SemeAd 2024
XXVII Seminários em Administração

05, 06, 07 e 08 de novembro de 2024
ISSN 2177-3866

Selecting the best mutual founds based on machine learning techiniques

PEDRO PAULO PORTELLA TELES
UNIVERSIDADE FEDERAL DE MINAS GERAIS (UFMG)

WANDERCI ALVES BITENCOURT
UNIVERSIDADE FEDERAL DE MINAS GERAIS (UFMG)

ROBERT ALDO IQUIAPAZA
UNIVERSIDADE FEDERAL DE MINAS GERAIS (UFMG)

Agradecimento à orgão de fomento:
We would like to thank the FAPEMIG, CNPQ and CAPES.

Selecting the best mutual funds based on Machine Learning techniques

1 Introduction

A fund is a collective investment instrument that has gained popularity among individual and institutional investors (Jones & Mo, 2021), making the fund industry a vital component of the global financial market. Its consistent growth can be attributed to several factors such as liquidity provision, diversification and low-cost professional management services, as noted by Cuthbertson et al. (2016) and Chua & Tam (2020).

When investors consider investing in an investment fund, their goal is to select a fund or a group of funds that will outperform the benchmark and create value (Kaniel et al., 2023). The rapid expansion of this instrument raises important questions about its ability to produce superior performance for investors, as well as how to identify the high-performing funds (Aggarwal & Jorion, 2010; Jones & Mo, 2021; Kaniel et al., 2023). Nonetheless, anticipating which funds will outperform ex-ante is deemed a challenging undertaking due to the numerous factors that impact fund performance (Bogle, 1992; DeMiguel et al., 2023). Therefore, there is a growing demand for research studies pertaining to this topic and the utilization of novel modeling techniques that can assist in accomplishing this task.

The existing literature on this subject can be categorized into three primary lines of investigation. Firstly, an extensive body of literature establishes a link between a fund's past performance and its future performance (Hendricks et al., 1993; Brown & Goetzmann, 1995; Carhart, 1997; Blake, 2015; Harvey & Liu, 2018). This line encompasses performance persistence, market timing, and volatility timing. Secondly, there are studies focus on understanding the impact of fund characteristics on performance (Chen et al., 2004; Yan, 2008; Gil-Bazo & Ruiz-Verdú, 2009; Pástor et al., 2015; Cuthbertson et al., 2016; Hu et al., 2016; Adams et al., 2018). Finally, the third line examines fund manager characteristics and the trading environment (Goetzmann et al., 2003; Wu et al., 2021).

Therefore, predicting future performance seems to involve a multitude of factors, it is unlikely that a single variable will be more efficient than a broad set of characteristics for predicting fund performance (DeMiguel et al., 2023). To address this issue, some recent studies have investigated the use of machine learning and artificial intelligence applications to help resolve it by accommodating an infinite number of variables and capturing non-linear relationships.

In a seminal article, Wu et al. (2021) employed four distinct machine learning algorithms, incorporating as predictors of idiosyncratic returns-based features and of macro-derivative features, to address the issue of identifying future hedge fund winners. They have demonstrated that these models consistently outperform the four-styled Hedge Fund Research Indices. Furthermore, they provided evidence that neural networks are the best perfor-

ming algorithm and that the combination of variables, such as kurtosis, can enhance the ability to predict hedge fund returns.

In a different approach, Li & Rossi (2021) evaluated several machine learning methods and found that by exploiting fund holdings and stock characteristics, one can build fund portfolios that earn significant alphas. It was found that the relationship between fund performance and its characteristics was nonlinear, and the characteristics related to trading frictions and momentum having the highest predictive power. Moreover, the authors have noted that the Boosted Regression Trees approach outperformed other models in predicting fund returns.

DeMiguel et al. (2023) expanded the scope of research into mutual funds by a comprehensive comparison of various long-only portfolio construction methods, including machine learning algorithms, Ordinary Least Squares, and naive strategies. Their analysis, which was conducted utilizing monthly data spanning from 1980 to 2020 and 17 characteristics of American mutual funds, revealed the superiority of machine learning algorithms in predicting fund alpha. Specifically, decision tree-based models demonstrated a remarkable ability to capture nonlinearities and interactions between variables, enabling the identification of small mutual funds that have skillful managers.

In a parallel study, Kaniel et al. (2023) employed a neural network to analyze the potential of a wide range of fund characteristics, stock characteristics, and macroeconomic variables in predicting abnormal fund returns. Unlike the study mentioned above, the authors focus on long-short portfolios of mutual funds end demonstrated that momentum factor and fund flow are the only necessary variables to differentiate funds with higher and lower future abnormal returns. Thus, the characteristics of the stocks held by the funds play a limited role in predicting abnormal returns. The study also emphasized the importance of the interaction between these fund characteristics and investor sentiment, highlighting the ability of machine learning techniques to capture these interaction effects.

Studies suggest that it is important to look at different markets because they have different characteristics that affect funds prediction (Dumitrescu & Gil-Bazo, 2018; Jones & Mo, 2021). However, the majority of studies are centered on the American market and there is a lack of comprehensive studies that explore the use of machine learning models in emerging markets, including the Brazilian market (Rubesam, 2022). A recent work identified is that of Sterenfeld (2023), who employed four machine learning models to identify outperforming funds. Results from this study indicated that non-linear models, especially Random Forest, have superior performance compared to linear models in fund selection.

Despite the relatively recent nature of these techniques in the financial literature, their integration with pre-existing literature on the identification of metrics able of anticipating future returns aims to overcome the limitations of conventional approaches. Thus, these studies contribute to the growing trend of applying machine learning to understand complex relationships among financial variables.

In light of the above, our objective is to assess the predictive potential of various machine learning algorithms in forecasting abnormal returns of Brazilian equity funds. We also aim to identify which variables are most relevant for this prediction task in distinguishing in advance between equity funds that will outperform and those that will underperform.

Our approach differs from existing literature in some aspects. First, although we present further details to XGBoost, we compare fourteen machine learning methods to predict

fund performance, categorized into three groups: linear models, ensemble models, and others. Secondly, using the methodology of momentum research, we conduct our analysis for all return-related variables across three distinct periods (short-term momentum, short-term reversal, and momentum).

Moreover, considering specific markets such as the Brazilian case can bring light to investors decision-making and help them take advantage of investment opportunities. As noted by Jones & Mo (2021), the distinct characteristics of each market have an impact on the predictability of fund performance and therefore, it is imperative to consider them. In addition, emerging markets typically possess distinct characteristics, and it is not unreasonable to assume that the findings of studies using US data will not automatically be applicable to these markets (Rubesam, 2022).

In our approach, we employ machine learning models to generate monthly forecasts for fund performance. To determine which funds underperform and which outperform, we use the abnormal return of Carhart's four-factor model as our dependent variable, similar to Kaniel et al. (2023). To evaluate the outcomes, we construct both long-only and long-short portfolios by deciles.

Our findings confirm the effectiveness of machine learning techniques in predicting the performance of mutual funds, in accordance with previous research. Furthermore, ensemble models have significantly outperformed conventional linear methods in this particular task, thus demonstrating the superiority of nonlinear models in capturing complex relationships and patterns in fund data. Unlike previous research, we found that metrics based on fund characteristics are not very relevant for predicting performance of Brazilian funds, but return-based metrics are, especially risk-based. These results may be attributed to the regulation of the Brazilian market and the widespread availability of information on equity funds (Sterenfeld, 2023), which may reduce the significance of fund characteristics in distinguishing their performance.

