Can Machine Learning Help to Select Portfolios of Mutual Funds?*

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

June 3, 2021

Abstract

Identifying outperforming mutual funds ex-ante is a notoriously difficult task. We use machine learning to exploit numerous fund characteristics and construct portfolios of equity funds that earn out-of-sample annual alpha of 4.2% net of costs. We show that such performance is the joint outcome of both exploiting multiple fund characteristics and allowing for flexibility in the relation between characteristics and performance. We demonstrate that even retail investors can benefit from investing in actively managed funds. The performance of our portfolios has declined over time, however, consistent with increased competition in asset markets and diseconomies of scale at the industry level.

Keywords: Mutual-fund performance; performance predictability; active management; elastic net; random forests; gradient boosting.

JEL classification: G23; G11; G17.

*Gil-Bazo, Department of Economics and Business, Universitat Pompeu Fabra and Barcelona GSE, 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, Big Data Institute, Universidad Carlos III de Madrid and Department of Economics, Universidade Federal de Santa Catarina, `andreportela@gmail.com`. We are grateful for comments from Joao Cocco, Francisco Gomes, Juan Imbet, Marcin Kacperczyk, Howard Kung, Narayan Naik, Jean Pauphilet, Anna Pavlova, Stephen Schaefer, Henri Servaes, Raman Uppal, and Paolo Zaffaroni, and seminar participants at CUNEF, London Business School, and Universidad Autónoma de Madrid. Javier Gil-Bazo acknowledges financial support from the Spanish Ministry of Economy and Competitiveness, through the Severo Ochoa Programme for Centres of Excellence in R&D (CEX2019-000915-S).

1 Introduction

Since August 2019, U.S. indexed domestic equity mutual funds and ETFs have more assets under management than actively managed mutual funds. Many interpret this victory of passive asset management as a consequence of the persistent inability of active managers to outperform cheaper passive alternatives (Gittelsohn, 2019). Indeed, mutual-fund research consistently shows that the average active fund earns negative risk-adjusted returns (alpha) after transaction costs, fees, and other expenses. Several recent studies, however, show that certain fund characteristics predict future performance. If investors can successfully exploit this performance predictability, then there is still room for active management in the fund industry. In this paper, we investigate whether investors can use machine learning combined with publicly available data to construct portfolios of mutual funds that deliver positive alpha net of all costs.

The underperformance of actively managed mutual funds is a pervasive finding in the empirical literature (Sharpe, 1966; Jensen, 1968; Gruber, 1996; Ferreira et al., 2013). One interpretation of this evidence is that managers lack the ability to generate alpha, and thus, active funds must necessarily underperform passive benchmarks due to transaction costs, fees, and other expenses. However, several studies document the existence of a subset of managers that outperform their benchmarks (Wermers, 2000; Kacperczyk et al., 2005, 2008; Kacperczyk and Seru, 2007; Barras et al., 2010; Fama and French, 2010; Kacperczyk et al., 2014; Berk and Van Binsbergen, 2015). Then, the relevant question is whether investors can benefit from active management by identifying the best managers ex-ante. To answer this question, researchers have investigated if future fund performance can be predicted by 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 fund historical returns. If there are diseconomies of scale in portfolio management, in equilibrium funds with positive past alpha attract more assets, and thus, offer the same expected net alpha as any other active fund: that of the alternative passive

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

benchmark (zero). However, the empirical evidence regarding diseconomies of scale in portfolio management is mixed (Chen et al., 2004; Reuter and Zitzewitz, 2010; Pástor et al., 2015; Zhu, 2018). Also, mutual-fund investors fail to appropriately adjust returns for risk, which suggests that investors' ability to judge mutual-fund performance is limited (Berk and Van Binsbergen, 2016; Barber et al., 2016; Evans and Sun, 2021). In addition, frictions may prevent investors' flows from driving fund performance towards 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. Typically in studies of performance predictability, funds are ranked every month or quarter on the basis of some mutual-fund characteristic. Funds are then allocated to quintile or decile portfolios based on their ranking, and portfolio-return alpha is computed every period. Despite many attempts, only a few studies, which we review below, are able to select portfolios of funds with positive alpha after transaction costs, fees, and other expenses.

In this paper, we also take on the challenge of identifying mutual funds with positive alpha. Our approach departs from the existing literature along three dimensions. First, our goal is not to *discover* a new predictor of fund performance. Instead, we aim to provide investors with a method that can help them exploit any predictability about fund performance that can be found in the data. Specifically, we consider a large set of mutual-fund characteristics that are either readily accessible to investors or can be easily computed from available data, and evaluate the ability of all the characteristics to jointly predict performance. By allowing for multiple variables to predict future performance, we account for the complex nature of the problem at hand. Fund performance is determined by a host of different factors including the manager's multifaceted ability, portfolio constraints, the resources allocated by the asset management firm to the fund manager, manager incentives and agency problems, the efficiency of the market in which the manager invests, competition among managers, as well as more direct determinants, such as trading costs, fees, and other expenses. In this setting, it seems unlikely that using a single variable to predict performance is as efficient as exploiting a large set of characteristics.

Second, we use three different machine-learning methods to forecast fund performance: elastic net, gradient boosting, and random forests. Machine learning is appropriate in this context because it can deal with complex relations between the variables. In particular, there is no a priori reason to believe that the relation between fund characteristics and performance is linear or even monotonic. Moreover, interactions between different fund characteristics may be important to predict mutual fund performance (e.g., Shen et al., 2021). Machine learning allows for a flexible mapping between performance and characteristics and can therefore help uncover predictability that would be missed by variable-sorting or linear methods. Also, machine learning is good at accommodating irrelevant predictors, so it allows one to consider multiple *potential* predictors with lower risk of overfitting than Ordinary Least Squares (OLS). Regularization methods such as elastic net and decision-tree-based methods, such as gradient boosting and random forests, have been applied to solve several problems in Finance (for instance, Rapach et al., 2013; Bryzgalova et al., 2019; Coulombe et al., 2020; Freyberger et al., 2020; Kozak et al., 2020). We choose these methods because they have been shown to outperform other machine-learning methods in forecasting economic and financial variables with structured (or tabular) data, as in our case (for intance, Medeiros et al., 2021; Gu et al., 2020). As a robustness test, in Section 5 we also use neural networks.

Third, our approach is dynamic and out-of-sample. The decision of whether and how to exploit a fund characteristic to identify outperforming funds is taken every time the relation between predictors and performance is reevaluated; that is, whenever the portfolio is rebalanced. Also, the decision is based exclusively on past data. By allowing for changes over time in the relation between characteristics and performance, our method can accommodate changes in the underlying determinants of fund performance due to investor learning or changes in market conditions and strategies of fund managers and management companies.

We implement our approach using monthly data on no-load actively managed US domestic equity mutual funds in the 1980-2018 period. We use the first 10 years of data to train the three machine-learning methods to predict one-year ahead net alpha. Specifically, we define our target variable as net alpha in each calendar year, which we estimate using the five-factor model of Fama and French (2015) augmented with momentum. As predictors, we use 17 mutual-fund characteristics in the previous year. We then use the methods to predict performance in the following year, form a long-only equally-weighted portfolio of the funds in the top decile of the predicted performance distribution, and compute the return of the portfolio in the following 12 months. For every remaining year, we expand the training sample forward one year, train the algorithms again, make new predictions for the following year, construct a new top-decile portfolio, and track its return for the next 12 months. This way, we construct a time series of monthly out-of-sample returns of the top-decile portfolio. Finally, we evaluate the net alpha of the portfolio over the whole out-of-sample period. We compare the performance of the portfolios constructed using the three machine-learning methods and OLS. We also consider the portfolios based on two naive strategies: equally weighted and asset-weighted portfolios of all available funds. The former is the portfolio of a mutual fund investor who does not believe that differences in performance across funds are predictable. The latter is the portfolio of an investor who relies on the aggregate revealed preferences of mutual fund investors.

