-4, and top-5 characteristics in terms of variable importance for each method and year. We also report the results obtained when all characteristics are included. Alphas are computed by regressing the out-of-sample excess monthly portfolio returns net of all costs against the Fama and French (1993) three-factor model augmented with momentum (FF3+MOM), the Fama and French (2015) five factors (FF5), and the FF5 model augmented with momentum (FF5+MOM) and with the liquidity risk factor of Pástor and Stambaugh (2003) (FF5+MOM+LIQ). The out-of-sample period spans from January 1991 to December 2020. We report standard errors with Newey-West adjustment for 12 lags in parentheses. One, two, and three asterisks indicate that the alpha is significant at the 10%, 5%, and 1% level, respectively.

	Gradient boosting				**Random forests**			
	FF3+MOM	FF5	FF5+MOM	FF5+MOM +LIQ	FF3+MOM	FF5	FF5+MOM	FF5+MOM +LIQ
top-2 regressors	0.091	0.124	0.118	0.124	0.105	0.152**	0.128*	0.133*
	(0.101)	(0.108)	(0.105)	(0.105)	(0.067)	(0.075)	(0.069)	(0.068)
top-3 regressors	0.081	0.129	0.103	0.105	0.126	0.187*	0.140*	0.143*
	(0.086)	(0.089)	(0.086)	(0.086)	(0.083)	(0.100)	(0.084)	(0.084)
top-4 regressors	0.137*	0.207**	0.166**	0.170**	0.131*	0.178**	0.130	0.133*
	(0.076)	(0.089)	(0.078)	(0.078)	(0.077)	(0.088)	(0.080)	(0.080)
top-5 regressors	0.128*	0.198**	0.156**	0.158**	0.128	0.193**	0.142*	0.146*
	(0.075)	(0.090)	(0.077)	(0.079)	(0.079)	(0.091)	(0.081)	(0.081)
all regressors	0.174**	0.216***	0.188***	0.187**	0.202**	0.248***	0.212**	0.213**
	(0.068)	(0.076)	(0.071)	(0.073)	(0.082)	(0.092)	(0.084)	(0.085)

Table 7: **Out-of-sample alpha of fund portfolios after removing the most important characteristics**

This table reports the monthly out-of-sample alphas (in %) net of all costs of the top-decile fund portfolios obtained after removing the most important characteristics (R^2) and the two most important characteristics (R^2 and market beta t-stat) from the set of predictors. Alphas are computed by regressing the out-of-sample excess monthly portfolio returns net of all costs against the Fama and French (1993) three-factor model augmented with momentum (FF3+MOM), the Fama and French (2015) five factors (FF5), and the FF5 model augmented with momentum (FF5+MOM) and with the liquidity risk factor of Pástor and Stambaugh (2003) (FF5+MOM+LIQ). The out-of-sample period spans from January 1991 to December 2020. We report standard errors with Newey-West adjustment for 12 lags in parentheses. One, two, and three asterisks indicate that the alpha is significant at the 10%, 5%, and 1% level, respectively.

	A: Without the most important predictor				B: Without the two most important predictors			
	FF3+MOM	FF5	FF5+MOM	FF5+MOM + LIQ	FF3+MOM	FF5	FF5+MOM	FF5+MOM + LIQ
Gradient boosting	0.129**	0.178**	0.165**	0.169**	0.081	0.096	0.102*	0.102*
	(0.066)	(0.073)	(0.071)	(0.071)	(0.055)	(0.060)	(0.060)	(0.061)
Random forests	0.187**	0.235***	0.206***	0.208***	0.097	0.119*	0.121*	0.123*
	(0.076)	(0.081)	(0.079)	(0.079)	(0.062)	(0.066)	(0.065)	(0.066)
Elastic net	0.028	0.058	0.067	0.073	0.018	0.047	0.055	0.061
	(0.061)	(0.063)	(0.066)	(0.067)	(0.061)	(0.063)	(0.066)	(0.067)
OLS	0.018	0.048	0.057	0.064	0.012	0.040	0.048	0.054
	(0.060)	(0.061)	(0.065)	(0.065)	(0.060)	(0.061)	(0.065)	(0.066)

Table 8: **Models and excluded predictors for post-publication decay analysis**

For each subperiod in our sample, this table lists the factor model used to compute the target alpha as well as the predicting variables excluded in our post-publication decay analysis. The first column lists the subperiod, the second column lists the factor model used to evaluate the target alpha, and the third column lists the predicting variables excluded in each subperiod.

Subperiod	Factor model	Excluded predictors
1980–1993	CAPM	value, size, momentum, investment, and profitability betas; value added; R^2
1994–1997	FF3	momentum, investment, and profitability betas; value added; R^2
1998–2013	FF3+MOM	investment and profitability betas; value added; R^2
2014–2015	FF3+MOM	investment and profitability betas; value added
2016–2020	FF5+MOM	none

Table 9: **Out-of-sample alpha of portfolios considering post-publication decay**

This table reports the monthly out-of-sample alphas (in %) net of all costs of the top-decile fund portfolios selected accounting for post-publication decay and using three machine-learning methods (gradient boosting, random forests, and elastic net), and Ordinary Least Squares (OLS). To account for post-publication decay, at each point in time we train the various prediction methods using only factor models and mutual-fund characteristics proposed in papers that have already been published, as listed in Table 8. In particular, at each point in time we compute the target alpha with respect to a published factor model and we only use as predictors mutual-fund characteristics that have already been published. Then, we evaluate the out-of-sample portfolio alphas by regressing the out-of-sample excess monthly portfolio returns net of all costs against the Fama and French (1993) three-factor model augmented with momentum (FF3+MOM), the Fama and French (2015) five factors (FF5), and the FF5 model augmented with momentum (FF5+MOM) and with the liquidity risk factor of Pástor and Stambaugh (2003) (FF5+MOM+LIQ). The out-of-sample period spans from January 1991 to December 2020. We report standard errors with Newey-West adjustment for 12 lags in parentheses. One, two, and three asterisks indicate that the alpha is significant at the 10%, 5%, and 1% level, respectively.

	FF3+MOM	FF5	FF5+MOM	FF5+MOM +LIQ
Gradient boosting	0.111	0.148**	0.127*	0.131*
	(0.071)	(0.072)	(0.068)	(0.067)
Random forests	0.187**	0.226***	0.202***	0.208***
	(0.081)	(0.084)	(0.077)	(0.077)
Elastic net	-0.030	-0.014	-0.012	-0.007
	(0.073)	(0.069)	(0.069)	(0.068)
OLS	-0.029	-0.021	-0.014	-0.009
	(0.074)	(0.070)	(0.070)	(0.069)

Table 10: **Out-of-sample alpha of fund portfolios based on alternative factor models**

This table reports the monthly out-of-sample alphas (in %) net of all costs of the top-decile fund portfolios obtained with three machine-learning methods (gradient boosting, random forests, and elastic net), with Ordinary Least Squares (OLS), and with two naive strategies (equally weighted and asset-weighted portfolios of all available funds). Alphas are computed by regressing the out-of-sample excess monthly portfolio returns net of all costs against the Cremers et al. (2013), Hou et al. (2015), and Stambaugh and Yuan (2017) factor models. The sample period of each regression varies depending on the available sample of factors returns. Cremers et al. (2013) monthly tradable factors were downloaded from the web page of Antti Petajisto and span the January 1991 to January 2014 period (277 months). Hou et al. (2015) monthly q-factors were downloaded from the data library at www.global-q.org and span the January 1991 to December 2020 period (360 months). Stambaugh and Yuan (2017) monthly mispricing factors were downloaded from the webpage of Robert Stambaugh and span the January 1991 to December 2016 period (312 months). We report standard errors with Newey-West adjustment for 12 lags in parentheses. One, two, and three asterisks indicate that the alpha is significant at the 10%, 5%, and 1% level, respectively.

