Finance Research Letters 58 (2023) 104419

Contents lists available at ScienceDirect

Finance Research Letters

journal homepage: www.elsevier.com/locate/frl

Fund performance evaluation with explainable artificial intelligence☆

Veera Raghava Reddy Kovvuri [a], Hsuan Fu [b,*], Xiuyi Fan [c], Monika Seisenberger [a]

[a] Swansea University, Swansea, Wales, United Kingdom
[b] Laval University, Quebec City, Quebec, Canada
[c] Nanyang Technological University, Singapore

ARTICLE INFO

JEL classification:
C52
C55
G11
G15

Keywords:
Global Open-Ended Funds
Country portfolios
Herfindahl–Hirschman Index
SHapley Additive exPlanations
Machine learning
eXtreme Gradient Boosting

ABSTRACT

We apply explainable artificial intelligence (xAI) to a large dataset of global equity funds. Our approach combines the XGBoost model with Shapley values; the former is a machine learning framework that enhances model fitness while the latter is an xAI method that provides informed explanations regarding the direction and significance of predictors. Based on macro-finance and fund-level factors, our fund performance evaluation of G10 countries uncovers novel insights into the diversification of country portfolios: both over- and under-diversification are associated with poor performance. Our analysis establishes consistency through a benchmark linear regression model and robustness at country level.

1. Introduction

Fund performance evaluation naturally differs from conventional analysis on cross-sectional stock returns, even though the same set of risk factors, as input variables,[1] may be under consideration. While evaluation models of stock returns are found to be mostly linear,[2] the impact of similar risk factors on equity fund performance is likely to be more complex and thus nonlinear. The literature has developed a number of solutions to overcome the challenges of fund evaluation. Some have focused on model inputs by designing performance measures tailored to equity funds (Kothari and Warner, 2001; Gil-Bazo and Ruiz-Verdú, 2009; McLemore et al., 2022), while others have attempted to improve their estimation strategy by incorporating insights from statistical modeling (Ferson and

☆ We thank the participants of the XTAI 2022 workshop in Swansea, Wales for helpful comments and Francel Dessou for his excellent research assistance. V.R.R. Kovvuri thanks Mitacs Inc. for the generous scholarship of the Globalink Research Award. H. Fu gratefully acknowledges financial support from the Social Sciences and Humanities Research Council of the Government of Canada and the AMF–GIRIF Fund at Laval University. H. Fu also thanks the grant from the Program of financial support to faculties for knowledge mobilization activities and the dissemination and promotion of research results, from Presses de l'Université Laval Development Fund. We all gratefully acknowledge financial support from the cooperation programme between the governments of Quebec and Wales, and honor the financial support we received for knowledge exchange in the Horizon 2020 project CID. Finally, we thank the two anonymous referees for their insightful feedback and constructive suggestions.

* Corresponding author.
E-mail addresses: v.r.r.kovvuri@swansea.ac.uk (V.R.R. Kovvuri), hsuan.fu@fsa.ulaval.ca (H. Fu), xyfan@ntu.edu.sg (X. Fan), m.seisenberger@swansea.ac.uk (M. Seisenberger).

[1] We use the terms risk factor, input variable, and explainer interchangeably, depending on which context we aim to highlight: finance, machine learning, or xAI.

[2] The factor models in the asset pricing literature usually provide fairly good explainability in the stock returns, e.g., Fama and French (1993), Carhart (1997).

https://doi.org/10.1016/j.frl.2023.104419
Received 31 May 2023; Received in revised form 1 August 2023; Accepted 1 September 2023
Available online 4 September 2023
1544-6123/© 2023 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Chen, 2021). This paper joins the latter stream of literature by using machine learning models to analyze the existing factors available for fund evaluation. Evidently, the machine learning techniques involved benefit from the nonlinear framework, which allows risk factors to determine performance in a different manner than with a typical linear regression model. In particular, we focus on explainable artificial intelligence (xAI) models in this paper for two reasons: first, the explainable feature provides the researcher with richer information on the relevance of each risk factor; second, xAI models also inform the direction of influence on the output variable, in a comparable way as with existing statistical models.

