Machine-Learning the Information Set of Mutual Fund Investors

Jung Jae Kim
Emory University

Jeong Ho (John) Kim
Florida State University

October 7, 2023

Abstract

We examine which information mutual fund investors make use of when they invest, using a machine learning method. We find that investors mostly consider fund characteristics including past flows and returns, but hardly respond to stock characteristics that a fund is holding although they are important to predict fund performance. Finally, we find that return predictability worsens if we only use the information that investors primarily consider.

1 Introduction

How investors allocate their capital within the market for mutual funds has been a long-standing question in financial economics. For a long time, a series of studies have documented that investor follows a naïve and simplistic return-chasing behavior: investors' flows in and out of mutual funds respond to past performance although it is not guaranteed to be persistent (Chevalier & Ellison, 1997; Hendricks, Patel, & Zeckhauser, 1993; Sirri & Tufano, 1998). In contrast, a growing literature argues that the flow-performance relation is a result of learning behaviors by rational investors. In a seminal paper, Berk and Green (2004) propose the rational expectation model where Bayesian agents learn a fund manager's skills of delivering positive risk-adjusted returns (alphas) using the information of past performance and reallocate their assets accordingly. According to the learning literature, when evaluating managerial skills, investors should consider any relevant information that can provide investment opportunities to give positive net alphas and either invest or divest the fund based on the information. Because the aggregate flows in and out of the fund reflect this behavior, one can infer that factors predicting future flows are important information believed by investors to give investment opportunities. In the spirit of this rationale, Berk and

Van Binsbergen (2016) and Barber, Huang, and Odean (2016) use fund flows to infer which asset pricing models investors take into consideration. However, as Berk and Van Binsbergen (2016) write, "To that end, the paper leaves as an unanswered question whether the unexplained part of flows results because investor investors use a superior, yet undiscovered risk model, or whether investors use other, non-risk-based criteria to make investment decision", few studies have investigated the relationship between fund flows and a large set of factors. This paper contributes to this literature by identifying whether potential factors that have been considered to relate to fund performance also predict fund flows. If a factor predicts fund returns (i.e., it is a useful signal of future performance), but it does not predict fund flows (i.e., it is a signal unaccounted for by investors), it would suggest that investors are leaving useful information on the table. Similarly, if a factor predicts fund flows, but does not predict fund returns, it would be puzzling as to why investors care about such fake signals.

We borrow a rich set of factors and econometric methods from recent literature on asset pricing. The literature has explored hundreds of potential factors whether they explain the cross-section of expected stock returns, bringing "Factor Zoo". However, as Harvey, Liu, and Zhu (2016) point out, data-snooping bias exists when multiple-testing the significance of each factor in the high-dimensional problem. Recently, machine learning methods such as principle components, the least absolute shrinkage and selection operator (LASSO), and neural networks have been leveraged to address the problem. (Chen, Pelger, & Zhu, 2023; Feng, Giglio, & Xiu, 2020; Freyberger, Neuhierl, & Weber, 2020; Gu, Kelly, & Xiu, 2020; Kozak, Nagel, & Santosh, 2020). These methods are also employed in the mutual fund literature, and multiple studies find that several factors have a significant impact on predicting a mutual fund's risk-adjusted returns (DeMiguel, Gil-Bazo, Nogales, & AP Santos, 2021; Kaniel, Lin, Pelger, & Van Nieuwerburgh, 2022; Li & Rossi, 2020). This finding leads us to our research question in which investors consider those factors when they invest.

We collect the following mutual fund characteristics: (1) stock characteristics based on stocks that a fund holds, (2) fund characteristics such as expense ratio, age, past flows, and momentum, and (3) family characteristics based on the management company. The methodology we adopt in this paper is the Boosted Regression Trees (BRT), which combines regression trees and boosting techniques. BRT has several advantages compared to the standard statistical method, e.g., the ordinary least squares (OLS). BRT can estimate the non-linear relation between predictors and response variables and also consider complex interactions among predictors. In addition, BRT works well in a high-dimensional problem and has been proven to have a decent predictive performance in various fields. Finally, the interpretability of the BRT can be easily achieved since it automatically performs a variable selection and computes a relative importance measure for each factor.

We start by presenting which factors are important to predict fund flows and risk-adjusted returns using the relative importance measure. We find that the majority of the factors that are important to predict fund flows are fund characteristics such as lagged flows, lagged returns, expense ratio, turnover ratio, and fund age, but the importance of stock characteristics is fairly low. In contrast, most of the stock characteristics are significant in predicting risk-adjusted returns. Consequently, it can be inferred that investors hardly consider stock characteristics although stock characteristics are important to predict risk-adjusted returns.

Next, we assess the credibility of our model by computing the out-of-sample R^2. If our model correctly estimates the relationship between the factors and fund flows, it should forecast out-of-sample future flows with the same factors used in the model. The averages out-of-sample R^2 of the BRT are from 15.11% to 23.43%, whereas those of the OLS is negative. This result confirms that the BRT can handle over-fitting risks in a high-dimensional problem and have a more stable predictive performance than the OLS.

Finally, we examine the fund return predictability of the model when we exclude some factors that investors do not respond to. We restrict the predictor space to the factors that are important to predict fund flows from the highest where the sum of the importance measure is 90%, 75%, and 50%. Then we construct a long-short portfolio based on the BRT predicted returns and find that the risk-adjusted return of the long-short portfolio monotonically falls as the predictor space is restricted.

This paper is organized as follows. Section 2 describes the data, fund flows and risk-adjusted returns being predicted, and a rich set of factors as predictors. Section 3 presents a pre-analysis using univariate sorts prior to the main analysis using the BRT. Section 4 introduces our model, BRT method, and how to implement it. Section 5 shows the result of our main analysis, and Section 6 concludes.

2 Data

Our data come from the CRSP Mutual Fund database and Thomson Reuters Mutual Fund Holdings database. Following the code of Doshi, Elkamhi, and Simutin (2015), we restrict our sample to domestic actively-managed equity mutual funds using CRSP funds' investment objectives code. Specifically, we exclude international, municipal bonds, bonds and preferred, and index funds. Our monthly data set includes 387,592 observations for a total of 3,156 mutual funds and 1,157 mutual funds by month on average. Our sample period is from January 1990 to November 2018 since the total net assets of mutual funds are reported monthly after 1990.

2.1 Fund Flow and Performance

Our main objects to predict with the information set are mutual fund flow and performance. Following van Binsbergen, Kim, and Kim (2021), we measure fund flow F over a horizon of length T as

$$F^T_{it+1} = \frac{AUM_{it+T} - AUM_{it}(1 + R_{it+T})}{AUM_{it}(1 + R_{it+T})} \tag{1}$$

where AUM_{it} and R_{it} are the asset under management and gross return of fund i at the end of month t, respectively. Throughout our analysis, we focus on $T = 1, 3, 6,$ and 12.

We measure fund performance with two different risk-adjusted returns. The first measure is the excess return defined as

$$R_{it+1}^{excess} = R_{it+1} - r_t^f \tag{2}$$

where r_t^f is the risk-free rate at the end of the month t. The second measure is the abnormal return relative to the CAPM. To get the abnormal return, we first estimate factor coefficients over the prior 36 months:

$$R_{it-35:t}^{excess} = \alpha_i + MKT_{t-35:t}\hat{\beta}_{it}$$

where MKT_t is the excess return on the market portfolio. Then the abnormal return relative to the CAPM can be computed as

$$R_{it+1}^{CAPM} = R_{it+1}^{excess} - MKT_t\hat{\beta}_{it} \tag{3}$$

Table 1 provides the summary statistics of our measures of flow and performance.

2.2 Stock, Fund, and Family Characteristics

We compute the stock characteristics of a mutual fund through weighted averages by the dollar amount of the fund's holding of stocks. Note that our sample is monthly frequency, whereas fund holdings data are quarterly frequency. Therefore, we impute monthly holdings data with the latest available holding data for each month. Stock characteristics are from Freyberger et al. (2020), covering 61 characteristics. Table 2 shows the characteristics by six categories.

We also construct 25 fund characteristics and 24 family characteristics shown in Table 3. In the fund momentum, fund 3-factor alpha and 4-factor alpha are the abnormal returns relative to Fama and French (1992) and Carhart (1997), respectively. The lagged fund flows are computed as equation (1). Following Kaniel et al. (2022), the fund family is identified by the management company code, and the characteristics are weighted by the total net assets of all funds in the family, excluding the fund itself.

Therefore, we have a total of 110 regressors as the information set and standardize both covariates and predicted variables cross-sectionally.

3 Pre-Analysis: Univariate Sorts

As a preliminary analysis prior to the main analysis, we test whether fund flows can be significantly predicted based on the value of each characteristic. We sort mutual funds into deciles based on the value of the characteristics and conduct a t-test of the fund flow difference between the top decile and bottom decile. Specifically, for each month t, mutual funds are sorted into deciles based on each value of x_{it}, out of 110 regressors. Then we compute the equal-weighted and value-weighted average of F^T_{it+1} for each decile and conduct a t-test of the difference between two extreme deciles using Newey-West standard errors with 12 lags. Note that this pre-analysis shows a simple univariate relation between regressors and fund flows as it ignores any non-linear relation or interaction effects between regressors.

Table 4 shows the t-test results for each of the 110 characteristics. The left panel shows the equal-weighted averages difference and the right shows the value-weighted averages difference between top and bottom deciles. Each panel reports the results for F^T_{it+1}, where $T = 1, 3, 6$, and 12. For equal and value-weighted flows, past fund flows are the most significant characteristics that predict 1-month inflows of 5.43% - 8.01%, where t-statistics are 17.76 - 29.03. The results are similar when predicting fund flows when $T = 3, 6$, and 12. Followed by past fund flows, fund momentum is an important characteristic to predict inflows to the funds, and Fama-French 3-factor momentum is the most significant among them. Other fund characteristics such as *exp_ratio*, *age*, and *log_real_tna* deliver outflows to the funds at a significant level. Most stock characteristics are insignificant to predict flows except past returns and *rel_to_high_price*. Finally, family characteristics are also important to predict flows, and the direction is similar to the counterpart of fund characteristics.

These results imply that investors mostly respond to the fund and family characteristics but hardly respond to stock characteristics. However, this pre-analysis only shows univariate sorts, and we need careful multivariate analysis to deeper understand investors' responses.

4 Method

Investors make use of the information set they have to make an investment decision. As an econometrician, we do not directly observe which information investors make use of and only observe aggregate fund flows ex-post. With a large number of characteristics, we then estimate which characteristics are important to predict aggregate fund flows, i.e., we can infer that investors respond to those characteristics on average when they invest. Formally, consider the following predictive regression problem:

$$F^T_{it+1} = g(\mathcal{I}_{it}) + \epsilon_{it+1} \tag{4}$$

where F^T_{it+1} denotes a fund flow defined in (1), $\mathcal{I}_{it}$ denotes a set of regressors at month t, and $g(\cdot)$ is a unknown function to be estimated. A natural candidate of $g(\cdot)$ is a linear function and is estimated by the ordinary least squares (OLS). However, OLS is vulnerable to over-fitting when the problem is high-dimensional and cannot consider complex non-linearities between fund characteristics and flow. To overcome the limitations of OLS, we use Boosted Regression Trees (BRT), similar to Gu et al. (2020) and Li and Rossi (2020).