	Cremers et al. factors	Hou et al. factors	Stambaugh and Yuan factors
Gradient boosting	0.161**	0.222**	0.168**
	(0.063)	(0.090)	(0.081)
Random forests	0.178**	0.254**	0.172*
	(0.074)	(0.109)	(0.093)
Elastic net	0.048	0.076	0.097
	(0.071)	(0.077)	(0.074)
OLS	0.050	0.076	0.100
	(0.070)	(0.077)	(0.073)
Equally weighted	0.020	-0.020	-0.017
	(0.038)	(0.039)	(0.048)
Asset weighted	-0.052**	-0.051	-0.026
	(0.026)	(0.034)	(0.038)

Table 11: **Out-of-sample alpha of retail share-class portfolios**

This table reports the monthly out-of-sample alphas (in %) net of all costs of the top-decile fund portfolios after excluding from our sample institutional share classes. Portfolios are obtained with three machine-learning methods (gradient boosting, random forests, and elastic net), with Ordinary Least Squares (OLS), and with two naive strategies (equally weighted and asset-weighted portfolios of all available funds). Alphas are computed by regressing the out-of-sample excess monthly portfolio returns net of all costs against the Fama and French (1993) three-factor model augmented with momentum (FF3+MOM), the Fama and French (2015) five factors (FF5), and the FF5 model augmented with momentum (FF5+MOM) and with the liquidity risk factor of Pástor and Stambaugh (2003) (FF5+MOM+LIQ). The out-of-sample period spans from January 1991 to December 2020. We report standard errors with Newey-West adjustment for 12 lags in parentheses. One, two, and three asterisks indicate that the alpha is significant at the 10%, 5%, and 1% level, respectively.

	FF3+MOM	FF5	FF5+MOM	FF5+MOM +LIQ
Gradient boosting	0.174**	0.225**	0.210**	0.209**
	(0.084)	(0.092)	(0.090)	(0.091)
Random forests	0.240***	0.277***	0.244***	0.244***
	(0.092)	(0.097)	(0.093)	(0.094)
Elastic net	0.013	0.044	0.054	0.059
	(0.064)	(0.064)	(0.066)	(0.066)
OLS	0.013	0.046	0.055	0.060
	(0.063)	(0.063)	(0.065)	(0.065)
Equally weighted	-0.012	0.000	-0.011	-0.011
	(0.047)	(0.047)	(0.046)	(0.047)
Asset weighted	-0.037	-0.025	-0.029	-0.028
	(0.038)	(0.037)	(0.037)	(0.038)

Table 12: **Out-of-sample alpha of top-5% and top-20% fund portfolios**

This table reports the monthly out-of-sample alphas (in %) net of all costs of the top-5% and top-20% fund portfolios obtained with three machine-learning methods (gradient boosting, random forests, and elastic net) and with Ordinary Least Squares (OLS). Alphas are computed by regressing the out-of-sample excess monthly portfolio returns net of all costs against the Fama and French (1993) three-factor model augmented with momentum (FF3+MOM), the Fama and French (2015) five factors (FF5), and the FF5 model augmented with momentum (FF5+MOM) and with the liquidity risk factor of Pástor and Stambaugh (2003) (FF5+MOM+LIQ). The out-of-sample period spans from January 1991 to December 2020. We report standard errors with Newey-West adjustment for 12 lags in parentheses. One, two, and three asterisks indicate that the alpha is significant at the 10%, 5%, and 1% level, respectively.

	Top-5% fund portfolios				**Top-20% fund portfolios**			
	FF3+MOM	FF5	FF5+MOM	FF5+MOM +LIQ	FF3+MOM	FF5	FF5+MOM	FF5+MOM +LIQ
Gradient boosting	0.178*	0.219**	0.190*	0.187*	0.115**	0.150**	0.129**	0.131**
	(0.096)	(0.105)	(0.100)	(0.102)	(0.058)	(0.064)	(0.061)	(0.061)
Random forests	0.194	0.250*	0.205*	0.205	0.106	0.136*	0.118*	0.120*
	(0.119)	(0.127)	(0.122)	(0.124)	(0.065)	(0.071)	(0.067)	(0.068)
Elastic net	0.058	0.106	0.124	0.133	0.031	0.049	0.061	0.067
	(0.084)	(0.089)	(0.089)	(0.089)	(0.056)	(0.057)	(0.060)	(0.060)
OLS	0.050	0.097	0.114	0.122	0.028	0.046	0.057	0.064
	(0.084)	(0.088)	(0.088)	(0.088)	(0.056)	(0.056)	(0.059)	(0.059)

Table 13: **Out-of-sample alpha of fund portfolios with 24-month and 36-month rebalancing frequencies**

This table reports the monthly out-of-sample alphas (in %) net of all costs of the top-decile fund portfolios with rebalancing frequencies of 24 months and 36 months. Portfolios are obtained with three machine-learning methods (gradient boosting, random forests, and elastic net) and with Ordinary Least Squares (OLS). Alphas are computed by regressing the out-of-sample excess monthly portfolio returns net of all costs against the Fama and French (1993) three-factor model augmented with momentum (FF3+MOM), the Fama and French (2015) five factors (FF5), and the FF5 model augmented with momentum (FF5+MOM) and with the liquidity risk factor of Pástor and Stambaugh (2003) (FF5+MOM+LIQ). The out-of-sample period spans from January 1991 to December 2020. We report standard errors with Newey-West adjustment for 12 lags in parentheses. One, two, and three asterisks indicate that the alpha is significant at the 10%, 5%, and 1% level, respectively.