This paper first implements a machine learning technique to improve the model's goodness of fit on fund performance and then presents an xAI application to address the following two points. First, we validate that the machine learning model findings are consistent with finance domain knowledge. Second, the results of the xAI model enable us to provide novel implications arising from an open question in the literature on fund performance and international diversification.

It has been demonstrated that machine learning models can generally outperform the conventional linear setting in asset pricing studies.[3] Our paper consists of two principal elements regarding machine learning. First, we employ eXtreme Gradient Boosting (XGBoost), a state-of-the-art model that usually outperforms other tree-based models such as random forest (Hastie et al., 2009). Second, in order to overcome the black-box nature of the machine learning models, we also incorporate an xAI technique,[4] namely SHapley Additive exPlanations (SHAP) as proposed by Lundberg and Lee (2017), to gain information about the significance and direction of risk factors' influence.

This paper considers two different types of risk factors as explanatory features. The first group captures the macro-finance dynamics at aggregate level, including stock returns, foreign exchange rate returns, and interest rates across countries. The second group contains more granular variables that describe each fund's past performance via an indicator as well as cross-country allocations of investment holdings by their Herfindahl–Hirschman Index (HHI) values.[5]

Our findings can be summarized in three dimensions. First, the XGBoost model significantly enhanced explanatory power with respect to a conventional linear regression benchmark. Second, the directional explanations provided by the SHAP method are consistent with the coefficient signs of the benchmark model, with statistical significance. Third, international diversification is generally advantageous but not always beneficial for fund performance. To the best of our knowledge, this is the first study that employs xAI techniques to examine the relationship of fund performance and international diversification on portfolio holdings.

There have been a number of studies on xAI application in finance, including credit risk management (Bussmann et al., 2021), cryptocurrency investments (Babaei et al., 2022), debt financing (Lin and Bai, 2022) and prediction of stock splits (Li et al., 2023). In a similar vein to our work, but applied to risk assessments for the stock market, Berger (2023) leveraged boosted trees combined with a Shapley values-based xAI model. See also Adadi and Berrada (2018), Molnar (2023); and Jung et al. (2023) for an overview of xAI applications in other domains, and Moncada-Torres et al. (2021) and Duell et al. (2021) for a comparison of xAI models.

This paper is organized as follows: Section 2 introduces the necessary xAI background. Section 3 describes the data preparation, while Section 4 presents the findings. Section 5 discusses the robustness tests, followed by concluding remarks in Section 6.

2. SHapley Additive exPlanations (SHAP)

In machine learning, advanced models such as eXtreme Gradient Boosting (XGBoost)[6] proposed by Chen and Guestrin (2016) are state-of-the-art models that usually outperform other tree-based models such as random forest; however, their complex nature leads to no insight into how predictions are made. Therefore, xAI techniques such as SHAP have been devised to provide additional information.

SHAP is based on Shapley values (Shapley, 1953), a concept from cooperative game theory. For a given instance x, Shapley values quantify the contribution ϕ_j of each feature x_j towards the original model f's prediction:

$$\phi_j = \sum_{S \subseteq F_{x_j}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left(v(S \cup x_j) - v(S)\right).$$

Here, F symbolizes the set of all features, and S ranges over all subsets of F excluding x_j; $v(S)$ is the predicted value for each 'coalition' S in the original model f; and $v(S \cup x_j) - v(S)$ describes the marginal contribution of feature x_j. SHAP then adds up the influence on the prediction of all features as

$$f(x) = \phi_0 + \sum_{j=1}^{m} \phi_j,$$

where ϕ_0 is the "average" prediction when none of the features in x is present.

[3] The application of machine learning models in finance has focused heavily on market returns (Gu et al., 2020).