Our results can be summarized as follows. Two of the three machine-learning methods, gradient boosting and random forests, are able to select a portfolio of funds that delivers positive and statistically significant net alpha. In particular, the portfolio constructed with gradient boosting earns annual alpha net of all fees, expenses, and transaction costs ranging from 3.5% to 4.2%, depending on the factor model employed to evaluate performance. If we use random forests to select funds, net alpha ranges from 2.4% to 3% per year.

The portfolios based on elastic net and OLS deliver positive alphas, but they are lower than that of the gradient-boosting portfolio and statistically indistinguishable from zero. However, both elastic net and OLS outperform the equally weighted and asset-weighted portfolios, which earn negative alpha. Therefore, 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.

Our findings are robust to whether or not we account for momentum to evaluate the performance of the fund portfolios. Our conclusions do not change either, if we include the liquidity risk factor of Pástor and Stambaugh (2003) or if we use other models to evaluate the performance of the fund portfolios, such as those of Carhart (1997), Cremers et al. (2013), Hou et al. (2015), and Stambaugh and Yuan (2017). Results are also robust to constructing portfolios consisting of funds in the top 5% or 20% of the predicted alpha distribution. Finally, 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 retail investors can also benefit from the predictability of fund performance by employing machine-learning methods. We also evaluate whether we can obtain improved prediction-based portfolios by resorting to neural networks. Specifically, we follow Gu et al. (2020) and Bianchi et al. (2021) and implement feed-forward neural networks with up to three layers. The neural-network portfolios deliver positive and statistically significant alphas—but systematically lower in comparison to those obtained with gradient boosting and similar to those obtained with random forests.

Second, we focus on the gradient-boosting portfolio and show that its performance is not driven by a single characteristic. More specifically, we analyze the importance of each characteristic and find that the second most important predictor has a very similar importance to that of the most important predictor. To further explore the performance of our multivariate approach, we obtain the gradient-boosting portfolio by including only the two, three, or four most important characteristics. We find that the performance of the resulting portfolio increases with the number of characteristics, but even with the four most important predictors, it remains well below the performance of the portfolio that exploits all fund characteristics. These findings suggest that attempts to exploit the predictive ability of a single fund characteristic to construct portfolios of funds are likely to be dominated by a multivariate approach.

Third, we show that the relative importance of different variables exhibits substantial variation over time. For instance, the importance of past performance as a predictor (relative to that of the most important predictor) in the gradient-boosting method varies from 14% to 86% throughout our evaluation period. Similar patterns appear in the vast majority of characteristics. Such variation in importance highlights the need for a dynamic approach, where the predictive relation between fund characteristics and performance is reevaluated every time the portfolio is rebalanced.

Finally, Jones and Mo (2020) analyze the out-of-sample performance of 27 characteristics that have been documented to forecast mutual-fund alphas. The authors provide evidence that the predictive ability of fund characteristics with respect to future fund performance has declined over time due to an increase in arbitrage activity and competition among mutual funds. Motivated by their finding, we evaluate the performance of the fund portfolios over rolling windows of five years. Our results indicate that the gradient-boosting portfolio consistently outperforms the OLS portfolio as well as the equally weighted and asset-weighted portfolios.

However, consistent with the findings of Jones and Mo (2020), alpha declines through the sample period for all portfolios, including the gradient-boosting portfolio. This result suggests that the best performing machine-learning method is able to extract alpha from the mutual-fund market, but only when there is any alpha to be extracted in the first place.

Our results are of great practical importance for investors, financial advisers, managers of funds of funds, and pension-plan administrators. The methods we propose are readily implementable and can be used to improve fund selection. Importantly, the data requirements are minimal, as all the information we employ is available in public registries and through commercial data vendors. Naturally, not all investors are equipped with the resources necessary to apply machine learning to select mutual funds. However, independent analysts, on which retail investors rely for their mutual-fund decisions, can use the same methods and data we employ in this paper to make their recommendations.

Our paper contributes to a large literature on the predictability of mutual-fund performance (Jones and Mo, 2020), which documents significant associations between a single mutual-fund characteristic and subsequent fund performance. However, constructing long-only portfolios of funds based on those characteristics does not necessarily enable investors to earn positive alphas. For instance, higher expense ratios are strongly and negatively associated with net fund alphas in the cross-section, but a portfolio that invests only in the cheapest funds does not outperform its passive benchmarks in net terms. In other words, the predictive ability of expense ratios with respect to performance helps investors to avoid expensive underperforming funds, but not to select funds with positive alphas. In fact, only seven of the 27 studies identified by Jones and Mo (2020) report positive and statistically significant 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 providing further evidence of out-of-sample predictability in positive net alphas. Instead of using a single characteristic, we exploit multiple characteristics and let the importance of each predictor vary over time as new information becomes available. Also, we introduce flexibility in the relation between fund characteristics and future performance by using machine learning to account for nonlinearities and interactions.

Our paper is related to recent research by Wu et al. (2021) and Li and Rossi (2021). Wu et al. (2021) apply machine learning to select *hedge funds*. In particular, they use hedge-fund characteristics constructed from fund historical returns to predict future hedge-fund alphas. Instead, we predict future *mutual-fund* alphas by exploiting both fund historical returns and observable *characteristics*. Li and Rossi (2021) use machine learning to select mutual funds by combining data on *fund holdings* and *stock characteristics*. In contrast, we construct portfolios of funds exploiting only *fund characteristics* that do not require the use of fund-holding or stock-characteristic data. Li and Rossi (2021) find that by exploiting fund holdings and stock characteristics one can build fund portfolios that earn significant alphas. Our findings complement theirs by showing that investors can alternatively select mutual funds with significant and positive net alpha by exploiting *solely* the information contained in fund characteristics.

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 in their portfolios to achieve better risk-return tradeoffs. Also, while those studies use a monthly rebalancing frequency, here we adopt a more realistic approach and allow investors to rebalance their portfolios annually.

Finally, our paper also contributes to a growing literature that employs machine learning to address empirical problems in Economics and Finance. These problems include: predicting global equity-market returns using lagged returns of all countries (Rapach et al., 2013); predicting consumer credit-card delinquencies and 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 factors and 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 relationship between multiple 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 different machine-learning methods to improve the classification of mutual funds in accordance to their investment category, but do not study fund performance. Chiang et al. (1996) and Indro et al. (1999) investigate the ability of neural networks to predict a 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, which we collect from CRSP. 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 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. Moreover, we restrict our analysis to share classes that charge no front-end or back-end load, and thus rebalancing our portfolios of mutual funds does not incur any costs. Our sample includes both institutional and retail share classes and spans the 1980–2018 period.

We apply a few filters common in the mutual-fund literature. First, we include only share classes of actively managed funds, therefore excluding ETFs and passive mutual funds. 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 with less than US$ 5 million of Total Net Assets (TNA) and 36 months of age (e.g. Elton et al., 2001; Evans, 2010).[2] Our final sample contains 6,216 unique share classes, of which 5,561 correspond to diversified equity funds (representing 94% of aggregate TNA in the

[2]We keep all the observations of a share class after it reaches US$ 5 million of TNA, even if TNA falls below that level.

sample) and 665 belong to sector funds.