	24-month rebalancing				**36-month rebalancing**			
	FF3+MOM	FF5	FF5+MOM	FF5+MOM +LIQ	FF3+MOM	FF5	FF5+MOM	FF5+MOM +LIQ
Gradient boosting	0.233***	0.304***	0.254***	0.255***	0.094	0.108*	0.094	0.095
	(0.077)	(0.096)	(0.085)	(0.084)	(0.065)	(0.065)	(0.065)	(0.065)
Random forests	0.314***	0.404***	0.342***	0.349***	0.148**	0.179**	0.148**	0.149**
	(0.092)	(0.114)	(0.099)	(0.096)	(0.072)	(0.075)	(0.072)	(0.072)
Elastic net	0.086	0.126*	0.123	0.133*	0.003	0.018	0.027	0.035
	(0.077)	(0.075)	(0.078)	(0.075)	(0.064)	(0.066)	(0.069)	(0.068)
OLS	0.080	0.125	0.120	0.130*	-0.005	0.014	0.021	0.030
	(0.078)	(0.077)	(0.080)	(0.076)	(0.065)	(0.069)	(0.071)	(0.070)

Table 14: **Out-of-sample alpha of fund portfolios obtained with neural networks**

This table reports the monthly out-of-sample alphas (in %) net of all costs of the top-decile fund portfolios obtained with feed-forward neural networks with one, two, and three hidden layers. Alphas are computed by regressing the out-of-sample excess monthly portfolio returns net of all costs against the Fama and French (1993) three-factor model augmented with momentum (FF3+MOM), the Fama and French (2015) five factors (FF5), and the FF5 model augmented with momentum (FF5+MOM) and with the liquidity risk factor of Pástor and Stambaugh (2003) (FF5+MOM+LIQ). The out-of-sample period spans from January 1991 to December 2020. We report standard errors with Newey-West adjustment for 12 lags in parentheses. One, two, and three asterisks indicate that the alpha is significant at the 10%, 5%, and 1% level, respectively.

	FF3+MOM	FF5	FF5+MOM	FF5+MOM +LIQ
1 layer (32 neurons)	0.111	0.126*	0.149**	0.154**
	(0.070)	(0.068)	(0.071)	(0.071)
2 layers (32-16 neurons)	0.097	0.109*	0.133*	0.138**
	(0.065)	(0.065)	(0.068)	(0.068)
3 layers (32-16-8 neurons)	0.069	0.086	0.099	0.105
	(0.075)	(0.076)	(0.075)	(0.076)

Table 15: **Out-of-sample alpha of fund portfolios with time-series cross validation**

This table reports the monthly out-of-sample alphas (in %) net of all costs of the top-decile fund portfolios obtained with three machine-learning methods (gradient boosting, random forests, and elastic net) when one uses the times-series cross validation method to select the corresponding hyper parameters of each method. The times-series cross validation method works as follows: at each estimation round, the first 70% of the estimation data is used as training samples and the subsequent 30% of data is used to pseudo out-of-sample evaluation. Alphas are computed by regressing the out-of-sample excess monthly portfolio returns net of all costs against the Fama and French (1993) three-factor model augmented with momentum (FF3+MOM), the Fama and French (2015) five factors (FF5), and the FF5 model augmented with momentum (FF5+MOM) and with the liquidity risk factor of Pástor and Stambaugh (2003) (FF5+MOM+LIQ). The out-of-sample period spans from January 1991 to December 2020. We report standard errors with Newey-West adjustment for 12 lags in parentheses. One, two, and three asterisks indicate that the alpha is significant at the 10%, 5%, and 1% level, respectively.

	FF3+MOM	FF5	FF5+MOM	FF5+MOM + LIQ
Gradient boosting	0.123*	0.163**	0.142**	0.142**
	(0.063)	(0.067)	(0.067)	(0.068)
Random forests	0.190**	0.238***	0.205**	0.207**
	(0.077)	(0.085)	(0.079)	(0.080)
Elastic net	0.036	0.064	0.082	0.090
	(0.065)	(0.067)	(0.068)	(0.068)

Table 16: **Out-of-sample alphas of one-share-class-per-fund portfolios**

This table reports the monthly out-of-sample alphas (in %) net of all costs of the top-decile fund portfolios containing only one share class per fund. In particular, when a fund in the top-decile portfolio has more than one share class, we include only the class with highest TNA. Portfolios are obtained with three machine-learning methods (gradient boosting, random forests, and elastic net), with Ordinary Least Squares (OLS), and with two naive strategies (equally weighted and asset-weighted portfolios of all available funds). Alphas are computed by regressing the out-of-sample excess monthly portfolio returns net of all costs against the Fama and French (1993) three-factor model augmented with momentum (FF3+MOM), the Fama and French (2015) five factors (FF5), and the FF5 model augmented with momentum (FF5+MOM) and with the liquidity risk factor of Pástor and Stambaugh (2003) (FF5+MOM+LIQ). The out-of-sample period spans from January 1991 to December 2020. We report standard errors with Newey-West adjustment for 12 lags in parentheses. One, two, and three asterisks indicate that the alpha is significant at the 10%, 5%, and 1% level, respectively.

	FF3+MOM	FF5	FF5+MOM	FF5+MOM +LIQ
Gradient boosting	0.183***	0.222***	0.197***	0.197***
	(0.066)	(0.075)	(0.071)	(0.072)
Random forests	0.214***	0.264***	0.227***	0.228***
	(0.082)	(0.092)	(0.083)	(0.085)
Elastic net	0.058	0.086	0.106	0.108
	(0.063)	(0.066)	(0.068)	(0.068)
OLS	0.059	0.090	0.108	0.107
	(0.063)	(0.066)	(0.067)	(0.067)

Table 17: **Out-of-sample alpha of fund portfolios under different market conditions**

This table reports the monthly out-of-sample alphas (in %) net of all costs for the top-decile fund portfolios obtained with gradient boosting and random forests under different market conditions. Alphas are computed by regressing the out-of-sample excess monthly portfolio returns net of all costs against the Fama and French (2015) five factors and momentum as well as indicator variables for expansions and recessions (Panel A), and high and low investor sentiment (Panel B). Expansions and recessions are defined following the NBER convention. The high (low) investor sentiment indicator equals one if investor sentiment, as defined in Baker and Wurgler (2006, 2007), is above (below) the median of the July 1965–December 2018 period. The out-of-sample period spans from January 1991 to December 2020 in Panel A and from January 1991 to December 2018 in Panel B because the sentiment indicator is only available until 2018. We report standard errors with Newey-West adjustment for 12 lags in parentheses. One, two, and three asterisks indicate that the alpha is significant at the 10%, 5%, and 1% level, respectively.

	Panel A. Business Cycle			Panel B. Investor Sentiment		
	Expansion	Recession	Expansion − Recession	High	Low	High − Low
Gradient boosting	0.176**	0.309	-0.133	0.233***	0.141	0.092
	(0.073)	(0.238)	(0.240)	(0.090)	(0.091)	(0.105)
Random forests	0.187**	0.464*	-0.277	0.251**	0.191*	0.060
	(0.083)	(0.250)	(0.241)	(0.101)	(0.114)	(0.124)

Figure 1: **Correlation matrix between the target variable and fund characteristics**

This figure reports correlation coefficients between the target variable (annual realized alpha) and fund characteristics used as predictors. Predictors are lagged one year with respect to the target variable.