This research contributes to the growing trend of using machine learning algorithms to improve fund performance prediction. Furthermore, conducting an empirical investigation into the Brazilian financial sector, an emerging and expanding market, can yield valuable insights that can aid investors and fund managers in making informed decisions regarding portfolio construction and investment strategies.

The rest of the paper is organized as follows. Section 2 presents the data and variables under consideration, as well as methodological procedures. Section 3 presents the results of our analysis, and section 4 concludes with remarks on possible future developments.

2 Data and methodology

2.1 Data

At this stage, we collected daily data spanning from February 2004 to February 2023. Data on equity mutual funds were sourced from Economatica, a Brazilian financial data provider, while the risk factors and the Brazilian risk-free rate were obtained from the Center for Research in Financial Economics of the University of São Paulo (NEFIN).

Although our dataset began in 2004, to generate the initial set of features, it is necessary to obtain 12 months of data. Therefore, our initial observation is for February 2005, and our predictive modeling begins in January 2010. Hence, we only used data from February 2005 to December 2009 for model training.

Establishing clear selection criteria is crucial for our analysis. In this manner, we only consider funds that have been active for at least 12 months and have data available for at least 95% of the trading days during the estimation period. To mitigate incubation bias (Evans, 2010), funds with less than 10 million reais (roughly equivalent to 2 million dollars as of February 2024) are excluded. Finally, we adopt the approach of Aggarwal & Jorion (2010) and account for the master-feeder structure commonly employed in the Brazilian mutual fund industry. Firstly, we exclude all master funds from our analysis, as they are inaccessible to investors. Secondly, in cases where a single master fund has several feeder funds beneath it, select the one with the largest asset under management if two funds from the same asset manager and has a correlation exceeding 0.99. Handling outliers in the funds' returns, we winsorize our data by replacing extreme values below the 1st percentile and above the 99th percentile with the values at these percentiles [1].

For a fair comparison between the models' predictions, it's crucial that they all be trained on the same data. If a model has access to more observations or features than the others, it becomes challenging to determine if differences in performance are due to differences in the data or the models' structure. Some algorithms, like linear regression, don't handle missing data natively. Therefore, we exclude observations with null values. Overall, around 400 observations are excluded out of a total of approximately 60,000 (less than 1%).

2.2 Variables

First, we formally define our dependent variable. As in Kaniel et al. (2023), this will be the fund's abnormal return ($R^{abn}_{i,t}$), However, in contrast this authors, factor loading are estimated over the prior 12 months:

$$R_{i,t-12:t-1} = \alpha_i + \hat{\beta}_{i,t-1} F_{t-12:t-1} + \varepsilon_{i,t-12:t-1}. \tag{1}$$

The initial regression (Equation 1) is used to estimate the factor loadings ($\hat{\beta}_{i,t-1}$) of the fund based on historical returns of the four factors Carhart (1997), namely Market, SMB, HML, and WML. As a result, we can compute the abnormal return of the fund i at the time t as follows:

$$R^{abn}_{i,t} = R_{i,t} - \hat{\beta}_{i,t-1} F_t. \tag{2}$$

In summary, the abnormal return of a fund at the time t is the difference between its realized return and its expected return. The expected return is calculated based on the fund's factor loadings from the previous periods ($t-12$ until $t-1$) and the factors' returns at the time t.

To better organize our analysis, we have divided our explanatory variables into two main groups. The first group comprises variables based on returns, and the second group comprises variables based on fund characteristics. Table 1 lists the all variables we consider.

The 11 fund returns-based characteristics we consider are associated with risk (Conditional VaR, kurtosis, and idiosyncratic volatility), indicators utilized in fund evaluation (tracking error, modified information ratio), and variables associated with the regression of the fund's return in relation to Carhart's four-factor model (t-stat alpha, t-stat betas for market, size, value, and momentum factors, and the regression adjustment).

Tabela 1: Explanatory variables categorized into returns-based and fund characteristics-based variables.

Category	Variable	Description
Returns	MIR	Modified Information Ratio
	CVaR	Conditional Value at Risk
	Track Error	Difference between the fund's return and the benchmark
	Kurtosis	Curtose
	t-stat Alpha	Alpha t-stat from model Carhart four-factors
	t-stat Market	Market beta t-stat from model Carhart four-factors
	t-stat Size	Size beta t-stat from model Carhart four-factors
	t-stat Value	Value beta t-stat from model Carhart four-factors
	t-stat Mom	Momentum beta t-stat from model Carhart four-factors
	Adjustament	R-squared from model Carhart four-factors
	IVol	Idiosyncratic Volatility
Characteristics	AUM	Assets Under Management (most recent available information)
	Inflows	Inflows within Last 12 Months
	Outflows	Outflows within Last 12 Months
	# Shareholders	Number of Shareholders
	% Flow	% net funding (inflow minus outflow) relative to AUM (start of the period)
	Leverage	If the fund can take on leverage positions (binary indicator)
	Redemption Period	Redemption Period
	FoF	If the fund is a Fund of Funds (binary indicator)
	Exclusive	If the fund is exclusive (binary indicator)
	Age (Years)	Fund Age (in years)
	Condo Type	If the investor is allowed to redeem the invested capital (binary indicator)

Source: the authors.

We adopt a similar methodology to Kaniel et al. (2023), dividing our analysis into three time frames based on momentum research, but we apply this approach to all return-based variables. Specifically, we consider three periods: (i) short-term momentum ($t-2$), (ii) short-term reversal ($t-1$), and (iii) momentum ($t-12$ until $t-3$). The first two periods are based on Jegadeesh & Titman (1993) and the third on Fama & French (1996).

We do not sort funds according to their estimated alpha. Mamaysky et al. (2007) found that sorting funds based on their estimated alpha does not reliably predict future winners or losers. Instead, the funds in the top and bottom deciles of estimated alpha tend to have the greatest estimation errors. To address this issue, as in DeMiguel et al. (2023), we scale the raw alpha and the betas by the standard error (t-stat), as this procedure better accounts for the estimation error. In addition, like this author, we use the R-squared from model Carhart four-factors as a predictor measure of fund activeness (low-R^2 funds track the benchmark less closely).

In the second group, which includes variables based on the characteristics of the funds, we also utilize eleven commonly used in fund literature. For further information, please

refer to the Cuthbertson et al. (2016).

2.3 Machine-learning models

In this paper, as presents in Table 2, we will consider fourteen machine learning algorithms, categorized into three groups: linear models, ensemble models, and others. Linear models are based on linear combinations of variables, ensemble models combine multiple other models in the prediction process, and the remaining algorithms form a separate category. Most implementations come from the Python package scikit-learn (Pedregosa et al., 2011), except for LightGBM and XGBoost, which have their own Python packages.

Tabela 2: Machine Learning Models Reference.

