Using Machine Learning to Separate Good and Bad Equity Mutual Fund Managers: Evidence from Brazil

2022

Abstract

We contribute to an emerging literature that shows that machine learning algorithms can discern between stock mutual funds that will outperform and underperform. In addition, we present evidence from the Brazilian equity mutual fund industry that using XGBoost, funds with the highest predicted abnormal returns outperformed funds with the lowest predicted abnormal returns by almost three times, while being less risky. Furthermore, we showed that CVaR is the most important feature for prediction and that return-based metrics greatly outperform characteristic-based ones. Moreover, we demonstrate that weighting methods that consider the predictions made by the model improve the strategy alpha by a significant amount. Finally, we also tested nine different ML algorithms and four classic methods. Our results provide evidence for the superiority of ML models. In specific, Light Gradient Boosting and Extra Trees were the best algorithms.

Keywords: Mutual Fund Performance, XGBoost, Machine Learning

Resumo

Contribuímos para uma literatura emergente que mostra que os algoritmos de aprendizagem de máquinas podem discernir entre fundos de ações que terão um desempenho superior e um desempenho inferior. Além disso, apresentamos provas para a indústria brasileira de fundos de ações que, utilizando XGBoost, os fundos com o maior retorno anormal previsto superam os fundos com o menor rendimento anormal previsto em quase três vezes, ao mesmo tempo que são menos arriscados. Além disso, mostramos que o CVaR é a característica mais importante para a previsão e que as métricas baseadas no retorno superam em muito as baseadas nas características. Além disso, demonstramos que os métodos de ponderação que consideram as previsões feitas pelo modelo melhoram significativamente o alfa da estratégia. Finalmente, testamos também nove algoritmos de ML e quatro métodos clássicos. Os nossos resultados fornecem provas da superioridade dos modelos de ML. Em particular, o Light Gradient Boosting e o Extra Trees foram os melhores algoritmos.

Keywords: Desempenho de Fundos de Ação, XGBoost, Aprendizado de Máquina

List of Figures

List of Tables

List of abbreviations and acronyms

ADA	Ada Boost Model
AI	Artificial Intelligence
B3	Brasil, Bolsa, Balcão
CVaR	Conditional Value-at-Risk
DNN	Deep Neural Network item [DUM] Dummy
EN	Elastic Net Model
ET	Extra Trees Model
FoF	Fund-of-Funds
GB	Gradient Boosting Model
IBrX	The Brazil 100 Index
KNN	K Nearest Neighborhood Model
LAS	LASSO Regression Model
LR	Linear Regression Model
LightGBM	Light Gradient Boosting Model
ML	Machine Learning
NEFIN	Brazilian Center for Research in Financial Economics of the University of São Paulo
RF	Random Forest Model
RID	Ridge Regresion Model
SVN	Suport Vector Machine Model
XGBoost	Extreme Gradient Boosting Model

Contents

1 Introduction

According to a report from the Brazilian Association of Financial and Capital Market Institutions ANBIMA (2022), combined, all the Brazilian investment funds had close to US$ 1.4 trillion in assets under management. From this total, about 6.5% are allocated to the 4000 existing equity mutual funds. Even though it is a small proportion of the whole industry, equity mutual funds attract the interest of investors who want to diversify their portfolios and have financial exposure to the stock market.

When an investor decides to buy a share of an equity mutual fund, he wishes to select the fund (or group of funds) that will deliver the higher return with the lowest possible risk. John Bogle, founder, and CEO of The Vanguard Group, in a 1992 paper Bogle (1992), wrote that, when selecting equity mutual funds, it is virtually impossible to pick the winners in advance. He also wrote that ”if (and I underscore the ”if") there is a systematic way to identify equity fund winners [...] it would surely be in this new era of the microcomputer". Thirty years after this statement, an emerging financial literature uses the recent developments in Machine Learning, Artificial Intelligence, and computational power to predict which equity mutual funds will deliver the best and worst performances in the future.

In a seminal paper, Wu et al. (2021) apply different machine learning algorithms to the problem of selecting future hedge fund winners. Using only features based on the fund’s past return, they show that, in most cases, these models significantly outperform the four-styled Hedge Fund Research Indices. In addition, they present evidence that neural networks are the top-performing algorithms and that kurtosis is the variable that has the greatest predictive power over the fund’s future return.

In addition, DeMiguel et al. (2021) note that machine learning algorithms deliver an edge for predicting a mutual fund’s five-factor alpha Fama e French (2015) because they allow for nonlinearities and interactions between the variables of interest. Moreover, they show that decision-tree methods (gradient boosting and random forests) deliver higher alphas when compared to linear methods (elastic net and OLS). Finally, they suggest that an approach that uses a single or just a few fund characteristics tends to be dominated by approaches that use multiple of them.

In contrast, using a feedforward neural network, Kaniel et al. (2022) show that fund momentum and flow are the only variables needed to differentiate funds with higher future Cahart abnormal returns (CARHART, 1997) from those with lower ones. Consequentially, the authors reveal that the characteristics of the stocks that funds hold, conditioned on fund momentum and fund flow, are not useful metrics to tell good and bad equity mutual fund managers apart. Furthermore, they show that these two metrics have much greater predictive power when investor sentiment is high. As they point out, linear models cannot grasp this kind of relationship.

In consonance with these previous works, Li e Rossi (2020) present evidence that indicates that boosted regression trees significantly outperform traditional linear methods. To support this claim, they construct long-short portfolios that buy (sell) the top 10% funds with the highest (lowest) predicted future performance. This strategy delivers an annual excess return of 6.68% and an even bigger risk-adjusted return of 7.46%, both statistically significant at the 1% level. The authors also find that out of the ten characteristics with the highest predictive power, seven are related to trading frictions and three to momentum.

These works are part of a bigger trend of applying machine learning to uncover different relationship structures between financial variables. Goodell et al. (2021) present an extensive review of the theme. As they point out, there are three main thematic structures of Artificial Intelligence (AI) and Machine Learning (ML) research in finance. Our paper is in the portfolio construction, valuation, and investor behavior category. The other two categories refer first to financial fraud and distress and then to sentiment inference, forecasting, and planning.

In this paper, as in previous works, we will focus our attention on trying to discern, in advance, equity fund managers that will outperform from those that will underperform. For that, we use a conventional stepwise chronological data split. This means that for every month between 2008-01-01 and 2021-12-31, we will train our XGBoost model (CHEN; GUESTRIN, 2016) on the data available until that month and then we will make predictions for the upcoming month. After that, we will rank the funds based on the predictions and create portfolios that go long (short) in the funds with the highest (lowest) predictions.

That explained, we need to justify why we are choosing XGBoost as our principal model over other Machine Learning Algorithms and what is our dependent variable - the metric that will define what is under and outperformace. First, we use XGBoost because it is computationally efficient (CHEN; GUESTRIN, 2016) and has been successfully used in various domains and ML

problems. Fauzan e Murfi (2018), for example, show that XGBoost gives better results than other methods like AdaBoost, Random Forest, and Neural Networks for insurance claim prediction. In addition, Giannakas et al. (2021) show that XGBoost performs better than a Deep Neural Network (DNN) with four hidden layers when predicting teams' performance. Finally, Zhang et al. (2020) shows that XGBoost outperforms Support Vector Machine, Random Forest, and Logistic Regression for transaction fraud detection. Even though we have a strong argument for using the XGBoost algorithm, we will also present the main result for other ML algorithms.

Second, the metric that will be used to define which funds underperformed and which outperformed is the Carhart four-factor abnormal return, as in Kaniel et al. (2022). This metric is the difference between the funds' realized return in month t and its expected return at the same time. The expected return is the inner product of the vector containing the factors' returns at month t and the vector containing the funds' exposure to each factor. The factor exposures are obtained from the regression of the funds returns in excess of the risk-free rate against the Cahart four factors (market, size, value, momentum) from $t-1$ to $t-12$.

Based upon the extensive literature regarding financial metrics that have predictive power over funds' future returns, we present our explanatory variables. Initially, we divide the independent variables into return-based and characteristics-based metrics. We selected eleven metrics for the first group: Carhart (1997) t-stat intercept and betas, CVaR, Modified Information Ratio (ISRAELSEN et al., 2005), Tracking Error, Kurtosis, R^2, Idiosyncratic Volatility. We applied them to three periods based on momentum literature (short-term reversal, short-term momentum, and momentum).

Furthermore, there are eleven variables in the characteristic-based group. These variables are AUM, three flow-related, number of shareholders, age, redemption notice periods, and dummies indicating if the fund is open - accepts inflows -, if it can take on leverage, if it is a Fund-of-Funds (FoF), and if it is an exclusive one - only one shareholder. In total, the are 44 independent variables.

The rest of the article is structured as follows: (i) first, we revise the literature about metrics with explanatory power over funds returns; (ii) next, we present the data and the explanatory variables that will be considered; (iii) after that, we present the basic idea of how XGBoost and the other ML models work; (iv) next, we show the results; (v) finally, we conclude and make remarks about possible future developments.