[4] There is another widely-used technique in the xAI domain named Local Interpretable Model-agnostic Explanations (LIME), developed by Ribeiro et al. (2016). Both LIME and SHAP belong to the class of additive feature attribution methods. The main difference between them lies in how they provide explanations. While LIME obtains the explanations by solving a penalized linear regression, SHAP considers all possible combinations of explanations in its estimation. Here, we are using SHAP because, by construction, LIME is more likely to suffer from the same problems faced in linear regression analysis (Molnar, 2023).

[5] The HHI is a concentration measure commonly applied in the finance studies (Camanho et al., 2022), defined as HHI = $\sum_i s_i^2$, where s_i represents the portfolio holding in the country i, and $\sum_i s_i = 1$. The HHI ranges between close to 0 and 1, where a higher value indicates that investment holdings are concentrated in a few or even one country where the HHI is equal to 1 assumes 100% holdings of the US stocks with $s_{US} = 100\%$ and $s_{non-US} = 0$. Conversely, a lower value represents a situation where portfolio holdings are well diversified across countries. For instance, HHI = 0.25 can be obtained from a portfolio whose holdings are equally weighted in four different countries.

[6] See Appendix A.1. for more details how we implement the XGBoost model in this paper.

Table 1
Descriptive statistics of fund data for G10 countries. The table presents the descriptive statistics of the data used in this paper. The input features include cross-country macro-finance indicators such as stock market return (ST), interest rate (IR), and exchange rate (ER), extracted using principal component analysis, and fund-level measures such as past performance (PPrfm), and the Herfindahl–Hirschman Index (HHI). The target feature Net Asset Value (NAV) is a binary variable, where a value of zero indicates a decrease and one indicates an increase compared with the previous quarter's value. The data at quarterly frequency cover the period from January 2017 to September 2021. In total, there are 18 quarters for each of the 4330 funds. We have also reported the metrics of the data distribution and temporal dependencies, including skewness, kurtosis, and the first-order autocorrelation AR(1).

	ST	IR	ER	PPrfm	HHI	NAV
AR(1)	−0.321	−0.235	0.879	0.695	0.935	−0.036
Skewness	1.582	−0.106	0.120	−0.410	−0.139	−0.503
Kurtosis	2.796	−1.575	−1.212	−0.465	−1.620	−1.746
mean	1.202	0.177	−0.739	2.485	0.526	0.622
Std. dev.	4.353	0.234	3.054	1.069	0.357	0.484
min.	−15.310	−0.220	−8.996	0.000	0.000	0.000
25%	0.520	0.012	−1.719	2.000	0.156	0.000
50%	2.050	0.146	−0.242	3.000	0.587	1.000
75%	3.530	0.410	1.011	3.000	0.880	1.000
max.	10.370	0.600	6.111	4.000	1.000	1.000

3. Data

3.1. Data collection

We collected data on Global Open-Ended Funds from the Morningstar Direct database. Although the data was available from as far back as 2000, the missing values in the portfolio holdings were substantial. In order to maintain a sufficiently large number of observations, we focused on the most recent years, from 2016 to 2021, in which we reserved the first year (four quarters) to construct an indicator of the fund's past performance. The holdings data was available at a monthly frequency but was transformed to quarterly frequency by keeping only the end-of-quarter observations,[7] as most funds conform to regulations by disclosing their holdings quarterly. We focused our study solely on funds domiciled in the G10 countries.[8] In short, our sample covered the period from 2017:Q1 to 2021:Q3,[9] encompassing 4,330 funds in total.