4.1 Boosted Regression Trees

BRT is a machine learning algorithm that combines regression trees and boosting techniques to perform regression tasks. Regression trees are a non-parametric supervised learning method allowing multi-way interactions between covariates. The method works by recursively partitioning the predictor space into smaller subsets using a tree structure, where each node of the tree represents a split in the data based on a selected feature and threshold value. The splitting process is based on minimizing the sum of squared errors between the predicted and actual values of the response variable. This sequential branching slices the space of predictors into rectangular partitions, and approximates the unknown function $g(\cdot)$ with the average value of the outcome variable within each partition. Formally, a regression tree can be defined by

$$g(x) = \sum_{j=1}^{J} w_j \mathbf{I}(x \in R_j) \tag{5}$$

where $R_j, j = 1, ..., J$ is the subset of the predictor space specified by the j'th node, $\mathbf{I}(\cdot)$ is an indicator function, and w_j is the predicted output for that node. We can easily estimate w_j as the average value in each partition R_j:

$$w_j = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} y_{it} \mathbf{I}(x_{it} \in R_j)}{\sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{I}(x_{it} \in R_j)}$$

To find optimal partitioned regions R_j, we need to minimize the following loss:

$$\mathcal{L}((R_j, w_j) : j = 1, ..., J) = \sum_{j=1}^{J} \sum_{x_{it} \in R_j} (y_{it} - w_j)^2$$

Due to the discrete tree structure, this loss function is not differentiable and finding the optimal partitions is NP-complete (Laurent & Rivest, 1976). Therefore, we use a greedy procedure, in which we iteratively grow the tree one node at a time. The procedure first considers a partitioning predictor p and a split threshold s, so the partitions are constructed as

$$R_1(p, s) = \{X | X_p \leq s\} \text{ and } R_2(p, s) = \{X | X_p > s\}$$

Then we choose p and s by solving

$$\min_{p,s} \left[\min_{w_1} \sum_{x_{it} \in R_1(p,s)} (y_{it} - w_1)^2 + \min_{w_2} \sum_{x_{it} \in R_2(p,s)} (y_{it} - w_2)^2 \right],$$

$$w_1 = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} y_{it} \mathbf{I}(x_{it} \in R_1)}{\sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{I}(x_{it} \in R_1)} \text{ and } w_2 = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} y_{it} \mathbf{I}(x_{it} \in R_2)}{\sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{I}(x_{it} \in R_2)}, \text{ for a given } p \text{ and } s$$

Given the optimal $R_1(p, s)$ and $R_2(p, s)$, we repeat the same splitting process for each of the partitions.

Note that the method performs automatic variable selection since predictors that are never used to split the predictor space do not affect the performance of the model. These non-parametric and sequential splits of the predictor space are likely to capture the non-linear relation between predictors and predicted variables, but over-fitting can be still problematic because fewer and fewer observations are used as trees grow further. To address this problem, we use the boosting technique, which is ensembles of trees.

Boosting is a method of building an ensemble of regression trees, where each subsequent tree is trained to correct the errors of the previous one. Suppose $\mathcal{T}(x; \{R_j, w_j\}_{j=1}^{J})$ is a regression tree defined in equation (5). Then boosted regression trees are the sum of regression trees:

$$g_B(x) = \sum_{b=1}^{B} \mathcal{T}_b(x; \{R_{b,j}, w_{b,j}\}_{j=1}^{J}) \tag{6}$$

where B is the number of boosting iterations and $\mathcal{T}_b(x; \{R_{b,j}, w_{b,j}\}_{j=1}^{J})$ is the regression tree in the b-th iteration. Let us define the error after $b-1$ boosting iterations:

$$e_{it,b-1} = y_t - g_{b-1}(x_{it})$$

Then the subsequent tree at step b can be estimated by solving

$$\min_{\{R_{b,j}, w_{b,j}\}_{j=1}^{J}} \sum_{i=1}^{N} \sum_{t=1}^{T} \left[e_{it,b-1} - \mathcal{T}_b(x; \{R_{b,j}, w_{b,j}\}_{j=1}^{J}) \right]^2,$$

$$w_{b,j} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} e_{it,b-1} \mathbf{I}(x_{it} \in R_{b,j})}{\sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{I}(x_{it} \in R_{b,j})}, \text{ for a given } R_{b,j}$$

4.2 Relative Importance Measure

As we discussed above, the BRT automatically selects characteristics as a tree grows. Therefore, we can see how important each characteristic is relative to other characteristics by summing up the empirical gains of each node where the characteristic is selected. Breiman, Friedman, Stone, and Olshen (1984) proposes a relative importance measure for each predictor variable X_l. For a single regression tree $\mathcal{T}$, the measure is defined as

$$I_l(\mathcal{T}) = \sum_{j=1}^{J-1} G_j \mathbf{I}(x_j = X_l) \tag{7}$$

where G_j is the reduction in squared empirical error at node j and x_j is the regressor selected at node j. If a regressor is selected more frequently for splitting and the gain is bigger, the measure is larger. On the other hand, if a regressor is never used for splitting, the measure is zero. By averaging over the number of boosted trees, we can get a more reliable importance measure:

$$I_l = \frac{1}{B} \sum_{b=1}^{B} I_l(\mathcal{T}_\lfloor)$$

Since the measure shows relative importance, we normalize the relative importance measure to be the total sum of 1.

4.3 Out-of-Sample R^2

If our method well uncovers the relationship between the characteristics and future flows by estimating the predictive regression model (4), the estimated model should be able to forecast flows using the same characteristics in the next period. Therefore, we can check the performance of the method by measuring out-of-sample R^2. Suppose we estimate the equation (4) with the BRT:

$$F_{it+1}^{T} = \hat{g}(\mathcal{I}_{it})$$

where $\hat{g}(\cdot)$ is the estimated function by the BRT. Then the model forecasts flows at $t+2$ using the information at $t+1$:

$$\widehat{F}^T_{it+2} = \hat{g}(\mathcal{I}_{it+1})$$

We can calculate the out-of-sample R^2 as follows

$$R^2_{\text{oos},t+1} = 1 - \frac{\sum_{i=1}^N \left(F^T_{it+2} - \widehat{F}^T_{it+2}\right)^2}{\sum_{i=1}^N \left(F^T_{it+2} - \bar{F}^T_{it+2}\right)^2}, \tag{8}$$

As we will use 1-month rolling windows, $R_{\text{oos},t+1}$ pools prediction errors across mutual funds at $t+1$, and we can see how it varies over time. When we estimate the equation (4) with the OLS, the model hardly forecasts flows, as most of $R_{\text{oos},t+1}$ are negative of around -20%. This confirms that the OLS is an inappropriate method when the predictor space is high-dimensional due to the over-fitting risk.

4.4 Implementation

For the implementation of the BRT model, we mainly follow Li and Rossi (2020)'s one-month rolling window specification, but we adopt two major modifications to their method.

First, we set a validation period to find the optimal number of boosting iterations. Specifically, we estimate the equation (4) by the BRT at each month t, evaluate the estimated model with the validation period at $t+1$ to find the optimal number of boosting iterations, and finally calculate R^2_{oos} at $t+2$. As the number of boosting iterations increases, the mean squared error in the training sample usually decreases since the boosting targets the errors of the previous tree. Therefore, too many boosting iterations may be exposed to the over-fitting risk. To address this problem, we stop the boosting iterations when the mean squared error evaluated at the validation sample stop decreasing. Actually, this modification significantly reduces the number of a negative R^2_{oos}, whereas simply setting the number of boosting iterations to 100 as in Li and Rossi (2020) produces many negative R^2_{oos}. We will discuss this more extensively later.

Second, we use the Huber robust objective function instead of the squared loss function when estimating the BRT model, similar to Gu et al. (2020). The Huber robust objective function is defined as

$$\mathcal{L}_H(\mathcal{T}(x)) = \sum_{t=1}^{T} H(y_t - \mathcal{T}(x), \xi),$$

where

$$H(x;\xi) = \begin{cases} x^2, & \text{if} \quad |x| \leq \xi \\ 2\xi|x| - \xi^2, & \text{if} \quad |x| > \xi \end{cases}$$

The Huber loss function is well-known in the machine learning literature for producing more stable predictions than the squared loss function in the presence of outliers. Since outliers are known to be common in financial returns and characteristics, we adopt the Huber loss function.

5 Results

5.1 Which Information Matters to Investors

In this section, we start by presenting the relative importance measure when predicting future fund flows. We rank the characteristics from the highest importance to the lowest and infer that investors make use of the highest- and lowest-ranked characteristic the most and the least, respectively. Since we estimate the model with one-month rolling windows, we have relative importance measures for every month in our sample period. Following Gu et al. (2020), we report the relative importance measure by averaging across the time. Figure 1 shows the relative importance measure for each characteristics when predicting F^T_{it+1} for $T = 1, 2, 3$, and 12. The result indicates that the 10 most important predictors are all fund characteristics for all F^T_{it+1}, including past flows, *log_real_tna*, *age*, *turn_ratio*, and long-term fund momentum. Especially, the importance of *flow_1_0* is greater than 10%, and the importance of *flow_2_1*, and *flow_12_2* is greater or similar to 5% for all F_{it+1}^{T}. Interestingly, long-term fund momentum turns out to be more important than short-term fund momentum, which implies that investors are not myopic but consider the fund's long-term performance when they decide to invest. The importance of stock characteristics is evenly dispersed around 1% for all F_{it+1}^{T}. Among them, the most important stock characteristic is *d_dgm_dsales*, which is in the profitability category. This highlights the importance of multivariate analysis as we recall that past returns and *rel_to_high_price* are significant in the univariate sorts. The least important characteristics are family characteristics, which indicates that investors hardly take the management company of the fund into account. Overall, fund characteristics including past flows and returns are the most important predictors as expected from previous research (Coval & Stafford, 2007), and stock and family characteristics are less important predictors.

Next, we estimate the model to predict future fund performance defined in (2) and (3) and check which characteristics are important. The left plot in Figure 2 and Figure 3 show the relative importance measures when predicting future excess returns and abnormal returns. Contrary to the previous result, many stock characteristics are ranked high in both figures. This result coincides with Li and Rossi (2020) who find that fund performance is largely exposed to 40-50 stock characteristics. Fund characteristics such as *exp_ratio*, *turn_ratio*, *log_real_tna*, and fund momentum turn are placed in the middle of stock characteristics, but family characteristics turn out to be less important.

We conclude that investors mostly respond to fund characteristics, but less consider stock characteristics although they are significant to predict future fund performance. We leave identifying the mechanism of the investor's behavior as a future study.

5.2 Model Evaluation

The reason why we leverage the BRT to estimate the model is that the OLS usually misleads the relationship between future flows and predictors due to the over-fitting problem in a high-dimensional setting. Then the BRT should be free of the over-fitting risk to make the results credible. Figure 4 shows the out-of-sample R^2 over time for all F^T_{it+1}. The green line is $R^2_{\rm oos}$ of the BRT with the validation sample to set the optimal number of boosting iterations, red is of the BRT with setting the number to 100, and the blue is of the OLS. For all F^T_{it+1}, the majority of $R^2_{\rm oos}$ of the OLS is negative, which indicates that the over-fitting problem is serious. For 1-month future flow, several $R^2_{\rm oos}$ of the BRT without the validation is negative, whereas most of $R^2_{\rm oos}$ of the BRT with the validation is greater than 0. This result implies that too many boosting iterations also result in the over-fitting problem. For 3,6, and 12-month future flows, both red and blue lines show a similar pattern where the green is slightly below the red but more stable with respect to the over-fitting.