2 Literature Revision

The investment fund industry plays an important role in the global financial market, as it represents a modality of collective investments that has been showing significant growth throughout the world over the years. Some of the reasons for that include the fact that they provide liquidity, diversification, and professional management at low costs Chua e Tam (2020).

The choices made by the manager in the selection and assembly of the portfolio will be decisive for the subsequent performance of the fund. From an investor's point of view, in general, to evaluate the performance of mutual funds, past performance, the Sharpe measure, the Treynor measure or the Jensen measure are used. However, interest in using machine learning techniques to assess fund performance appears to be a promising option, as past performance, fund characteristics, fund dynamics and fund flow may be the most important predictors of future fund performance (DEMIGUEL et al., 2021; KANIEL et al., 2022).

Although the literature on ML and AI applications in finance is relatively new, there is a well-developed literature that finds metrics that have predictive power over the future returns of mutual funds and hedge funds. Thus, financial econometrics has been widely used in empirical research in financial and economic studies (LEE, 2021).

At the same time, with the development of modeling techniques, new propositions about the coherence of risk measures used by fund managers emerge (ARTZNER et al., 1999), making the discussion on asset allocation in constant evolution. Liang e Park (2007), for example, presents evidence that risk measures such as Expected Shortfall (CVaR) and Tail Risk explain the cross-section of hedge fund realized returns, while semi-deviation and Value at Risk do not, calling into question the traditional assessment measures.

In line with that, there is an extensive literature that establishes a relationship between mutual funds' past performance and future performance. Hendricks, Patel e Zeckhauser (1993), Brown e Goetzmann (1995), Carhart (1997), for example, find that funds with lowest returns tend to underperform in the following months. Also, Carhart (1997) shows that the performance of past mutual fund winners tend to reverse after one year. In addition, Harvey e Liu (2018) states that past fund performance has low predictive power over future returns because it is too noisy. However, they show that, using a random effects framework, they can improve alpha forecasts.

Furthermore, still exploring the relationship between past performance and future returns, Vidal-García et al. (2019), using UK equity mutual fund data, show that there is a negative relationship between idiosyncratic risk and returns. Kacperczyk, Sialm e Zheng (2005) also present evidence that more concentrated funds outperform, which may indicate that tracking error may be a relevant feature for mutual fund return prediction.

In parallel, there is also a relevant literature that uses fund holding characteristics to predict future returns. Cremers e Petajisto (2009), for example, introduce Active Share and show that this new metric predicts fund performance. Moreover, a metric that captures the impact of unobserved trading, return gap, also has predictive power over funds future returns (KACPERCZYK; SIALM; ZHENG, 2008). Finally, other metrics such as active weights (DOSHI; ELKAMHI; SIMUTIN, 2015), and risk shifting (HUANG; SIALM; ZHANG, 2011), have been shown to be able to separate good and bad equity mutual fund managers.

Added to that, Titman e Tiu (2011) show, for the hedge fund industry, that funds with lower R^2 with respect to systematic factors tend to have higher alphas and a superior risk-return relationship. In consonance, for the mutual fund industry, Amihud e Goyenko (2013) present evidence that a portfolio that goes long (short) funds with lower (higher) R^2 produce a statistically significant alpha of 3.8%.

In a moral hazard paradigm, Wu et al. (2021), based on Goetzmann, Jr e Ross (2003), reason that a fund manager may increase the fund's risk level if it falls under the high-water mark. For them, this means that the fund's current drawdown may be a good proxy for the risk of a fund manager's moral hazard. They also show that kurtosis is the most important variable when making predictions about a fund's future returns.

Additionally, Aragon (2007) show that there is a positive, concave relationship between hedge fund returns and share restrictions, such as lockup period restrictions, redemption notice periods, and minimum investment amounts. They hypothesize that this may be because funds with longer lockup periods can handle illiquid assets more efficiently, being able to capture a liquidity premium. These findings are corroborated by Agarwal, Daniel e Naik (2009).

Plus, Chen et al. (2004), Yan (2008) show that a negative relationship between return and fund size (AUM) exists. Both of them connect this phenomenon to liquidity. In contrast, Adams, Hayunga e Mansi (2018) present evidence that outliers are responsible for the existence of this relationship. Furthermore, examining the data manually, they track these extreme observations

to bad data and show that removing them makes the relationship between fund size and return economically and statistically insignificant. In consonance, Pástor, Stambaugh e Taylor (2015) note that, after correcting econometric biases, there is an insignificant relationship between the variables. However, they present evidence of decreasing returns to scale at the industry level.

In addition, GRUBER (1996), Zheng (1999) present evidence of the smart money effect: fund investors can discern between skilled and nonskilled equity mutual fund managers and are able to allocate resources to these good funds that will, subsequently, receive more inflows and tend to outperform in the future. Also, Zheng (1999) points out that the effect is short-lived. In contradiction, Frazzini e Lamont (2008)'s results indicate that mutual fund investors' reallocation tends to reduce their wealth on average. The explanation is that the "dumb money" effect dominates the "smart money" effect, since the last exists only on short horizons, whereas the first is long-lived.

Finally, Gil-Bazo e Ruiz-Verdú (2009) investigate the relationship between fee and performance and reveal a puzzling phenomenon: funds with worse before-fee performance tend to charge higher fees. They hypothesize that this is the result of strategic fee-setting. However, Hu, Chao e Lim (2016) demonstrates that investor sentiment is a better explanation of the effect.

3 Methodological procedures

3.1 Data

Our data regarding equity mutual funds were extracted from Economatica, a Brazilian financial data provider. In addition, we get data for factor portfolios (market, size, value, and momentum) and the Brazilian risk-free rate from NEFIN-USP. Finally, from Bloomberg, we extract data about IBrX, a Brazilian market index that tracks the stock performance of 100 large companies listed on B3, the Brazilian stock exchange. All this data is in daily frequency and starts on 2004-02-01 and ends on 2021-12-31. It is also valid to state that the fund's returns are net of fees.

Even though our data start at the beginning of 2004, we only start making predictions for 2008. We do that to ensure we have enough data to train our model properly. In addition, we need 12 months of data to create the first set of features. In the end, the data from February 2005 to December 2007 is used only for model training. In total, the predictions for January 2008 use more than 2750 observations, ensuring that, from the first prediction, the model has enough data to properly learn.

It is also essential to define the criteria for selecting a fund for our analysis. The first is that it needs to be active for at least 12 months. In addition, during the estimation and evaluation period, it must have data for at least 90% of the trading days. Finally, we eliminate funds with less than 10 million reais (EVANS, 2010), which is close to 2 million dollars in the end of 2022.

Because we have some outliers in the funds' returns, we winsorize our return data in the to 1st percentile, meaning that each day the extreme observations - the ones below the 1st and above the 99th percentile - are replaced by the values corresponding to these percentiles.

3.1.1 Dependent Variable

First, we formally define our dependent variable. As in Kaniel et al. (2022), this will be the fund's abnormal return ($R_{i,t}^{abn}$). We begin by writing,

Table 1 – Data Summary Statistics

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
# Funds	45	189	484	469.3	605.5	1272
Return-based						
Abnormal Return	-0.34	-0.01	0.01	0.01	0.02	0.35
MIR (STM)	0	0	0.01	0.09	0.14	1.13
CVaR (STM)	-0.25	-0.03	-0.02	-0.02	-0.01	0
Track Error (STM)	0	0	0.01	0.01	0.01	0.08
Kurtosis (STM)	1.29	2.29	2.73	3.03	3.36	19.73
Alpha (STM)	-93.53	-0.6	0.06	0.06	0.72	6.58
Beta-Market (STM)	-9.9	4.47	7.64	8.92	11.6	92.21
Beta-Size (STM)	-6.42	0.01	0.89	0.97	1.84	10.54
Beta-Value (STM)	-8.2	-1.5	-0.53	-0.61	0.41	7.28
Beta-Momentum (STM)	-7.46	-0.52	0.38	0.42	1.28	9.28
R^2 (STM)	0	0.73	0.87	0.79	0.94	1
IVol (STM)	0	0	0	0	0	0.04
MIR (Mom.)	0	0	0	0.03	0.05	0.37
CVaR (Mom.)	-0.21	-0.04	-0.03	-0.04	-0.02	0
Track Error (Mom.)	0	0.01	0.01	0.01	0.01	0.04
Kurtosis (Mom.)	2.21	3.51	4.08	6.75	7.2	89.56
Alpha (Mom.)	-108.85	-0.53	0.21	0.26	1.01	7.07
Beta-Market (Mom.)	-7.81	18.2	28.63	32.6	41.04	216.76
Beta-Size (Mom.)	-12.01	1.4	3.25	3.38	5.3	20.27
Beta-Value (Mom.)	-12.84	-3.46	-1.43	-1.67	0.31	13.48
Beta-Momentum (Mom.)	-11.14	-0.37	1.3	1.62	3.5	13.16
R^2 (Mom.)	0	0.71	0.85	0.77	0.92	1
IVol (Mom.)	0	0	0	0.01	0.01	0.02
MIR (STR)	0	0	0.01	0.08	0.14	1.13
CVaR (STR)	-0.25	-0.03	-0.02	-0.02	-0.01	0
Track Error (STR)	0	0	0.01	0.01	0.01	0.08
Kurtosis (STR)	1.29	2.28	2.71	3.02	3.35	19.73
Alpha (STR)	-20.93	-0.59	0.07	0.07	0.73	6.58
Beta-Market (STR)	-9.9	4.5	7.62	8.92	11.56	92.21
Beta-Size (STR)	-6.42	0.02	0.9	0.97	1.86	10.54
Beta-Value (STR)	-8.38	-1.55	-0.55	-0.65	0.39	7.94
Beta-Momentum (STR)	-7.46	-0.53	0.37	0.41	1.27	9.28
R^2 (STR)	0	0.73	0.87	0.8	0.94	1
IVol (STR)	0	0	0	0	0.01	0.04
Fund's Characteristics						
AUM	10000.9	29596.4	73535.7	224579	194415	12198579
Inflows	0	191.3	9740	72139.76	47519.4	5881765
Outflows	0	400	9000	49837.16	38496.7	2881490
% Flow	-2.37	-0.1	0	1.23	0.33	8065.4
# Shareholders	0	2	8	572.84	65	149767
Fees	0	0.14	1	1.06	1.9	6
Leveradge	0	0	1	0.52	1	1
Open	0	1	1	0.98	1	1
FoF	0	0	1	0.52	1	1
Exclusive	0	0	0	0.1	0	1
Age	1	2.44	4.73	5.84	8.08	41.91