3.2. Macro-finance and fund-level variables

The targeted output of this study is fund performance as measured by the fund's net asset value (NAV); we use a binary variable denoted as *NAV* taking the values 1 and 0 to indicate whether or not the net asset value is higher than in the previous period. In addition, we gather a set of input features containing two groups of risk factors. The first group incorporates the macro-finance dynamics, including stock market returns (ST), interest rates (IR), and exchange rate returns (ER). For each factor, we employ a principal component analysis to reduce the dimension of a multi-country panel data into one single time series. The second group contains two pieces of fund-specific information: past performance (PPrfm), ranging from 0 to 4 calculated as the sum of the last four NAV values; and the Herfindahl–Hirschman index over the country portfolios, defined as $HHI_j = \sum_i s_{i,j}^2$ where $s_{i,j}$ is the equity investment in the ith country held by the fund j.[10]

Table 1 presents the whole-sample descriptive statistics for the variables used in this paper, including autocorrelation with one lag to verify the degree of time-series correlation. Generally, the fund-specific factors, namely PPrfm and HHI, are found to have higher persistence than the macro-finance variables.

Finally, we refer to Fig. 1 for a detailed view of the proposed model's architectural workflow, from data preprocessing to prediction interpretation using SHAP.

4. Results

We start with our benchmark model of a probit regression, comparing it with our XGBoost model in Section 4.1. In Section 4.2, we apply the xAI technique, further examining the implications of international diversification in Section 4.3.

[7] Since regulations vary between countries, quarterly holdings are not always recorded at the end of the quarter. For missing data, we employed a forward-filling strategy by using the most recent observation recorded in the previous two months. If there is no observation in those two months, we leave the data as missing in this particular quarter.

[8] Despite its name, the G10 actually includes 11 countries, namely Belgium (BEL), Canada (CAN), France (FRA), Germany (DEU), Italy (ITA), Japan (JPN), the Netherlands (NLD), Sweden (SWE), Switzerland (CHE), the United Kingdom (GBR), and the United States (USA).

[9] While the whole sample spans 19 quarters, our analysis includes only 18 due to one-period lag applied to the input variables.

[10] Further details on these features and how they are computed can be found in Appendix B.

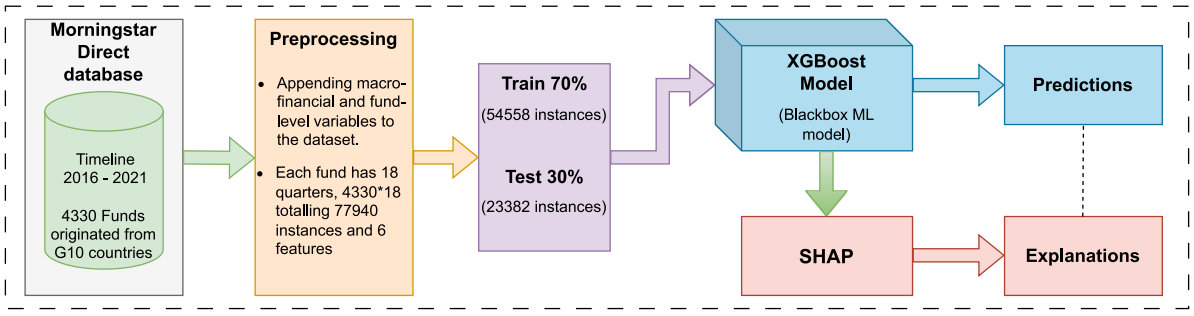


Fig. 1. Architecture of the proposed model's workflow. The process starts with the Morningstar Direct dataset, to which macro-financial and fund-level variables are added during preprocessing. The enriched dataset is then split into a 70% training and 30% testing set. These sets are subsequently used as input to the XGBoost model. The model's output is interpreted using the SHAP model, providing comprehensible explanations for the predictions.

Table 2
Regression results. The table presents the results of a probit regression analysis of the relationship between various factors, namely Stock Market, Interest Rates, Exchange Rates, PPrfm, and HHI. The table displays the coefficients and standard errors for six regression models. Additionally, it reports the pseudo R-squared values for each regression model, which measure the goodness of fit of the model. The regression analysis reveals key trends: a rise in Stock Market returns and PPrfm, representing past fund performance, correlates with improved fund performance, while conversely increased Interest and Exchange Rates are linked to decreased fund performance. Moreover, a higher HHI value, signifying market concentration, is associated with increased fund performance. The standard errors are robust to heteroskedasticity and autocorrelation-consistent (HAC) and are reported in parentheses. The significance levels for the coefficients are denoted as *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$. The data cover the period from January 2017 to September 2021.