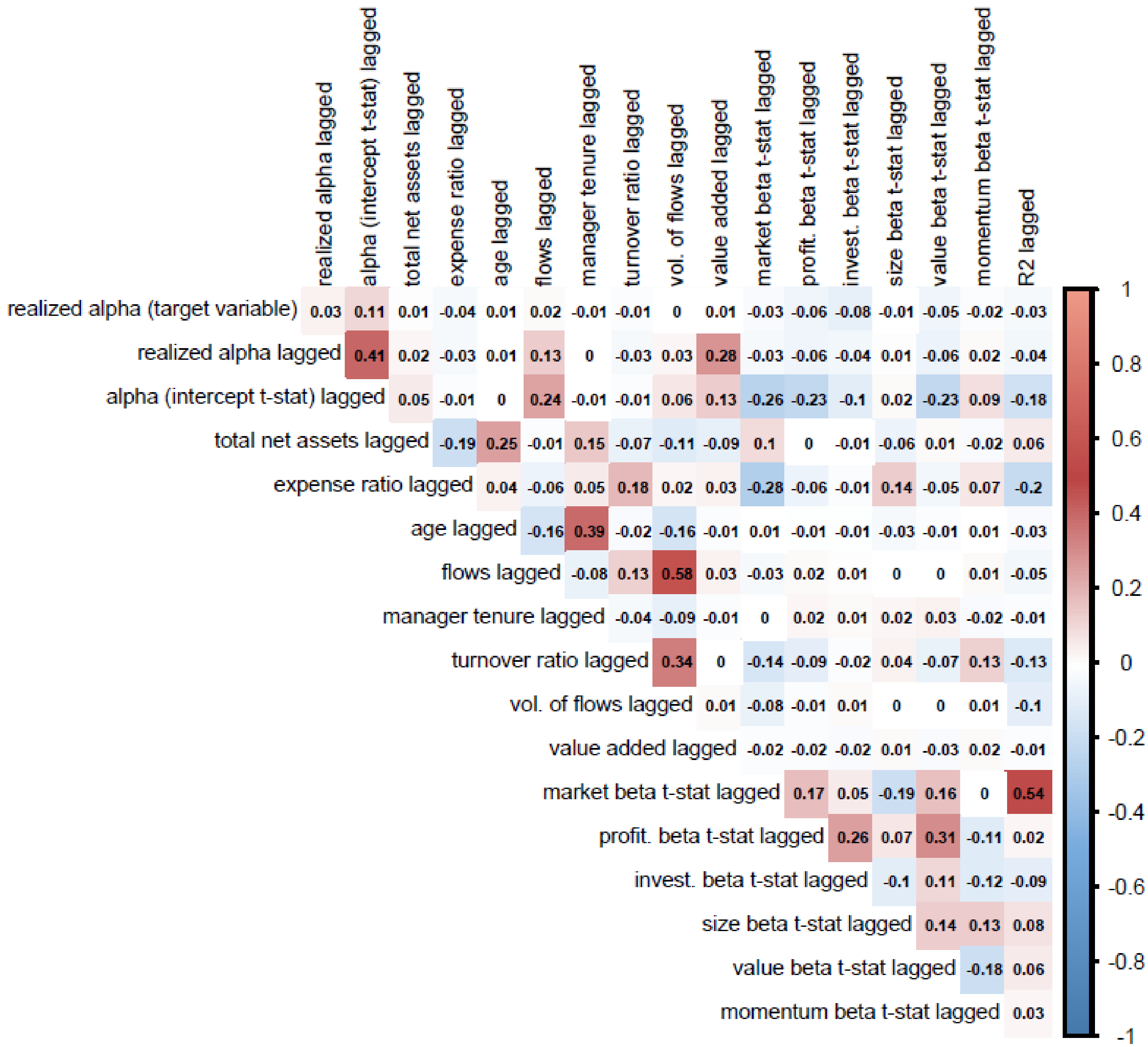

Figure 2: **Characteristic importance**

This figure reports the *relative* importance of each characteristic, ranging from zero for the least important characteristics to 100 for the most important characteristic, and for the gradient boosting (GB), random-forest (RF), elastic net (EN), and OLS portfolios. We report relative importance for the last estimation window, which spans the 1980–2019 period.

	GB	RF	EN	OLS
realized alpha	49	54	6	6
alpha (intercept t−stat)	17	26	100	100
total net assets	13	30	4	5
expense ratio	8	10	50	54
age	0	0	7	9
flows	0	3	8	8
manager tenure	4	3	5	6
turnover	9	20	5	5
vol. of flows	1	1	0	1
value added	13	14	1	0
market beta (t−stat)	84	78	24	27
profit. beta (t−stat)	34	46	22	23
invest. beta (t−stat)	29	45	52	58
size beta (t−stat)	16	34	15	20
value beta (t−stat)	13	30	40	44
momentum beta (t−stat)	11	36	54	63
R2	100	100	89	89

Figure 3: **Nonlinearity in the relation between fund characteristics and performance**

This figure displays partial-dependence plots for the three most important characteristics—R^2 (left graph), market beta t-stat (center graph), and realized alpha (right graph)—for the gradient-boosting (red line) and random-forest (blue line) methods. For each partial-dependence plot, the horizontal axis shows the characteristic cross-sectionally standardized to lie between -1.0 and 1.0 and the vertical axis graphs the prediction of a given machine-learning method conditional on the value of the characteristic averaged across all observations of the other characteristics; see Friedman (2001) and Hastie et al. (2009). Estimates are for the last estimation window spanning the period from 1980 to 2019.