Acronymous	Algorithm	Type	Reference
XGB	XGBoost	Ensemble	(Chen & Guestrin, 2016)
SVM	Suport Vector Machine	Other	(Cortes & Vapnik, 1995)
RID	Ridge Regresion	Linear	(Hoerl & Kennard, 1970)
RF	Random Forest	Ensemble	(Breiman, 2001)
LR	Linear Regression	Linear	(Seal, 1967)
LGB	Light Gradient Boosting	Ensemble	(Ke et al., 2017)
LASSO	LASSO Regression	Linear	(Tibshirani, 1996)
KNN	K Nearest Neighborhood	Other	(Fix & Hodges, 1989; Altman, 1992)
GB	Gradient Boosting	Ensemble	(Friedman, 2001)
ET	Extra Trees	Ensemble	(Geurts et al., 2006)
EN	Elastic Net	Linear	(Zou & Hastie, 2005)
DUM	Dummy	Other	(Miller & Erickson, 1974)
DT	Decision Tree	Other	(Gordon et al., 1984)
ADA	Ada Boost	Ensemble	(Freund & Schapire, 1997)

Source: the authors.

It is also valid to state that each month we normalize the features in both the training and test sets. To avoid data leakage, we estimate the mean and standard deviation using only the training set. We follow this procedure because some models rely on the calculation of distances, which is sensitive to the feature's scale.

Similar to DeMiguel et al. (2023), we organize our data in a panel structure, and we start our analysis with the traditional econometric regression model estimated via OLS (pooled regression), which we use as a benchmark to identify the most significant variables for prediction. Although machine-learning methods often outperform OLS, this comparative analysis is useful to evaluate and discuss differences between linear models and methods that capture nonlinearities.

For the specific task of identifying the importance of variables in predicting abnormal returns of funds and building portfolios, we have chosen the XGBoost algorithm as our primary model, and we have compelling justifications for this decision. Firstly, XGBoost is renowned for its computational efficiency (Chen & Guestrin, 2016). Moreover, XGBoost has consistently demonstrated outstanding performance across diverse domains and machine learning tasks (Fauzan & Murfi, 2018; Zhang et al., 2020; Giannakas et al., 2021).

The XGBoost algorithm was also the model used to detail and discuss the allocation of funds to long-only and long-short strategies (classifying fund forecasts based on features monthly) and and evaluate the performance of machine learning models. However,

to ensure a comprehensive comparison of the performance of various machine learning algorithms, we will present key results for all models.

For a complete explanation, we refer the reader to the papers in Table 2 and for a detailed description of the methods Coqueret & Guida (2020).

2.4 Portfolio construction and allocation

In this section, we describe in detail the procedure for Portfolio construction and allocation to evaluate the performance of the resulting strategies. All portfolios are based on the deciles of the predicted-performance distribution by Machine Learning models, with rebalance every month. We proposed one long-only and three long-short portfolios.

First, we construct an long-only equally weighted portfolio that invested in each fund within each decile of the predicted-performance distribution by the machine learning algorithm. Although equal-weighted $(1/n)$ approach might initially appear overly simplistic or "naive,"it has demonstrated out-of-sample superiority over other methodologies in prior studies (DeMiguel et al., 2009; Plyakha et al., 2012; Malladi & Fabozzi, 2017)

This portfolio is rebalanced monthly. Thus, for each successive month, we extend the sample forward one period, train the algorithm again on the expanded sample, make new predictions, construct a new decile portfolio, and track its return. In this way, we construct a time series of monthly out-of-sample returns of the decile portfolio from February 2010 to February 2022 (144 months).

Based on this series of returns, we evaluate the performance of the portfolios considering the measures of annualized return, standard deviation, alpha, t(alpha), beta, Information Ratio, Sharpe Ratio, Track Error, CVaR, and Maximum Drawdown. Additionally, we analyze the differences among the characteristics of these funds.

Despite the fact that long-short strategies are not a common practice among fund managers due to the inherent characteristics of this financial instrument, it is worth noting that this strategy serves as a viable benchmark for evaluating the relative efficacy of the best versus the worst funds and, consequently, of machine learn models. Therefore, we present three distinct strategies for developing a long-short strategy based on machine learn predictions.

Each of these portfolios long-short adheres to the following structure: a 100% long position is taken in the top 30% of funds predicted to have the highest abnormal return, a 100% short position is taken in the bottom 30% of funds anticipated to yield the lowest abnormal return, and finally, a 100% long position is allocated towards the risk-free rate. The sole variance across these portfolios resides in the method utilized for determining the weight of each fund within the long and short portfolios.

The first portfolio implements the equal-weighted $(1/n)$ method. The composition of the second portfolio is determined by a ranking-based methodology. Unlike (Kaniel et al., 2023), we also introduce results based on ranking weights, represented by the following equation 3:

$$w_{i,t} = \frac{i_t}{\sum_i i_t}, \tag{3}$$

where $i = 1, ..., n$ corresponds to the index of each fund in the predictive ascending ranking, while $w_{i,t}$ means the final weight assigned to each respective fund within the

portfolio.

While the third and final portfolio is constructed utilizing raw predictions, where we calculate the portfolio weights for each tercile based on the predictions as equation 4:

$$w_{i,t} = \frac{|\mu_{i,t}|}{\sum_i |\mu_{i,t}|}, \tag{4}$$

where $\hat{\mu}_{i,t}$ are the machine learn model predictions, and $w_{i,t}$ are the final weights.

In each instance, we integrate additional information into our portfolio weighting scheme. The baseline case, which is the equal-weighted approach, does not integrate any of our predictive data into the weighting scheme. The second method incorporates only the ranking information; thus, even if a fund is projected to have twice the abnormal return, it will have a comparable weight to another fund provided they are near to the ranking. The final weighting scheme, which considers the raw predictions, avails itself of all available information concerning the cross-sectional distribution of expected returns for each fund.

3 Empirical results

One concern when using machine learning models is that they require a considerable amount of data to be effective (Yao, 2021). Because we consider in our analysis an extended period and evaluate a developing market, valid concerns can be raised about the appropriateness of our approach. Therefore, we begin our results by presenting the Table 3, which presents descriptive statistics for the features under consideration, wherein returns-based variables are measured by varying time lags. We observe that, on average, around 290 funds meet our pre-processing criteria, which effectively eliminates this concern.