2.2 Mutual-fund characteristics

We construct a dataset of 17 share-class characteristics. For the ith share class in the mth month, we download from CRSP the *return* in excess of the risk-free rate and net of expenses and transaction costs ($r_{i,m}$), the *total net assets* ($TNA_{i,m}$), the *expense ratio* ($ER_{i,m}$), and the *turnover* ratio. In addition, we compute the class's *age* as the number of months since the class's inception date; we estimate the monthly *flows* as the relative growth in the class's 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.[3]

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.[4] 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). We also compute the monthly realized alpha for the ith share class in the mth month ($\alpha_{i,m}$) as:

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

[3]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.

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

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 five Fama-French and momentum 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. Following Berk and Van Binsbergen (2015), we define value added as

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

This variable captures the dollar value extracted by the share-class's manager from the asset market.[5]

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, size, and momentum factors. The average R^2 is 0.9, 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 503,521 to 592,493.

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 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

[5] In their study, Berk and Van Binsbergen (2015) estimate before-fee alpha by regressing funds' 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.

other characteristics, we use their values in December of each year.

Second, similar to 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 are used as predictors when training the machine-learning methods. Our target variable is the share-class's realized alpha in the calendar year. This choice is consistent with our goal to exploit any information in share-class characteristics about the manager's ability to generate positive alpha, regardless of the source of alpha. In contrast, Li and Rossi (2021) 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.[6] 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 (59%).

3 Machine-learning methods

We use well-known `R` packages to implement the different machine-learning methods. The interested reader can refer to the `R`-package documentation for a detailed description of the methods.[7] Gu et al. (2020) also provide a detailed description of various machine-learning methods in the context of asset pricing. In the remainder of this section, we provide a brief

[6] 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.

[7] 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.

description of the methods we consider and we discuss how we calibrate 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 an 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 highly interpretable prediction, machine-learning methods often outperform OLS for data that exhibit high variance, non-linearities, 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 more accurate predictions. To capture non-linearities and interactions between the characteristics, we consider two types of ensembles of decision trees (*random forests* and *gradient boosting*), which often outperform the linear methods on structured (or tabular) data like ours; see, for instance, Medeiros et al. (2021).[8]

Another popular machine-learning method is neural networks, which tend to perform well on non-structured data or highly non-linear structured data. To capture these non-linearities, neural networks require estimating a larger number of parameters, and hence, they require a large number of observations to deliver accurate estimates. Therefore, neural networks are likely to underperform random forests and gradient boosting for our mutual-fund database. Nonetheless, as a robustness check we evaluate the performance of feed-forward neural networks with up to three hidden layers in Section 5.[9]

[8] Li and Rossi (2021) use the same methods we employ, but they rely on a completely different set of predictors and they use excess returns instead of realized alpha as the target variable.

[9] 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).

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.[10] We explain in Section 3.4 how we tune 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 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.

Figure 2 depicts a decision tree based on three mutual-fund characteristics (market beta t-stat, R^2, and realized alpha). At the root node, market beta t-stat is the characteristic that better explains the sample data, and thus, the tree splits the sample depending on whether observations have a *standardized* market beta t-stat smaller or greater than -0.57.

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

For observations with standardized market beta t-stat smaller than -0.57, the characteristic that better explains the data is R^2, and thus, the decision tree splits the observations in this node depending on whether their standardized R^2 is smaller or greater than -0.28. For observations with standardized market beta t-stat greater than -0.57, on the other hand, it is best to split the data depending on whether their standardized realized alpha is smaller or greater than 1.70.

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 that, random forests use bootstrap to randomly select the observations for each tree and the 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 non-linear 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 how fast the ensemble learns from its most recent tree; that is, how much 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, on the depth and number of nodes of each tree, and on the minimum number of observations in a leaf node.

3.4 Cross validation of hyper parameters

We tune the hyper parameters of the elastic net, random forests, and gradient boosting methods using five-fold cross-validation; 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 cross validation that accounts for the time-series properties of the data is *pseudo out-of-sample* evaluation, which reserves a section at the end of the training sample for evaluation. We do not use pseudo out-of-sample evaluation because the empirical and theoretical results in Bergmeir et al. (2018) and Coulombe et al. (2020) show that cross validation tends to perform better.

4 Empirical strategy and main results

We now describe the procedure we use to select share classes and evaluate the performance of the resulting portfolios. Then, we discuss the main results of the paper. 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. We then use the 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, we assume that 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 return during the next 12 months. This way, we construct a time series of monthly out-of-sample returns of the top-decile fund portfolio that spans from January 1991 to December 2018 (346 months).

To evaluate the out-of-sample performance of the top-decile fund portfolio, we compute its alpha by running a time-series regression of the 346 out-of-sample monthly portfolio returns on contemporaneous risk-factor returns. 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.

Table 3 reports the out-of-sample alpha of the top-decile fund portfolios selected by the three machine-learning methods—gradient boosting, random forests, and elastic net—and by Ordinary Least Squares (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.

Two important findings emerge from Table 3. First, all prediction-based algorithms, including OLS, allow investors to construct portfolios with positive out-of-sample alphas. In contrast, naive portfolios earn (in almost all cases) negative, albeit insignificant, alphas according to the four performance-attribution models considered. Interestingly, the asset-weighted portfolio underperforms the equally weighted portfolio, which implies that the average dollar invested in active funds earns lower risk-adjusted returns than the average fund.

Second, both gradient boosting and random forests select portfolios of mutual funds with positive alphas that are significant both statistically and economically. In particular, risk-adjusted returns of the gradient-boosting portfolio range from 29.4 bp per month (3.5% per year) according to the FF3+MOM model, to 34.8 bp per month (4.2% per year) according to the FF5 model. Interestingly, performance is slightly higher with respect to the FF5 model, which ignores momentum, than with respect to the FF5+MOM model (3.8% per year), despite the fact that our target variable is alpha with respect to the FF5+MOM model. If we include exposure to aggregate liquidity risk, performance reduces only marginally (3.9%). Overall, our findings are robust to the choice of performance-attribution model. The alpha of the random-forest fund portfolio is lower than that of the gradient-boosting portfolio, but still positive and statistically significant, ranging from 20.3 bp per month (2.4% per year) to 25 bp per month (3% per year). In contrast, neither the elastic-net fund portfolio nor the OLS fund portfolio achieve statistically significant alphas. This lack of significance appears to be due both to higher standard errors and lower estimated alphas.

Our gradient-boosting fund portfolio earns an alpha with respect to the FF3+MOM model of 3.5% per year, which is very similar to that of the best top-decile portfolio of Li and Rossi (2021), 2.88%. This is somewhat surprising given that the studies use two disjoint sets of predictors: fund characteristics in our case, and stock characteristics combined with fund holdings in the study of Li and Rossi (2021). Thus, our empirical findings complement those of Li and Rossi (2021) by showing that just like manager portfolio holdings, fund traits contain information that is relevant for fund risk-adjusted performance.[11]

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

[11] Li and Rossi (2020, Subsections 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.

gradient-boosting and OLS portfolios is positive and significant, ranging from 21 bp to 25 bp per month (2.5% to 3% per year). A similar conclusion holds for the random-forest portfolio. 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 select portfolios of funds with positive alpha net of all costs, so that investors can combine them with passive portfolios to achieve better risk-return tradeoffs. However, investors may choose to invest only in active funds, so it is interesting to study how the various portfolio 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 returns 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 (2021). 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: 0.184 and 0.169 for gradient boosting and random forests, respectively, followed closely by the equally weighted portfolio (0.166). 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.