Table 5 shows the average, minimum, and maximum of $R^2_{\rm oos}$, and the proportion of the negative value across the time for each model. Not surprisingly, the average of $R^2_{\rm oos}$ of OLS is from -16.94% to -23.38%, and the proportion of the negative value is all above 80%. Now we focus on the 1-month future flow. The mean of $R^2_{\rm oos}$ of BRT without validation is 6.23% and the proportion is 22.46% while BRT with validation shows 15.11% and the proportion drops dramatically to 1.8%. Moreover, the minimum of the former is -31.98%, whereas the latter is only -2.41%. Therefore, using the validation sample to set the optimal number of boosting iterations helps to address the over-fitting problem and produce stable predictions for 1-month future flow. The proportion of negative values also improves for 3,6, and 12-month future flows, but the averages slightly decrease. This might be because we use the information at $t+1$ for validation, and only use the trained model with the information at t to forecast the value at $t+2$. Although there is a disadvantage due to the information loss, stable predictions without the over-fitting risk should be emphasized for credible results.

5.3 Predicting Fund Returns based on Investor's Information Set

In this section, we construct a long-short portfolio based on the predicted fund returns similar to Li and Rossi (2020) and Kaniel et al. (2022), but the predictor space is restricted to the characteristics that are important to predict 3-month future flow from the highest where the sum of importance is 90%, 75%, and 50%. The rationale behind this restriction is to test how important the characteristics investors do not consider are important to predict future performance. The right plot in Figure 2 and Figure 3 is the relative importance measure to predict future excess and abnormal returns when restricting the predictor space to the sum of the measure for 3-month future flow to be 50%. The number of predictors is only 19 out of 110 regressors, and there are only 5 stock characteristics: *rel_to_high_price*, *d_ceq*, *suv*, *noa*, and *d_dgm_dsales*.

With the restricted predictor space, we sort funds into deciles based on the predicted excess and abnormal returns. For each decile, we compute the average of realized excess and abnormal returns with either equal weights or value weights by the predicted value. We then construct a long-short portfolio by holding the funds in the top decile and selling the funds in the bottom decile. Table 6 reports the excess and abnormal returns of the long-short portfolio and their t-statistics computed using Newey-West standard errors with 12 lags. For both equal- and value-weighted long-short portfolios, the average of the excess returns monotonically decreases from 0.57-0.58% to 0.46-0.47% as the predictor space is restricted further. The average of the abnormal returns is not exactly a monotone decrease, but it decreases from 0.5% using all predictors to 0.46-0.47% using the 19 predictors.

6 Conclusions

In this paper, we shed light on mutual fund investors' responsiveness to the information set by leveraging the machine learning method. We divide the information set into three groups: (1) stock characteristics, (2) fund characteristics, and (3) family characteristics. We show that important characteristics to predict future flows are mostly fund characteristics, whereas stock characteristics are far less important even though they are important to predict future fund performance. If we restrict the predictor space to the characteristics ranked in order from the highest importance to predict future flows where the sum is 90%, 75%, and 50%, the performance of the long-short portfolio based on the predicted fund performance decreases monotonically. We also confirm that the predictability of our model is stable over time, as evidenced by the out-of-sample R^2.

The natural next step for future research is to identify the mechanism of the investor's behavior. The possible reason why the investor does not respond to the stock characteristics might be the costly information acquisition, and the recent advance of rational inattention literature can help model the investor's learning behavior. The other possible strand of future research is about a policy implication of our results. It might be socially desirable if mutual fund managers disclose the stock characteristics they hold so that the information is easily accessible to the investor.

References

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

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

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

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. (1984). *Classification and regression trees*. CRC press.

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

Chen, L., Pelger, M., & Zhu, J. (2023). Deep learning in asset pricing. *Management Science*.

Chevalier, J., & Ellison, G. (1997). Risk taking by mutual funds as a response to incentives. *Journal of Political Economy*, *105*(6), 1167–1200.

Coval, J., & Stafford, E. (2007). Asset fire sales (and purchases) in equity markets. *Journal of Financial Economics*, *86*(2), 479–512.

DeMiguel, V., Gil-Bazo, J., Nogales, F. J., & AP Santos, A. (2021). Machine learning and fund characteristics help to select mutual funds with positive alpha. *Proceedings of Paris December 2021 Finance Meeting EUROFIDAI - ESSEC, Available at SSRN 3768753*.

Doshi, H., Elkamhi, R., & Simutin, M. (2015). Managerial activeness and mutual fund performance. *The Review of Asset Pricing Studies*, *5*(2), 156–184.

Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. *the Journal of Finance*, *47*(2), 427–465.

Feng, G., Giglio, S., & Xiu, D. (2020). Taming the factor zoo: A test of new factors. *The Journal of Finance*, *75*(3), 1327–1370.

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

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223–2273.

Harvey, C. R., Liu, Y., & Zhu, H. (2016). … and the cross-section of expected returns. *The Review of Financial Studies*, *29*(1), 5–68.

Hendricks, D., Patel, J., & Zeckhauser, R. (1993). Hot hands in mutual funds: Short-run persistence of relative performance, 1974–1988. *The Journal of finance*, *48*(1), 93–130.

Kaniel, R., Lin, Z., Pelger, M., & Van Nieuwerburgh, S. (2022). *Machine-learning the skill of mutual fund managers* (Tech. Rep.). National Bureau of Economic Research.

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

Laurent, H., & Rivest, R. L. (1976). Constructing optimal binary decision trees is np-complete. *Information processing letters*, *5*(1), 15–17.

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

Sirri, E. R., & Tufano, P. (1998). Costly search and mutual fund flows. *The Journal of Finance*, *53*(5), 1589–1622.

van Binsbergen, J. H., Kim, J. H. J., & Kim, S. (2021). Capital allocation and the market for mutual funds: Inspecting the mechanism. *Available at SSRN 3462749*.

Table 1: Summary statistics of fund flow and performance

Statistic	N	Mean	Median	Std.Dev	Min	5%	95%	Max
Flow_1month	387,174	0.0002	-0.0048	0.0582	-0.4705	-0.0532	0.0636	2.7285
Flow_3month	385,362	0.0043	-0.0160	0.1725	-0.7123	-0.1352	0.1836	13.4980
Flow_6month	381,207	0.0185	-0.0325	0.3822	-0.7162	-0.2368	0.3778	73.5192
Flow_12month	371,774	0.0679	-0.0640	0.8502	-0.7944	-0.3883	0.7984	117.8057
Excess reutrn	387,183	0.0061	0.0100	0.0510	-0.3113	-0.0824	0.0797	0.4026
CAPM alpha	387,183	0.0003	-0.0002	0.0240	-0.2443	-0.0350	0.0367	0.3755

This table reports summary statistics of the fund flows and risk-adjusted returns. The sample period is from 1990/01 to 2018/11

Table 2: Stock Characteristics by category

	Past Returns	
(1)	r1_0	Return 1 month before prediction
(2)	r6_2	Return from 6 to 2 month before prediction
(3)	r12_2	Return from 12 to 2 month before prediction
(4)	r12_7	Return from 12 to 7 month before prediction
(5)	r36_13	Return from 36 to 13 month before prediction
	Investment	
(6)	Investment	% change in AT
(7)	dCEQ	% change in BE
(8)	dPI2A	Change in PP&E and inventory over lagged AT
(9)	IVC	Change in inventory over average AT
(10)	NOA	Net-operating assets over lagged AT
	Profitability	
(11)	ATO	Sales to lagged net operating assets
(12)	CTO	Sales to lagged total assets
(13)	d(dGM-dSales)	d(% change in gross margin and % change in sales)
(14)	EPS	Earnings per share
(15)	IPM	Pretax income over sales
(16)	PCM	Sales minus costs of goods sold to sales
(17)	PM	OI after depreciation over sales
(18)	PM_adj	Profit margin - mean PM in Fama-French 48 industry
(19)	Prof	Gross profitability over BE
(20)	RNA	OI after depreciation to lagged net operating assets
(21)	ROA	Income before extraordinary items to lagged AT
(22)	ROC	Size + longterm debt - total assets to cash
(23)	ROE	Income before extraordinary items to lagged BE
(24)	ROIC	Return on invested capital
(25)	S2C	Sales to cash
(26)	SAT	Sales to total assets
(27)	SAT_adj	SAT - mean SAT in Fama-French 48 industry
	Intangibles	
(28)	AOA	Absolute value of operating accurals
(29)	OL	Costs of goods solds + SG&A to total assets
(30)	Tan	Tangibility
(31)	OA	Operating accurals
	Value	
(32)	A2ME	Total assets to Size
(33)	BEME	Book to market ratio
(34)	BEME_adj	BEME - mean BEME in Fama-French 48 industry
(35)	C	Cash to AT
(36)	C2D	Cash flow to total liabilities
(37)	dSO	Log change in split-adjusted shares outstanding
(38)	Debt2P	Total debt to Size
(39)	E2P	Income before extraordinary items to Size
(40)	Free CF	Free cash flow to BE
(41)	LDP	Trailing 12-months dividens to price
(42)	NOP	Net payouts to Size
(43)	O2P	Operating payouts to market cap
(44)	Q	Tobin's Q
(45)	S2P	Sales to price
(46)	Sales_g	Sales growth
	Trading frictions	
(47)	AT	Total assets
(48)	Beta	Correlation x ratio of vols
(49)	Beta daily	CAPM beta using daily returns
(50)	DTO	De-trended Turnover - market Turnover
(51)	Idio vol	Idio vol of Fama-French 3 factor model
(52)	LME	Price times shares outstanding
(53)	LME_adj	Size - mean size in Fama-French 48 industry
(54)	Lturnover	Last month's volume to shares outstanding
(55)	Rel-to_high_price	Price to 52 week high price
(56)	Ret_max	Maximum daily return
(57)	Spread	Average daily bid-ask spread
(58)	Std turnover	Standard deviation of daily turnover
(59)	Std volume	Standard deviation of daily volume
(60)	SUV	Standard unexplained volume
(61)	Total vol	Standard deviation of daily returns

This table report 61 stock characteristics from Freyberger et al. (2020) sorted into six categories.

