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

Abstract

We contribute to an emerging literature that shows that Machine Learning algorithms can discern between equity mutual funds that will outperform and underperform. In addition, we present evidence from the Brazilian equity mutual fund industry and show that using XGBoost, funds with higher predicted abnormal returns outperformed the funds with lower predicted abnormal returns by almost four times while being 15% less risky. Finally, we also test nine different ML algorithms and four classical methods and present evidence of ML models' superiority.

Keywords— Mutual Fund Performance, XGBoost, Machine Learning

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.

DeMiguel et al. (2021) note that machine learning algorithms deliver an edge for predicting the five-factor alpha (Fama and French (2015)) because they allow for nonlinearities and interactions between the variables of interest. In addition, 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 fund momentum and flow 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 and 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 and Guestrin (2016)) on the data available until that month and then we will make predictions for the upcoming month. Then, we will rank the funds based on the predictions and create portfolios that go long (short) in the funds with the highest (lowest) predictions. For robustness, we will also present the portfolio results for different holding periods.

That explained, we need to justify why we are choosing XGBoost 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 and Guestrin (2016)) and has been successfully used in various domains and ML problems. Fauzan and Murfi (2018), for example, show that XGBoost gives better results than other methods like AdaBoost, Random Forest, Stochastic GB, 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 Cahart 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 $t-12$.

With our dependent variable already defined, we present our explanatory variables. Initially, we divide the independent variables into return-based and characteristics-based metrics. We selected eight metrics for the first group (Alpha, CVaR, Modified Information Ratio, Tracking Error, and the four Carhart (1997) betas). We applied them to three periods based on momentum literature (short-term reversal, short-term momentum, and momentum). Furthermore, there are ten variables in the characteristic-based group. These variables are AUM, flow-related, number of shareholders, age, and dummies indicating if the fund is open, if it can take on leverage, if it is a Fund-of-Funds (FoF), and if it is an exclusive one. In total, the are 34 independent variables.

The rest of the article is structured as follows: (i) first, we present the data and features used; (ii) then, we present the basic idea of how XGBoost and the other ML models considered work; (iii) next, we show the results; (iv) finally, we conclude and make remarks about possible improvements.

2 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 almost 3900 observations.

It is also essential to define the criteria for selecting a fund for our analysis. The first is that it needs to have existed 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.

Because we have some outliers in the funds' returns, we apply a simple rule: if the return (in a single day) is smaller than -80% or bigger than 80%, we transform this return into a missing value. Out of more than three million observations, we transform 79 in missing values. In future works, more robust methods for outlier detection could be used.

Table 1: Data Summary Statistics

	1st Qu.	Median	Mean	3rd Qu.
# Funds	315	650	590.68	779
Return-based				
Abnormal Return	-0.01	0.01	0.01	0.02
MIR (STM)	0	0	0.08	0.14
CVaR (STM)	-0.03	-0.02	-0.02	-0.01
Track Error (STM)	0	0.01	0.01	0.01
Alpha (STM)	0	0	0	0
Beta-Market (STM)	0.55	0.77	0.73	0.95
Beta-Size (STM)	-0.01	0.12	0.15	0.28
Beta-Value (STM)	-0.23	-0.06	-0.07	0.09
Beta-Momentum (STM)	-0.1	0.04	0.03	0.18
MIR (Mom.)	0	0	0.03	0.05
CVaR (Mom.)	-0.04	-0.03	-0.04	-0.02
Track Error (Mom.)	0.01	0.01	0.01	0.01
Alpha (Mom.)	0	0	0	0
Beta-Market (Mom.)	0.6	0.78	0.75	0.94
Beta-Size (Mom.)	0.05	0.13	0.15	0.23
Beta-Value (Mom.)	-0.14	-0.05	-0.06	0.03
Beta-Momentum (Mom.)	-0.03	0.04	0.04	0.12
MIR (STR)	0	0	0.08	0.14
CVaR (STR)	-0.03	-0.02	-0.02	-0.01
Track Error (STR)	0	0.01	0.01	0.01
Alpha (STR)	0	0	0	0
Beta-Market (STR)	0.56	0.77	0.74	0.95
Beta-Size (STR)	-0.01	0.12	0.15	0.28
Beta-Value (STR)	-0.23	-0.06	-0.07	0.09
Beta-Momentum (STR)	-0.1	0.04	0.03	0.18
Fund's Characteristics				
AUM	13047.28	44552.84	168107.83	141380.7
Inflows	10	4062.01	52572.17	30761.47
Outflows	155.94	5000	38986.6	28268.1
% Flow	-0.14	0	24570.03	0.25
# Shareholders	2	8	1226.06	63
Leveradge	0	1	0.51	1
Open	1	1	0.98	1
FoF	0	0	0.49	1
Exclusive	0	0	0.09	0
Age	2.26	4.25	5.72	7.6