Taken together, the results in this section suggest that investors can use observable fund characteristics to improve significantly upon the performance of the average or the asset-weighted average active share class. 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 positive alpha ex-ante. Only methods that allow for non-linearities and interactions in the relationship between fund characteristics and subsequent performance, namely gradient boosting and random forests, can detect funds with large and significant positive alphas. Moreover, the resulting portfolios have the highest Sharpe and Sortino ratios among all the portfolios considered.

5 Robustness checks

We now show that our main findings are robust to: (i) considering alternative cut-off points to select funds; (ii) using alternative models to measure risk-adjusted performance; (iii) building portfolios of only *retail* mutual-fund share classes; and (iv) using neural networks to obtain prediction-based portfolios.

First, 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 6 shows that our findings are robust to considering different cut-off points: the out-of-sample alpha of the gradient-boosting and random-forest portfolios of the top 5% and 20% funds remain positive and significant. Moreover, just like for the top-decile portfolios, neither elastic net nor OLS are able to select a portfolio of funds with positive and significant alpha regardless of the threshold employed. Comparing the portfolios of the top 5%, 10%, and 20% of funds, we observe that the out-of-sample alpha declines as the cut-off point increases from 5% to 20%, but the standard error of the alpha increases. In other words, performance is higher on average but less reliable if we invest only in the top-5% funds.

Second, we check if our results are robust to using alternative factor models for evaluating performance (not for selecting funds). More specifically, in addition to the four different models considered in Table 3, we also estimate the risk-adjusted performance of the prediction-based portfolios using the models of Cremers et al. (2013), Hou et al. (2015) and Stambaugh and Yuan (2017). Results are qualitatively similar to those in Table 3. Gracient boosting and random forests yield the best results with the top-decile portfolio earning positive and significant alphas. 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. The only noteworthy difference with respect to Table 3 is the reduced statistical significance of the performance of the top-decile portfolio selected by gradient boosting and random forests when we use the risk factors of Stambaugh and Yuan (2017) to evaluate performance.

Third, our sample includes both institutional and retail share classes. It is therefore unclear whether the machine-learning methods considered are simply picking institutional share classes, which usually charge lower costs and are subject to more stringent monitoring by investors. To answer this question, we exclude institutional share classes from the sample and repeat the analysis. Table 8 shows that the risk-adjusted performance of the portfolios of retail funds selected by gradient boosting and random forests is as good, and in most cases 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 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 mirror those in Table 3.

Finally, we investigate the performance of neural networks. We follow Gu et al. (2020) and Bianchi et al. (2021) and implement feed-forward neural networks with up to three hidden layers.[12] In particular, 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. 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. Table 9 shows that the neural-network fund portfolios achieve positive and significant out-of-sample alpha for the architectures with one and two hidden layers, but their alphas are systematically lower in comparison to those obtained with the gradient-boosting portfolios. Moreover, we find that

[12]Gu et al. (2020) implement feed-forward neural networks with up to five hidden layers. We refrain from implementing neural networks with more than three layers since our results suggest that including additional layers is not associated to better performance in terms of net portfolio alpha; see Table 9.

single-layer networks yield prediction-based portfolios with higher alpha in comparison to multi-layer networks, which suggests that shallow learning is more appropriate than deep learning in this particular context. This is consistent with Gu et al. (2020), who find that neural-network performance peaks at just three hidden layers and then declines as more layers are added.

6 Which mutual-fund characteristics matter?

In this section, we study which mutual-fund characteristics matter for each of the four prediction methods. We also study whether the favorable performance of machine-learning methods requires using many characteristics or it suffices to use only a few characteristics.

We first quantify the relative importance of each characteristic for the four prediction methods. Like Gu et al. (2020), we compute the relative importance of each characteristic in the gradient-boosting and random-forest methods as 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.

Figure 3 reports characteristic importance for the gradient-boosting (GB), random-forest (RF), elastic-net (EN), and OLS methods for the last estimation window, which spans the 1980–2017 period. To facilitate interpretation, we report *relative* importance ranging from zero for the least important to 100 for the most important characteristic. The figure shows that no single characteristic dominates for any of the methods. For gradient boosting, the second and third most important characteristics are almost as important as the first one, and the fourth and fifth are half as important. For random forests, the first and second characteristics are almost equally important. Elastic net and OLS are very similar in terms of characteristic importance, with four characteristics dominating the others. Interestingly, R^2 is among the top characteristics for all methods. However, the methods differ sharply in the importance of other predictors; gradient boosting relies heavily on realized alpha in the previous year, which is less important for random forests, and almost ignored by elastic net and OLS. The predictions of elastic net and OLS are, instead, strongly influenced by the fund's three-year precision-adjusted alpha. Therefore, the ability of recent performance to improve forecasts of future performance is only apparent when we allow for non-linearities and interactions between variables. Gradient boosting and random forests exploit the fund's precision-adjusted market beta to select funds, but this variable is much less important in linear methods, which rely, instead, on the fund's precision-adjusted beta with respect to momentum. While linear models exploit funds' expense ratios, their predictive ability is subsumed by other fund characteristics in non-linear models. These differences highlight the importance of allowing for non-linearities and interactions in the relation between predictors and fund performance.

Given that non-linear and linear methods rely on different fund characteristics to predict performance, it is interesting to study how funds selected by different methods differ in terms of their characteristics. To address this question, we cross-sectionally standardize fund characteristics and define the top-decile portfolio characteristics at the end of each year as the equally weighted average of the fund characteristics across funds in the top-decile portfolio. Figure 4 reports the time-series average of each portfolio characteristic. Interestingly, selected funds are more similar across methods than suggested by the variable importance chart in Figure 3. In particular, all methods tend to select funds with realized alpha in the previous year between 0.5 and 0.7 standard deviations above the average and precision-adjusted three-year alpha between 0.8 and 1.1 standard deviations above the average. As expected, all methods select funds with below-average R^2, although the portfolios selected by gradient boosting and random forests are much more skewed towards this feature. All methods select funds with above-average flows, turnover, and value added, although the methods do not clearly rely on these characteristics to select funds. All methods tend to select funds with below-average betas. However, funds selected by gradient boosting and random forests differ more from the average fund in terms of market beta and linear methods are particularly skewed towards funds with low investment betas. Interestingly, although OLS and elastic net rely on expense ratios to choose funds, selected funds are only 0.1 standard deviations cheaper than the average fund. In contrast, gradient boosting and random forests select funds that are about 0.3 standard deviations *more expensive* than the average fund.

Our results suggest that allowing for flexibility in the relation between characteristics and performance can help investors select actively managed equity funds that earn positive alphas. We now study whether the remarkable performance of the machine-learning methods is driven by flexibility alone or by flexibility combined with the multivariate approach, which exploits the predictive ability of multiple predictors. To investigate the extent to which very few characteristics are responsible for the performance of the gradient-boosting method in selecting mutual funds, we repeat the analysis using only the two, three, and four most important characteristics selected in each estimation round. Table 10 shows that, when only the two most important fund characteristics are used to predict performance, the top-decile portfolio of mutual funds selected by the gradient-boosting algorithm earns negative alpha according to all factor models considered, except for the Fama-French 5-factor model. However, alpha is statistically indistinguishable from zero. If we include also the third most important characteristic, performance becomes positive for all models, except for the Fama-French 3-factor model, but insignificant. Finally, if we include the fourth most important characteristic, the performance of the top-decile portfolio increases substantially, and even becomes significant, although in all cases it remains below the performance of the top-decile portfolio that exploits all characteristics by more than 10 bp per month. These results further support the notion that flexibility is not enough to explain the performance of the gradient-boosting approach in selecting portfolios of mutual funds. The method exploits the predictability contained in many different fund characteristics and their interactions.