Table 3: Fund and family characteristics by category

	Fund momentum	
(1)	F_r1_0	Fund return 1 month before prediction
(2)	F_r2_1	Fund return from 2 to 1 month before prediction
(3)	F_r12_2	Fund return from 12 to 2 month before prediction
(4)	F_excess_r1_0	Fund excess return 1 month before prediction
(5)	F_excess_r2_1	Fund excess return from 2 to 1 month before prediction
(6)	F_excess_r12_2	Fund excess return from 12 to 2 month before prediction
(7)	F_mar1_0	Fund market-adjusted return 1 month before prediction
(8)	F_mar2_1	Fund market-adjusted return from 2 to 1 month before prediction
(9)	F_mar12_2	Fund market-adjusted return from 12 to 2 month before prediction
(10)	F_capm1_0	Fund CAPM alpha 1 month before prediction
(11)	F_capm2_1	Fund CAPM alpha from 2 to 1 month before prediction
(12)	F_capm12_2	Fund CAPM alpha from 12 to 2 month before prediction
(13)	F_3F_alpha_1_0	Fund 3-factor alpha 1 month before prediction
(14)	F_3F_alpha_2_1	Fund 3-factor alpha from 2 to 1 month before prediction
(15)	F_3F_alpha_12_2	Fund 3-factor alpha from 12 to 2 month before prediction
(16)	F_4F_alpha_1_0	Fund 4-factor alpha 1 month before prediction
(17)	F_4F_alpha_2_1	Fund 4-factor alpha from 2 to 1 month before prediction
(18)	F_4F_alpha_12_2	Fund 4-factor alpha from 12 to 2 month before prediction
	Fund flow	
(19)	Flow_1_0	Fund flow 1 month before prediction
(20)	Flow_2_1	Fund flow from 2 to 1 month before prediction
(21)	Flow_12_2	Fund flow from 12 to 2 month before prediction
	Fund characteristics	
(22)	Age	Fund age
(23)	Log_real_TNA	Log of inflation-adjusted total net assets
(24)	Exp_ratio	Fund expense ratio
(25)	Turnover_ratio	Fund turnover ratio

	Fund family characteristics	
(1)	family_ret_1_0	
(2)	family_ret_2_1	
(3)	family_ret_12_2	
(4)	family_excess_ret_1_0	
(5)	family_excess_ret_2_1	
(6)	family_excess_ret_12_2	
(7)	family_MAR_1_0	
(8)	family_MAR_2_1	
(9)	family_MAR_12_2	
(10)	family_CAPM_1_0	
(11)	family_CAPM_2_1	
(12)	family_CAPM_12_2	Fund-level counter parts weighted by TNA in family
(13)	family_3F_alpha_1_0	
(14)	family_3F_alpha_2_1	
(15)	family_3F_alpha_12_2	
(16)	family_4F_alpha_1_0	
(17)	family_4F_alpha_2_1	
(18)	family_4F_alpha_12_2	
(19)	family_flow_1_0	
(20)	family_flow_2_1	
(21)	family_flow_12-2	
(22)	family_age	
(23)	family_log_real_tna	
(24)	family_no	Number of funds in family

This table shows 25 fund characteristics and 24 family characteristics

Table 4: Univariate analysis of mutual fund flows

	Equal Weighted								Value Weighted							
Characteristics	flow_1month	t-stat	flow_3month	t-stat	flow_6month	t-stat	flow_12month	t-stat	flow_1month	t-stat	flow_3month	t-stat	flow_6month	t-stat	flow_12month	t-stat
exp_ratio	-0.0043	-3.67	-0.0124	-3.30	-0.0234	-2.84	-0.0455	-2.23	-0.0059	-4.26	-0.0165	-3.80	-0.0322	-3.55	-0.0705	-3.30
turn_ratio	-0.0029	-1.36	-0.0051	-0.59	-0.0006	-0.03	0.0295	0.36	-0.0023	-0.93	-0.0051	-0.57	-0.0022	-0.09	0.0130	0.22
age	-0.0167	-9.16	-0.0562	-6.93	-0.1230	-5.41	-0.2781	-4.99	-0.0161	-7.70	-0.0540	-6.05	-0.1177	-4.76	-0.2667	-4.48
log_real_tna	-0.0032	-2.57	-0.0221	-3.69	-0.0749	-4.13	-0.2594	-4.61	-0.0028	-2.11	-0.0185	-3.60	-0.0648	-4.40	-0.2290	-5.27
flow_lag1	0.0602	29.03	0.1873	25.71	0.3668	20.40	0.6974	17.21	0.0801	18.87	0.2484	16.25	0.4974	13.44	0.9688	11.90
flow_lag2_lag1	0.0589	27.20	0.1774	23.57	0.3455	19.95	0.6551	17.86	0.0725	17.76	0.2230	15.68	0.4459	12.77	0.8758	11.93
flow_lag12_lag2	0.0543	22.17	0.1572	19.68	0.3030	14.26	0.5590	12.21	0.0647	19.09	0.1877	16.53	0.3620	13.01	0.6732	13.02
F_ret_1_0	0.0164	12.60	0.0491	11.88	0.1021	11.80	0.2187	10.50	0.0195	11.07	0.0578	10.00	0.1189	9.55	0.2596	7.89
F_ret_2_1	0.0219	12.72	0.0674	11.82	0.1419	10.55	0.2933	9.40	0.0266	12.45	0.0822	11.08	0.1674	10.19	0.3596	7.91
F_ret_12_2	0.0403	15.70	0.1282	12.52	0.2552	10.49	0.4914	9.97	0.0498	14.40	0.1546	11.12	0.3021	9.48	0.5744	8.69
F_excess_ret_1_0	0.0164	12.60	0.0491	11.88	0.1021	11.80	0.2187	10.50	0.0195	11.07	0.0578	10.00	0.1189	9.55	0.2596	7.89
F_excess_ret_2_1	0.0219	12.70	0.0674	11.81	0.1421	10.54	0.2934	9.40	0.0266	12.45	0.0822	11.07	0.1676	10.19	0.3598	7.91
F_excess_ret_12_2	0.0403	15.70	0.1283	12.50	0.2553	10.50	0.4918	9.97	0.0498	14.40	0.1546	11.12	0.3021	9.48	0.5745	8.69
F_MAR_1_0	0.0164	12.60	0.0491	11.88	0.1021	11.80	0.2187	10.50	0.0195	11.07	0.0578	10.00	0.1189	9.55	0.2596	7.89
F_MAR_2_1	0.0218	12.59	0.0672	11.79	0.1416	10.47	0.2920	9.31	0.0265	12.37	0.0822	11.08	0.1672	10.08	0.3576	7.79
F_MAR_12_2	0.0401	15.47	0.1257	12.44	0.2495	10.60	0.4771	9.78	0.0495	14.40	0.1529	10.89	0.2981	9.25	0.5641	8.49
F_CAPM_1_0	0.0170	12.24	0.0518	11.84	0.1065	11.47	0.2230	10.43	0.0202	12.17	0.0618	10.78	0.1299	10.20	0.2791	8.57
F_CAPM_2_1	0.0224	13.31	0.0693	12.09	0.1425	11.39	0.2966	10.46	0.0277	12.87	0.0861	10.99	0.1758	10.02	0.3742	8.27
F_CAPM_12_2	0.0404	15.75	0.1272	12.80	0.2529	11.25	0.4899	10.06	0.0486	14.13	0.1505	11.72	0.2956	10.32	0.5677	9.22
F_3F_alpha_1_0	0.0157	14.64	0.0474	14.75	0.0997	13.22	0.2125	11.44	0.0194	11.51	0.0566	11.13	0.1204	10.68	0.2647	8.59
F_3F_alpha_2_1	0.0209	13.89	0.0652	13.38	0.1398	12.43	0.2935	10.68	0.0262	13.35	0.0804	12.10	0.1713	10.94	0.3685	8.58
F_3F_alpha_12_2	0.0392	18.14	0.1219	14.36	0.2429	13.11	0.4723	11.41	0.0475	15.95	0.1467	12.85	0.2894	11.63	0.5555	10.58
F_4F_alpha_1_0	0.0149	15.51	0.0436	15.33	0.0921	13.96	0.2013	11.59	0.0181	12.41	0.0535	12.10	0.1157	10.73	0.2577	8.44
F_4F_alpha_2_1	0.0197	16.23	0.0606	15.34	0.1307	13.32	0.2844	11.46	0.0249	15.02	0.0758	12.69	0.1627	10.77	0.3591	8.29
F_4F_alpha_12_2	0.0376	18.18	0.1183	14.08	0.2390	11.81	0.4797	10.83	0.0469	14.96	0.1458	10.65	0.2920	9.81	0.5760	9.43
lme_weighted	-0.0010	-0.47	-0.0038	-0.40	-0.0086	-0.38	-0.0250	-0.50	-0.0019	-0.81	-0.0070	-0.73	-0.0127	-0.54	-0.0361	-0.68
lturnover_weighted	0.0007	0.22	-0.0008	-0.06	-0.0032	-0.09	0.0029	0.04	0.0001	0.03	-0.0033	-0.27	-0.0093	-0.29	-0.0090	-0.13
ldp_weighted	0.0014	0.63	0.0076	0.80	0.0160	0.70	0.0304	0.64	0.0019	0.66	0.0089	0.77	0.0217	0.68	0.0442	0.73
beme_weighted	0.0012	0.44	0.0079	0.62	0.0213	0.55	0.0571	0.57	0.0004	0.14	0.0076	0.58	0.0198	0.57	0.0493	0.57
at_weighted	-0.0010	-0.63	-0.0026	-0.37	-0.0040	-0.21	-0.0143	-0.27	-0.0012	-0.64	-0.0048	-0.68	-0.0068	-0.34	-0.0207	-0.43