Tabela 3: Data Summary Statistics.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
# Funds	28	141	346	290	379	712
Return-based						
Abnormal Return	-0.34	-0.01	0	0	0.01	0.34
MIR_{t-2}	0	0	0	0.08	0.14	1.11
CVaR_{t-2}	-0.24	-0.03	-0.02	-0.02	-0.01	0
Track Error_{t-2}	0	0	0.01	0.01	0.01	0.07
Kurtosis_{t-2}	-1.86	-0.71	-0.29	0.05	0.34	18.05
$\text{t-stat Alpha}_{t-2}$	-7171.94	-0.63	0.04	-0.49	0.7	6.41
$\text{t-stat Market}_{t-2}$	-15.55	3.73	6.76	8.31	10.96	87.85
t-stat Size_{t-2}	-5.12	-0.09	0.77	0.85	1.71	10.54
$\text{t-stat Value}_{t-2}$	-8.36	-1.25	-0.35	-0.41	0.53	7.68
t-stat Mom_{t-2}	-7.89	-0.61	0.25	0.28	1.13	9.28
Adjustment_{t-2}	0	0.68	0.86	0.77	0.94	1
IVol_{t-2}	0	0	0	0	0.01	0.05
$\text{MIR}_{t-12:t-3}$	0	0	0	0.03	0.05	0.4
$\text{CVaR}_{t-12:t-3}$	-0.21	-0.03	-0.03	-0.03	-0.02	0
$\text{Track Error}_{t-12:t-3}$	0	0	0.01	0.01	0.01	0.03
$\text{Kurtosis}_{t-12:t-3}$	-0.89	0.48	1.02	3.13	2.69	147.18
$\text{Alpha}_{t-12:t-3}$	-202.18	-0.61	0.13	0.15	0.92	7.04
$\text{t-stat Market}_{t-12:t-3}$	-14.64	15.1	25.16	29.35	37.86	212.88
$\text{t-stat Size}_{t-12:t-3}$	-11.93	1.1	2.87	3.01	4.78	20.33
$\text{t-stat Value}_{t-12:t-3}$	-13	-2.58	-0.85	-1.1	0.7	13.39
$\text{t-stat Mom}_{t-12:t-3}$	-11.55	-0.62	0.93	1.07	2.58	13.16
$\text{Adjustment}_{t-12:t-3}$	0	0.63	0.82	0.74	0.91	1
$\text{IVol}_{t-12:t-3}$	0	0	0	0.01	0.01	0.02
MIR_{t-1}	0	0	0	0.08	0.14	1.11
CVaR_{t-1}	-0.24	-0.03	-0.02	-0.02	-0.01	0
Track Error_{t-1}	0	0	0.01	0.01	0.01	0.07
Kurtosis_{t-1}	-1.8	-0.71	-0.29	0.05	0.34	18.05
Alpha_{t-1}	-4660.4	-0.64	0.04	-0.46	0.7	6.41
$\text{t-stat Market}_{t-1}$	-15.55	3.75	6.8	8.32	10.97	87.85
t-stat Size_{t-1}	-5.94	-0.09	0.78	0.86	1.71	10.54
$\text{t-stat Value}_{t-1}$	-8.36	-1.27	-0.36	-0.42	0.53	7.68
t-stat Mom_{t-1}	-7.89	-0.61	0.24	0.28	1.12	9.28
Adjustment_{t-1}	0	0.68	0.86	0.77	0.94	1
IVol_{t-1}	0	0	0	0	0.01	0.05
Characteristics-based						
AUM	10000.93	24187.81	60126.4	192562.79	167145.3	9718218.61
Inflows	0	208.71	7667.66	59177.43	40546.63	14539912.7
Outflows	0	414.7	8310.04	47376.61	37171.34	9950092.02
# Shareholders	0	2	11	1403.31	103	877812
% Flow	-38695.67	-0.12	0	2.04	0.25	42526.88
Redemption Period	0	4	5	33.78	57	3601
Age (Years)	1	2.52	4.73	5.96	8.1	36.81

Note: "MIR"represents the Modified Information Ratio. The "t-stat"prefix indicates Carhart model coefficients. The "Adjustment"is the coefficient of determination of the Carhart model. "IVOL"represents Idiosyncratic Volatility.
Source: the authors.

Other observations that can be made about the data in the Table 3 are that upon analyzing the distributions of the variables, it can be noted that the median value of AUM is approximately $60 million, and the average age of the funds is six years. The number of shareholders shows a relatively low median of only eleven, with a mean value close to 1400, indicating that a few funds hold the majority of shareholders.

Looking at the frequency of the dummy variables, we identified that around half of the funds can take on leveraged positions, with around 41% being Funds of Funds (FoFs). Most funds are open, whereas exclusive funds constitute only approximately 11% of the total.

3.1 Importance of Predictors

This section is dedicated to presenting the results for the Pooled Regression and XGboost algorithm, where we will discuss the degree of importance of the variables and the performance of its predictions for equity funds.

First of all, we analyzed the results of the pooled regression (Table 4), estimating 45 parameters. Among them, 22 were statistically significant at the 5% level. Due to space limitations, variables that did not exhibit statistical significance at the 10% level were excluded from the presentation.

Tabela 4: Pooled Regression.

	Dependent variable:
	Abnormal Return
MIR_{t-2}	0.002^{*} (0.001)
CVaR_{t-2}	-0.045^{***} (0.013)
t-stat Value_{t-2}	0.001^{***} (0.0001)
t-stat Momentum_{t-2}	0.0005^{***} (0.0001)
Adjustment_{t-2}	-0.005^{***} (0.001)
IVol_{t-2}	-0.375^{***} (0.110)
$\text{MIR}_{t-12:t-3}$	0.010^{***} (0.003)
$\text{Kurtosis}_{t-12:t-3}$	0.0002^{***} (0.00002)
$\text{Alpha}_{t-12:t-3}$	0.001^{***} (0.0001)
t-stat $\text{Market}_{t-12:t-3}$	-0.0001^{***} (0.00001)
t-stat $\text{Size}_{t-12:t-3}$	-0.0003^{***} (0.0001)
t-stat $\text{Value}_{t-12:t-3}$	-0.001^{***} (0.0001)
t-stat $\text{Momentum}_{t-12:t-3}$	-0.001^{***} (0.0001)
$\text{Adjustment}_{t-12:t-3}$	0.005^{***} (0.001)
$\text{IVol}_{t-12:t-3}$	-0.467^{***} (0.106)
MIR_{t-1}	0.002^{*} (0.001)
Kurtosis_{t-1}	-0.001^{***} (0.0001)
t-stat Size_{t-1}	-0.0003^{***} (0.0001)
t-stat Momentum_{t-1}	0.0003^{***} (0.0001)
# Shareholders	-0.000^{*} (0.000)
% Flow	0.00000^{***} (0.0000)
Redemption Period	-0.00000^{***} (0.0000)
Condo Type	0.002^{**} (0.001)
Exclusive	-0.001^{***} (0.0004)
Constant	0.004^{***} (0.001)
Observations	59,456
R^2	0.017
Adjusted R^2	0.017
Residual Std. Error	0.028 (df = 59411)
F Statistic	23.686^{***} (df = 44; 59411)

Note: "MIR"represents the Modified Information Ratio. The "t-stat"prefix indicates Carhart model coefficients. The "Adjustment"is the coefficient of determination of the Carhart model. "IVOL"represents Idiosyncratic Volatility. The subscripts represent different time lags for the variables, where $t-2$ represents short-term momentum, $t-1$ represents short-term reversal and $t-12:t-3$ represents momentum. *, ** and *** represents the significance levels of 1%, 5% and 10%, respectively. *Source:* the authors.

Results indicate that return-based metrics provided more valuable information for predicting future abnormal returns compared to feature-based metrics. Specifically, nearly 50% of the return-based features were statistically significant in our model, whereas only around 35% of the characteristics based ones were. This strongly supports the use of return-based metrics. It's worth noting that age and AUM were not significant, despite a vast literature that links these variables to future performance.

Furthermore, upon examining the coefficients, we observed a positive relationship between risk and abnormal return for shorter terms, as evidenced by the negative statistically significant coefficients of CVar_{t-1} and idiosyncratic volatility_{t-1}, and the positive statistically

significant coefficients of t-stat Value$_{t-1}$ and t-stat Momentum$_{t-1}$. This finding aligns with the principles of modern finance as outlined by (Markowitz, 1952; Sharpe, 1964). However, when we looked at more extended periods ($t-12$ until $t-3$), the results were mixed, with all t-stat betas indicating a higher abnormal return for lower scaled betas. This observation is consistent with recent literature that emphasizes the superior performance of less risky assets relative to riskier ones (Blitz & Van Vliet, 2007; Houweling & van Zundert, 2017).