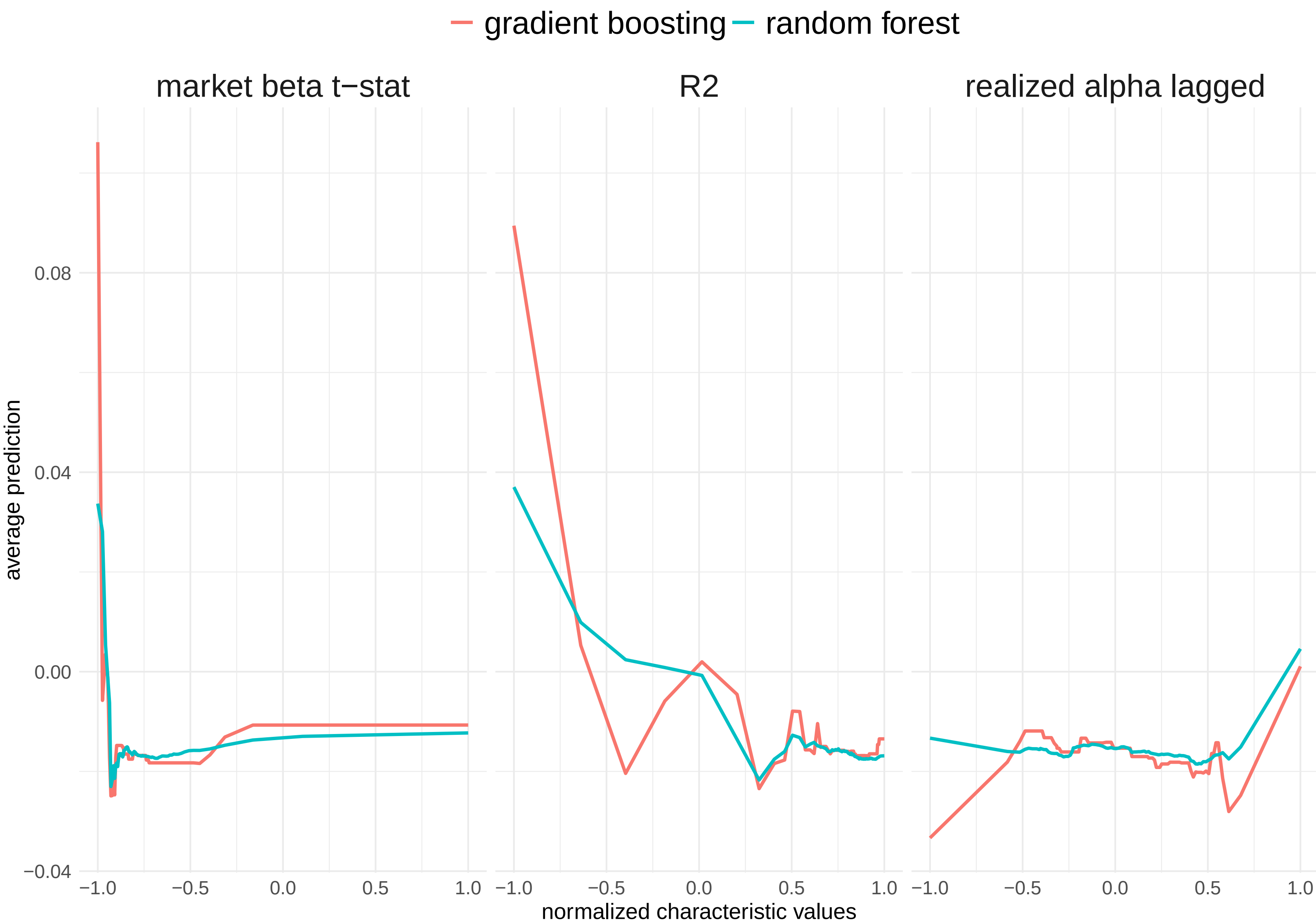

Figure 4: **Interaction importance**

This figure plots the interaction strength of the 20 most important interactions for the gradient boosting and random forests methods and for the last estimation window, which spans the period from 1980 to 2019. We calculate interaction strength using the approach of Greenwell et al. (2018).

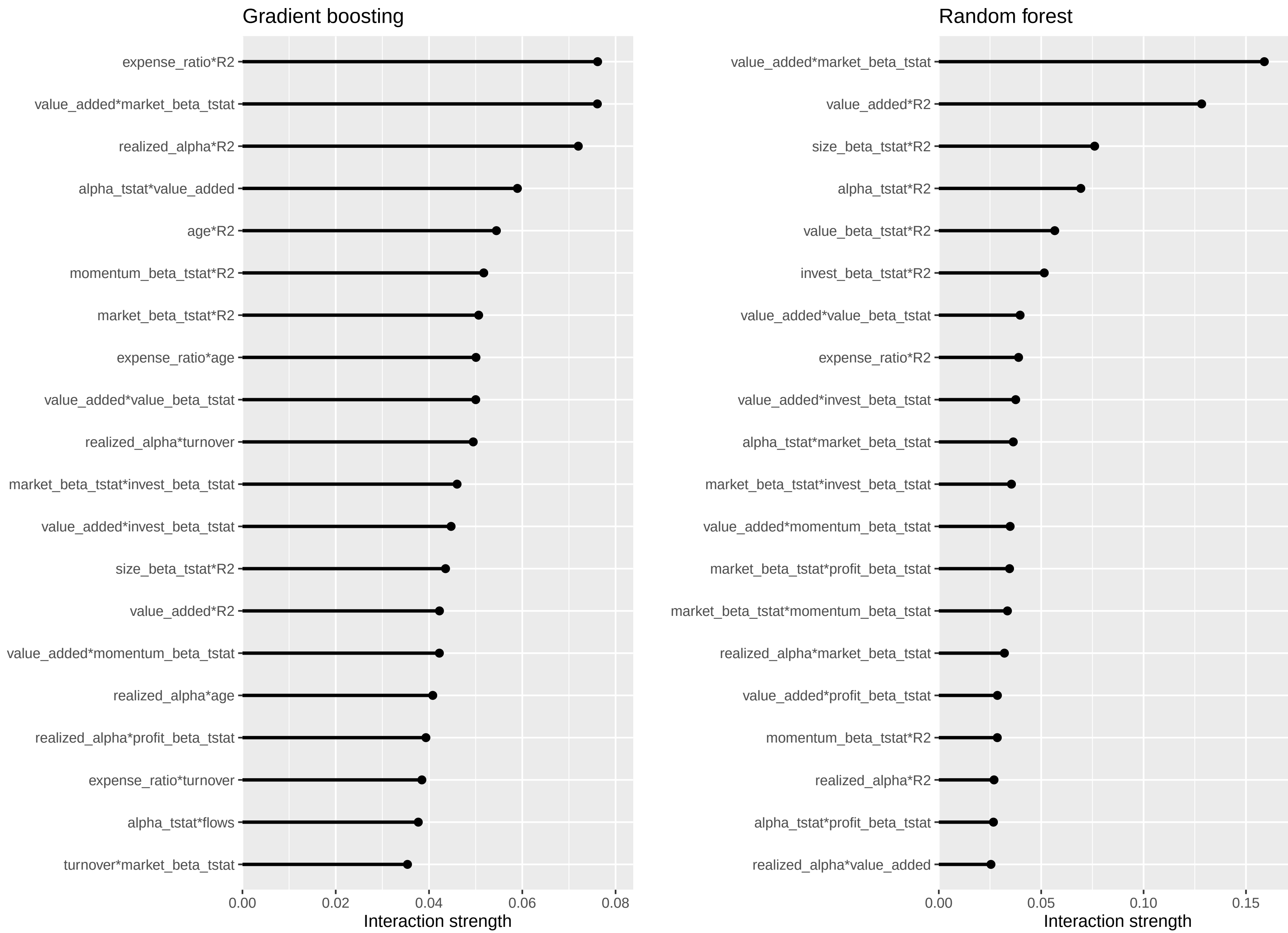

Figure 5: **Tree-based visualization of the interaction between value added and R-squared**

This figure depicts a four-level decision tree that exploits value added and R-squared to predict future alpha. Predictors are cross-sectionally standardized to have zero mean and unit standard deviation. Estimates are for the last estimation window spanning from 1980 to 2019. Each node shows predicted alpha in decimal units and the percentage of observations in the associated partition. At the root node, value added is the characteristic that best explains the sample data, and thus, the tree splits the sample at a *standardized* value added of 0.16. For observations with standardized value added above 0.16, the decision tree splits the observations at a standardized R^2 of -1.1. For observations with standardized value added below 0.16, the characteristic that best explains the data is again value added, and thus, the decision tree splits the observations at a standardized value added of -0.2. For observations with standardized value added below -0.2, the decision tree splits the observations at a standardized R^2 of -0.48. The decision tree illustrates the interaction between fund activeness and past performance because it predicts a higher future alpha for more active funds ($R^2 \leq -1.1$) conditional on their having good past performance (value added ≥ 0.16), but for less active funds ($R^2 \geq -0.48$) conditional on poor past performance (value added < -0.2).