c_weighted	0.0008	0.28	-0.0005	-0.05	0.0015	0.04	0.0111	0.20	0.0008	0.28	0.0002	0.01	0.0059	0.16	0.0171	0.21
ol_weighted	0.0005	0.33	0.0008	0.14	-0.0027	-0.18	-0.0071	-0.19	-0.0001	-0.08	-0.0006	-0.08	-0.0027	-0.14	-0.0071	-0.18
pcm_weighted	-0.0006	-0.25	-0.0038	-0.36	-0.0084	-0.30	-0.0182	-0.29	-0.0007	-0.27	-0.0032	-0.29	-0.0028	-0.12	-0.0013	-0.02
prof_weighted	0.0006	0.37	0.0018	0.30	0.0064	0.36	0.0088	0.14	0.0008	0.42	0.0042	0.61	0.0130	0.66	0.0353	0.49
roe_weighted	0.0009	0.45	0.0021	0.24	0.0045	0.18	0.0181	0.33	0.0013	0.62	0.0040	0.42	0.0124	0.42	0.0450	0.62
investment_weighted	-0.0010	-0.46	-0.0039	-0.40	-0.0103	-0.50	-0.0174	-0.39	-0.0007	-0.30	-0.0044	-0.46	-0.0089	-0.39	-0.0038	-0.08
oa_weighted	-0.0018	-1.25	-0.0086	-1.84	-0.0221	-1.83	-0.0390	-1.36	-0.0005	-0.26	-0.0031	-0.44	-0.0125	-0.66	-0.0309	-0.88
free_cf_weighted	0.0007	0.38	0.0023	0.28	0.0049	0.22	0.0123	0.23	0.0017	0.81	0.0081	0.87	0.0216	0.74	0.0617	0.79
noa_weighted	-0.0011	-0.74	-0.0053	-1.02	-0.0136	-1.29	-0.0201	-0.99	-0.0010	-0.60	-0.0061	-1.06	-0.0113	-0.87	-0.0136	-0.53
roa_weighted	0.0001	0.04	-0.0009	-0.13	-0.0012	-0.06	-0.0031	-0.06	0.0009	0.54	0.0033	0.51	0.0100	0.57	0.0310	0.68
debt2p_weighted	0.0010	0.44	0.0061	0.67	0.0155	0.59	0.0407	0.62	0.0028	1.03	0.0150	1.20	0.0373	1.07	0.0932	0.99
s2p_weighted	0.0016	0.59	0.0076	0.66	0.0154	0.48	0.0392	0.59	0.0006	0.19	0.0058	0.43	0.0129	0.40	0.0292	0.40
d_so_weighted	0.0002	0.07	-0.0030	-0.27	-0.0051	-0.19	0.0062	0.12	-0.0001	-0.05	-0.0035	-0.33	-0.0035	-0.13	0.0085	0.16
a2me_weighted	0.0022	0.84	0.0111	0.95	0.0258	0.71	0.0533	0.62	0.0029	0.97	0.0147	1.08	0.0352	0.92	0.0822	0.85
e2p_weighted	0.0019	0.64	0.0082	0.64	0.0125	0.44	0.0282	0.37	0.0040	1.14	0.0162	1.08	0.0355	1.08	0.0822	1.07
eps_weighted	-0.0007	-0.29	-0.0030	-0.31	-0.0086	-0.31	-0.0202	-0.26	-0.0007	-0.28	-0.0041	-0.40	-0.0103	-0.36	-0.0179	-0.24
o2p_weighted	0.0019	0.71	0.0084	0.70	0.0216	0.56	0.0447	0.52	0.0028	0.98	0.0133	1.09	0.0323	1.13	0.0751	1.09
nop_weighted	0.0015	0.57	0.0074	0.67	0.0202	0.62	0.0462	0.74	0.0018	0.58	0.0103	0.84	0.0285	0.99	0.0596	0.93
dpi2a_weighted	0.0001	0.05	-0.0007	-0.09	-0.0048	-0.25	-0.0113	-0.26	0.0003	0.15	-0.0003	-0.03	-0.0044	-0.21	-0.0138	-0.32
ivc_weighted	-0.0016	-1.02	-0.0082	-1.56	-0.0243	-1.54	-0.0567	-1.53	-0.0010	-0.70	-0.0055	-1.08	-0.0148	-1.02	-0.0326	-1.09
rna_weighted	0.0010	0.62	0.0021	0.31	0.0023	0.14	0.0031	0.08	0.0029	1.84	0.0086	1.37	0.0165	1.08	0.0437	1.08
pm_weighted	-0.0009	-0.47	-0.0035	-0.43	-0.0059	-0.25	-0.0058	-0.08	-0.0013	-0.67	-0.0039	-0.47	-0.0038	-0.15	0.0122	0.16
ato_weighted	0.0006	0.41	0.0006	0.11	-0.0037	-0.27	-0.0162	-0.48	0.0030	2.06	0.0087	1.77	0.0181	1.52	0.0418	1.36
cto_weighted	0.0012	0.82	0.0025	0.48	0.0033	0.24	0.0129	0.46	0.0013	0.87	0.0035	0.55	0.0076	0.50	0.0258	0.98
tan_weighted	-0.0011	-0.54	-0.0040	-0.51	-0.0066	-0.33	-0.0264	-0.52	-0.0018	-0.79	-0.0082	-0.96	-0.0127	-0.50	-0.0255	-0.39
s2c_weighted	-0.0002	-0.06	0.0007	0.07	0.0002	0.01	0.0051	0.09	-0.0002	-0.09	0.0024	0.23	0.0034	0.13	0.0136	0.24
c2d_weighted	-0.0001	-0.10	-0.0025	-0.43	-0.0062	-0.40	-0.0070	-0.23	0.0008	0.55	0.0026	0.43	0.0095	0.57	0.0408	1.23
sales_g_weighted	-0.0004	-0.18	-0.0050	-0.57	-0.0103	-0.39	-0.0203	-0.38	-0.0007	-0.30	-0.0057	-0.61	-0.0071	-0.31	0.0040	0.09
d_dgm_dsales_weighted	0.0006	0.49	0.0009	0.17	-0.0031	-0.21	-0.0207	-0.58	-0.0002	-0.12	-0.0016	-0.28	-0.0058	-0.34	-0.0036	-0.10
d_ceq_weighted	-0.0010	-0.43	-0.0059	-0.57	-0.0098	-0.30	-0.0227	-0.38	-0.0016	-0.68	-0.0069	-0.68	-0.0076	-0.28	0.0041	0.08
roc_weighted	-0.0033	-1.31	-0.0153	-1.29	-0.0340	-1.13	-0.0881	-1.32	-0.0032	-1.18	-0.0143	-1.13	-0.0304	-0.88	-0.0822	-1.18
aoa_weighted	-0.0005	-0.25	-0.0017	-0.21	-0.0049	-0.24	-0.0016	-0.03	-0.0007	-0.33	-0.0011	-0.14	-0.0041	-0.20	0.0004	0.01
roic_weighted	0.0001	0.05	-0.0019	-0.22	-0.0073	-0.27	-0.0088	-0.11	0.0018	0.74	0.0052	0.56	0.0122	0.46	0.0375	0.51
ipm_weighted	-0.0009	-0.50	-0.0045	-0.59	-0.0084	-0.37	-0.0125	-0.27	-0.0014	-0.75	-0.0060	-0.73	-0.0096	-0.36	-0.0158	-0.23

sat_weighted	0.0012	0.88	0.0033	0.65	0.0045	0.36	0.0034	0.11	0.0004	0.26	0.0016	0.22	0.0048	0.26	0.0089	0.25
q_weighted	-0.0020	-0.74	-0.0101	-0.80	-0.0245	-0.64	-0.0643	-0.69	-0.0017	-0.65	-0.0074	-0.62	-0.0155	-0.49	-0.0443	-0.49
spread_mean_weighted	-0.0002	-0.07	-0.0009	-0.07	-0.0036	-0.13	-0.0069	-0.11	-0.0016	-0.61	-0.0054	-0.45	-0.0110	-0.40	-0.0252	-0.41
rel_to_high_price_weighted	0.0123	6.39	0.0388	5.63	0.0776	5.23	0.1556	4.95	0.0135	5.32	0.0439	4.64	0.0914	4.44	0.1809	4.50
cum_return_1_0_weighted	0.0116	10.90	0.0339	10.04	0.0683	9.54	0.1434	9.54	0.0140	8.67	0.0408	8.95	0.0835	8.38	0.1848	8.01
cum_return_12_7_weighted	0.0089	3.75	0.0253	2.86	0.0473	2.11	0.0806	2.37	0.0087	3.45	0.0264	2.83	0.0521	2.20	0.0958	2.23
cum_return_12_2_weighted	0.0147	5.74	0.0433	4.32	0.0857	3.60	0.1652	3.26	0.0146	5.05	0.0436	3.71	0.0878	3.07	0.1748	2.68
cum_return_36_13_weighted	0.0028	1.05	0.0043	0.42	0.0067	0.23	0.0103	0.17	0.0030	1.11	0.0060	0.55	0.0101	0.36	0.0233	0.36
cum_return_6_2_weighted	0.0148	7.04	0.0446	5.72	0.0934	5.71	0.1839	5.42	0.0152	6.46	0.0469	5.15	0.0988	5.17	0.1992	4.95
beta_weighted	-0.0036	-1.34	-0.0131	-1.11	-0.0277	-0.88	-0.0633	-1.09	-0.0037	-1.22	-0.0135	-0.97	-0.0268	-0.62	-0.0671	-0.82
dto_weighted	0.0011	1.50	0.0023	0.90	0.0058	1.14	0.0144	1.52	0.0017	1.90	0.0051	1.55	0.0153	2.18	0.0370	2.86
suv_weighted	0.0010	1.52	0.0019	0.84	0.0047	0.99	0.0099	0.79	0.0013	1.26	0.0034	1.17	0.0093	1.52	0.0259	1.56
ret_max_weighted	0.0001	0.06	-0.0013	-0.15	-0.0015	-0.08	0.0029	0.07	-0.0008	-0.37	-0.0052	-0.67	-0.0095	-0.55	-0.0109	-0.27
beta_daily_weighted	0.0006	0.20	-0.0008	-0.09	-0.0035	-0.18	-0.0132	-0.32	0.0006	0.19	-0.0014	-0.15	-0.0009	-0.05	0.0001	0.00
idio_vol_weighted	-0.0008	-0.35	-0.0042	-0.52	-0.0084	-0.42	-0.0124	-0.29	-0.0025	-1.33	-0.0108	-1.52	-0.0212	-1.22	-0.0354	-0.98
total_vol_weighted	-0.0005	-0.19	-0.0038	-0.43	-0.0077	-0.35	-0.0140	-0.30	-0.0020	-0.88	-0.0104	-1.21	-0.0199	-0.98	-0.0338	-0.83
std_volume_weighted	-0.0027	-1.02	-0.0097	-0.87	-0.0172	-0.73	-0.0350	-0.69	-0.0032	-1.13	-0.0099	-0.94	-0.0145	-0.55	-0.0296	-0.48
std_turn_weighted	0.0001	0.04	-0.0021	-0.23	-0.0056	-0.26	-0.0045	-0.09	-0.0007	-0.26	-0.0064	-0.71	-0.0130	-0.61	-0.0170	-0.38
lme_adj_weighted	-0.0014	-0.72	-0.0049	-0.55	-0.0087	-0.42	-0.0190	-0.44	-0.0021	-0.98	-0.0076	-0.83	-0.0142	-0.63	-0.0341	-0.76
beme_adj_weighted	0.0007	0.32	0.0078	0.86	0.0186	0.73	0.0457	0.75	-0.0007	-0.28	0.0026	0.26	0.0049	0.20	0.0270	0.43
pm_adj_weighted	-0.0018	-0.61	-0.0086	-0.67	-0.0182	-0.50	-0.0271	-0.42	-0.0005	-0.13	-0.0030	-0.19	-0.0042	-0.11	-0.0092	-0.10
at_adj_weighted	0.0014	0.88	0.0057	0.97	0.0131	1.01	0.0261	0.80	-0.0001	-0.03	0.0019	0.25	0.0093	0.50	0.0260	0.55
family_ret_1_0	0.0084	9.97	0.0252	12.57	0.0534	10.91	0.1109	9.15	0.0105	8.31	0.0308	9.57	0.0661	8.85	0.1374	8.39
family_ret_2_1	0.0120	11.78	0.0364	13.11	0.0762	10.71	0.1558	9.66	0.0149	12.46	0.0461	11.16	0.0919	10.04	0.1848	8.59
family_ret_12_2	0.0209	14.81	0.0667	11.22	0.1324	8.97	0.2635	8.83	0.0274	14.77	0.0854	12.22	0.1675	9.85	0.3388	9.24
family_excess_ret_1_0	0.0084	9.97	0.0252	12.57	0.0534	10.91	0.1109	9.15	0.0105	8.31	0.0308	9.57	0.0661	8.85	0.1374	8.39
family_excess_ret_2_1	0.0120	11.78	0.0364	13.11	0.0762	10.71	0.1558	9.66	0.0149	12.46	0.0461	11.16	0.0919	10.04	0.1848	8.59
family_excess_ret_12_2	0.0209	14.81	0.0667	11.22	0.1324	8.97	0.2635	8.83	0.0274	14.77	0.0854	12.22	0.1675	9.85	0.3388	9.24
family_MAR_1_0	0.0084	9.97	0.0252	12.57	0.0534	10.91	0.1109	9.15	0.0105	8.31	0.0308	9.57	0.0661	8.85	0.1374	8.39
family_MAR_2_1	0.0119	12.21	0.0358	13.61	0.0754	11.25	0.1553	10.06	0.0148	12.39	0.0461	11.08	0.0917	10.08	0.1850	8.74
family_MAR_12_2	0.0208	14.46	0.0664	10.84	0.1316	8.85	0.2620	8.42	0.0273	14.73	0.0848	12.11	0.1662	9.83	0.3366	9.11
family_CAPM_1_0	0.0091	9.36	0.0270	11.12	0.0560	10.20	0.1175	8.73	0.0120	9.50	0.0340	9.51	0.0724	9.27	0.1528	8.43
family_CAPM_2_1	0.0117	10.96	0.0358	11.29	0.0742	10.32	0.1549	9.74	0.0151	11.33	0.0461	9.92	0.0912	9.05	0.1867	8.47
family_CAPM_12_2	0.0207	15.16	0.0630	12.82	0.1245	11.58	0.2373	11.63	0.0272	14.94	0.0820	11.72	0.1617	10.47	0.3254	9.83