Finally, it is worth noting that only one out of the three fund flow metrics, namely % Flow, was found to be statistically significant. The lack of statistical significance of inflow contradicts a large body of literature, as discussed earlier, which establishes a link between fund inflows and future performance (Gruber, 1996; Zheng, 1999; Keswani & Stolin, 2008). However, it is possible to argue that % Flow incorporates all relevant information related to fund flows.

The pooled panel regression, while offering a traditional modeling approach, exhibits limited explanatory power and identifies certain variables as significant predictors that may not be highlighted by the XGBoost model. Panel models and XGBoost represent distinct modeling methods, and while direct comparison between them may not be appropriate, it is valuable to consider their complementarity. Thus, presenting the results from this traditional model serves as a baseline for comparison with those obtained through the XGBoost algorithm. This comparison allows us to evaluate the performance of the machine learning approach and assess whether it offers insights beyond those provided by conventional statistical methods. Additionally, it is important to note that while some features may not demonstrate significance in conventional statistical methods, they can still play a crucial role in machine learning algorithms, capturing nonlinear relationships and complexities in interactions between variables.

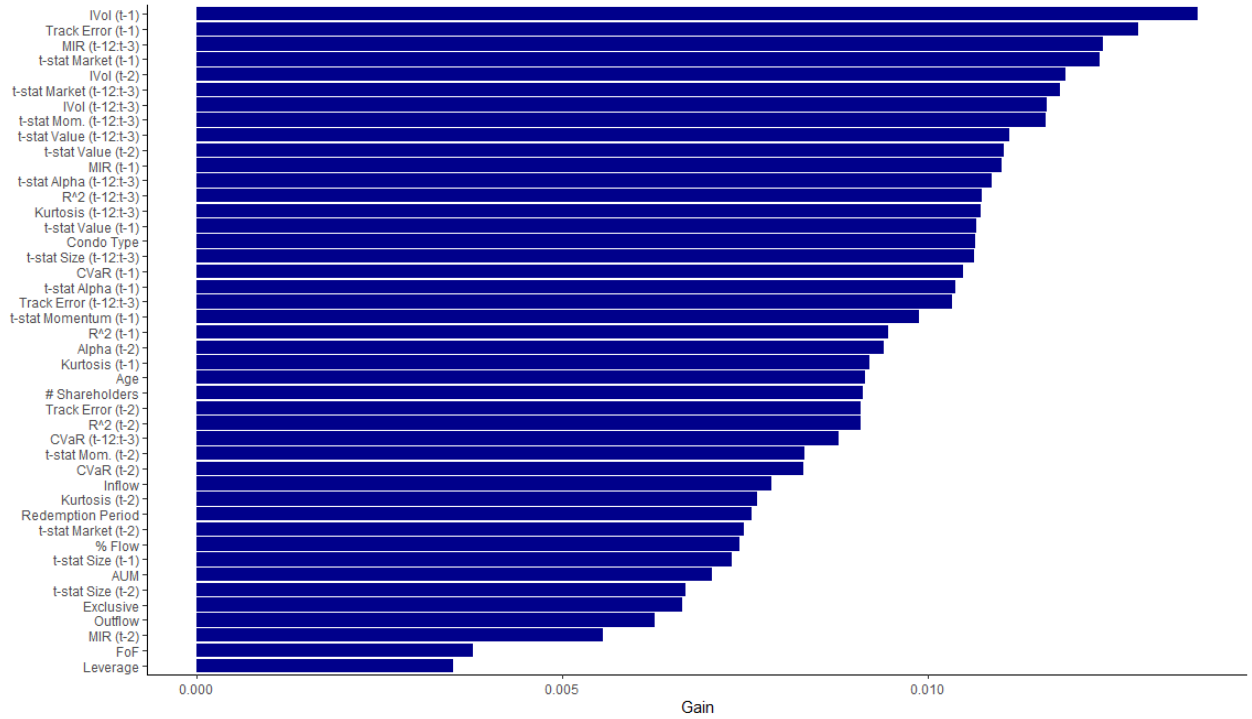
In Figure 1, we present the results to XGboost algorithm and observe that idiosyncratic volatility for the short-term reversal ($t-2$) time frame is the most important variable for predicting abnormal return[2]. Furthermore, IVol$_{t-1}$ and IVol$_{t-12:t-3}$ are the fifth and seventh most crucial features, respectively, which suggests that incorporating multiple time frames can substantially enhance the model's performance.

Another interesting finding is the prevalence of the periods $t-12:t-3$ and $t-1$ periods over $t-2$. Only two out of the top twenty most important features pertain to the $t-2$ time frame. Moreover, risk-related metrics dominate the list of the most significant features. Out of the top twenty variables, only four are not directly related to risk, which are the Modified Information Ratio ($t-1$ and $t-12:t-3$), t-stat Alpha ($t-12:t-3$), and the dummy variable indicating if the investor can redeem their money (Condo Type).

As with pooled regression, return-based metrics are in fact highly valuable for predicting abnormal returns for equity funds. However, the low importance given to resource-based metrics is puzzling given that past research has suggested that metrics such as fund flows, assets under management, lock-in period and age can strongly predict a fund's future returns. Interestingly, a study by (Kaniel et al., 2023) identified time and cash flow as the most important predictors, which contrasts with our findings.

However, it is crucial to recognize that the XGboost algorithm's emphasis on return-related metrics, especially idiosyncratic volatility, does not invalidate the use of feature-based metrics. In our analysis, condo type and fund age are the resource-based variables that rank highest in importance.

Figura 1: XGBoost Feature Importance



3.2 Portfolio performance

After our analysis of the Importance of Predictors, to assess the efficacy of the model in distinguishing between equity funds with good versus bad relative future performance, we construct one long-only portfolio and using three approaches by long-short portfolios.

3.2.1 Portfolio long-only

Long-only strategies are based an equally weighted portfolio between funds in the same decile, resulting in ten portfolios.

Table 5 Panel A shows how effective the XGBoost model was in achieving the task mentioned earlier. The results indicate that the first decile had a 4.75 times higher return than the last decile. Moreover, the first decile had a 13% lower risk compared to the last decile. It's worth noting that during the same period, the Brazilian market index (IBrX) had an annualized return of 7% and an annualized volatility of 23.52%. Therefore, the first decile significantly outperformed the market in terms of both total and risk-adjusted returns, even after accounting for fees.

Vardharaj et al. (2004) notes that when an active manager takes positions that deviate substantially from the benchmark, the manager will generate significant active returns, whether positive or negative. The results presented in Table 5 demonstrate this parabolic relationship: the extreme deciles exhibit greater tracking errors and significantly higher absolute returns. Conversely, the deciles in the middle display lower tracking errors and lower absolute returns.

Moreover, it is worth noting the alpha of each decile. As anticipated, the first decile had the highest (numerical) four-factor alpha, while the last decile had the lowest. Surprisin-

gly, only the tenth decile had a significant and negative alpha, indicating that it destroyed value, while the other deciles neither created nor destroyed any value. One possible explanation for the insignificant alpha of the first few deciles may be that we utilized after-fee returns (Fama & French, 2010).