	Dependent variable: Net asset value (NAV)					
	(1)	(2)	(3)	(4)	(5)	(6)
Stock Market	0.012***				0.021***	0.021***
	(0.000)				(0.000)	(0.000)
Exchange Rate		−0.066***			−0.051***	−0.050***
		(0.002)			(0.002)	(0.002)
PPrfm			0.121***		0.155***	0.128***
			(0.002)		(0.003)	(0.003)
Interest Rate				−0.217***	−0.549***	−0.633***
				(0.018)	(0.023)	(0.024)
HHI						0.190***
						(0.014)
Pseudo R-squared	−0.032	−0.025	0.004	−0.042	0.049	0.050
Observations	54,558	54,558	54,558	54,558	54,558	54,558

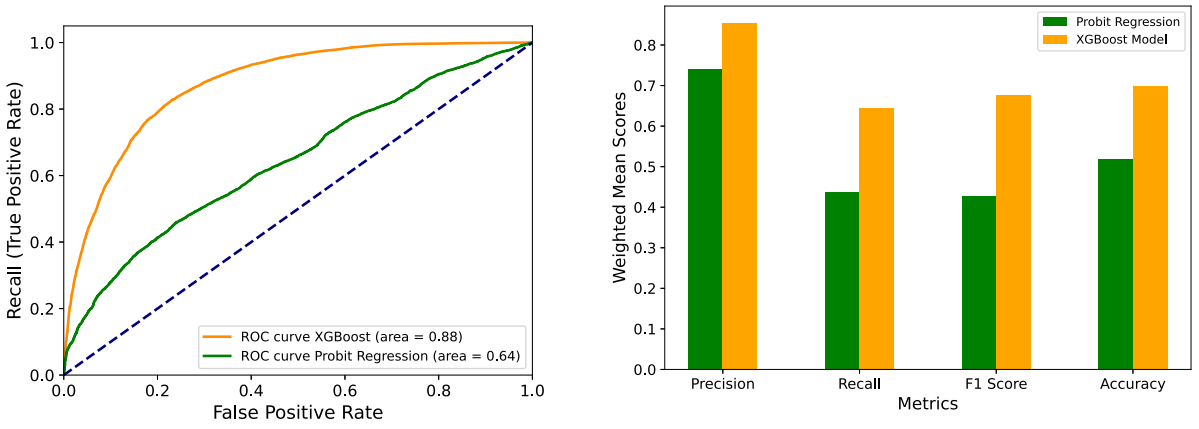
Table 3
Performance metrics of XGBoost model. The table presents the evaluation results of a binary classification model on a synthesized dataset of 28,886 instances, with two classes, 0 and 1. The metrics shown include support, precision, recall, f1 score, and accuracy. For class 1, the model achieved a precision of 0.82, a recall of 0.88, and an f1 score of 0.85. This means that out of all instances predicted as class 1, 82% were actually class 1, and the model was able to correctly identify 88% of all instances that actually belong to class 1. The f1 score, which is a harmonic mean of precision and recall, was 0.85. For class 0, the values are computed analogously. The data covers the period from 2017 January to 2021 September.

Support	Precision	Recall	f1 score	Model accuracy
0: 14394	0: 0.87	0: 0.80	0: 0.83	0.84
1: 14492	1: 0.82	1: 0.88	1: 0.85	

4.1. Probit regression versus the XGBoost model

The probit regression model assesses the binary classification problem as found in the XGBoost model by using a linear structure. We report the results of both univariate and multivariate analyses in Table 2. The pseudo R-squared is small or even negative, implying poor model fitness. Alternatively, the input features show strong correlations to the fund performance. The coefficient signs serve as our benchmark, which can be used to verify the reliability of the xAI application.