STM, Mom, and STR refer to the time frame division explained in section 2.2.1. The acronymous means short-term momentum, momentum, and short-term reversal, respectively. In addition, MIR and MSR refer, respectively, to the modified Information and Sharpe Ratio proposed by Israelsen et al. (2005)

2.1 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,

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

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:

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

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.

2.2 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.

2.2.1 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 just 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 and Titman (1993) and the third on Fama and 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) and the betas related to the market, size, value, and momentum factors. In addition, we have the Conditional VaR (Rockafellar et al. (2000); Bali et al. (2007)), tracking error, and the modified information ratio (Israelsen et al. (2005)).

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 thing that deserves observation is that most funds have positive exposure to size and momentum and negative exposure to value.

2.2.2 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).

Analyzing the distributions of these variables from Table 1, we can see that the median fund has close to ten 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 1200, indicating that few funds hold the majority of shareholders. This fact is also consistent with the incubation bias. In addition, half of the funds can have leveraged positions, and a similar amount is Funds of Funds. Moreover, the vast majority are open, and close to 10

3 XGBoost and other ML models

Table 2: Machine Learning Models Reference

Acronymous	Algorithm	Type	Reference
XGB	XGBoost	Ensemble	Chen and Guestrin (2016)
SVM	Suport Vector Machine	Other	Cortes and Vapnik (1995)
RID	Ridge Regresion	Linear	Hoerl and 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	-
GB	Gradient Boosting	Ensemble	Friedman (2001)
ET	Extra Trees	Ensemble	Geurts et al. (2006)
EN	Elastic Net	Linear	Zou and Hastie (2005)
DUM	Dummy	Other	-
DT	Decision Tree	Other	-
ADA	Ada Boost	Ensemble	Freund and Schapire (1997)

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 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, more than 500 funds that meet our criteria. In fact, the month with the least amount of data has 78 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 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.

3.1 Linear Models

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:

$$\hat{\beta}^{OLS} = \underset{\beta \in \mathbb{R}^k}{\operatorname{argmin}}(||y - X\beta||_2^2) \tag{1}$$

where $||.||_2$ denotes the ℓ_2 norm.

The other linear models considered are regularized regression methods. In Ridge regression (Hoerl and 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. Mathematically,

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

LASSO (Tibshirani (1996)) is very similar to Ridge regression. However, while Ridge considers the square of the coefficients, LASSO considers their absolute value. In addition, unlike Ridge, which can only shrink a coefficient toward zero, LASSO can shrink the coefficient to 0, leading to a sparse solution. Again, mathematically,

$$\hat{\beta}^{LAS} = \underset{\beta \in \mathbb{R}^k}{\operatorname{argmin}}(||y - X\beta||_2^2 + \lambda||\beta||_1) \tag{3}$$

where $||.||_1$ denotes the ℓ_1 norm.

Finally, Elastic Net (Zou and 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 4, 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.

$$\hat{\beta}^{EN} = \underset{\beta}{\operatorname{argmin}}(||y - X\beta||^2 + \lambda_1||\beta||_1 + \lambda_2||\beta||_2^2) \tag{4}$$

3.2 Others

Now we present a basic idea of the models that are not in either of the categories defined previously. Readers interested in a deeper understanding of the models should reference the cited papers in Table 2.

First, the Support Vector Machine (Cortes and Vapnik (1995)) searches for a hyperplane in N-dimensional space (N = number of features) with the maximum number of points.

In turn, the k-nearest neighbors algorithm receives an arbitrarily defined k parameter and the training data and returns, for each prediction, the average of the k closest observations.

Next, as the name indicates, the decision tree model 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).

Finally, 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 real simple: its predictions are equal to the average value of the dependent variable in the training data.

3.3 Ensemble Methods

All the ensemble methods presented here will blend multiple models (usually weak learners) to improve out-of-sample results.

Random Forest (Breiman (2001)), for example, will combine the output of various decision trees to make a single prediction. Similar to RF, Extra Trees (Geurts et al. (2006)) will combine different decision trees, but this model has an additional bias-variance analysis. In addition, AdaBoost (Freund and Schapire (1997)) will follow the same procedure as a Random Forest does, but instead of using decision trees, it will use a decision stump (a decision tree with a single stump).