$$R_{i,t-12:t-1} = \alpha_i + \beta_i' F_{t-12:t-1} + \varepsilon_{i,t-12:t-1} \tag{3.1}$$

In this case, $F_{t-12:t-1}$ is the matrix containing the daily returns of the Carhart (1997) factors (Market, SMB, HML, WML), and β_i is the vector containing the fund's i factor loadings. $R_{i,t-12:t-1}$ is the fund's after-fee returns.

Finally, the abnormal return of the fund i at time t will be:

$$R_{i,t}^{abn} = R_{i,t} - \beta_i F_t \tag{3.2}$$

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

3.1.2 Independent Variables

We can divide our explanatory variables into two main groups: the ones based on the returns and the others based on fund characteristics. Summary statistics for all these variables are presented in Table 1.

3.1.2.1 Return Based

First, following a similar procedure used by Kaniel et al. (2022), we consider three time frames based on the momentum literature. However, unlike Kaniel et al. (2022), that only used this time frame for the variables related to momentum, every return-based metric will have one version for each time frame. These periods are: (i) short-term momentum ($t-2$); (ii) short-term reversal ($t-1$); momentum ($t-12$ until $t-3$). The first two periods are based on Jegadeesh e Titman (1993) and the third on Fama e French (1996).

Now that we have established the time frames, we present the return-based variables. First, there are those related to the regression of the fund's return against the Cahart four-factor model (CARHART, 1997); these are the alpha (intercept), betas related to the market, size, value, and momentum factors, and the regression's R^2. In addition, we have the Conditional VaR (ROCKAFELLAR; URYASEV et al., 2000; BALI; GOKCAN; LIANG, 2007), tracking error, modified information ratio (ISRAELSEN et al., 2005), kurtosis, and Idiosyncratic Volatility.

Mamaysky, Spiegel e Zhang (2007) present evidence that, when sorting funds based on the estimated alpha, the funds in the top (bottom) decile will not be the future winners (losers). In fact, the funds in these deciles are those with the greatest estimation error. For that reason, as in DeMiguel et al. (2021), we use the raw alpha scaled by the standard error (t-stat) to account for this estimation error.

Table 1 presents the summary statistics of the variables. First, it is interesting to see that, unlike returns, the abnormal return has a mean different from 0. In fact, both the mean and the median round to 1% per month. Another point that deserves observation is the fact that most funds have positive exposure to size and momentum and negative exposure to value.

3.1.2.2 Funds' Characteristics

We consider ten different variables related to the funds themselves. These are: (i) last available information about assets under management (AUM); (ii) inflows in the last twelve months (Inflows); (iii) outflows in the last twelve months (Outflows); (iv) ratio between net funding (inflow - outflow) and AUM at the beginning of the period (% Flows); (v) number of shareholders (# Shareholders); (vi) dummy variable indicating if the fund is allowed to take on leverage positions (leveraged); (vii) dummy variable indicating if the shareholders are allowed to redeem the invested capital (Open); (viii) dummy indicating if the fund is exclusive - can have only one investor (Exclusive); (ix) and the age of the fund (Age); and (x) lockup period.

Analyzing the distributions of these variables from Table 1, we can see that the median fund has close to fifteen million dollars in AUM and has experienced close to zero net flows in the sample. Furthermore, it has just eight shareholders, whereas the mean number of shareholders in the sample is close to 570, indicating that few funds hold the majority of shareholders. This fact is also consistent with the incubation bias (EVANS, 2010). In addition, half of the funds can have leveraged positions, and a similar amount are Funds of Funds. Moreover, the vast majority are open, and close to 10% are exclusive. Finally, the average fund is five years and nine months old.

3.2 Machine Learning Models

Machine Learning models demand a considerable amount of data to be effective (YAO, 2021). Because we consider an extended time frame in our analysis and the Brazilian capital

Table 2 – Machine Learning Models Reference

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

Source: the authors.

market is still in development, one might raise concerns about the validity of our approach. As we can see in Table 1, there are, on average, close to 500 funds that meet our criteria. In fact, the month with the least amount of data has 45 funds, but we only include these observations in the training data. With this concern dismissed, we can present the ML models that will be considered.

For this paper, we will consider a total of fourteen machine learning algorithms that will be grouped into two categories: linear and ensemble models. Algorithms that do not fit in either will be grouped in a separate category. Linear models are linear combinations of the independent variables, and ensemble models, in turn, combine multiple other models in the prediction process. With exception of LightGBM and XGBoost, the implementation of the algorithms come from the Python package scikit-learn (PEDREGOSA et al., 2011). The first two algorithms come from the homonymous Python packages.

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 must follow this procedure because some models rely on the calculation of distances, which is sensitive to the feature's scale.

3.2.1 XGBoost

Taking into account that XGBoost is our main algorithm, we present a high-level explanation of the model's inner workings. For a complete explanation, we direct the reader to the original article (CHEN; GUESTRIN, 2016).

The first difference between XGBoost and other ensemble algorithms is the way it builds trees. Unlike Random Forest, for example, which splits nodes using Gini or Entropy, XGBoost splits based on Similarity Score (SS):

$$SS_j = \frac{(\sum_{i=1}^{n} y_i - \hat{y}_i)^2}{n + \lambda} \tag{3.3}$$

where y_i is the observed value, $\hat{y}_i$ is the predict value, n is the number of residuals in leaf j, and λ is a regularization parameter. It is important to note that the numerator is the squared residual sum, not the typical residual sum of squares. If the similarity score is small, the residuals are dissimilar - they cancel each other out; in contrast, if the value is big, the similarity between residuals is greater.

One last step before starting the algorithm is stating that XGBoost is a recursive model that will keep building trees until it reaches a user-defined maximum number of trees or the gain for adding a new tree is sufficiently small.

The first step in the process is to compute the average of the dependent value and consider it as the first prediction ($f(x)_0$). After that, the first tree leaf is populated with the residuals of this prediction. The next step is splitting the root, which results in a new branch (new left and right leaf). For that, we use the Similarity Score and the Information Gain, which is defined as:

$$Gain = [SS_{left} + SS_{right} - SS_{root}] - \gamma \tag{3.4}$$

Calculating the Information Gain for different possible thresholds and features, we split our data based on the feature and threshold combination that results in the greatest Gain. In reality, we split only if the Gain is positive, meaning that γ works as a pruning parameter: depending on the value defined, the tree may grow more or less. As with the λ parameter, γ reduces the chance of overfitting the training dataset. The tree will grow until the Gain is negative. With the complete tree we calculate the output value for each leaf as:

$$Output\ Value_j = \frac{\sum_{i \in I_j} y_i - \hat{y}_i}{n + \lambda} \tag{3.5}$$

As we can see, if $\lambda = 0$, a jth leaf output value will be, simply, the average of the residuals in that leaf. Having the complete tree as well as the output values, we make new predictions based on the formula below:

$$f(x)_t = f(x)_{t-1} + \eta\, Output\ Value(x) \tag{3.6}$$

where η is the learning rate, which controls how quickly the algorithm adapts to the training set.

3.2.2 Other ML algorithms

The first linear model that we will consider is linear regression. This model will minimize the sum of squared errors. Mathematically, the objective function is:

$$\hat{\beta}^{OLS} = \underset{\beta \in \mathbb{R}^k}{\operatorname{argmin}}(||y - X\beta||_2^2) \tag{3.7}$$

where $||.||_2$ denotes the ℓ_2 norm.