To further examine the characteristics of the deciles, we now analyze their average attributes. Table 5 Panel B reveals that the funds with higher predicted abnormal returns tend to have certain characteristics on average. Specifically, they tend to have a larger asset under management, be younger in age, have fewer shareholders, and receive more inflows. The observation about higher AUM for funds with higher predicted abnormal returns contradicts the literature (Chen et al., 2004; Yan, 2008). However, the finding that higher predicted returns correspond to higher inflows is consistent with the concept of the "smart money"effect (Gruber, 1996; Zheng, 1999).

Finally, as we have shown, funds with higher predicted abnormal returns tend to have larger assets under management and fewer shareholders. This suggests that a small group of more financially capitalized investors may be better able to identify funds with good or bad future performance. Conversely, a group with more members but less financial capital tends to select funds with lower predicted abnormal returns. It would be interesting for future research to investigate whether there is a correlation between these groups and institutional and retail investors.

Tabela 5: Deciles Return Statistics and Average Characteristics

Panel A: Deciles Return Statistics

	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
Annual. Return	9.88	9.23	8.05	9.58	8.76	7.41	8.52	7.28	5.44	2.08
Std. Deviation	15.32	16.73	17.35	17.58	17.83	17.86	18.22	18.13	18	17.61
Alpha	1.22	0.36	-0.59	0.87	0.2	-0.88	0.14	-0.78	-2.15	-5.04
t(alpha)	0.87	0.3	-0.5	0.76	0.18	-0.8	0.12	-0.67	-1.69	-3.14
Beta	0.66	0.74	0.77	0.78	0.8	0.8	0.81	0.81	0.8	0.76
Info. Ratio	0.1	0.08	0	0.16	0.06	0	0.03	0	0	0
Sharpe Ratio	0.14	0.11	0.05	0.13	0.09	0.02	0.08	0.01	0	-0.01
Track Error	9.5	7.92	7.38	6.97	6.68	6.58	6.42	6.59	7.11	8.56
CVaR	-3.4	-3.7	-4.29	-4.48	-4.53	-4.51	-4.34	-3.88	-3.97	-3.56
Max. Drawdown	35.57	41.46	42.11	42.12	42.75	42.57	42.18	43.44	43.91	43.29

Panel B: Average Characteristics

	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
AUM	225696.55	195822.14	199612.77	197739.88	192806.16	187126.19	178707.18	180533.48	193438.97	223344.34
Inflows	75768.43	64515.07	63942.30	65398.61	63814.87	58245.97	55918.04	54544.83	56849.67	57297.01
Outflows	59053.90	48399.30	49197.72	49073.77	47864.02	46190.83	46304.01	41840.85	46903.76	55479.41
# Shareholders	853.48	598.11	600.86	575.53	569.22	581.56	712.76	1075.07	1253.86	2012.51
Leverage	0.52	0.53	0.52	0.53	0.51	0.51	0.48	0.49	0.49	0.46
FoF	0.40	0.45	0.45	0.43	0.43	0.43	0.43	0.43	0.40	0.32
Exclusive	0.08	0.09	0.09	0.11	0.11	0.12	0.12	0.13	0.12	0.12
Age (Years)	5.58	5.85	6.04	6.08	6.14	6.01	6.08	6.05	6.11	6.01
Redemption Period	38.65	41.55	36.08	33.39	32.22	31.88	30.94	33.81	34.72	43.35

Note: Market is the Market Factor, calculated by NEFIN, represents the difference between the value-weighted daily return of the market portfolio and the daily risk-free rate. The daily risk-free rate is determined using the 30-day DI Swap. IBrX is the Brazilian market index (IBrX), used in this study as a market reference, its annualized return, standard deviation, Sharpe ratio, CVaR and maximum drawdown were 7.4, 22.2, 0.05, -4, 82 and 44.99, respectively. *Source:* the authors.

3.2.2 Portfolio long-short

In line with the approach used by Kaniel et al. (2023), in this section we will analyze the effects of the weighting method on the final portfolio. Through this analysis, we aim to determine the efficacy of our predictive model in distinguishing between exceptional (or disastrous) mutual funds within a data set that already comprises good (or bad) ones.

In Subsection behind, we have demonstrated that our predictive model are efficient at distinguishing between top-tier and bottom-tier performers. Following this, our current focus is to ascertain whether the rankings derived from our predictions, as well as the predictions themselves, can enhance the value of a preexisting portfolio composition.

Table 6 reveals a positive monotonic relationship between the alpha and the extent of information derived from our predictions. In this spectrum, the equal-weighted portfolio, utilizing the least amount of predictive information, yields the lowest alpha. However, it's worth noting that this alpha remains positive and is statistically significant.

Tabela 6: Long-short Portfolios Weighting Schemes.

	Equal Weighted	**Rank Weighted**	**Prediction Weighted**
Annual. Return	12.72	14.54	16.84
Std. Deviation	3.37	4.25	5.21
Alpha	2.98	4.63	6.89
t(alpha)	3.27	3.94	4.66
Beta	-0.05	-0.06	-0.07
Info. Ratio	0.1	0.17	0.26
Sharpe Ratio	1.09	1.26	1.45
Track Error	23.61	23.98	24.34
CVaR	-0.42	-0.52	-0.61
Max. Drawdown	3.87	4.69	6.02

Note: The annualized risk-free rate considered was 8.8.
Source: the authors.

The ranking-based portfolio incorporates a larger proportion of the predictive information and exhibits a substantially higher alpha, improving upon the equal-weighted alpha by over 55%. Lastly, the portfolio based on raw predictions, which fully employs all available predictive information, exhibits an alpha that is more than double that of the equal-weighted portfolio and 1.48 times that of the rank-weighted portfolio.

These findings strongly indicate that our abnormal return predictions hold substantial value, providing key insights into future returns that significantly exceed those necessary for tercile/decile construction. This constitutes an intriguing revelation, implying that these predictions can be effectively leveraged for both fund selection and weight definition concurrently.

In conclusion, it is crucial to highlight that all the long-short portfolios generated a positive alpha, which was both statistically and economically significant at a 5% level. In addition, the portfolios demonstrated a robust annualized return coupled with low volatility, resulting in a highly favorable Sharpe Ratio of 0.7.

From a risk perspective, the portfolios exhibited low Conditional Value-at-Risk (CVaR) and Maximum Drawdown, further enhancing their investment appeal. Although mutual funds cannot be shorted in reality, these results compellingly suggest that investors would

be well-advised to sidestep the funds predicted to perform poorly and, conversely, favor those predicted to be high performers.

3.3 Comparison of machine learning algorithms

As mentioned earlier, we chose XGBoost to evaluate the effectiveness of using machine learning algorithms in predicting abnormal returns for equity funds. Regardless of this a priori choice, it is relevant to include a comparison between the different ML models in the study. This comparison will provide information on the performance of the evaluated models, allowing the investor to choose other algorithms that present an expected performance equal to or greater than that of XGBoost. For this, we refer to Figure 2, where the x-axis represents the out-of-sample R^2 for predictions made by each ML model.