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.

4 Results

4.1 Panel Regression

First, we present the results using traditional statistical tools. The pooled regression (Table 3) shows that out of 35 estimated parameters, 23 are statistically significant at the 5% level. In addition, only five return-based metrics were not significant, with Tracking Error's coefficient being significant only for the short-term momentum time frame. In contrast, most characteristics-based metrics were not statistically significant, with outflows, the dummy indicating if the fund is open, and the age of the fund being an exception.

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-Market (STM), and Beta-Market (STR)). 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.) and Beta-Market (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 and Van Vliet (2007); Houweling and van 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 and 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. This phenomenon might be linked to career concerns, in which older mutual fund managers tend to be less risk-averse than younger ones (Chevalier and Ellison (1999)). This, in turn, might be detrimental to the fund's performance (Blitz and Van Vliet (2007); Houweling and van Zundert (2017)).

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 that relates fund inflow to future performance (GRUBER (1996); Zheng (1999); KESWANI and STOLIN (2008)).

4.2 XGBoost Deciles

In this subsection, 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

Table 3: Pooled Regression

	Dependent variable:
	Abnormal Return
MIR (STM)	-0.004^{***} (0.001)
CVaR (STM)	-0.027^{***} (0.007)
Track Error (STM)	-0.125^{***} (0.032)
Alpha (STM)	0.760^{***} (0.061)
Beta-Market (STM)	0.002^{***} (0.0005)
Beta-Size (STM)	0.0001 (0.0003)
Beta-Value (STM)	0.004^{***} (0.0003)
Beta-Momentum (STM)	0.003^{***} (0.0003)
MIR (Mom.)	0.004 (0.003)
CVaR (Mom.)	0.060^{***} (0.004)
Track Error (Mom.)	-0.027 (0.032)
Alpha (Mom.)	3.034^{***} (0.225)
Beta-Market (Mom.)	-0.006^{***} (0.001)
Beta-Size (Mom.)	0.004^{***} (0.001)
Beta-Value (Mom.)	-0.0001 (0.001)
Beta-Momentum (Mom.)	-0.009^{***} (0.001)
MIR (STR)	0.00002 (0.001)
CVaR (STR)	0.030^{***} (0.007)
Track Error (STR)	0.112^{***} (0.031)
Alpha (STR)	0.427^{***} (0.060)
Beta-Market (STR)	0.002^{***} (0.0005)
Beta-Size (STR)	-0.001^{***} (0.0003)
Beta-Value (STR)	0.002^{***} (0.0003)
Beta-Momentum (STR)	-0.001^{**} (0.0003)
AUM	-0.000 (0.000)
Inflows	-0.000 (0.000)
Outflows	0.000^{**} (0.000)
% Flow	-0.000 (0.000)
# Shareholders	-0.000 (0.000)
Leveraged	-0.0002 (0.0002)
Open	0.002^{***} (0.001)
FoF	0.0003 (0.0002)
Exclusive	-0.00003 (0.0004)
Age	-0.0001^{***} (0.00002)
Constant	0.009^{***} (0.001)
Observations	118,801
R^2	0.015
Adjusted R^2	0.015
Residual Std. Error	0.039 (df = 118766)
F Statistic	52.442^{***} (df = 34; 118766)
Note:	*p<0.1; **p<0.05; ***p<0.01

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, we can see a perfectly monotonically relationship between the deciles order and the information about risk and return, as measured by the annualized volatility and return: the first decile is less risky and has a greater return than the second, while the second decile is less risky and has a greater return than the third, and so on.

Furthermore, the magnitudes are also impressive. The first decile has an almost four times bigger return than the last one. More impressive, the first decile also carries 15% 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 twelve times bigger than that of the market.

Vardharaj et al. (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 biggest (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 alpha at 10%. This suggests that none of the portfolios generated or destroyed value. This fact may be (partially) explained by the fact that we work with after-fee returns (FAMA and 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 and have fewer shareholders.