One important feature of our approach is that we do not advocate in favor of a single characteristic and, instead, reevaluate the model as new information becomes available. This feature is an advantage if the predictive ability of some characteristics changes with time as investors learn to exploit their predictive content, or if market conditions or manager strategies change. To investigate this possibility, in Figure 5, we plot the importance of each predictor in each year of the out-of-sample period for the gradient-boosting portfolio. The figure confirms that the importance of some of the key characteristics varies substantially over time. For instance, the relative importance of market beta ranges from 65% to 100%, the relative importance of R^2 ranges from 50% to 100%; and the importance of realized alpha is particularly unstable, ranging from about 20% to 80%.

Our findings suggest that flexibility in the forecasting method alone is not sufficient to achieve the outstanding performance of the gradient-boosting portfolio. Instead, it is also 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 peformed dynamically over time.

7 Has alpha declined over time?

As mentioned in the introduction, Jones and Mo (2020) provide evidence that the ability of fund characteristics to predict performance has declined over time. They further show evidence that this decline is driven by an increase in arbitrage activity and competition among mutual funds. Jones and Mo (2020) use OLS to predict performance, so it is unclear whether their conclusions extend to portfolios of funds selected by machine-learning methods.

To explore this possibility, we evaluate the out-of-sample performance of the top-decile portfolio over five-year rolling windows for the gradient-boosting, OLS, equally weighted and asset-weighted portfolios. Figure 6 shows that the top-decile portfolio selected by gradient boosting consistently beats the equally weighted and asset-weighted portfolios and by a wide margin during most of the sample period. The gradient-boosting portfolio of funds also beats the OLS portfolio in every single 5-year period up to the late 2000s. Since then, however, the performance of the gradient-boosting and OLS portfolios has been very similar. Moreover, since 2015 all four portfolios have converged in terms of performance, with negative alphas characterizing the last years of the sample. Such decline in the performance of prediction-based portfolios of funds is consistent with the findings of Jones and Mo (2020). We may therefore conclude that the best performing machine-learning algorithm is able to extract alpha from the mutual-fund market, but only when there is any alpha to be extracted in the first place.

8 Conclusions

The question of whether mutual fund investors can benefit from active asset management has received much attention from academics, practitioners, and regulators. In this paper, we posit that the pessimistic results that dominate the literature could be a consequence of the methods employed to exploit predictability in fund performance. We contribute to the literature by showing that machine-learning methods can use information contained in multiple fund characteristics to select funds that earn economically and statistically significant positive risk-adjusted returns net of fees and transaction costs. Such positive performance is robust to the model employed to evaluate performance and can be attained by both institutional and retail investors. In contrast, linear forecasting models can help investors only to avoid negative alphas. Therefore, our results demonstrate that investors, including retail investors, can benefit from investing in actively managed funds.

References

Amihud, Y. and R. Goyenko (2013). Mutual fund's R2 as predictor of performance. *Review of Financial Studies 26*(3), 667–694.

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

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* (Forthcoming).

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. *Available at SSRN 3350138*.

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 in 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*, 341–367.

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

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

Freyberger, J., A. Neuhierl, and M. Weber (2020). Dissecting characteristics nonparametrically. *Review of Financial Studies 33*(5), 2326–2377.

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

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.

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.

Kacperczyk, M. and A. Seru (2007). Fund manager use of public information: New evidence on managerial skills. *Journal of Finance 62*(2), 485–528.

Kacperczyk, M., C. Sialm, and L. Zheng (2005). On the industry concentration of actively managed equity mutual funds. *Journal of Finance 60*(4), 1983–2011.

Kacperczyk, M., C. Sialm, and L. Zheng (2008). Unobserved actions of mutual funds. *Review of Financial Studies 21*(6), 2379–2416.

Kozak, S., S. Nagel, and S. Santosh (2020). Shrinking the cross-section. *Journal of Financial Economics 135*(2), 271–292.

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 (2021). 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. *arXiv preprint: https://arxiv.org/abs/2012.12802*.

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. *Availalbe in 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.

Pástor, L., R. F. Stambaugh, and L. A. Taylor (2015). Scale and skill in active management. *Journal of Financial Economics 116*(1), 23–45.

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.

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.

Reuter, J. and E. Zitzewitz (2010). How much does size erode mutual fund performance? A regression discontinuity approach. Technical report, National Bureau of Economic Research.

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. *Review of Financial Studies* (Forthcoming).

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.

Shen, K., L. Tong, and T. Yao (2021). Heterogeneous turnover-performance relations. *Journal of Banking & Finance*, 106054.

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* (Forthcoming).

Zhu, M. (2018). Informative fund size, managerial skill, and investor rationality. *Journal of Financial Economics 130*(1), 114–134.

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: Definition**

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 gives 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 share-class'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 1980-2018 period.

	Mean	Median	Standard deviation	Class-month observations
monthly return	0.72%	1.11%	5.01%	592,483
monthly realized alpha	-0.12%	-0.13%	2.24%	553,311
alpha t-stat	-0.423	-0.418	1.229	553,620
TNA (USD mill.)	628.0	86.6	2,462.5	592,988
expense ratio	1.16%	1.07%	0.65%	589,816
age (months)	143.1	113.0	111.7	592,988
flows	0.005	-0.003	0.042	589,735
manager tenure (years)	7.929	6.753	5.211	547,146
turnover ratio	0.854	0.590	1.277	588,217
volatility of flows	0.055	0.028	0.078	589,735
value added	-0.182	-0.011	11.120	503,521
market beta t-stat	16.111	14.330	10.590	553,620
profitability beta t-stat	-0.102	-0.103	1.450	553,620
investment beta t-stat	-0.452	-0.489	1.506	553,620
size beta t-stat	1.575	0.751	3.834	553,620
value beta t-stat	-0.018	-0.083	2.135	553,620
momentum beta t-stat	0.099	0.100	1.905	553,620
R^2	0.904	0.941	0.123	553,620

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

This table reports the monthly out-of-sample alphas (in %) of the excess returns 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 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.

	FF3+MOM	FF5	FF5+MOM	FF5+MOM +LIQ
Gradient boosting	0.294**	0.348***	0.319**	0.325***
	(0.121)	(0.133)	(0.123)	(0.124)
Random forests	0.203**	0.250**	0.211**	0.213**
	(0.086)	(0.100)	(0.089)	(0.091)
Elastic net	0.069	0.098	0.104	0.114
	(0.066)	(0.069)	(0.071)	(0.071)
OLS	0.070	0.099	0.105	0.115
	(0.066)	(0.070)	(0.072)	(0.071)
Equally weighted	-0.019	-0.013	-0.022	-0.020
	(0.047)	(0.046)	(0.046)	(0.046)
Asset weighted	-0.045	-0.038	-0.041	-0.039
	(0.037)	(0.036)	(0.037)	(0.037)

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

This table reports the monthly out-of-sample alphas (in %) of excess returns net of all costs of the long-short fund portfolio that goes long in the funds included by each of the methods we consider (gradient boosting, random forests, elastic net, equally weighted, asset weighted) and short the funds included in the OLS portfolio. 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 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.