family_3F_alpha_1_0	0.0078	12.32	0.0230	14.82	0.0481	13.83	0.1002	11.61	0.0111	10.32	0.0314	12.64	0.0648	11.49	0.1396	9.86
family_3F_alpha_2_1	0.0110	12.82	0.0319	12.99	0.0667	11.73	0.1361	9.77	0.0150	12.82	0.0426	11.35	0.0860	11.65	0.1796	9.13
family_3F_alpha_12_2	0.0194	16.03	0.0587	13.73	0.1139	12.45	0.2098	9.49	0.0262	16.99	0.0782	14.34	0.1516	11.86	0.2923	8.83
family_4F_alpha_1_0	0.0072	9.90	0.0207	11.52	0.0435	11.58	0.0903	9.27	0.0101	8.62	0.0290	10.39	0.0602	10.12	0.1270	8.72
family_4F_alpha_2_1	0.0103	12.55	0.0294	12.20	0.0603	10.93	0.1250	9.52	0.0145	12.02	0.0402	11.03	0.0811	10.31	0.1709	8.26
family_4F_alpha_12_2	0.0191	15.06	0.0576	13.35	0.1137	12.04	0.2256	11.07	0.0263	16.00	0.0775	14.08	0.1516	12.21	0.3077	10.60
family_no	0.0012	0.90	0.0087	1.50	0.0208	1.31	0.0567	1.50	0.0010	0.73	0.0081	1.39	0.0194	1.23	0.0520	1.37
family_log_real_tna	0.0040	2.07	0.0192	2.22	0.0453	1.89	0.1192	1.72	0.0041	1.92	0.0134	1.49	0.0243	0.75	0.0469	0.36
family_flow_lag1	0.0271	20.22	0.0843	19.04	0.1638	16.85	0.3103	14.98	0.0426	11.11	0.1296	12.29	0.2507	11.70	0.4939	9.82
family_flow_lag2_lag1	0.0265	17.10	0.0821	18.68	0.1585	16.29	0.3006	14.71	0.0393	13.94	0.1204	13.15	0.2364	10.52	0.4751	8.49
family_flow_lag12_lag2	0.0280	15.92	0.0821	14.22	0.1619	11.23	0.3184	9.79	0.0398	8.14	0.1073	9.49	0.2121	6.86	0.4111	5.71
family_age	-0.0011	-0.53	-0.0009	-0.12	-0.0009	-0.04	0.0085	0.12	-0.0014	-0.74	-0.0014	-0.22	-0.0031	-0.18	-0.0020	-0.04

This table shows the result of univariate analysis based on each of the 110 characteristics. We sort mutual funds into deciles based on the value of each characteristic at month t and compute equal- and value-weighted averages of fund flows at month $t + 1$ for each decile. Then we conduct a t-test of the difference between the bottom and top decile using Newey-West standard errors with 12 lags.

Table 5: Summary statistics of Out of Sample R-sqaured

	Mean	Min	Max	Proportion of negative R-sq
BRT_flow_1month	0.0623	-0.3198	0.2968	22.46%
BRT_flow_3month	0.2721	-0.1787	0.5416	1.78%
BRT_flow_6month	0.3184	-0.0563	0.5736	0.90%
BRT_flow_12month	0.3128	-0.0938	0.5716	0.90%
BRT_v_flow_1month	0.1511	-0.0241	0.3369	1.80%
BRT_v_flow_3month	0.2143	-0.0927	0.4540	0.60%
BRT_v_flow_6month	0.2343	-0.1015	0.4863	0.30%
BRT_v_flow_12month	0.2158	0.0406	0.4448	0%
OLS_flow_1month	-0.2338	-0.8608	0.1700	95.51%
OLS_flow_3month	-0.1857	-0.7846	0.4793	87.13%
OLS_flow_6month	-0.1733	-0.8621	0.6490	85.03%
OLS_flow_12month	-0.1694	-0.8285	0.7110	85.03%

This table reports the summary statistics of out-of-sample R-squared for each model we use. "BRT_v" indicates the BRT model using the validation sample to set the optimal number of boosting iterations

Table 6: Mutual Fund Portfolios Using Predicted Values with Restricted Predictor Space Sorted

Panel A: Equal Weighted

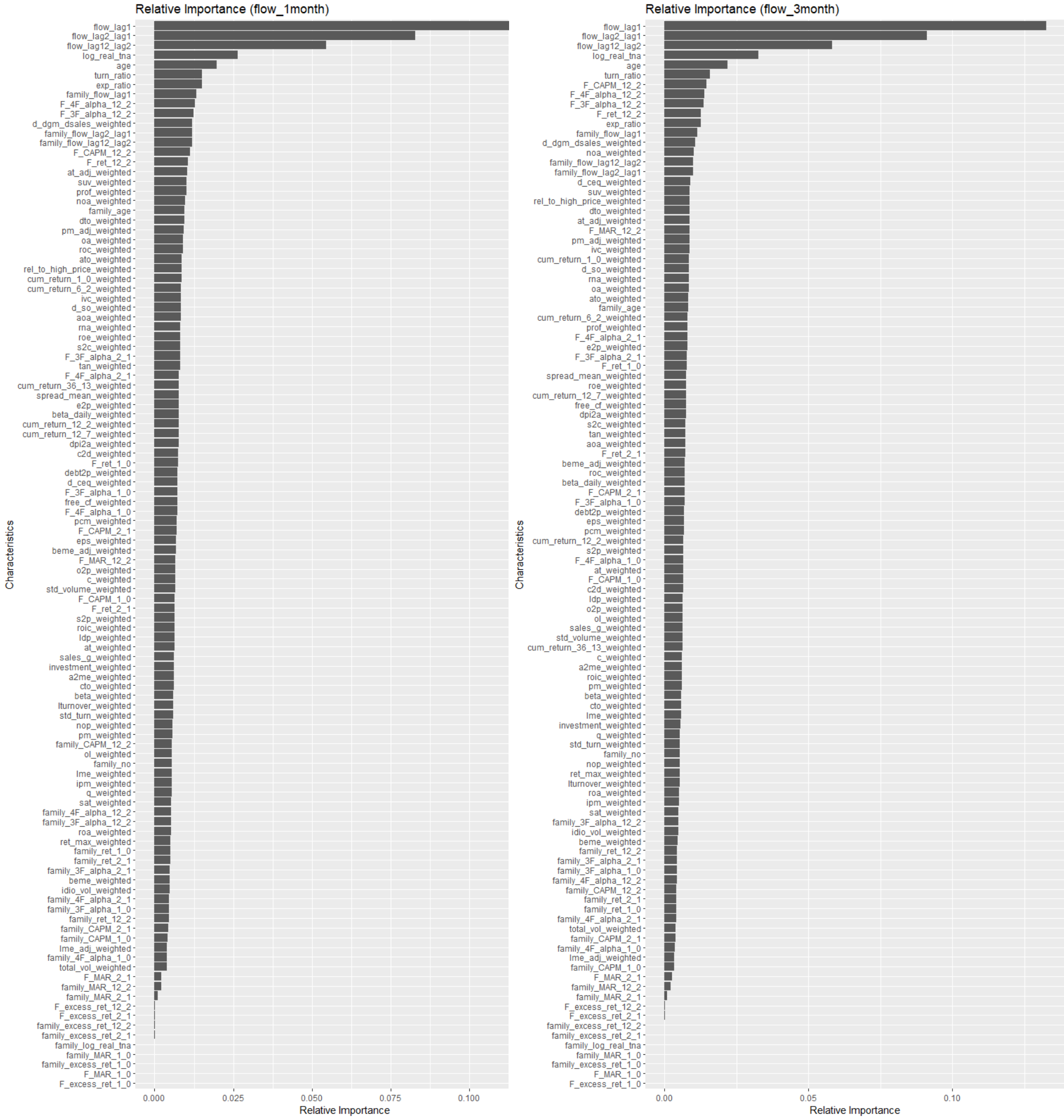
	All Predictor				**90% Predictor**				**75% Predictor**				**50% Predictor**			
Decile	Excess Ret	t-stat	Capm Alpha	t-stat	Excess Ret	t-stat	Capm Alpha	t-stat	Excess Ret	t-stat	Capm Alpha	t-stat	Excess Ret	t-stat	Capm Alpha	t-stat
Bottom	0.0042	1.49	-0.0021	-1.90	0.0044	1.56	-0.0018	-1.71	0.0046	1.64	-0.0019	-1.81	0.0047	1.68	-0.0019	-1.91
2	0.0055	2.14	-0.0010	-1.20	0.0056	2.15	-0.0010	-1.13	0.0057	2.19	-0.0011	-1.37	0.0057	2.17	-0.0010	-1.33
3	0.0059	2.28	-0.0004	-0.63	0.0059	2.30	-0.0004	-0.57	0.0060	2.36	-0.0007	-1.05	0.0060	2.31	-0.0007	-1.13
4	0.0064	2.51	-0.0003	-0.56	0.0065	2.57	0.0000	-0.03	0.0063	2.46	-0.0001	-0.25	0.0065	2.56	0.0000	0.02
5	0.0068	2.70	0.0000	0.03	0.0067	2.67	0.0001	0.11	0.0066	2.67	-0.0001	-0.15	0.0066	2.64	0.0001	0.10
6	0.0069	2.76	0.0002	0.48	0.0068	2.71	0.0001	0.22	0.0070	2.84	0.0003	0.61	0.0071	2.84	0.0005	0.75
7	0.0074	3.00	0.0006	1.01	0.0075	3.11	0.0007	1.16	0.0072	2.94	0.0008	1.35	0.0073	2.95	0.0006	0.94
8	0.0078	3.09	0.0011	1.46	0.0077	3.09	0.0010	1.24	0.0077	3.07	0.0011	1.54	0.0078	3.12	0.0009	1.11
9	0.0089	3.49	0.0020	1.94	0.0087	3.38	0.0017	1.71	0.0087	3.36	0.0016	1.62	0.0084	3.32	0.0018	1.75
Top	0.0099	3.54	0.0029	2.02	0.0097	3.50	0.0027	1.96	0.0097	3.45	0.0029	2.11	0.0093	3.42	0.0027	1.92
Top-Bottom	0.0057	3.03	0.0050	2.69	0.0053	2.97	0.0046	2.57	0.0051	2.86	0.0048	2.91	0.0046	2.50	0.0046	2.62