The other linear models considered are regularized regression methods. In Ridge regression (HOERL; KENNARD, 1970), for example, we abandon the requirement of an unbiased estimator and minimize the residual sum of squares plus a penalty term on the betas (L2 regularization). The idea of applying a penalty function goes back to the bias-variance trade-off in machine learning: with an increase in the bias from the penalty function, the model tends to generalize better on the test data. Mathematically,

$$\hat{\beta}^{RID} = \underset{\beta \in \mathbb{R}^k}{\operatorname{argmin}}(||y - X\beta||_2^2 + \lambda ||\beta||_2^2) \tag{3.8}$$

LASSO (TIBSHIRANI, 1996) is very similar to Ridge regression: both have a regularization term in the objective function and are robust to multicollinearity. However, while Ridge considers the square of the coefficients, LASSO considers their absolute value (L1 regularization). In addition, unlike Ridge, which can only shrink a coefficient toward zero, LASSO can shrink the coefficient all the way to 0, leading to a sparser solution. Again, mathematically,

$$\hat{\beta}^{LAS} = \underset{\beta \in \mathbb{R}^k}{\operatorname{argmin}}(||y - X\beta||_2^2 + \lambda||\beta||_1) \tag{3.9}$$

where $||.||_1$ denotes the ℓ_1 norm.

Finally, Elastic Net (ZOU; HASTIE, 2005) overcomes the LASSO's limitations related to situations with many features and few observations. To do that, Elastic Net adds a quadratic part to the LASSO penalty. It is also interesting to notice that the Elastic Net can be interpreted as a generalization of the previously discussed linear algorithms. From Equation 3.10, we can see that if $\lambda_1 = \lambda_2 = 0$, we have the classical linear regression objective function; if $\lambda_1 = 0$, we have ridge regression; finally, if $\lambda_2 = 0$, we have LASSO.

$$\hat{\beta}^{EN} = \underset{\beta}{\operatorname{argmin}}(||y - X\beta||^2 + \lambda_1||\beta||_1 + \lambda_2||\beta||_2^2) \tag{3.10}$$

Exiting the world of linear models, we present, first, the Support Vector Machine (CORTES; VAPNIK, 1995), which searches for a hyperplane in N-dimensional space (N = number of features) with the maximum number of points. Differently from the linear models that minimize a cost function that includes the sum of squared errors, SVM will minimize the L2-norm of the coefficient vector, subjected to all residuals having a value less than ε (arbitrarily defined). This model is not used very often for regression problems, but has some advantages, like the fact that it is robust to outliers, can be easily updated and usually generalizes well. Opposed to that, this algorithm may not be appropriate when dealing with large datasets, tend to not perform well with overlapping target classes, and when there are few degrees of freedom.

In turn, the k-nearest neighbors algorithm (FIX; HODGES, 1989; ALTMAN, 1992) receives an arbitrarily defined k parameter and the training data and returns, for each prediction, the average of the k closest observations. The definition of "closest" comes from the Euclidean distance. As consequence of the model's simplicity, we can easily generalize the algorithm considering n dimensions (number of features) and m observations. Supposing we want to make a prediction for a new observation x_{m+1}, we first calculate the distance from this point to every other point available in the training set ($i = 1, 2, ..., m$):

$$d(x_{m+1}, x_i) = \sqrt{\sum_{j=1}^{n} (x_{m+1,j} - x_{i,j})^2} \tag{3.11}$$

After that, we sort the resulting list in ascending order and average the first k observations. This algorithm has advantages, like not requiring a training period and being easy to implement and update. Despite that, it is sensitive to outliers - uses the square of the difference - and doesn't scale well - because the euclidean algorithm requires squaring every difference, as the dataset gets bigger, the performance deteriorates rapidly. By default, $k = 5$ in the sckit-learn implementation.

Next, as the name indicates, the decision tree algorithm uses a tree-like decision model. The process involves doing recursive binary splitting, in which every feature is considered, and the split (decision) is done by minimizing a cost function - usually Gini or Entropy.

After that, we analyze the dummy model. The idea is not to use this model for making predictions but to have a baseline to compare the other models. What it does is really simple: its predictions are equal to the average value of the dependent variable in the training data. Because the prediction is the same for every observation, the model can not discern between good and bad mutual fund managers.

It is not easy to determine the best choice among Machine Learning tools. An ensemble is a solution to this situation, as it is a combination of several algorithms with the objective of trying to extract the best from each technique (ZHOU, 2012). All the ensemble methods presented here will blend multiple models (usually weak learners) to improve out-of-sample results.

Another point in ensemble models is the difference between bagging (BREIMAN, 1996) and boosting (FREUND; SCHAPIRE et al., 1996) algorithms. In the last, trees are created sequentially in such a way that the next tree learns from the mistakes made by the previous one and updates the residual error based on that information. Usually, these trees consist of weak learners: models that tend to perform just slightly better than random guessing. In contrast, bagging algorithms work through a voting scheme: multiple full-sized trees are grown and the model's final prediction is the average of the predictions made by all trees. Random Forests is an example of a bagging algorithm, while Gradient Boost and XGBoost are examples of boosting algorithms. Finally, it's also valid to note that boosting algorithms tend to outperform (QUINLAN et al., 1996).

Random Forest, for example, will combine the output of various decision trees to make a single prediction. The original theoretical properties of random forests are demonstrated

in Breiman (2001) for classification trees. Similar to RF, Extra Trees (GEURTS; ERNST; WEHENKEL, 2006) will combine different decision trees, but this model has an additional bias-variance analysis.

In random forests the classification is taken by a vote, hence each tree votes for a particular class and the class with the most votes wins, according to function

$$mg = M^{-1}\Sigma_{m=1}^{M} 1_{\{h_m(x)=y\}} - \max_{j \neq y}(M^{-1}\Sigma_{m=1}^{M} 1_{\{h_m(x)=j\}}) \tag{3.12}$$

where the left part is the average number of votes based on the M trees h_m for the correct class. The right part is the maximum average for any other class. The mg is the margin and reflects the confidence that the aggregate forest will classify properly.

In the random forest, diversification among many trees was expected to improve the overall quality of the model. In boosting, we seek to iteratively improve the model whenever a new tree is added. This work used four boosting models: Gradient Boosting, Light Gradient Boosting, Ada Boost and XGBoost.

AdaBoost (FREUND; SCHAPIRE, 1997) improves the learning process by progressively focusing on the instances that yield the largest errors. For that, the algorithm starts by assigning equal weights for each observation. After that, using the Gini Index, a stump - a decision tree with just one node - is built. Next, the initial weights are updated in a way that the next tree is penalized if it commits the same mistakes as the previous ones. Finally, we normalize the new sample weights to ensure that they sum up to one. This process is repeated until a low training error is achieved.

Finally, in Gradient Boosting (FRIEDMAN, 2001), decision trees are generally also used. Today, this algorithm is considered a generalization of AdaBoost. Light Gradient Boosting (KE et al., 2017) is a more computationally efficient implementation of Gradient Boosting.

To ensure that the comparison between the predictions of the above models is fair, we must guarantee that all models are trained in the same data. Allowing a model to train on more observations or features makes the comparison useless, as we can not decompose the differences in performance between differences in the data and differences in the models' structure. Linear regression and some other algorithms don't handle missing data natively. In view of this fact, we drop the observations that contain missing data. In total, we drop 343 observations out of 95268.

4 Results

4.1 Pooled Regression

First, we present the results using traditional statistical tools. The pooled regression (Table 3) shows that out of 45 estimated parameters, 28 are statistically significant at the 5% level. Even though many of our features were not significant, they still might be important for Machine Learning Algorithms, as they explore nonlinearities and interactions between the variables. We omitted variables that were not statistically significant at the 10% level for space-related reasons.

In addition, we can see that the return-based metrics seem to carry more information about future abnormal returns when compared to characteristics-based metrics: in our model, close to 80% of the return-based features were significant, while only close to 50% of the characteristics-based ones were. In this case, only assets under management (AUM), outflows, fees, and the dummies indicating if the fund is open and if it can take on leverage were significant. However, with exception of age, the characteristics-based metrics that weren't significant didn't have support in the literature regarding their predictive power over mutual funds' returns.

Furthermore, by analyzing the coefficients, it is possible to see that there seems to be a positive relationship between risk and abnormal return for shorter terms (CVar (STM), Beta-Size (STM), Beta-Value (STM)). This fact is consistent with the fundamentals of modern finance (MARKOWITZ, 1952; SHARPE, 1964). However, when we analyze more extended periods (Mom.), the relation is inverted (CVar (Mom.), Beta-Market (Mom.), Beta-Size (Mom.), Beta-Value (Mom.), Beta-Momentum (Mom.)). This fact is consistent with a more recent literature that highlights the out-performance of less risky assets compared to more risky ones (BLITZ; VLIET, 2007; HOUWELING; ZUNDERT, 2017).

In addition, one might expect older funds to have more significant abnormal returns than newer ones after controlling for AUM, due to decreasing returns to scale (HARVEY; LIU, 2021). This is a reasonable expectation since one can imagine that an older fund should have a more structured investment process and a more experienced management team. However, our regression shows a negative relationship between abnormal return and age. Even though it seems counter-intuitive, this phenomenon is well documented in the literature (see Stafylas, Anderson

e Uddin (2016) for revision) and might be linked to career concerns, in which older mutual fund managers tend to be less risk-averse than younger ones (CHEVALIER; ELLISON, 1999). This, in turn, might be detrimental to the fund's performance.