	FF3+MOM	FF5	FF5+MOM	FF5+MOM +LIQ
Gradient boosting minus OLS	0.225**	0.249**	0.214**	0.210**
	(0.097)	(0.105)	(0.094)	(0.093)
Random forests minus OLS	0.133***	0.151**	0.106**	0.098*
	(0.051)	(0.063)	(0.051)	(0.051)
Elastic net minus OLS	0.000	-0.001	-0.001	-0.001
	(0.007)	(0.007)	(0.007)	(0.007)
Equally weighted minus OLS	-0.089*	-0.112**	-0.127**	-0.135***
	(0.046)	(0.050)	(0.050)	(0.049)
Asset weighted minus OLS	-0.114**	-0.137**	-0.146***	-0.154***
	(0.048)	(0.053)	(0.053)	(0.052)

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

For each fund portfolio, this table reports the following 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.

	Mean	Standard deviation	Sharpe ratio	Sortino ratio	Maximum drawdown	VaR 99%
Gradient boosting	0.91%	4.91%	0.184	0.282	52.2%	11.0%
Random forests	0.82%	4.87%	0.169	0.256	53.1%	12.0%
Elastic net	0.71%	4.78%	0.149	0.219	59.3%	12.0%
OLS	0.72%	4.78%	0.150	0.221	59.2%	11.8%
Equally weighted	0.70%	4.24%	0.166	0.242	51.7%	10.0%
Asset weighted	0.65%	4.30%	0.150	0.216	53.1%	10.6%

Table 6: **Out-of-sample alpha of top-5% and top-20% fund portfolios**

This table reports the monthly out-of-sample alphas (in %) of the excess returns 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 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.

	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	0.365**	0.439**	0.388**	0.393**	0.179**	0.219**	0.200**	0.206**
	(0.158)	(0.179)	(0.161)	(0.162)	(0.091)	(0.099)	(0.093)	(0.094)
Random forests	0.304***	0.349***	0.295***	0.296***	0.139*	0.178**	0.150**	0.153**
	(0.108)	(0.119)	(0.109)	(0.111)	(0.074)	(0.084)	(0.076)	(0.077)
Elastic net	0.096	0.133	0.146	0.155*	0.030	0.046	0.056	0.065
	(0.082)	(0.090)	(0.091)	(0.089)	(0.057)	(0.058)	(0.062)	(0.061)
OLS	0.086	0.120	0.135	0.145	0.029	0.046	0.055	0.063
	(0.086)	(0.094)	(0.095)	(0.093)	(0.060)	(0.061)	(0.064)	(0.064)

Table 7: **Out-of-sample alpha of portfolios based on alternative factor models**

This table reports the monthly out-of-sample alphas (in %) of the excess returns 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 factors returns 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 factors returns were downloaded from the *q*-factors data library at www.global-q.org and span the January 1991 to December 2018 period (336 months). Stambaugh and Yuan (2017) monthly factor returns 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	0.285**	0.332**	0.243*
	(0.134)	(0.143)	(0.131)
Random forests	0.151*	0.237**	0.149
	(0.077)	(0.112)	(0.100)
Elastic net	0.061	0.113	0.095
	(0.068)	(0.082)	(0.075)
OLS	0.064	0.112	0.098
	(0.070)	(0.083)	(0.077)
Equally weighted	0.020	-0.008	-0.017
	(0.038)	(0.035)	(0.048)
Asset weighted	-0.052**	-0.038	-0.026
	(0.026)	(0.031)	(0.038)

Table 8: **Out-of-sample alphas of retail share-class portfolios**

This table reports the monthly out-of-sample alphas (in %) of the excess returns 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 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.

	FF3+MOM	FF5	FF5+MOM	FF5+MOM +LIQ
Gradient boosting	0.347**	0.408***	0.367***	0.371***
	(0.134)	(0.147)	(0.136)	(0.136)
Random forests	0.223**	0.240**	0.223**	0.226**
	(0.088)	(0.093)	(0.090)	(0.090)
Elastic net	0.048	0.080	0.083	0.090
	(0.068)	(0.069)	(0.071)	(0.070)
OLS	0.045	0.077	0.080	0.088
	(0.068)	(0.068)	(0.071)	(0.070)
Equally weighted	-0.005	0.002	-0.007	-0.006
	(0.049)	(0.049)	(0.048)	(0.048)
Asset weighted	-0.032	-0.024	-0.027	-0.026
	(0.039)	(0.038)	(0.038)	(0.039)

Table 9: **Out-of-sample alpha of fund portfolios obtained with neural networks**

This table reports the monthly out-of-sample alphas (in %) of the excess returns 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 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.

	FF3+MOM	FF5	FF5+MOM	FF5+MOM +LIQ
Neural network (1 layer)	0.176**	0.196**	0.201***	0.211***
	(0.073)	(0.078)	(0.077)	(0.074)
Neural network (2 layers)	0.173**	0.192**	0.198**	0.209***
	(0.077)	(0.080)	(0.080)	(0.078)
Neural network (3 layers)	0.096	0.115	0.131*	0.140*
	(0.069)	(0.073)	(0.075)	(0.075)

Table 10: **Out-of-sample alpha of portfolios using only most important predictors**

This table reports the monthly out-of-sample alphas (in %) of the excess returns net of all costs of the top-decile fund portfolios obtained with the gradient-boosting method when only a subset of fund characteristics are used to predict performance. Specifically, portfolios are obtained when only the top-2, top-3, and top-4 characteristics in terms of their importance for the gradient-boosting method are included each year. 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 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.

	FF3+MOM	FF5	FF5+MOM	FF5+MOM +LIQ
top-2 characteristics	-0.024	0.026	-0.004	-0.011
	(0.264)	(0.274)	(0.259)	(0.263)
top-3 characteristics	0.015	0.055	0.022	0.022
	(0.095)	(0.102)	(0.099)	(0.101)
top-4 characteristics	0.194**	0.254**	0.201**	0.202**
	(0.095)	(0.114)	(0.099)	(0.101)

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.