Panel B: Value Weighted

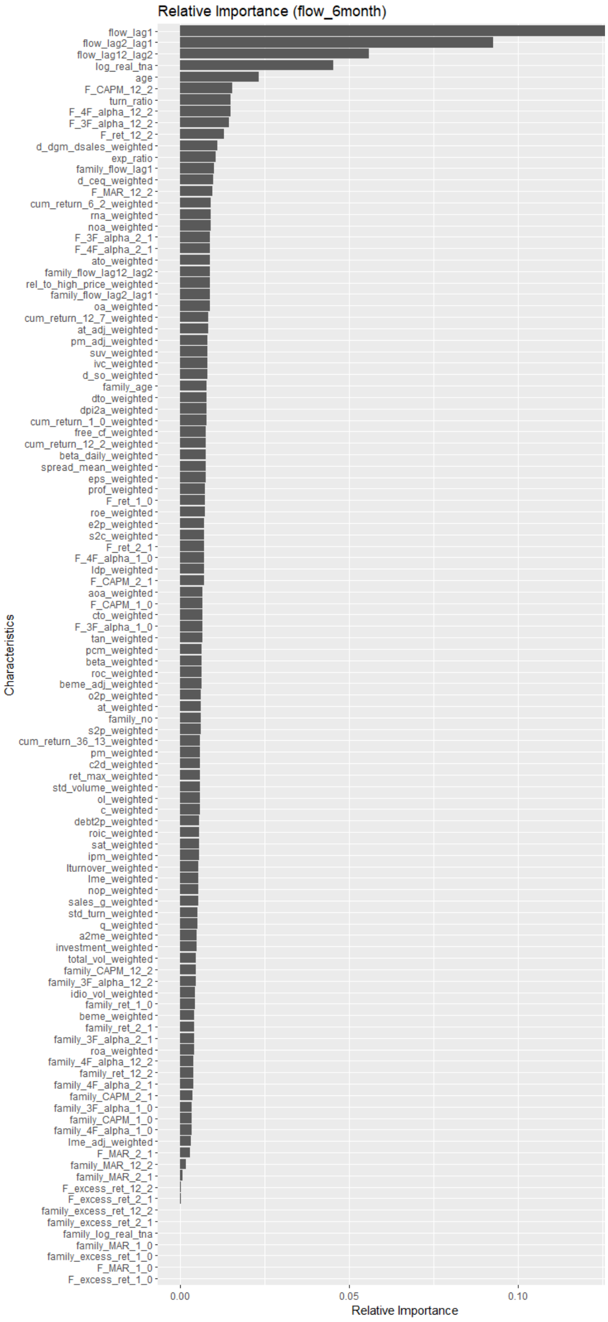
	All Predictor				**90% Predictor**				**75% Predictor**				**50% Predictor**			
Decile	Excess Ret	t-stat	Capm Alpha	t-stat	Excess Ret	t-stat	Capm Alpha	t-stat	Excess Ret	t-stat	Capm Alpha	t-stat	Excess Ret	t-stat	Capm Alpha	t-stat
Bottom	0.0045	1.60	-0.0017	-1.65	0.0047	1.69	-0.0016	-1.59	0.0048	1.72	-0.0017	-1.67	0.0049	1.77	-0.0017	-1.81
2	0.0055	2.13	-0.0010	-1.20	0.0057	2.18	-0.0009	-1.08	0.0057	2.20	-0.0011	-1.40	0.0058	2.21	-0.0009	-1.33
3	0.0058	2.27	-0.0003	-0.47	0.0060	2.34	-0.0002	-0.37	0.0061	2.39	-0.0005	-0.71	0.0061	2.34	-0.0006	-0.94
4	0.0065	2.58	-0.0003	-0.50	0.0065	2.58	0.0001	0.25	0.0062	2.43	0.0000	-0.06	0.0066	2.61	0.0001	0.13
5	0.0068	2.73	0.0001	0.27	0.0066	2.67	0.0000	0.06	0.0066	2.66	-0.0001	-0.19	0.0067	2.72	0.0000	0.04
6	0.0070	2.83	0.0003	0.63	0.0069	2.78	0.0001	0.26	0.0071	2.86	0.0003	0.64	0.0072	2.86	0.0004	0.74
7	0.0075	3.06	0.0005	0.85	0.0076	3.15	0.0008	1.24	0.0073	2.97	0.0009	1.39	0.0074	2.99	0.0007	1.00
8	0.0078	3.09	0.0012	1.44	0.0079	3.16	0.0010	1.21	0.0077	3.09	0.0013	1.67	0.0078	3.15	0.0010	1.25
9	0.0092	3.57	0.0023	2.13	0.0088	3.37	0.0018	1.66	0.0089	3.46	0.0018	1.71	0.0087	3.39	0.0019	1.76
Top	0.0103	3.57	0.0033	2.08	0.0099	3.45	0.0031	1.89	0.0099	3.39	0.0032	2.00	0.0096	3.40	0.0031	2.01
Top-Bottom	0.0058	2.88	0.0050	2.49	0.0052	2.72	0.0047	2.35	0.0052	2.60	0.0049	2.62	0.0047	2.45	0.0047	2.58

This table shows average excess returns and CAPM alphas for each portfolio sorted using BRT predicted values. Panel A and B present equal- and value-weighted average returns, respectively. We restrict the predictor space to the characteristics that are important to predict 3-month future flows from the highest where the sum of importance is 90%, 75%, and 50%. “Top-Bottom” indicates the long-short portfolio, together with t-statistics using Newey West standard errors with 12 lags.

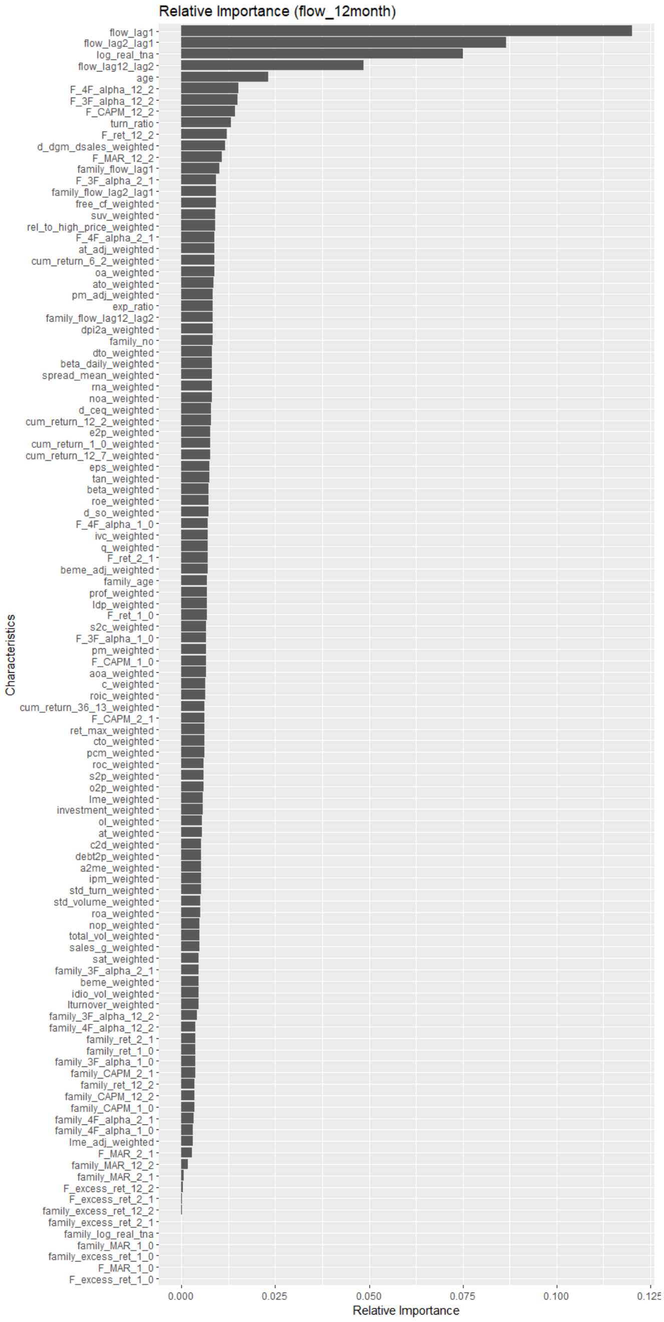
Figure 1: Relative Importance Plot to Predict Flows in the BRT model



(a) 1-month flow **(b)** 3-month flow



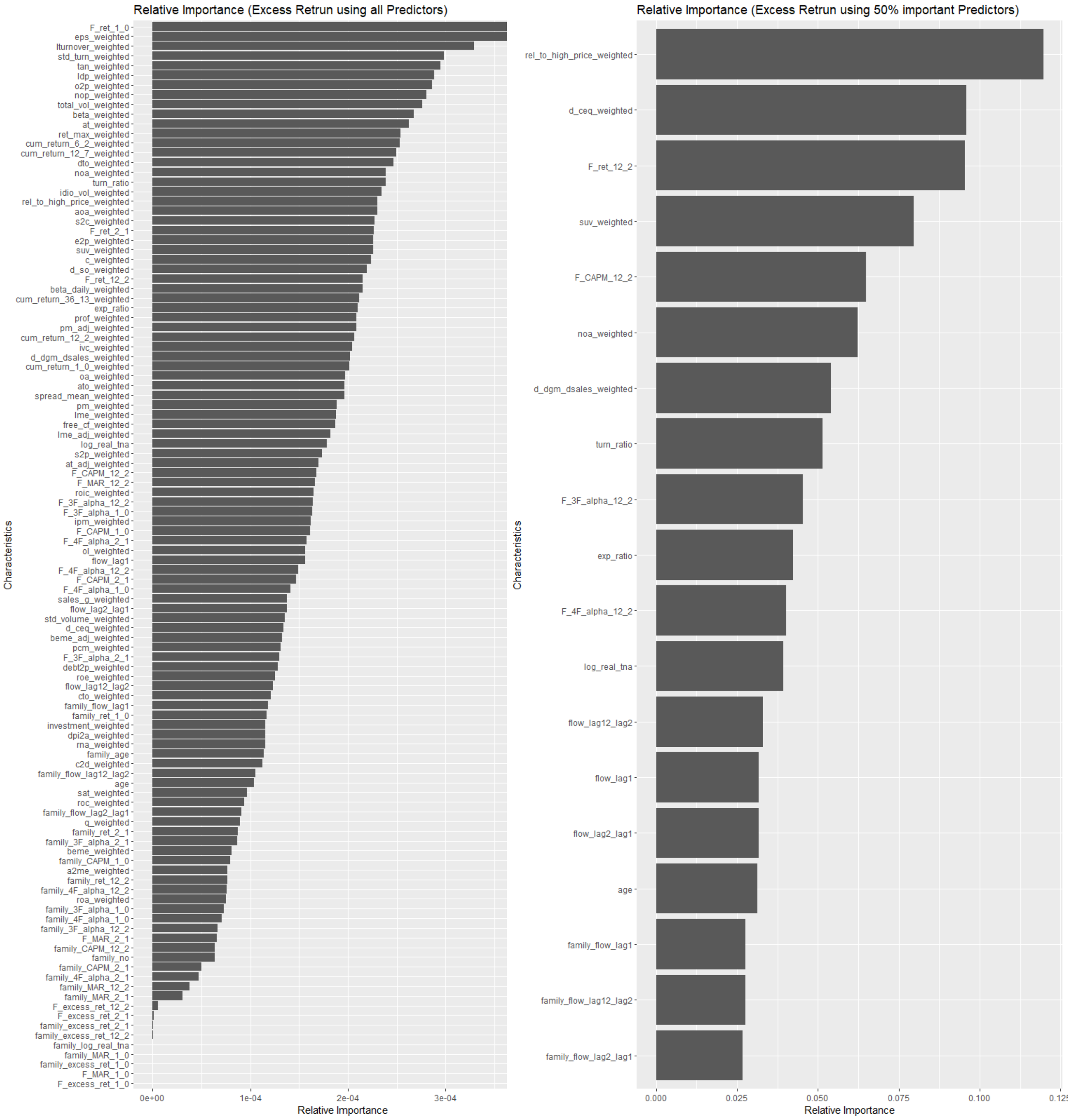
(c) 6-month flow



(d) 12-month flow

This figure shows the relative importance measure when predicting 1, 3, 6, and 12-month flows in the BRT model. The y axis denotes 110 characteristics, and the x axis denotes each regressor's relative importance measure. The sum of relative importance measure across all covariates is 1.