Finally, it is interesting to notice that out of three metrics related to fund flow, only outflow was significant. The fact that inflow was not statistically significant goes against an extensive literature, revised above, that relates fund inflow to future performance (GRUBER, 1996; ZHENG, 1999; KESWANI; STOLIN, 2008).

4.2 XGBoost Deciles

In this section, we explore how effectively the XGBoost model separated the equity mutual funds with good from those with bad relative future performance. For that, for every month from February 2008 to December 2021, we rank the funds based on the predictions made by the XGBoost model. After that, we divide the funds into deciles and simulate an equal-weighted portfolio that goes long in every fund in each decile.

Table 4 Panel A allows us to see how effective the XGBoost model was in the task specified above. First of all, the first decile has an almost three times higher return than the last one. More impressively, the first decile also carries 12% less risk. For comparison reasons, the Brazilian market index (IBrX), in the same period, had an annualized return of 5.47% and an annualized volatility of 27.29%. This leads to the first decile's modified Sharpe Ratio (ISRAELSEN et al., 2005) being close to twenty times bigger than that of the market.

Vardharaj, Fabozzi e Jones (2004) points out that when an active manager takes positions that deviate a lot from the benchmark, he or she will have significant active returns, either positive or negative. From the results in Table 4, we can see precisely this parabolic relationship: the extreme deciles have higher tracking errors while also having significant returns. In contrast, the deciles in the middle have lower tracking errors and lower returns in absolute terms.

Moreover, another point of interest is the alpha of each decile. As expected, the highest (numerically) four-factor alpha is in the first decile, while the lowest is in the last decile. However, none of the portfolios had an intercept statistically different from 0, considering a 5% significance level, and only the tenth decile had a significant and negative alpha at 10%. This suggests that none of the portfolios generated or destroyed value. This fact may be (partially) explained by the

Table 3 – Pooled Regression

	Dependent variable:
	Abnormal Return
MIR (STM)	0.001^{*} (0.001)
CVaR (STM)	-0.022^{**} (0.010)
Kurtosis (STM)	-0.0002^{***} (0.0001)
Beta-Size (STM)	0.0002^{*} (0.0001)
Beta-Value (STM)	0.001^{***} (0.0001)
Beta-Momentum (STM)	0.0004^{***} (0.0001)
R^2 (STM)	-0.004^{***} (0.001)
IVol (STM)	-0.373^{***} (0.096)
MIR (Mom.)	0.019^{***} (0.003)
CVaR (Mom.)	0.050^{***} (0.006)
Track Error (Mom.)	0.298^{***} (0.052)
Kurtosis (Mom.)	0.0002^{***} (0.00002)
Alpha (Mom.)	0.0002^{***} (0.00004)
Beta-Market (Mom.)	-0.0001^{***} (0.00001)
Beta-Size (Mom.)	-0.001^{***} (0.00005)
Beta-Value (Mom.)	-0.001^{***} (0.00005)
Beta-Momentum (Mom.)	-0.001^{***} (0.00004)
R^2 (Mom.)	0.007^{***} (0.001)
IVol (Mom.)	-0.529^{***} (0.091)
MIR (STR)	0.002^{**} (0.001)
CVaR (STR)	0.017^{*} (0.010)
Track Error (STR)	0.138^{***} (0.042)
Kurtosis (STR)	-0.001^{***} (0.0001)
Beta-Size (STR)	-0.0003^{***} (0.0001)
Beta-Value (STR)	0.0003^{***} (0.0001)
Beta-Momentum (STR)	0.0002^{**} (0.0001)
AUM	-0.000^{***} (0.000)
Outflows	0.000^{*} (0.000)
Fees	-0.001^{***} (0.0002)
Leveraged	-0.000^{**} (0.000)
Open	0.003^{***} (0.001)
Intercept	0.010^{***} (0.001)
Observations	76,549
R^2	0.030
Adjusted R^2	0.029
Residual Std. Error	0.026 (df = 76504)
F Statistic	52.855^{***} (df = 44; 76504)
Note:	*p<0.1; **p<0.05; ***p<0.01

Source: the authors.

fact that we work with after-fee returns (FAMA; FRENCH, 2010).

After analyzing the deciles' returns statistics, we now analyze the deciles' average characteristics. Table 4 Panel B shows that the funds that the model predicts higher abnormal returns tend to be, on average, bigger (AUM), younger, have fewer shareholders, and a slightly greater management fee and lock-up period. In contradiction with the literature, the decile containing the fund with highest predictions tend to be bigger (CHEN et al., 2004; YAN, 2008) and have a bigger management fee (GIL-BAZO; RUIZ-VERDÚ, 2009). However, the fact that they tend to be younger (WEBSTER, 2002) and have a longer lock-up period (ARAGON, 2007) are in agreement with the existing literature.

Also, the fact that mutual funds in the first decile tend to charge more is a great indication that the strategy is actually able to select funds with top gross and net future returns. If the average management fee difference between the first and last decile was significant, the model could be discerning between high and low-cost funds, which is not what we aim to do.

Finally, one last fact deserves attention. As we show, the funds for which the model predicts higher abnormal returns tend to be, on average, bigger (AUM) and have fewer shareholders. This seems to point towards a small group of more capitalized investors having a greater ability to discern funds with future good and bad abnormal returns. In contrast, a group with a higher number of members but less capitalized tends to be on the opposite side: they select the funds with lower abnormal future returns. Future works could investigate if there is a correlation between these groups and institutional and retail investors.

Table 4 – Deciles Return Statistics and 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	11	10.3	10.19	7.78	9.21	8.81	7.41	8.3	7.41	4.08
Std. Deviation	17.11	18.75	19.68	20.41	20.34	20.65	21.3	20.9	20.91	19.63
Alpha	2.37	2.03	2.12	-0.17	1.31	1.06	-0.2	0.67	0.08	-3.02
t(alpha)	1.38	1.35	1.55	-0.13	0.99	0.83	-0.15	0.48	0.05	-1.79
Beta	0.65	0.73	0.77	0.8	0.8	0.81	0.83	0.81	0.81	0.75
Info. Ratio	0.21	0.22	0.25	0.06	0.19	0.17	0.04	0.12	0.03	0
Sharpe Ratio	0.19	0.15	0.15	0.04	0.11	0.09	0.03	0.07	0.03	-0.01
Track Error	13.67	11.51	10.31	9.59	9.67	9.28	8.97	9.25	9.43	11.52
CVaR	-2.97	-3.78	-1.7	-2.82	-3.43	-3.91	-2.48	-2.46	-2.57	-3.65
Max. Drawdown	51.37	48.76	53.57	56.66	52.59	54.5	58.82	52.88	53.59	51.27
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	194072	189723	178254	166280	160407	167531	154104	158101	160302	173478
Inflows	60497.2	57953.1	57189.8	52612.4	51404.6	52075.4	47720.4	49477.7	50551.3	52207.5
Outflows	48727.6	43570.9	41835	39846.2	39194.8	40187	37710.4	40288.8	38910.8	38580.2
Mangt. Fee (%)	1.13	1.1	1.03	1.05	1.06	1.07	1.06	1.07	1.07	1.08
# Shareholders	339.64	298.76	540.01	469.99	425.29	308.04	925.15	774.45	1425.45	2270.65
Leveraged	0.51	0.51	0.5	0.51	0.49	0.47	0.47	0.46	0.46	0.44
FoF	0.48	0.49	0.5	0.49	0.48	0.48	0.48	0.46	0.45	0.42
Exclusive	0.07	0.09	0.1	0.1	0.11	0.12	0.12	0.12	0.12	0.11
Age	4.65	4.75	4.79	4.81	4.93	4.91	5	4.98	5.14	5.11
Lock-up Period	38.51	37.03	36.48	33.25	32.91	31.51	30.56	31.42	32.05	37.46

Source: the authors.

4.3 Weighting Schemes

As in Kaniel et al. (2022), we also want to study how the weighting method impacts the final portfolio. Based on this analysis, we can uncover how effective our predictions are in separating great (disastrous) mutual funds in a sample already populated with good (bad) ones. In subsection 4.2, we showed that the predictions can efficiently discern between the top and bottom performers. Now, we investigate if the ranks based on the predictions [1] and the predictions themselves [2] can add value to the already defined portfolio.

For that, we build three Long & Short portfolios. All of them go 100% long in the 30% funds with the highest predicted abnormal return, 100% short in the 30% funds with the lowest predicted abnormal return, and 100% long in the risk-free rate. The only difference between them is related to how the weights of each fund inside the long and the short portfolios are determined. The first portfolio uses the equal-weighted ($\frac{1}{n}$) method. Even tough it's a really simple ("naive") way of determining the portfolio weights, it has shown out-of-sample superiority over other methods (DEMIGUEL; GARLAPPI; UPPAL, 2009; PLYAKHA; UPPAL; VILKOV, 2012). The second portfolio is based on the ranking, and the last based on the raw predictions.