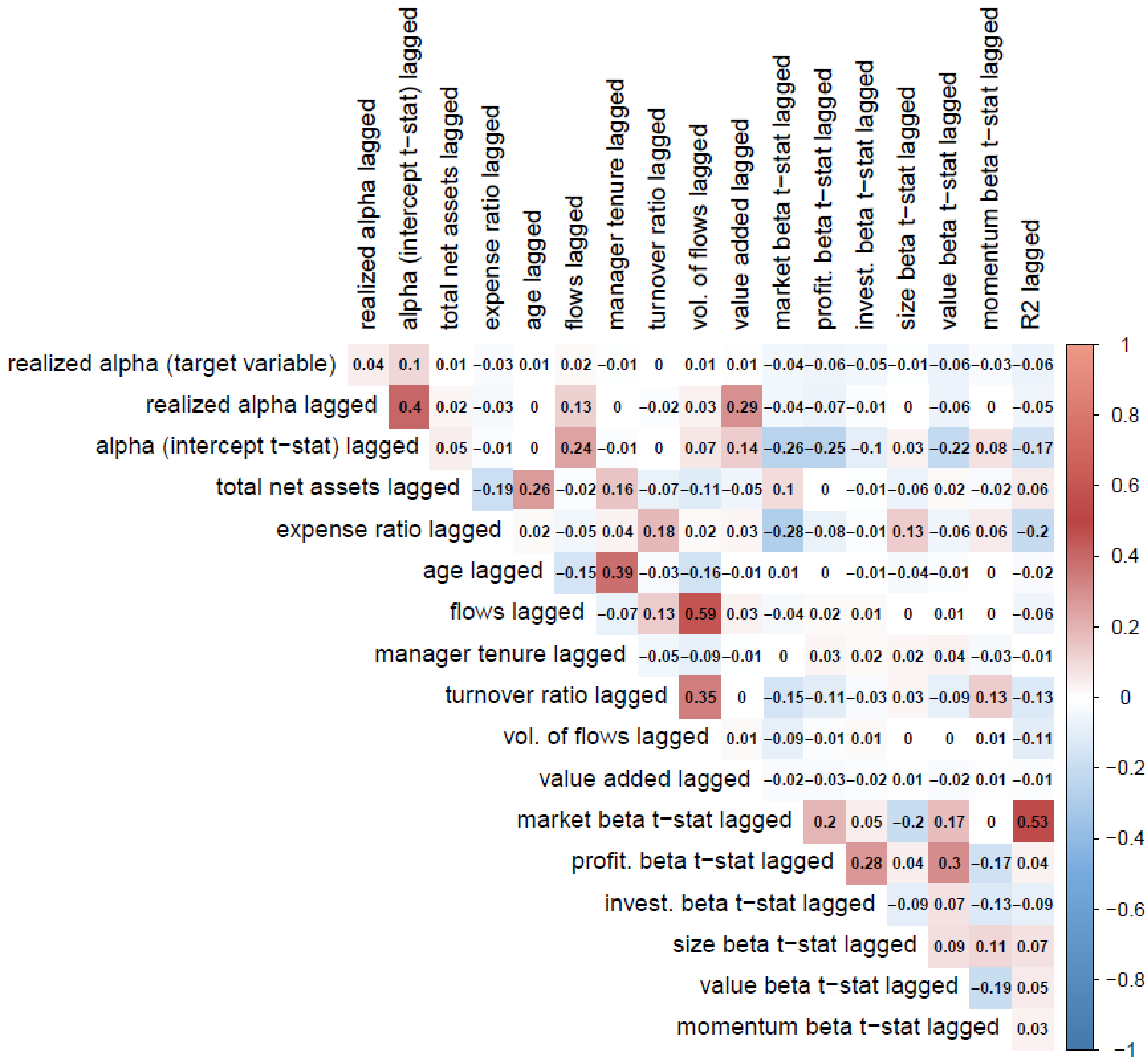

Figure 2: **Example of decision tree**

This figure plots an example of a decision tree that uses three share-class characteristics (market beta t-stat, R^2, and realized alpha) to split the sample into four categories represented by the orange leaf nodes.

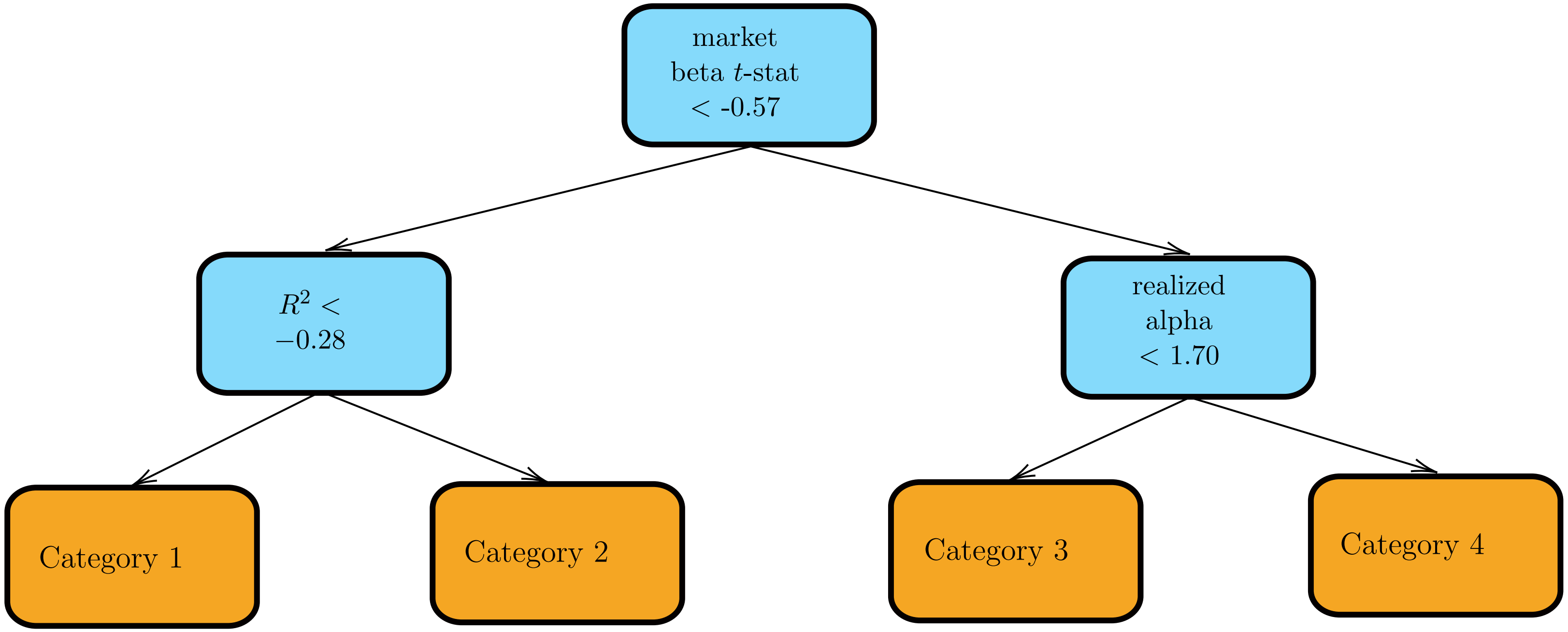

Figure 3: **Characteristic importance**

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 1980–2017 period.

	GB	RF	EN	OLS
realized alpha	86	44	5	6
alpha (intercept t−stat)	34	25	100	100
total net assets	27	29	3	7
expense ratio	12	10	57	65
age	0	0	4	7
flows	6	4	7	6
manager tenure	5	3	7	10
turnover	9	18	2	4
vol. of flows	2	1	0	0
value added	12	15	6	9
market beta (t−stat)	100	94	16	21
profit. beta (t−stat)	30	41	39	41
invest. beta (t−stat)	52	40	27	32
size beta (t−stat)	26	26	0	3
value beta (t−stat)	22	32	60	68
momentum beta (t−stat)	57	32	73	87
R2	99	100	96	100

Figure 4: **Portfolio characteristics**

This figure reports the time-series average of the top-decile portfolio characteristics. We cross-sectionally standardize the characteristics so that they have zero mean and unit standard deviation and define the top-decile portfolio characteristics at the end of each year as the equally weighted average of the fund characteristics across funds in the top-decile portfolio. The figure reports the time-series average of each standardized portfolio characteristic.