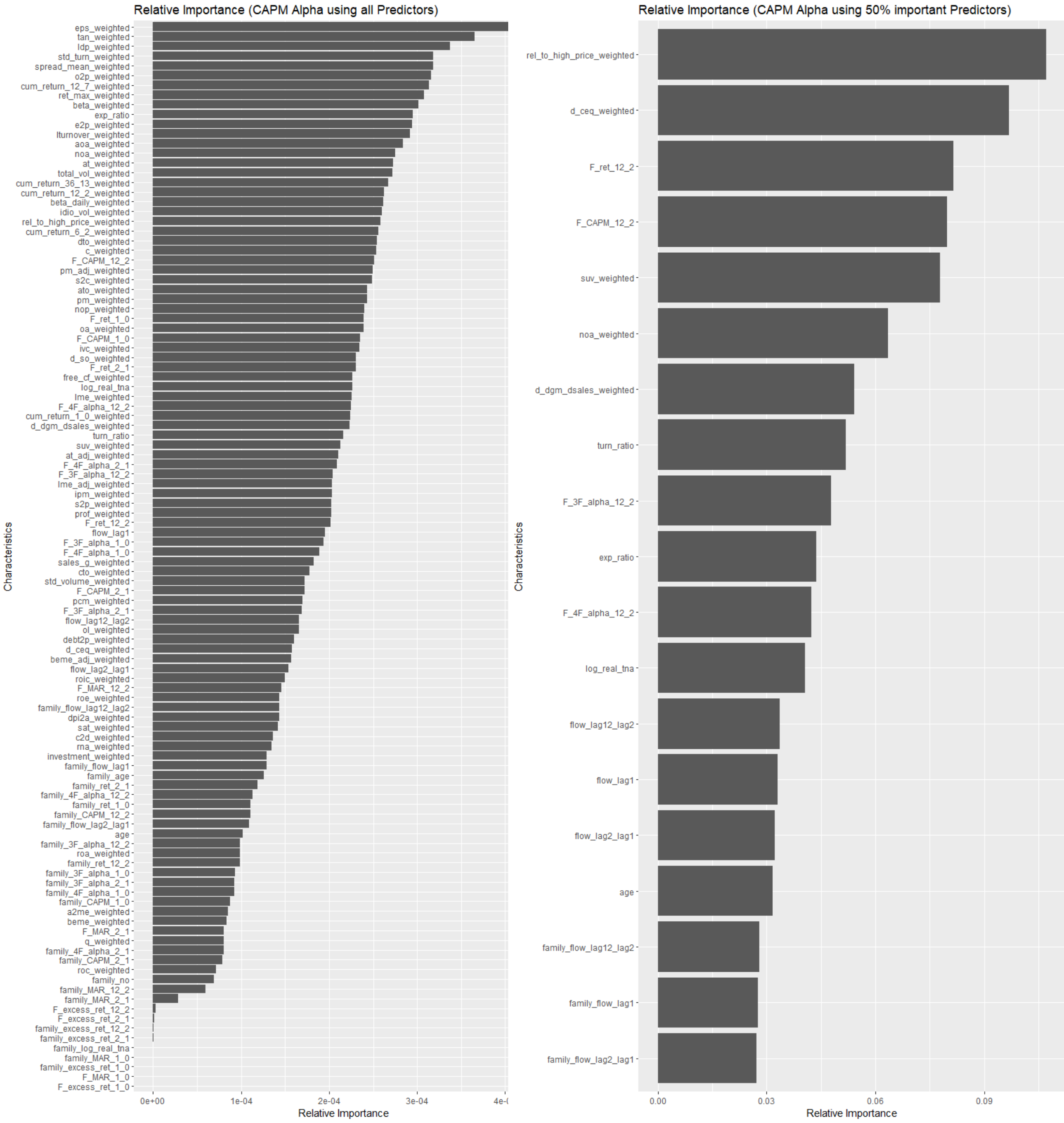
Figure 2: Relative Importance Plot to Predict Excess Returns in the BRT model



(a) All Predictor **(b)** 50% Predictor

This figure shows the relative importance measure when predicting excess returns using either all predictors or predictors that are important to predict 3-month future flows from the highest where the sum of importance is 50% in the BRT model. The y axis denotes 110 characteristics, and the x axis denotes each regressor's relative importance measure.

Figure 3: Relative Importance Plot to Predict CAPM Alphas in the BRT model

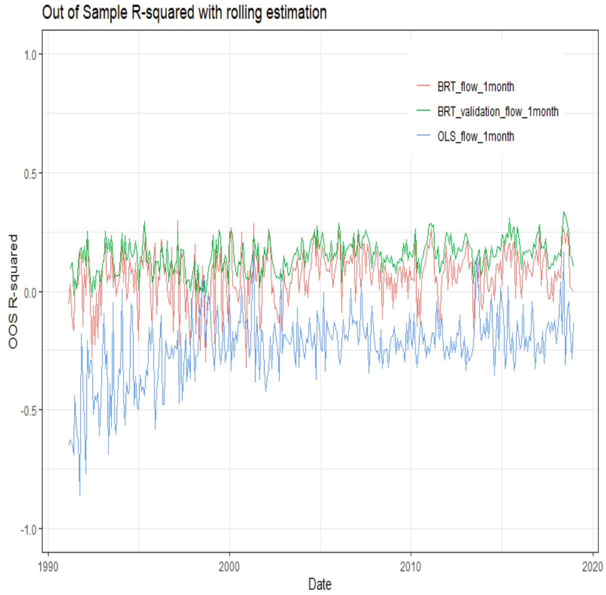


(a) All Predictor

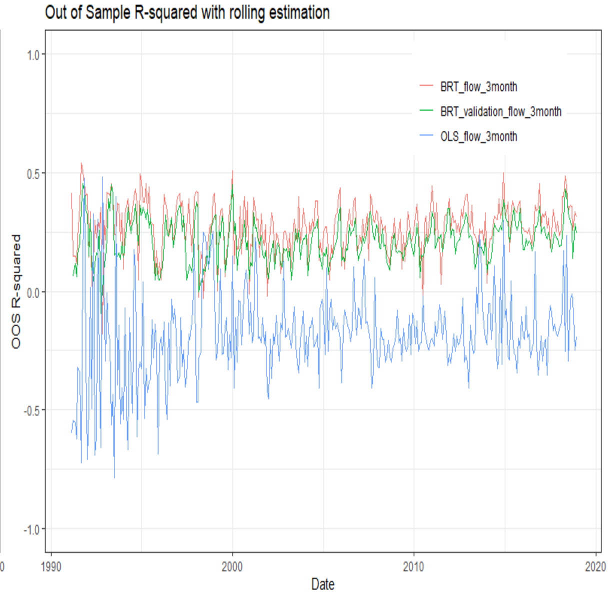
(b) 50% Predictor

This figure shows the relative importance measure when predicting CAPM alphas using either all predictors or predictors that are important to predict 3-month future flows from the highest where the sum of importance is 50% in the BRT model. The y axis denotes 110 characteristics, and the x axis denotes each regressor's relative importance measure.

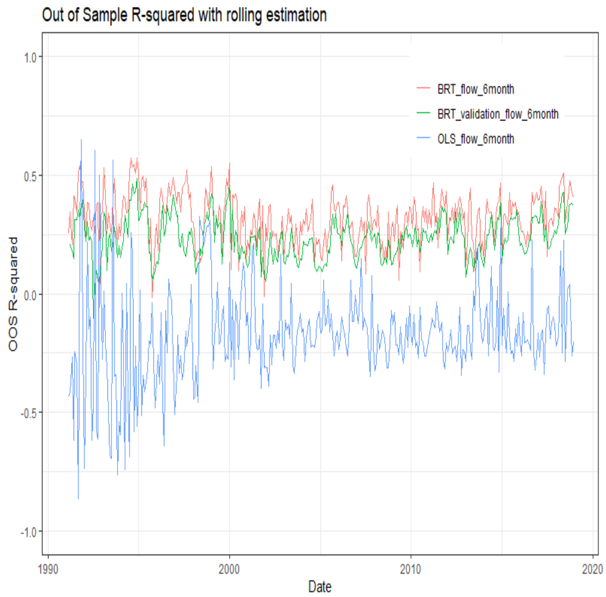
Figure 4: Out of Sample R-squared over Time



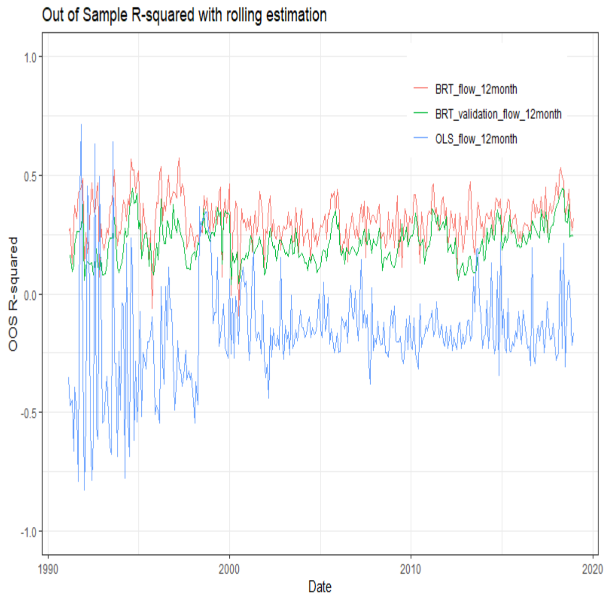
(a) 1-month flow



(b) 3-month flow



(c) 6-month flow



(d) 12-month flow

This figure presents the time-series plot of the out-of-sample R-squared in the 1-month rolling window estimation predicting 1, 3, 6, and 12-month flows. The red line indicates the BRT without the validation sample, the green indicates the BRT with the validation sample, and the blue is the OLS. The y axis denotes out-of-sample R-squared, and the x axis denotes the date.