As we can see from Table 5, there is a positive monotonic relationship between the alpha and the degree of information used from the predictions: the equal-weighted portfolio uses the least amount of information and has the lowest alpha; the ranking-based portfolio uses a bit more of the information contained inside the predictions and has the second highest alpha; finally, the portfolio based on the raw predictions, which uses the complete information, has an alpha more than 40% greater than the equal-weighted portfolio. These results show that the abnormal return predictions carry essential information about future returns, way beyond the one required for

[1] Unlike Kaniel et al. (2022), we also present the results based on ranking weights:

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

where $i = 1, ..., n$ represents the index of each fund in the prediction ranking, and $w_{i,t}$ is the final weight of each fund in the portfolio.

[2] As in Kaniel et al. (2022), we calculate the portfolio weights based on the predictions as:

$$\text{Top portfolio: } \mu_{i,t} = \hat{\mu}_{i,t} - min(\hat{\mu}_{i,t})$$

$$\text{Bottom portfolio: } \mu_{i,t} = \hat{\mu}_{i,t} - max(\hat{\mu}_{i,t})$$

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

where $\hat{\mu}_{i,t}$ are the XGBoost predictions, and $w_{i,t}$ are the final weights

decile construction.

In addition, we must point out the fact that all the Long & Short portfolios generated a positive, statistically, and economically significative alpha at 5%. Furthermore, the portfolios also presented an expressive annualized return with low volatility, resulting in portfolios with a very high Shape Ratio (0.7). On the risk side, we can also observe the CVaR and Maximum Drawdown were very low. Even though it's not possible to short a mutual fund, these results strongly suggest that investors should avoid the funds with the lowest predictions and, in turn, give preference to the high-ranking ones.

Table 5 – Long & Short Portfolios Weighting Schemes

	Equal Weighted	Rank	Prediction
Annual. Return	12.18	12.86	13.43
Std. Deviation	4.22	4.99	5.92
Alpha	2.41	2.95	3.43
t(alpha)	2.23	2.32	2.24
Beta	0.05	0.06	0.06
Info. Ratio	0.1	0.12	0.14
Sharpe Ratio	0.69	0.71	0.7
CVaR	-0.84	-0.97	-1.1
Max. Drawdown	9.69	11.61	12.97

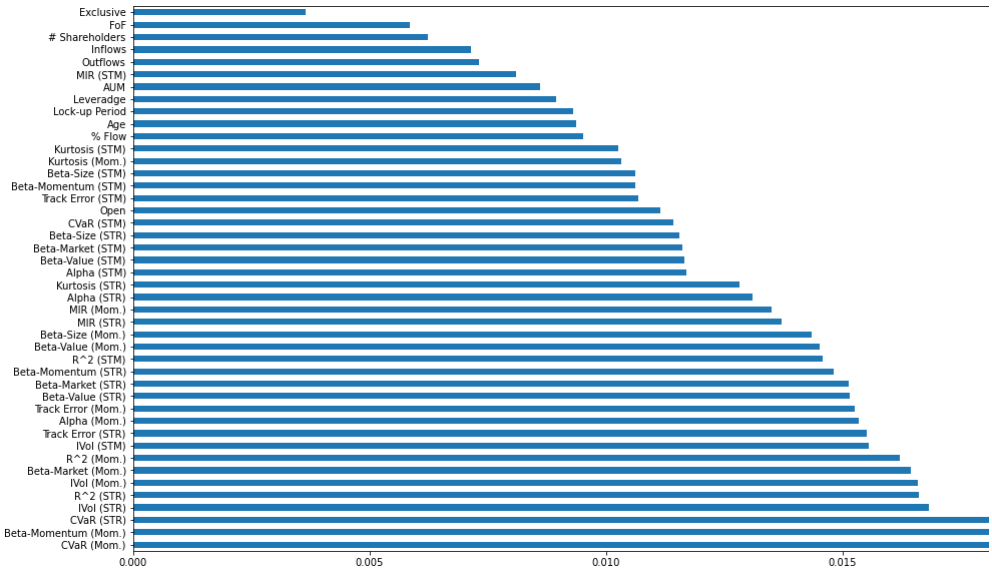
4.4 Feature Importance

From Figure 1, we can see that CVaR (Mom.) was the single most important variable for prediction [3]. In addition, CVaR (STR) was the third most important feature, indicating that using multiple time frames can add a lot of value to the model. Another interesting point is the dominance of the Momentum and Short Term Reversals periods over the Short Term Momentum: out of the fifteen most important features, only one referred to the STM time frame. Furthermore, we can see a clear prevalence of metrics related to risk at the top of the most important features. Again, out of the fifteen most relevant variables, only three are not directly risk related ($R^2(STR)$, $R^2(Mom.)$, and Alpha (Mom.)).

Moreover, there was also clear domination of the return-based metrics over the characteristic-based ones, which is intriguing since the literature indicated some of these metrics as carrying considerable prediction power over the funds' future returns. This is the case for flows, AUM,

[3] To analyze XGBoost feature importance, we use information gain - average gain (Equation 3.4) of splits which use the feature.

Figure 1 – XGBoost Feature Importance



lockup period, and age, for example. On top of that, (KANIEL et al., 2022) noted that fund momentum and fund flow were the most important predictors, contrasting greatly with our results.

4.5 Comparison of machine learning algorithms

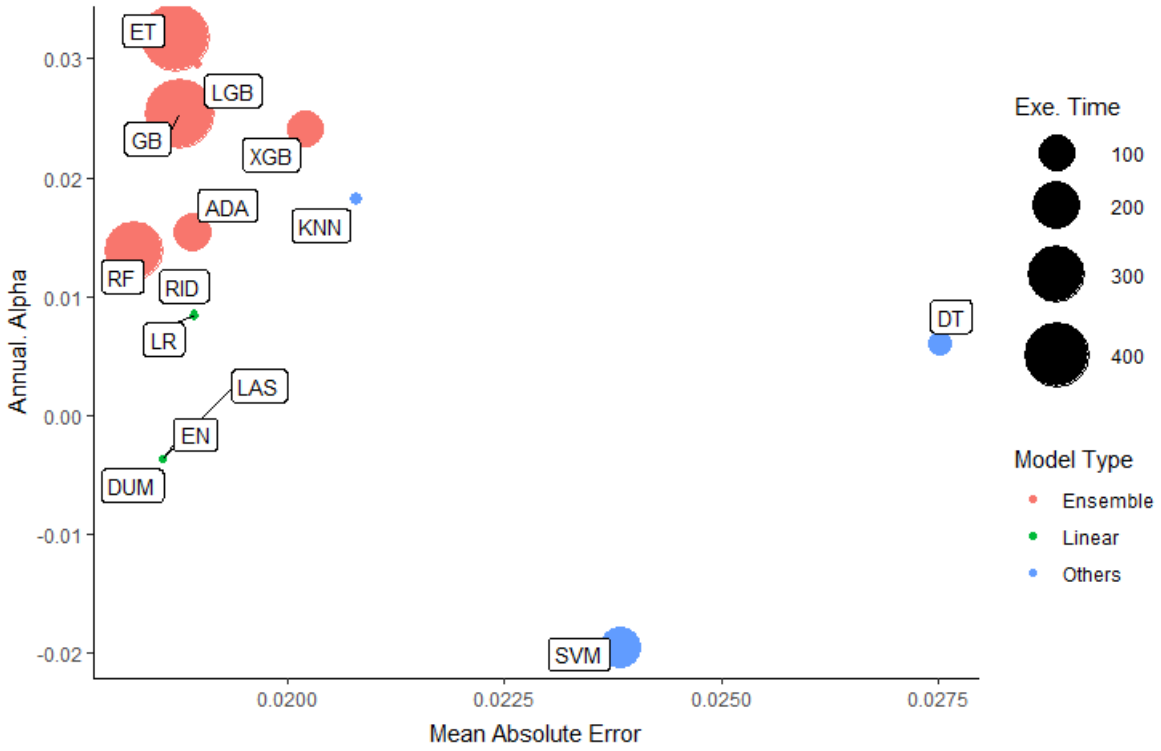
Before running any tests, we needed to choose a Machine Learning algorithm to o decrease the likelihood of our results being biased (multiple testing, Prado (2015)). We chose the XGBoost because of its high performance in various Machine Learning problems and because it is computationally efficient. In this section, we evaluate if we chose the best model and compare its performance against other ML models.

To accomplish that, we will rely upon Figure 1. The x-axis in this figure presents the Mean Absolute Error (MAE) for the predictions made by each ML model. We use MAE instead of other metrics like Mean Squared Error (MSE) because it is less sensitive to outliers.

The y-axis, in turn, presents information about the four-factor alpha for the Long & Short portfolio based on the predictions of each ML algorithm. To construct this portfolio, every month, we sort the funds based on the predictions made by each model. Then, we create an equal-weighted L&S portfolio that goes long the 30% funds with the best predictions and short the 30% funds with the worse.

In addition, we scale the points based on the time (in seconds) it takes for the model to

Figure 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 November 2021 and predict December 2021. Refer to Table 2 for the acronymous meanings.

train on the data from February 2005 to November 2021 and predict December 2021. We do that to understand the trade-off between performance and cost. Finally, the points' colors indicate the model type (refer to Table 2).