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 the point towards a small group of more capitalized investors having more 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	9.6	9.49	9.58	8.89	7.93	8.54	6.45	7.46	6.13	2.5
Std. Deviation	18.72	19.27	19.46	20.26	20.42	20.66	21.03	21.42	21.48	22.01
Alpha	1.55	1.24	1.41	1.16	0.34	1.1	-0.77	0.28	-0.79	-3.52
t(alpha)	0.85	0.86	1.08	0.87	0.27	0.88	-0.6	0.2	-0.52	-1.73
Beta	0.7	0.75	0.76	0.79	0.8	0.81	0.83	0.84	0.83	0.83
Info. Ratio	0.15	0.18	0.2	0.16	0.07	0.14	0	0.05	0	0
Sharpe Ratio	0.12	0.12	0.12	0.09	0.05	0.08	0	0.04	0	-0.01
Track Error	12.63	10.83	10.18	9.49	9.21	9.03	8.96	8.93	9.31	11.11
CVaR	-4.27	-3.4	-3.43	-2.84	-2.6	-3.23	-3.91	-3.43	-3.1	-3.48
Max. Drawdown	52.09	56.53	53.18	55.89	55.63	53.11	53.36	52.87	55.06	54.88
Panel B: Deciles Average Characteristics										
	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
AUM	162791.14	159179.52	152518.56	141293.97	131863	132182.82	124254.92	129847.56	127856.47	138696.45
Inflows	49766.04	51447.38	49559.57	43612.47	40334.89	41408.04	40767.24	41702.36	42353.18	44868.36
Outflows	38771.33	38874.03	38959.37	33694.86	31776.34	32016.37	30486.3	30347.46	30417.46	32085.41
# Shareholders	525.52	418.19	246.6	257.46	349.97	438.51	408.76	877.1	1079.52	3230.27
Leveraged	0.52	0.52	0.53	0.52	0.51	0.5	0.49	0.48	0.47	0.48
FoF	0.44	0.5	0.51	0.51	0.49	0.48	0.47	0.45	0.43	0.38
Exclusive	0.07	0.07	0.08	0.1	0.1	0.1	0.1	0.11	0.1	0.09
Age	4.59	4.58	4.66	4.75	4.74	4.89	4.83	4.95	5.17	5.42

4.3 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, De 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 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).

Before diving into the comparison, we point out the inverse relationship between the four-factor alpha and MAE (Adjusted R^2 of 33%). On average, the models with better (worse) predictions were also the models that generated higher (lower) alphas. This may seem obvious, but it shows that a fund's abnormal return carries information about its four-factor alpha. If it did not, the MAE could be equal to zero (perfect prediction), and the model's ability to discern between good and bad managers would be low.

Even though the association between four-factor alpha and the MAE is clear, it is essential to state that a low MAE does not mean that a model's predictions will produce a high alpha. This becomes clear when we analyze the performance of the dummy model, which predicts that every fund will have an abnormal return equal to the mean abnormal return in the training data. In this case, because the model predictions cannot distinguish funds, the ranking is random, and the portfolio generates a negative alpha.

A comparison between model types shows that the ensemble methods did remarkably well. Except for Ada Boost, this group generated high alphas with a low MAE. The ensemble outperformed the linear models, presenting additional evidence that nonlinear relationships and interactions between the variables exist. However, in defense of the linear models, we must state that they delivered a good alpha based on the computational power required to run them.