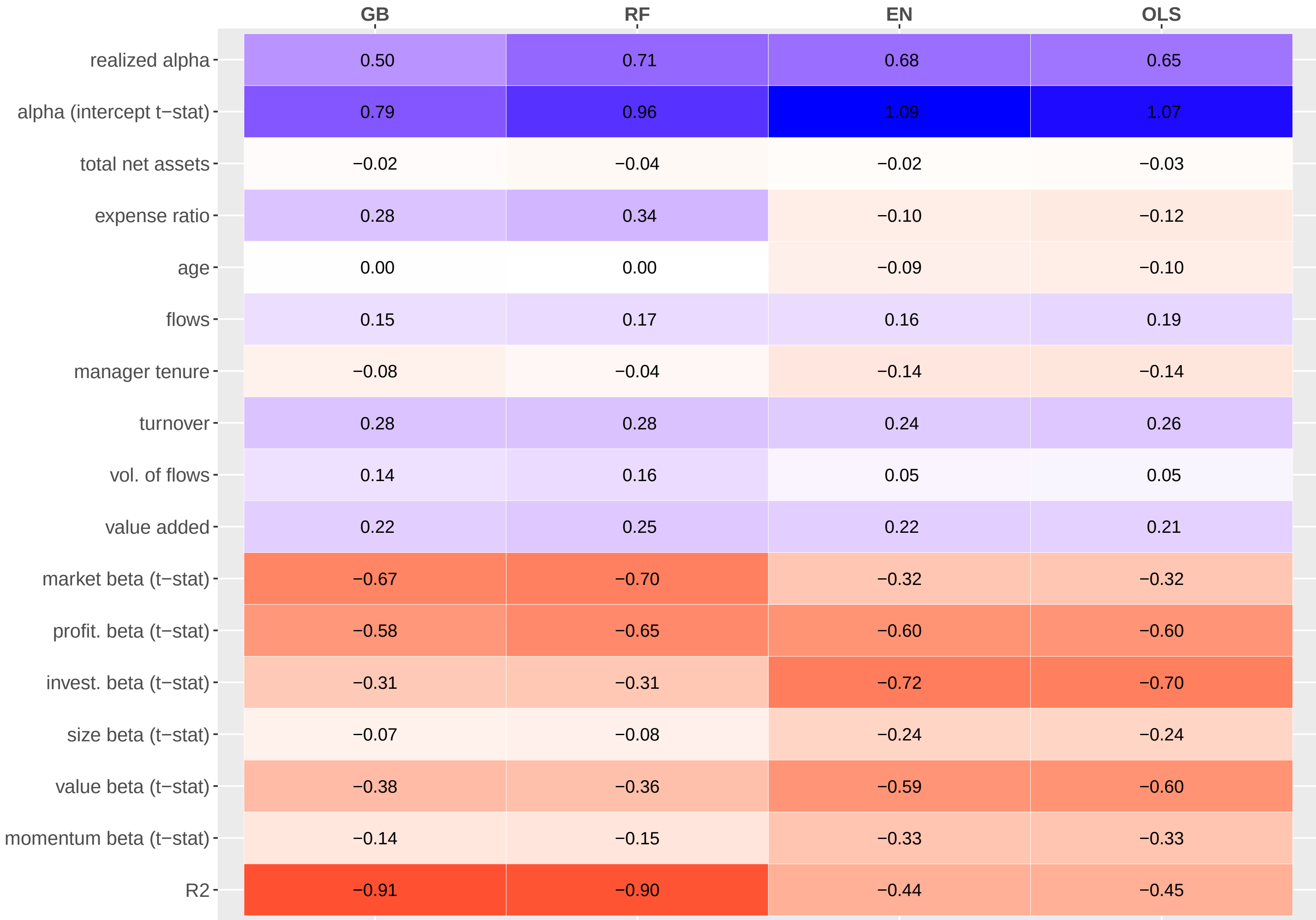

Figure 5: **Time series of variable importance for the gradient-boosting method**

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 1980–2017.

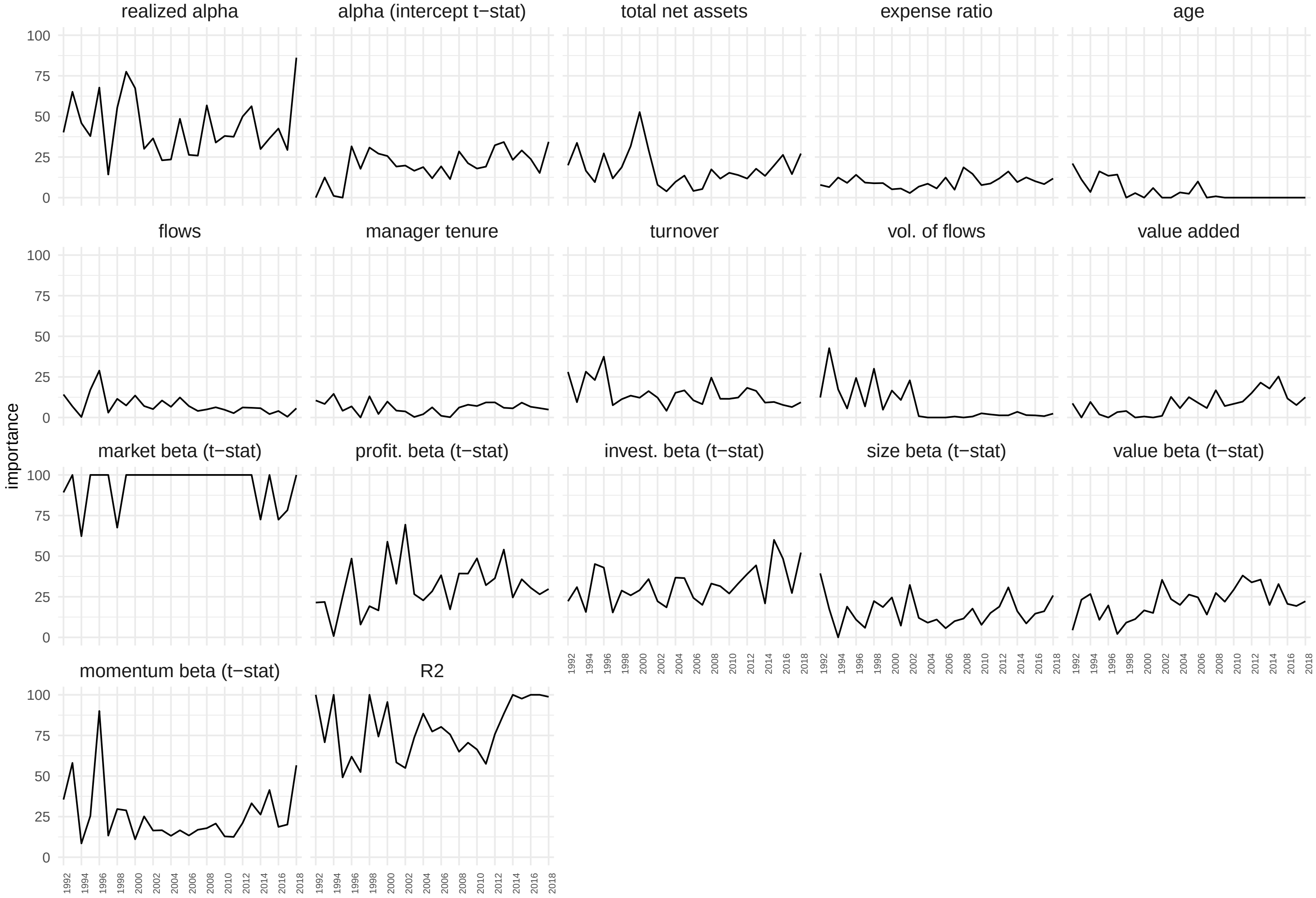

Figure 6: **Five-year rolling window alphas**

This table reports the monthly out-of-sample alphas (in %) of the excess returns net of all costs of the top-decile fund portfolios for five-year rolling-window regressions on the five Fama-French factors augmented with momentum (FF5+MOM). Portfolios are obtained with gradient boosting (GB), Ordinary Least Squares (OLS), and with two naive strategies (equally weighted and asset-weighted portfolios of all available funds).