Table 3 presents the performance metrics of our XGBoost application. For tuning the hyperparameters, we employ a 70/30 split between the training and testing sets in the XGBoost model. Note that Classes 0 and 1 represent bad and good fund performance, respectively. Counting the percentage of true positives in the XGBoost model, the precision metric for Class 0 (1) indicates that 87% (82%) of predicted of bad (good) performances were actually bad (good). Similarly to precision, the recall metric accommodating the concept of false negatives indicates the fraction that is correctly identified in each class. Last, the F1 score, an average between the previous two metrics, is balanced as a metric between false negatives and false positives. As our values of precision and recall



(i) ROC Curves. (ii) Metrics Scores Comparison

Fig. 2. Model performance comparison between Probit Regression vs XGBoost Model. (i) Receiver Operating Characteristic (ROC) curve analysis. The x- and y-axes represent the false and true positive rates, respectively. The dashed line represents a random classifier. The orange line indicates the ROC curve for the XGBoost model, with an AUC of 0.87 demonstrating superior predictive accuracy compared to the probit regression model. The green line represents the ROC curve for the Probit Regression model, with an AUC of 0.70. **(ii) Metrics Scores Comparison.** The bar plots represent the weighted mean scores of precision, recall, F1, and accuracy metrics. The probit regression and XGBoost models are represented in green and orange, respectively. Across all metrics, the XGBoost model consistently outperforms the probit regression model.

metrics are equally high, it is natural to obtain well-balanced f1 scores across classes.[11] This result establishes the validity of the machine learning model which is essential for the xAI application in the next step.

Finally, in Fig. 2, we compare the performance of the two models, Probit regression and XGBoost, using an ROC curve[12] and metric comparison chart.[13] XGBoost proves superior in capturing these highly non-linear relationships in explaining NAV changes, making it a more suitable choice for our xAI analysis going forward.

4.2. xAI results based on input features

We compute SHAP values based on the XGBoost model to analyze the importance and directional influence of the input features on fund performance. Fig. 3 depicts the results. Each data point signifies a fund data instance, with a color gradient that marks each feature value: copper for low and black for high. A high-value feature in black with a positive SHAP value suggests a strong positive influence on fund performance. On the y-axis, the explanatory features are ranked according to their importance. Our xAI analysis reveals a strong positive impact of Stock Market (Oh and Parwada, 2007), followed by a positive impact of Fund's Past Performance and a negative influence of the Interest Rate (Lipton and Buetow, 2000). The last two features are Exchange Rate, exhibiting a negative correlation, and HHI with an ambiguous relationship to fund performance.

4.3. Influence of international diversification

Building on the previous section, where we identified an ambiguous relationship between HHI and fund performance, we aim to further elucidate this ambiguity. To do so, we subdivided the fund dataset into four quartiles based on HHI values (see Fig. 4)[14] and then applied the XGBoost classifier to each subsample. The model showcased robust performance, with an accuracy ranging between 0.83 and 0.84.[15] Fig. 5 illustrates the SHAP results applied to each quartile. In the first quartile, where the funds are well-diversified, low HHI values lead to a lower fund performance. Conversely, in the fourth quartile (Fig. 5(iv)), high HHI values are negatively correlated with performance. With regard to the other explanatory features, the relationships are largely consistent with Section 4.2.

Overall, our results indicate that fund performance varies across different HHI groups, with a moderate level of diversification having the potential to enhance fund performance.

[11] For details of all the performance metrics used in Table 3, see Appendix A.2.

[12] The ROC curve visualizes the performance of a binary classifier, showing the trade-off between sensitivity (true positive rate) and 1-specificity (false positive rate) as the classification threshold changes.

[13] See Figure B.3 in Appendix B for a detailed analysis of how these performance metrics vary with the decision threshold for each model. This additional analysis provides a more comprehensive understanding of the performance dynamics of the models across a wide range of thresholds.

[14] The associated summary statistics for these subsamples can be found in Appendix C.