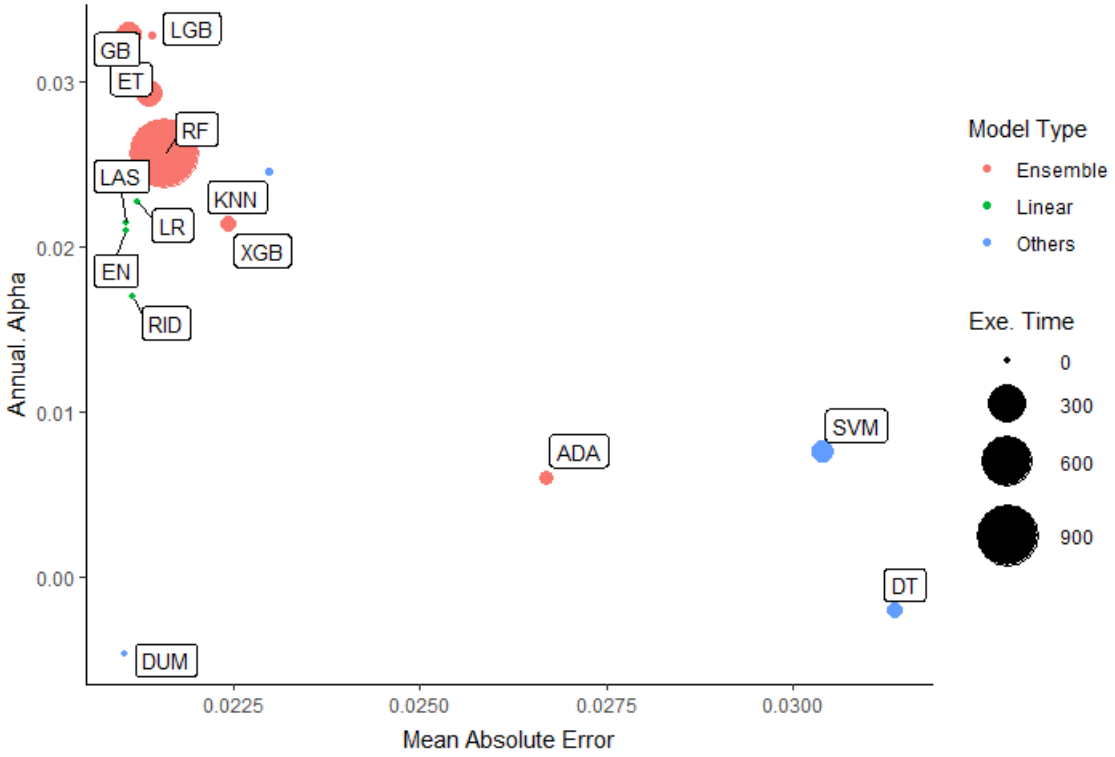
Finally, it is safe to say that the model that offered the best performance-cost relationship is the LGB. This model generated the biggest alpha while being more than 150 times faster to train than the second best performing algorithm (Gradient Boosting). XGBoost, our initial choice, did not perform 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) was the model with the highest capacity to select future winners and identify future losers.

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 four times while

Figure 1: ML Model Comparison



"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.

being 15% 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 (LGB) model generated close to 45% more alpha when compared to the best linear model (Linear Regression) while being computationally super efficient.

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. Third, we could check how the alpha decays as we make the holding period longer. Fourth, we could compare the equal-weighted portfolio to one that gives more weight, within deciles, to the funds with a higher expected abnormal return (Kaniel et al. (2022)). Finally, we could investigate which features are more critical for making good predictions.

References

ANBIMA. Boletim de fundos de investimento. ANBIMA URL, 2022. [Online; accessed 04-August-2022].

T. G. Bali, S. Gokcan, and B. Liang. Value at risk and the cross-section of hedge fund returns. *Journal of Banking & Finance*, 31(4):1135–1166, 2007.

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

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

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

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

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

J. Chevalier and G. Ellison. Career concerns of mutual fund managers. *The Quarterly Journal of Economics*, 114(2):389–432, 1999.

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

M. L. De Prado. The future of empirical finance. *The Journal of Portfolio Management*, 41(4):140–144, 2015.

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

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

E. F. FAMA and K. R. FRENCH. Luck versus skill in the cross-section of mutual fund returns. *The Journal of Finance*, 65(5):1915–1947, 2010. doi: https://doi.org/10.1111/j.1540-6261.2010.01598.x.

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

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

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

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

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

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

J. W. Goodell, S. Kumar, W. M. Lim, and D. Pattnaik. Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *Journal of Behavioral and Experimental Finance*, 32:100577, 2021.

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

C. R. Harvey and Y. Liu. Decreasing returns to scale, fund flows, and performance. *Fund Flows, and Performance (June 21, 2021)*, 2021.

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

P. Houweling and J. van Zundert. Factor investing in the corporate bond market. *Financial Analysts Journal*, 73(2):100–115, 2017. doi: 10.2469/faj.v73.n2.1.

C. L. Israelsen et al. A refinement to the sharpe ratio and information ratio. *Journal of Asset Management*, 5(6):423–427, 2005.

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

R. Kaniel, Z. Lin, M. Pelger, and S. Van Nieuwerburgh. Machine-learning the skill of mutual fund managers. Technical report, National Bureau of Economic Research, 2022.

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

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

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

H. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952. doi: https://doi.org/10.1111/j.1540-6261.1952.tb01525.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1952.tb01525.x`.

R. T. Rockafellar, S. Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

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

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

R. Vardharaj, F. J. Fabozzi, and F. J. Jones. Determinants of tracking error for equity portfolios. *The Journal of Investing*, 13(2):37–47, 2004. doi: 10.3905/joi.2004.412305.

F. Yao. Machine learning with limited data, 2021. URL `https://arxiv.org/abs/2101.11461`.

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

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

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