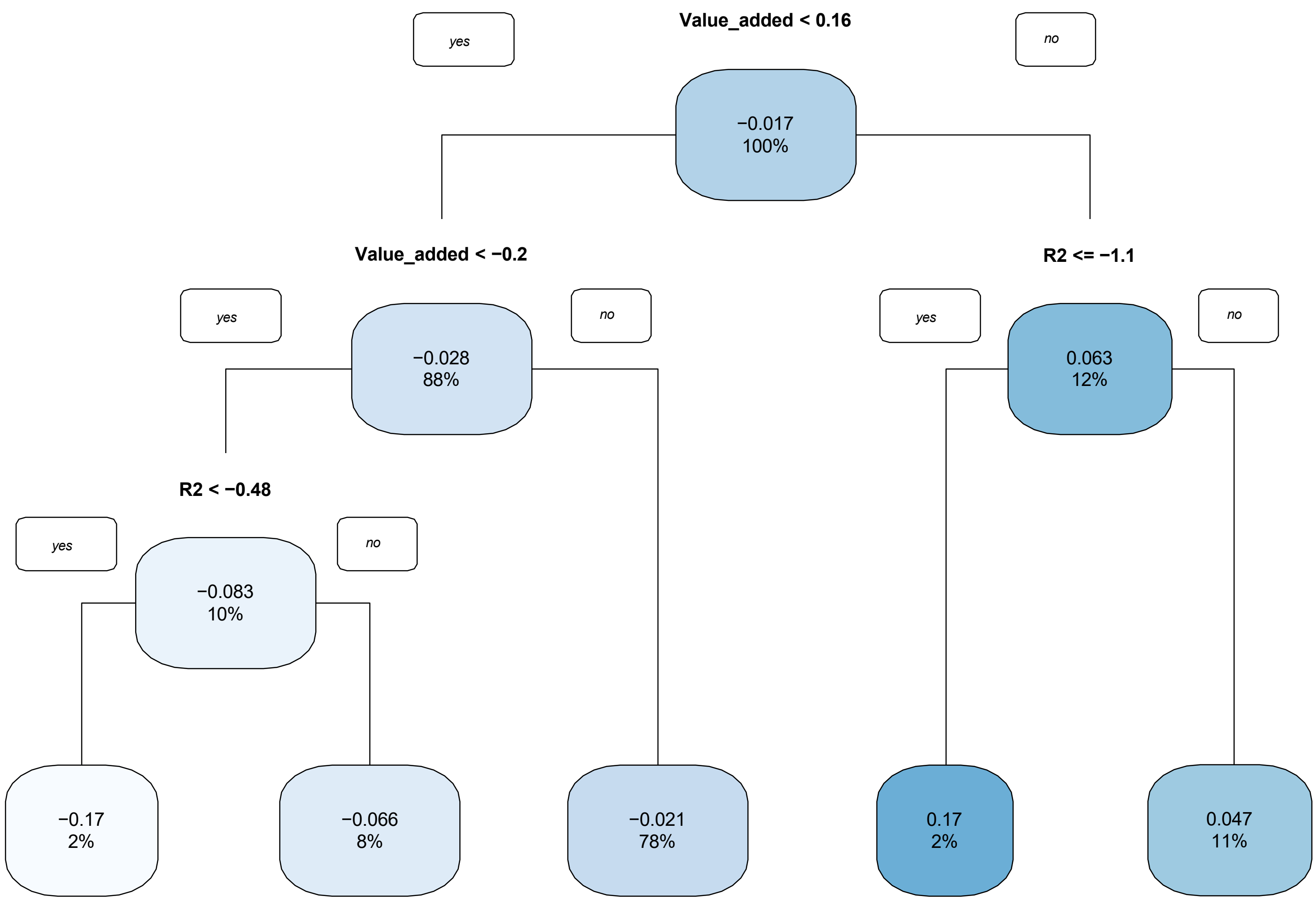

Figure 6: **Tree-based visualization of the interaction between value added and market beta *t*-stat**

This figure illustrates the interaction between value added and market beta *t*-stat by depicting a four-level decision tree that exploits the two characteristics to predict alpha. Predictors are cross-sectionally standardized to have zero mean and unit standard deviation. Estimates are for the last estimation window spanning the period from 1980 to 2019.