In order to assess the performance of the portfolios resulted from the various ML algorithms in discriminating between high-performing and low-performing equity mutual funds, we implement the equal-weighted long-short methodology delineated in the preceding subsection (refer to subsection 3.2). More explicitly, each month we rank the funds based on the predictions generated by each model. Subsequently, we create multiple equal-weighted long-short portfolios that take long positions in the top 30% of funds based on these predictions and short positions in the bottom 30% of funds.

The y-axis in Figure 2 illustrates the four-factor alpha (Carhart, 1997) of these portfolios for each corresponding ML algorithm. This provides a comparative view of the alpha generated by each algorithm, offering insights into their relative effectiveness in predicting mutual fund performance.

Additionally, we have scaled the points on the graph based on the average training time (in seconds) for each model on every chronological data split. This allows us to examine the relationship between model performance and computational cost. Finally, the points are color-coded based on the model type, as listed in Table 2.

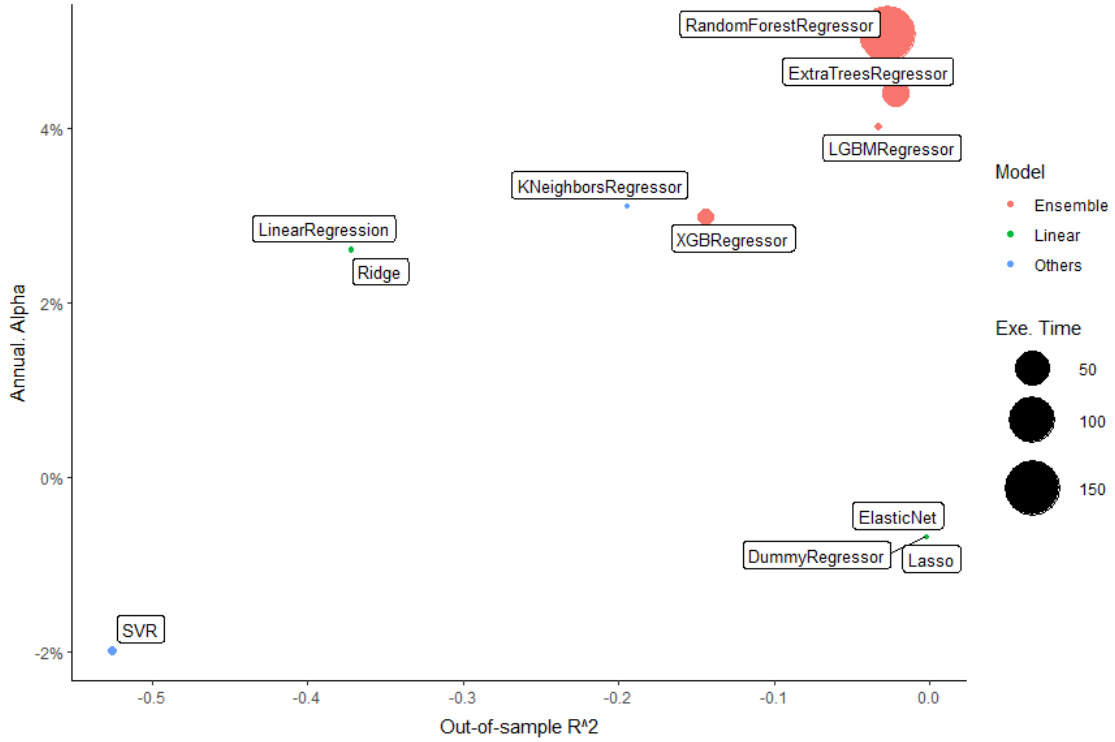
Our comparison between model types reveals that the ensemble methods performed remarkably well, producing high alphas with varying levels of out-of-sample R^2. In particular, the ensemble models outperformed the linear models, providing additional evidence that nonlinear relationships and interactions between the variables are present. Additionally, we observed that Support Vector Machines performed much worse than the baseline model (Dummy), suggesting that SVR may not be suitable for this task, and Decision Trees work much better in an ensemble model (Random Forest).

Furthermore, Elastic Net and Lasso models were unable to make predictions that differed from the average abnormal return observed in the training set. As a result, these algorithms produced the same portfolio as the Dummy model, which is a portfolio that assigns equal weight to every mutual fund in our investment universe. This indicates that the regularization term in the equation for Lasso may be too large, causing the added loss from this term to outweigh the reduction in error.

Moreover, we can observe that despite having models that generated a high alpha, all of them had a negative out-of-sample R^2 score. The reason for this result can be explored in future research, as there appears to be no correlation between alpha generation and metrics used to evaluate prediction performance.

In conclusion, the Light Gradient Boosting model stands out as the best-performing model in terms of the performance-cost trade-off. Despite ranking third in terms of alpha

Figura 2: ML Model Comparison



Note:"Annual. Alpha"is the Carhart (1997) alpha annualized over 252 days of the long-short portfolio. "Execution time"is the time (seconds) for the model to train on the data from February 2005 to January 2022 and predict February 2022. Refer to Table 2 for the acronymous meanings.
Source: the authors.

generation, LGB required less than 2% of the time needed by the top two algorithms on average. This finding is consistent with the literature, which has highlighted the superior performance of LGB in several domains ((Li & Rossi, 2021; Ke et al., 2017)). Although our initial choice, XGBoost, did not perform as well as LGB and the Random Forest algorithm, it was still able to accurately differentiate between good and bad equity mutual funds (as shown in Table 5).

4 Conclusion

In this study, we contribute to the existing literature by providing further evidence of the effectiveness of machine learning models in distinguishing equity mutual funds performance. Our results demonstrate that the Ensemble Models group outperforms other groups of models. Among the Ensemble Models, Light Gradient Boosting (LGB) performs the best in terms of predicting future winners and losers, while considering the balance between predictive power and computational resources required. However, if we focus solely on the portfolio alpha, Random Forest is the most suitable algorithm.

Although our initial choice of model (XGBoost) did not perform as well as LGB and Random Forest, it still enabled us to rank the mutual funds based on predicted abnormal returns. This ranking revealed a stark contrast between the performance of the first decile, characterized by higher predicted abnormal returns, and the last decile, marked by lower predicted abnormal returns. Specifically, the first decile outperformed the last decile by almost five times in terms of returns, while also exhibiting 13% lower risk. Additionally, it is noteworthy that the last decile generated a negative and statistically significant alpha.

Furthermore, our study delved into the use of long-short portfolio construction strategies. Though these strategies were purely theoretical assumptions, they offered an objective means to visualize the behavior of machine learning models in predicting stock fund performance. This provide additional evidence in favor of the notion that Machine Learning algorithms hold greater predictive power compared to traditional statistical methods like linear models. Specifically, our best-performing ML model (Random Forest) generated an alpha almost twice as high as the best linear model (Linear Regression).

Our investigation revealed that the most critical predictors for fund performance were return-based metrics, particularly risk-based ones, with idiosyncratic volatility ranking as the first, fifth, and seventh most important variable. Interestingly, contrary to previous research, we discovered that metrics based on the fund's characteristics, such as assets under management, flows, age, and redemption period, were not very relevant for predicting performance.