[15] Key metrics (precision, recall, F1 score, accuracy), are provided in Table C.2 in Appendix C. Specifically, the model attained an accuracy of 0.83 for Quartile 1 (HHI< 0.16), 0.84 for Quartile 2 (HHI in the range 0.16–0.58) and Quartile 3 (HHI in the range 0.58–0.88), and 0.83 for Quartile 4 (HHI> 0.88).

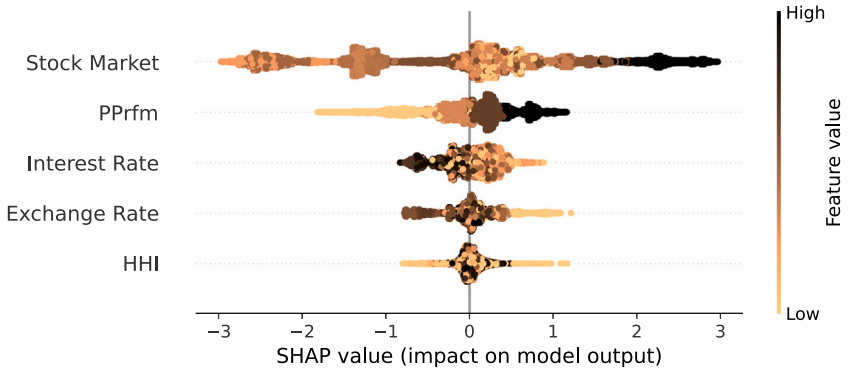


Fig. 3. SHAP summary plot. The data covers the timeline from January 2017 to September 2021, illustrating feature importance and relationships with respect to fund performance. The plot showcases the impact of each feature on the model's prediction, represented by its position along the x-axis, while the left y-axis displays feature names sorted by importance. The right y-axis depicts a color gradient indicating feature values, ranging from copper to black. Analysis of the fund data reveals significant relationships such as a positive link between stock market returns and fund performance, a positive correlation with historical fund performance, a negative association between interest rates and fund performance, and a negative impact of exchange rates on fund performance. The relationship with HHI, however, remains ambiguous, as suggested by the copper color on both sides of the plot.

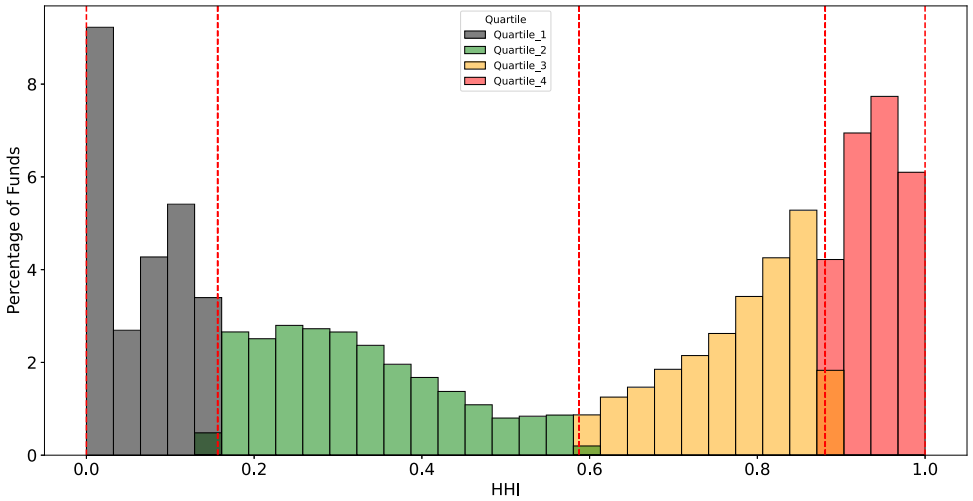


Fig. 4. Histogram of HHI quartiles. This figure plots Herfindahl–Hirschman Index (HHI) values on funds' portfolio holdings as a histogram. For the first quartile, the HHI value are lower than 0.156. For the second and third quartiles, the ranges of HHI values are (0.156, 0.587] and (0.587, 0.881], respectively. Finally, the fourth quartile contains HHI values exceeding 0.881. The x-axis represents HHI value, while the y-axis represents the percentage of funds. Quartiles are separated by vertical dashed lines.