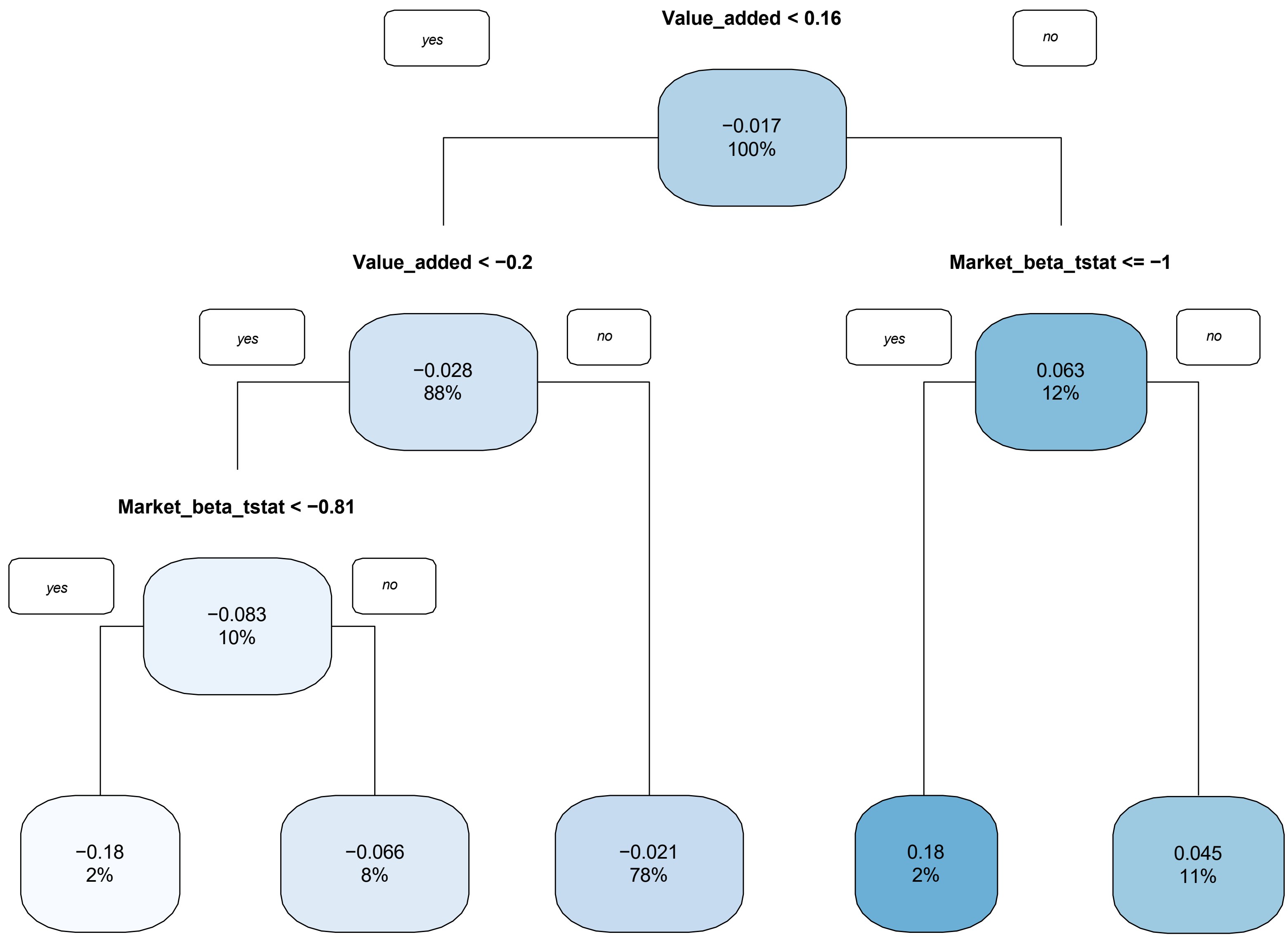

Figure 7: **Tree-based visualization of the interaction between alpha t-stat and R-squared**

This figure illustrates the interaction between alpha t-stat and R-squared by depicting a four-level decision tree that exploits the two characteristics to predict alpha. Predictors are cross-sectionally standardized to have zero mean and unit standard deviation. Predictions are for the last estimation window, which spans the period from 1980 to 2019.

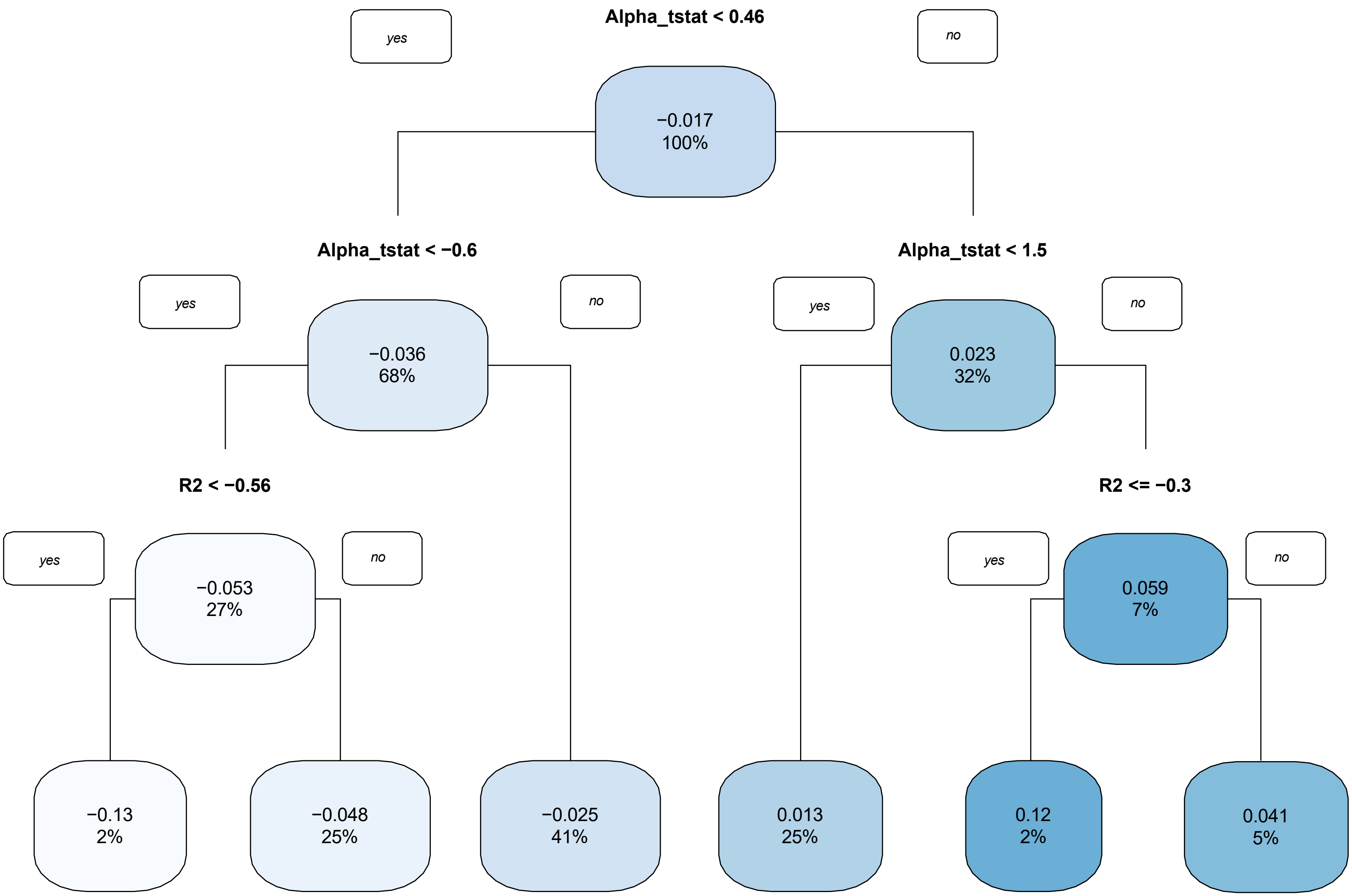

Figure 8: **Time series of variable importance for the gradient-boosting method**

This figure plots the time evolution of the relative importance of each characteristic for the gradient-boosting method, where the *relative* importance ranges from zero for the least important characteristics to 100 for the most important characteristic. The relative importance is computed for each year from 1980–2019.