This work has the potential to benefit society in many ways. Given that many Brazilians have financial exposure to the market through personal savings, Funds of Funds, and retirement plans, it is crucial to have a systematic way to identify future winners and avoid future losers. By doing so, we can improve the investment process for a large group of people and institutions, making it more robust and reliable. Moreover, this method can make the market more efficient by rewarding skilled managers and penalizing unskilled ones. In the future, this type of analysis is likely to become commonplace, leading to a more efficient and developed market with greater benefits for society.

Finally, we offer some suggestions for future research. First, we recommend hyperparameter tuning in a validation set before making predictions. Second, we suggest employing more reliable methods for identifying and addressing outliers. In addition, it would be worthwhile to examine how alpha decreases over longer holding periods. Lastly, we recommend exploring alternative models such as Neural Networks and techniques like conformal prediction to evaluate uncertainty.

Notes

[1]Additionally, we manually identify and correct approximately 50 observations, transforming their values to missing. These corrections were mostly necessary due to the initial net asset value (NAV) being set to 1 for some funds on their first day and readjusting to another base, such as 10, on their second day.

[2]To analyze XGBoost feature importance, we use information gain - average gain (Equation **??**) of splits which use the feature.

Referências

Adams, J. C., Hayunga, D. K., & Mansi, S. (2018). Diseconomies of scale in the actively-managed mutual fund industry: What do the outliers in the data tell us? *Available at SSRN 3194005*.

Aggarwal, R. K. & Jorion, P. (2010). The performance of emerging hedge funds and managers. *Journal of financial economics*, 96(2), 238–256.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3), 175–185.

Blake, D. (2015). Portfolio performance measurement. *Wiley Encyclopedia of Management*, (pp. 1–6).

Blitz, D. C. & Van Vliet, P. (2007). The volatility effect. *The Journal of Portfolio Management*, 34(1), 102–113.

Bogle, J. C. (1992). Selecting equity mutual funds. *Journal of Portfolio Management*, 18(2), 94.

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

Brown, S. J. & Goetzmann, W. N. (1995). Performance persistence. *The Journal of finance*, 50(2), 679–698.

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

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

Chen, T. & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Chua, A. K. P. & Tam, O. K. (2020). The shrouded business of style drift in active mutual funds. *Journal of Corporate Finance*, 64, 101667.

Coqueret, G. & Guida, T. (2020). *Machine learning for factor investing: R version*. CRC Press.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273–297.

Cuthbertson, K., Nitzsche, D., & O'Sullivan, N. (2016). A review of behavioural and management effects in mutual fund performance. *International Review of Financial Analysis*, 44, 162–176.

DeMiguel, V., Garlappi, L., & Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The review of Financial studies*, 22(5), 1915–1953.

DeMiguel, V., Gil-Bazo, J., Nogales, F. J., & Santos, A. A. (2023). Machine learning and fund characteristics help to select mutual funds with positive alpha. *Journal of Financial Economics*, 150(3), 103737.

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

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

Fama, E. F. & French, K. R. (1996). Multifactor explanations of asset pricing anomalies. *The journal of finance*, 51(1), 55–84.

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

Fauzan, M. A. & Murfi, H. (2018). The accuracy of xgboost for insurance claim prediction. *Int. J. Adv. Soft Comput. Appl*, 10(2), 159–171.

Fix, E. & Hodges, J. L. (1989). Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, 57(3), 238–247.

Freund, Y. & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1), 119–139.

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

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1), 3–42.

Giannakas, F., Troussas, C., Krouska, A., Sgouropoulou, C., & Voyiatzis, I. (2021). Xgboost and deep neural network comparison: The case of teams' performance. In *International Conference on Intelligent Tutoring Systems* (pp. 343–349).: Springer.

Gil-Bazo, J. & Ruiz-Verdú, P. (2009). The relation between price and performance in the mutual fund industry. *The Journal of Finance*, 64(5), 2153–2183.

Goetzmann, W. N., Ingersoll Jr, J. E., & Ross, S. A. (2003). High-water marks and hedge fund management contracts. *The Journal of Finance*, 58(4), 1685–1718.

Gordon, A., Breiman, L., Friedman, J., Olshen, R., & Stone, C. J. (1984). Classification and regression trees. *Biometrics*, 40(3), 874.

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

Harvey, C. R. & Liu, Y. (2018). Detecting repeatable performance. *The Review of Financial Studies*, 31(7), 2499–2552.

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.

Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67.

Houweling, P. & van Zundert, J. (2017). Factor investing in the corporate bond market. *Financial Analysts Journal*, 73(2), 100–115.

Hu, M., Chao, C.-C., & Lim, J. H. (2016). Another explanation of the mutual fund fee puzzle. *International Review of Economics & Finance*, 42, 134–152.

Jegadeesh, N. & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of finance*, 48(1), 65–91.

Jones, C. S. & Mo, H. (2021). Out-of-sample performance of mutual fund predictors. *The Review of Financial Studies*, 34(1), 149–193.

Kaniel, R., Lin, Z., Pelger, M., & Van Nieuwerburgh, S. (2023). Machine-learning the skill of mutual fund managers. *Journal of Financial Economics*, 150(1), 94–138.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Keswani, A. & Stolin, D. (2008). Which money is smart? mutual fund buys and sells of individual and institutional investors. *The Journal of Finance*, 63(1), 85–118.

Li, B. & Rossi, A. G. (2021). Selecting mutual funds from the stocks they hold: A machine learning approach. *Working Paper*.

Malladi, R. & Fabozzi, F. J. (2017). Equal-weighted strategy: Why it outperforms value-weighted strategies? theory and evidence. *Journal of Asset Management*, 18, 188–208.

Mamaysky, H., Spiegel, M., & Zhang, H. (2007). Improved forecasting of mutual fund alphas and betas. *Review of Finance*, 11(3), 359–400.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1), 77–91.

Miller, J. L. & Erickson, M. L. (1974). On dummy variable regression analysis: A description and illustration of the method. *Sociological Methods & Research*, 2(4), 409–430.

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Plyakha, Y., Uppal, R., & Vilkov, G. (2012). Why does an equal-weighted portfolio outperform value-and price-weighted portfolios? *Available at SSRN 2724535*.

Rubesam, A. (2022). Machine learning portfolios with equal risk contributions: Evidence from the brazilian market. *Emerging Markets Review*, 51, 100891.

Seal, H. L. (1967). Studies in the history of probability and statistics. xv the historical development of the gauss linear model. *Biometrika*, 54(1-2), 1–24.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, 19(3), 425–442.

Sterenfeld, D. (2023). *Aplicando aprendizado por máquina na seleção de fundos de ações brasileiros*. PhD thesis, Fundação Getulio Vargas-São Paulo.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.

Vardharaj, R., Fabozzi, F. J., & Jones, F. J. (2004). Determinants of tracking error for equity portfolios. *The Journal of Investing*, 13(2), 37–47.

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

Yan, X. S. (2008). Liquidity, investment style, and the relation between fund size and fund performance. *Journal of Financial and Quantitative Analysis*, 43(3), 741–767.

Yao, F. (2021). Machine learning with limited data.

Zhang, Y., Tong, J., Wang, Z., & Gao, F. (2020). Customer transaction fraud detection using xgboost model. In *2020 International Conference on Computer Engineering and Application (ICCEA)* (pp. 554–558).: IEEE.

Zheng, L. (1999). Is money smart? a study of mutual fund investors' fund selection ability. *The Journal of Finance*, 54(3), 901–933.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2), 301–320.