A comparison between model types shows that the ensemble methods did remarkably well. This group generated high alphas with varying levels of MAE. The ensemble outperformed the linear models, presenting additional evidence that nonlinear relationships and interactions between the variables exist. In addition, the performance of Support Vector Machines, which was much worse than the baseline model (Dummy), and Decision Trees indicate that SVM may not be suitable for the task and that DT works much better in an ensemble model.

Finally, it is safe to say that the model that offered the best performance-cost relationship is the LGB. This model generated the second biggest alpha while being more than 200 times faster to train than the best performing algorithm (Extra Tress). The outperformance of this model is in line with the literature (LI; ROSSI, 2020; KE et al., 2017). XGBoost, our initial choice, did not perform as well but could still differentiate good and bad equity mutual funds with high precision (see Table 4).

5 Conclusion

We contribute to the literature by presenting additional evidence of the ability of machine learning models to discern between equity mutual funds that will outperform and underperform. Furthermore, we tested many ML algorithms and showed that Light Gradient Boosting (LGB) is the model with the highest capacity to select future winners and identify future losers, when we balance predictive power and computational resources required. If we consider only the resulting portfolio alpha, Extra Trees is the most suitable algorithm.

Even though our previously selected model (XGBoost) did not perform as well, the predictions made by this model allowed us to sort the funds in deciles in such a way that the first decile (higher predicted abnormal return) outperformed the last decile (lower predicted abnormal return) by almost three times while being less risky.

In addition, we could also provide additional evidence of the greater predictive power of Machine Learning algorithms compared to the traditional statistical methods (linear models). The best ML (ET) model generated an alpha 3.67 times greater when compared to the best linear model (Ridge Regression).

Furthermore, we showed that a portfolio may benefit if the predictions are also used for defining the weight of each fund in the portfolio, instead of just ranking for decile separation. The strategy that follows this procedure delivered an alpha more than 40% greater than the naive equal-weighted portfolio. In addition, all Long & Short portfolios generated a positive, statistically, and economically significative alpha.

Moreover, we presented evidence that the risk-related metrics were the most important for predictions. CVaR was the first and third most important variable. Contrary to previous research, we also found that metrics based on characteristics of the fund, like Assets Under Management, Flows, Age, and Lockup Period weren't very relevant.

This work may benefit society through many different channels. A lot of Brazilians have financial exposure to the financial market, via personal savings, Funds of Funds, and retirement plans, for example. In light of this fact, it is clear that having a systematic way to identify future winners and avoid future losers has the potential to improve the quality and robustness of the investment process of a large group of people and institutions. Another advantage is that this

method allows the market to be more efficient, providing an easy and fast way to reward skilled managers and penalize unskilled ones. In the coming years, this kind of analysis will become commonplace, and society will be able to harvest the benefits of a more efficient and developed market.

Finally, we present some possible future developments for interested researchers. First, we could do some hyper-parameter tuning in a validation set before making the predictions. Second, we could use more robust methods for outlier detection and treatment. Lastly, we could check how the alpha decays as we make the holding period longer.

Bibliography

ADAMS, J. C.; HAYUNGA, D. K.; MANSI, S. Diseconomies of scale in the actively-managed mutual fund industry: What do the outliers in the data tell us? *Available at SSRN 3194005*, 2018. Citado na página 11.

AGARWAL, V.; DANIEL, N. D.; NAIK, N. Y. Role of managerial incentives and discretion in hedge fund performance. *The Journal of Finance*, Wiley Online Library, v. 64, n. 5, p. 2221–2256, 2009. Citado na página 11.

ALTMAN, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, Taylor & Francis, v. 46, n. 3, p. 175–185, 1992. Citado 2 vezes nas páginas 17 and 20.

AMIHUD, Y.; GOYENKO, R. Mutual fund's r 2 as predictor of performance. *The Review of Financial Studies*, Society for Financial Studies, v. 26, n. 3, p. 667–694, 2013. Citado na página 11.

ANBIMA. *Boletim de Fundos de Investimento*. 2022. ANBIMA URL. [Online; accessed 04-August-2022]. Citado na página 7.

ARAGON, G. O. Share restrictions and asset pricing: Evidence from the hedge fund industry. *Journal of financial economics*, Elsevier, v. 83, n. 1, p. 33–58, 2007. Citado 2 vezes nas páginas 11 and 26.

ARTZNER, P. et al. Coherent measures of risk. *Mathematical finance*, Wiley Online Library, v. 9, n. 3, p. 203–228, 1999. Citado na página 10.

BALI, T. G.; GOKCAN, S.; LIANG, B. Value at risk and the cross-section of hedge fund returns. *Journal of Banking & Finance*, v. 31, n. 4, p. 1135–1166, 2007. Citado na página 15.

BLITZ, D. C.; VLIET, P. V. The volatility effect. *The Journal of Portfolio Management*, Institutional Investor Journals Umbrella, v. 34, n. 1, p. 102–113, 2007. Citado na página 23.

BOGLE, J. C. Selecting equity mutual funds. *Journal of Portfolio Management*, Pageant Media, v. 18, n. 2, p. 94, 1992. Citado na página 7.

BREIMAN, L. Bagging predictors. *Machine learning*, Springer, v. 24, n. 2, p. 123–140, 1996. Citado na página 21.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado 2 vezes nas páginas 17 and 22.

BROWN, S. J.; GOETZMANN, W. N. Performance persistence. *The Journal of finance*, Wiley Online Library, v. 50, n. 2, p. 679–698, 1995. Citado na página 10.

CARHART, M. M. On persistence in mutual fund performance. *The Journal of finance*, Wiley Online Library, v. 52, n. 1, p. 57–82, 1997. Citado 5 vezes nas páginas 8, 9, 10, 15, and 31.

CHEN, J. et al. Does fund size erode mutual fund performance? the role of liquidity and organization. *American Economic Review*, v. 94, n. 5, p. 1276–1302, 2004. Citado 2 vezes nas páginas 11 and 26.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. [S.l.: s.n.], 2016. p. 785–794. Citado 3 vezes nas páginas 8, 17, and 18.

CHEVALIER, J.; ELLISON, G. Career concerns of mutual fund managers. *The Quarterly Journal of Economics*, MIT Press, v. 114, n. 2, p. 389–432, 1999. Citado na página 24.

CHUA, A. K. P.; TAM, O. K. The shrouded business of style drift in active mutual funds. *Journal of Corporate Finance*, Elsevier, v. 64, p. 101667, 2020. Citado na página 10.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, Springer, v. 20, n. 3, p. 273–297, 1995. Citado 2 vezes nas páginas 17 and 20.

CREMERS, K. M.; PETAJISTO, A. How active is your fund manager? a new measure that predicts performance. *The review of financial studies*, Oxford University Press, v. 22, n. 9, p. 3329–3365, 2009. Citado na página 11.

DEMIGUEL, V.; GARLAPPI, L.; UPPAL, R. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The review of Financial studies*, Oxford University Press, v. 22, n. 5, p. 1915–1953, 2009. Citado na página 28.

DEMIGUEL, V. et al. Machine learning and fund characteristics help to select mutual funds with positive alpha. In: *Proceedings of Paris December 2021 Finance Meeting EUROFIDAI-ESSEC*. [S.l.: s.n.], 2021. Citado 3 vezes nas páginas 7, 10, and 16.

DOSHI, H.; ELKAMHI, R.; SIMUTIN, M. Managerial activeness and mutual fund performance. *The Review of Asset Pricing Studies*, Oxford University Press, v. 5, n. 2, p. 156–184, 2015. Citado na página 11.

EVANS, R. B. Mutual fund incubation. *The Journal of Finance*, Wiley Online Library, v. 65, n. 4, p. 1581–1611, 2010. Citado 2 vezes nas páginas 13 and 16.

FAMA, E. F.; FRENCH, K. R. Multifactor explanations of asset pricing anomalies. *The journal of finance*, Wiley Online Library, v. 51, n. 1, p. 55–84, 1996. Citado na página 15.

FAMA, E. F.; FRENCH, K. R. Luck versus skill in the cross-section of mutual fund returns. *The Journal of Finance*, v. 65, n. 5, p. 1915–1947, 2010. Citado na página 26.

FAMA, E. F.; FRENCH, K. R. A five-factor asset pricing model. *Journal of financial economics*, Elsevier, v. 116, n. 1, p. 1–22, 2015. Citado na página 7.

FAUZAN, M. A.; MURFI, H. The accuracy of xgboost for insurance claim prediction. *Int. J. Adv. Soft Comput. Appl*, v. 10, n. 2, p. 159–171, 2018. Citado na página 9.

FIX, E.; HODGES, J. L. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, JSTOR, v. 57, n. 3, p. 238–247, 1989. Citado 2 vezes nas páginas 17 and 20.

FRAZZINI, A.; LAMONT, O. A. Dumb money: Mutual fund flows and the cross-section of stock returns. *Journal of financial economics*, Elsevier, v. 88, n. 2, p. 299–322, 2008. Citado na página 12.

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, Elsevier, v. 55, n. 1, p. 119–139, 1997. Citado 2 vezes nas páginas 17 and 22.