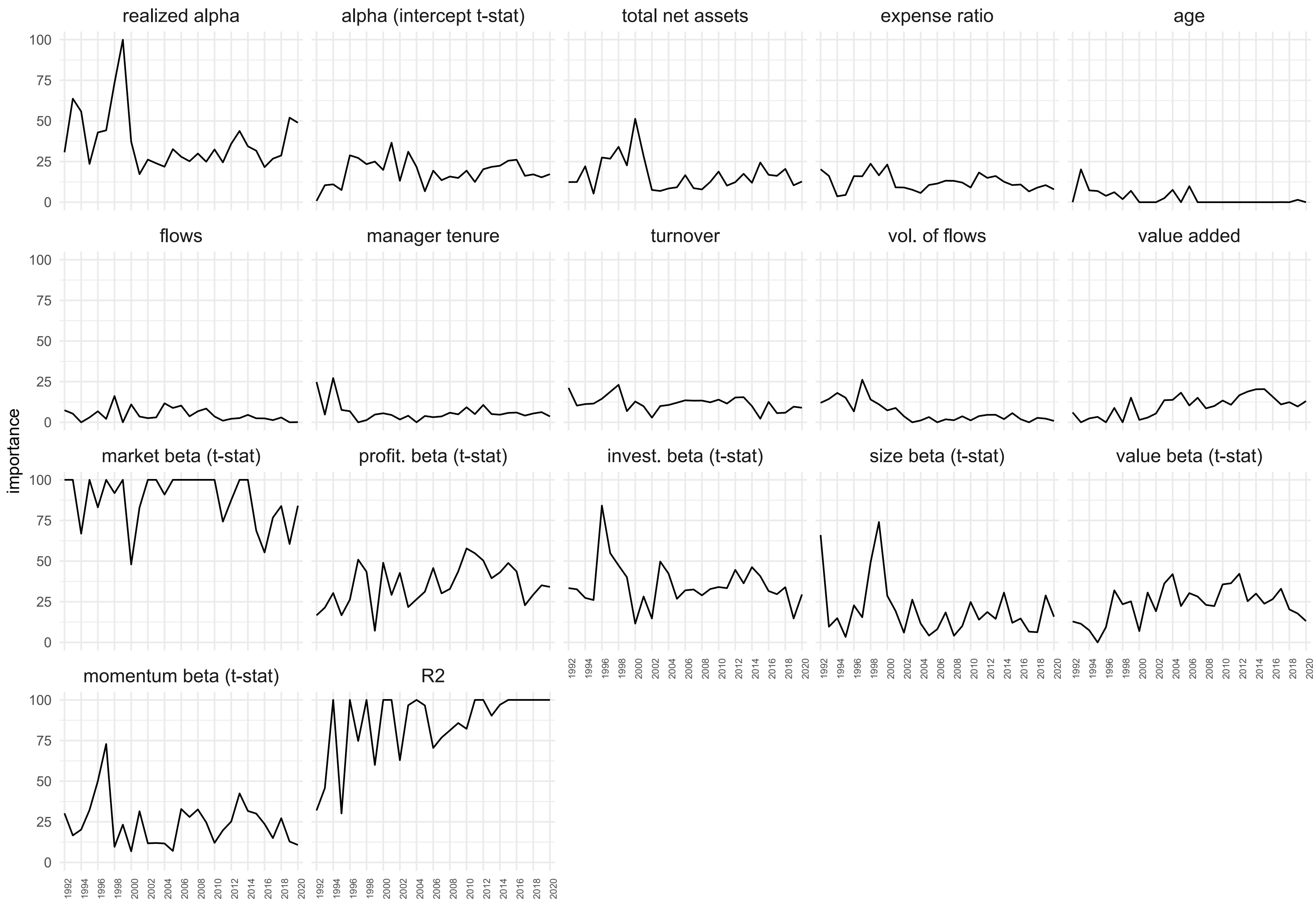

Figure 9: **Time series of variable importance for the random forest method**

This figure plots the time evolution of the relative importance of each characteristic for the random method, where the *relative* importance ranges from zero for the least important characteristics to 100 for the most important characteristic. The relative importance is computed for each year from 1980–2019.

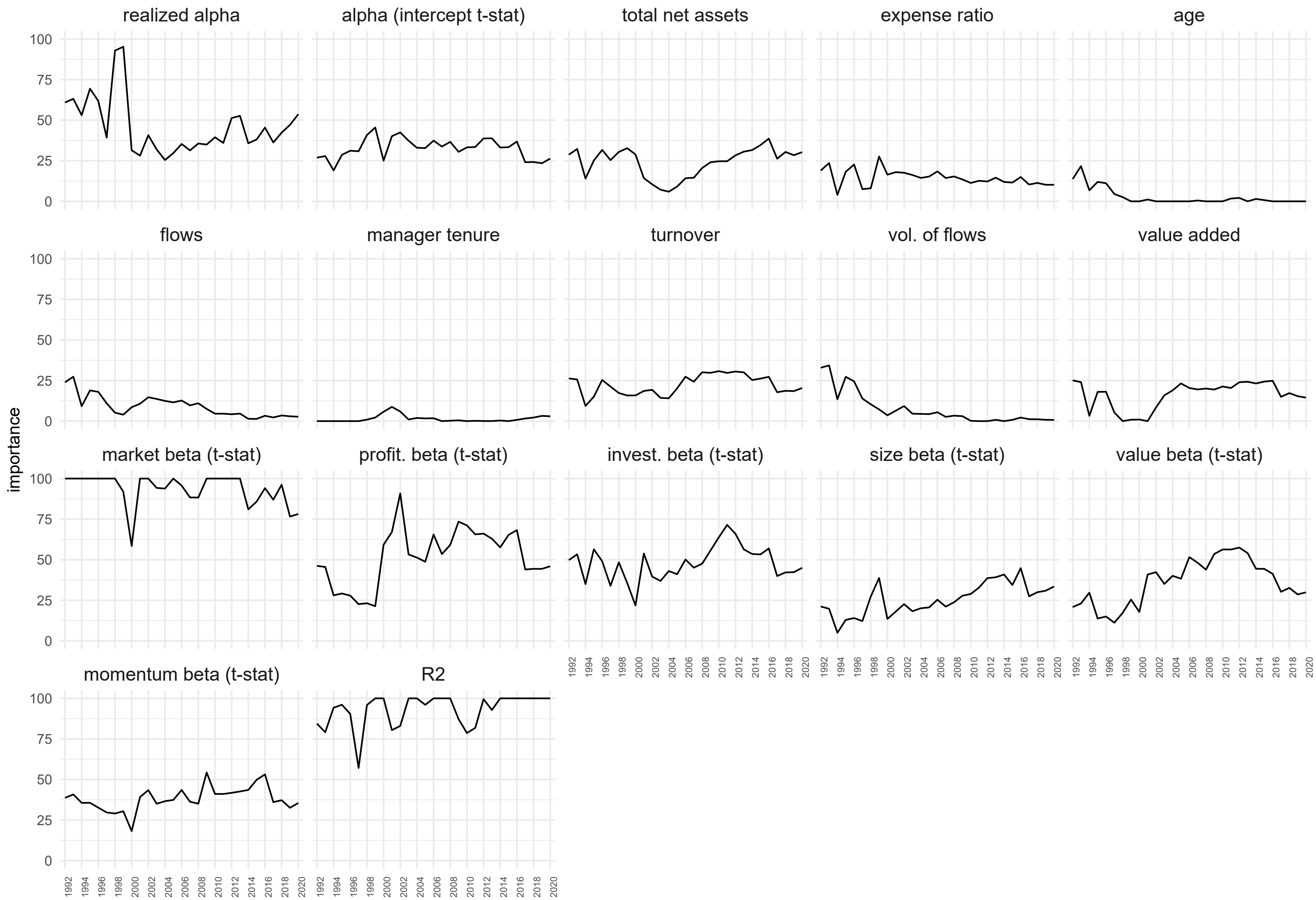

Figure 10: **Cumulative portfolio alpha**

This figure plots the time series of cumulative out-of-sample portfolio realized alphas of the excess returns net of all costs of the top-decile fund portfolios. Realized portfolio alphas are based on the regressions on the five Fama-French factors augmented with momentum (FF5+MOM). Portfolios are obtained with gradient boosting (GB), random forests (RF), OLS, and with two naive strategies (equally weighted (EW) and asset-weighted (AW) portfolios of all available funds).