FREUND, Y.; SCHAPIRE, R. E. et al. Experiments with a new boosting algorithm. In: CITESEER. *icml*. [S.l.], 1996. v. 96, p. 148–156. Citado na página 21.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, JSTOR, p. 1189–1232, 2001. Citado 2 vezes nas páginas 17 and 22.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. *Machine learning*, Springer, v. 63, n. 1, p. 3–42, 2006. Citado 2 vezes nas páginas 17 and 22.

GIANNAKAS, F. et al. Xgboost and deep neural network comparison: The case of teams performance. In: SPRINGER. *International Conference on Intelligent Tutoring Systems*. [S.l.], 2021. p. 343–349. Citado na página 9.

GIL-BAZO, J.; RUIZ-VERDÚ, P. The relation between price and performance in the mutual fund industry. *The Journal of Finance*, Wiley Online Library, v. 64, n. 5, p. 2153–2183, 2009. Citado 2 vezes nas páginas 12 and 26.

GOETZMANN, W. N.; JR, J. E. I.; ROSS, S. A. High-water marks and hedge fund management contracts. *The Journal of Finance*, Wiley Online Library, v. 58, n. 4, p. 1685–1718, 2003. Citado na página 11.

GOODELL, J. W. et al. Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *Journal of Behavioral and Experimental Finance*, Elsevier, v. 32, p. 100577, 2021. Citado na página 8.

GRUBER, M. J. Another puzzle: The growth in actively managed mutual funds. *The Journal of Finance*, v. 51, n. 3, p. 783–810, 1996. Citado 2 vezes nas páginas 12 and 24.

HARVEY, C. R.; LIU, Y. Detecting repeatable performance. *The Review of Financial Studies*, JSTOR, v. 31, n. 7, p. 2499–2552, 2018. Citado na página 10.

HARVEY, C. R.; LIU, Y. Decreasing returns to scale, fund flows, and performance. *Fund Flows, and Performance (June 21, 2021)*, 2021. Citado na página 23.

HENDRICKS, D.; PATEL, J.; ZECKHAUSER, R. Hot hands in mutual funds: Short-run persistence of relative performance, 1974–1988. *The Journal of finance*, Wiley Online Library, v. 48, n. 1, p. 93–130, 1993. Citado na página 10.

HOERL, A. E.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, Taylor & Francis, v. 12, n. 1, p. 55–67, 1970. Citado 2 vezes nas páginas 17 and 19.

HOUWELING, P.; ZUNDERT, J. van. Factor investing in the corporate bond market. *Financial Analysts Journal*, Routledge, v. 73, n. 2, p. 100–115, 2017. Citado na página 23.

HU, M.; CHAO, C.-C.; LIM, J. H. Another explanation of the mutual fund fee puzzle. *International Review of Economics & Finance*, Elsevier, v. 42, p. 134–152, 2016. Citado na página 12.

HUANG, J.; SIALM, C.; ZHANG, H. Risk shifting and mutual fund performance. *The Review of Financial Studies*, Oxford University Press, v. 24, n. 8, p. 2575–2616, 2011. Citado na página 11.

ISRAELSEN, C. L. et al. A refinement to the sharpe ratio and information ratio. *Journal of Asset Management*, v. 5, n. 6, p. 423–427, 2005. Citado 3 vezes nas páginas 9, 15, and 24.

JEGADEESH, N.; TITMAN, S. Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of finance*, Wiley Online Library, v. 48, n. 1, p. 65–91, 1993. Citado na página 15.

KACPERCZYK, M.; SIALM, C.; ZHENG, L. On the industry concentration of actively managed equity mutual funds. *The Journal of Finance*, Wiley Online Library, v. 60, n. 4, p. 1983–2011, 2005. Citado na página 11.

KACPERCZYK, M.; SIALM, C.; ZHENG, L. Unobserved actions of mutual funds. *The Review of Financial Studies*, Society for Financial Studies, v. 21, n. 6, p. 2379–2416, 2008. Citado na página 11.

KANIEL, R. et al. *Machine-learning the skill of mutual fund managers*. [S.l.], 2022. Citado 7 vezes nas páginas 8, 9, 10, 13, 15, 28, and 30.

KE, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, v. 30, 2017. Citado 3 vezes nas páginas 17, 22, and 31.

KESWANI, A.; STOLIN, D. Which money is smart? mutual fund buys and sells of individual and institutional investors. *The Journal of Finance*, v. 63, n. 1, p. 85–118, 2008. Citado na página 24.

LEE, C. F. Introduction to financial econometrics, mathematics, statistics, and machine learning. In: *HANDBOOK OF FINANCIAL ECONOMETRICS, MATHEMATICS, STATISTICS, AND MACHINE LEARNING*. [S.l.]: World Scientific, 2021. p. 1–99. Citado na página 10.

LI, B.; ROSSI, A. G. Selecting mutual funds from the stocks they hold: A machine learning approach. *Available at SSRN 3737667*, 2020. Citado 2 vezes nas páginas 8 and 31.

LIANG, B.; PARK, H. Risk measures for hedge funds: a cross-sectional approach. *European financial management*, Wiley Online Library, v. 13, n. 2, p. 333–370, 2007. Citado na página 10.

MAMAYSKY, H.; SPIEGEL, M.; ZHANG, H. Improved forecasting of mutual fund alphas and betas. *Review of Finance*, Oxford University Press, v. 11, n. 3, p. 359–400, 2007. Citado na página 16.

MARKOWITZ, H. Portfolio selection. *The Journal of Finance*, v. 7, n. 1, p. 77–91, 1952. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1952.tb01525.x>. Citado na página 23.

PÁSTOR, L.; STAMBAUGH, R. F.; TAYLOR, L. A. Scale and skill in active management. *Journal of Financial Economics*, Elsevier, v. 116, n. 1, p. 23–45, 2015. Citado na página 12.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado na página 17.

PLYAKHA, Y.; UPPAL, R.; VILKOV, G. Why does an equal-weighted portfolio outperform value-and price-weighted portfolios? *Available at SSRN 2724535*, 2012. Citado na página 28.

PRADO, M. L. D. The future of empirical finance. *The Journal of Portfolio Management*, Institutional Investor Journals Umbrella, v. 41, n. 4, p. 140–144, 2015. Citado na página 30.

QUINLAN, J. R. et al. Bagging, boosting, and c4. 5. In: *Aaai/Iaai, vol. 1*. [S.l.: s.n.], 1996. p. 725–730. Citado na página 21.

ROCKAFELLAR, R. T.; URYASEV, S. et al. Optimization of conditional value-at-risk. *Journal of risk*, Citeseer, v. 2, p. 21–42, 2000. Citado na página 15.

SHARPE, W. F. Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, Wiley Online Library, v. 19, n. 3, p. 425–442, 1964. Citado na página 23.

STAFYLAS, D.; ANDERSON, K.; UDDIN, M. Recent advances in hedge funds' performance attribution: Performance persistence and fundamental factors. *International Review of Financial Analysis*, Elsevier, v. 43, p. 48–61, 2016. Citado na página 24.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 58, n. 1, p. 267–288, 1996. Citado 2 vezes nas páginas 17 and 19.

TITMAN, S.; TIU, C. Do the best hedge funds hedge? *The Review of Financial Studies*, Society for Financial Studies, v. 24, n. 1, p. 123–168, 2011. Citado na página 11.

VARDHARAJ, R.; FABOZZI, F. J.; JONES, F. J. Determinants of tracking error for equity portfolios. *The Journal of Investing*, Institutional Investor Journals Umbrella, v. 13, n. 2, p. 37–47, 2004. Citado na página 24.

VIDAL-GARCÍA, J. et al. Idiosyncratic risk and mutual fund performance. *Annals of Operations Research*, Springer, v. 281, n. 1, p. 349–372, 2019. Citado na página 11.

WEBSTER, D. Mutual fund performance and fund age. *Available at SSRN 1764543*, 2002. Citado na página 26.

WU, W. et al. A cross-sectional machine learning approach for hedge fund return prediction and selection. *Management Science*, INFORMS, v. 67, n. 7, p. 4577–4601, 2021. Citado 2 vezes nas páginas 7 and 11.

YAN, X. S. Liquidity, investment style, and the relation between fund size and fund performance. *Journal of Financial and Quantitative Analysis*, Cambridge University Press, v. 43, n. 3, p. 741–767, 2008. Citado 2 vezes nas páginas 11 and 26.

YAO, F. *Machine learning with limited data*. arXiv, 2021. Disponível em: <https://arxiv.org/abs/2101.11461>. Citado na página 16.

ZHANG, Y. et al. Customer transaction fraud detection using xgboost model. In: IEEE. *2020 International Conference on Computer Engineering and Application (ICCEA)*. [S.l.], 2020. p. 554–558. Citado na página 9.

ZHENG, L. Is money smart? a study of mutual fund investors' fund selection ability. *The Journal of Finance*, v. 54, n. 3, p. 901–933, 1999. Citado 2 vezes nas páginas 12 and 24.

ZHOU, Z.-H. *Ensemble methods: foundations and algorithms*. [S.l.]: CRC press, 2012. Citado na página 21.

ZOU, H.; HASTIE, T. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, Wiley Online Library, v. 67, n. 2, p. 301–320, 2005. Citado 2 vezes nas páginas 17 and 20.