5. Robustness tests

Our empirical analysis provides consistent results for fund performance evaluation. In this section, we will examine whether the model is robust in the cross-section and time series. Specifically, we check if the findings are driven by specific countries or time periods.

5.1. Country-level robustness

To investigate whether the results are driven by one single country, we designed two exercises. First, we retrained the XGBoost model with subsamples of 10 out of 11 countries. The performance metrics, namely precision, recall, and F1-score, remain fairly stable across different subsets of countries. Note that the model's accuracy is remarkably stable, ranging over the narrow range between 0.83 and 0.84.[16] Additionally, we calculate the Pearson correlation between fund performance and features for each country.

[16] Comprehensive details of these metrics can be found in Appendix D.

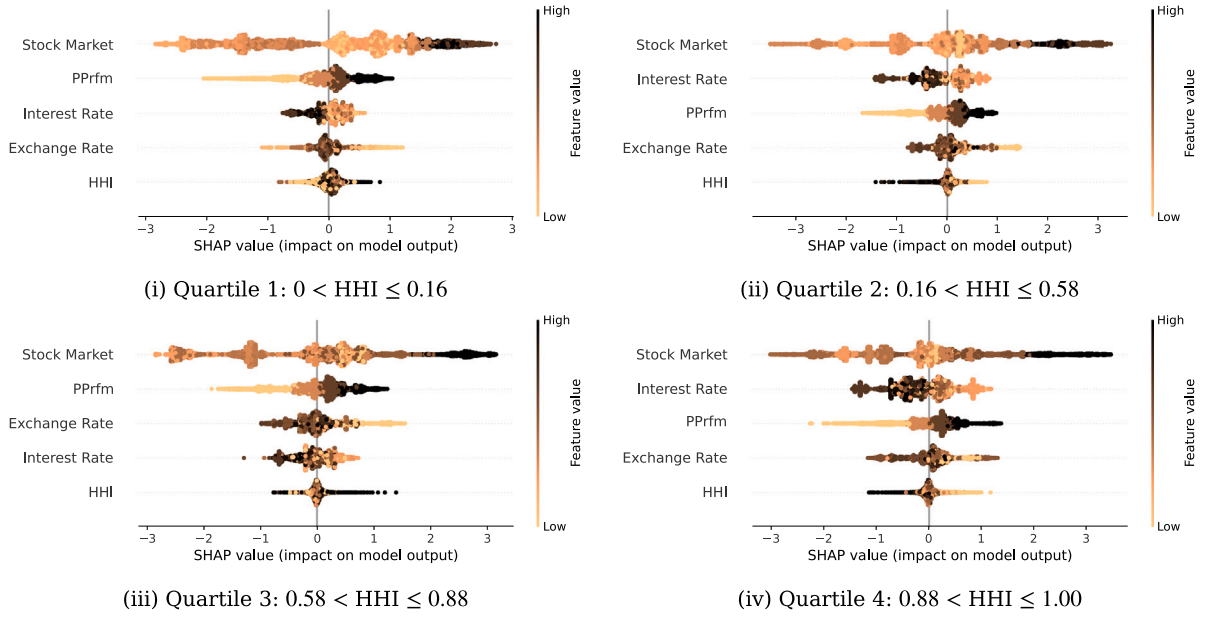


Fig. 5. SHAP summary plot in HHI quartiles. This figure illustrates subsamples of funds by the HHI quartiles from January 2017 to September 2021. The high (low) HHI values within each quartile are marked by dark (light) color. In the first quartile Fig. 5(i), the light tail on the left and dark tail on the right shows a positive correlation between the HHI values and fund performance, which is found to be reversed in the forth quartile in Fig. 5(iv), implying underperformance of funds with extreme HHI values – that are too close to 0 or 1. Note that relationships for the other features are consistent with Section 4.2.

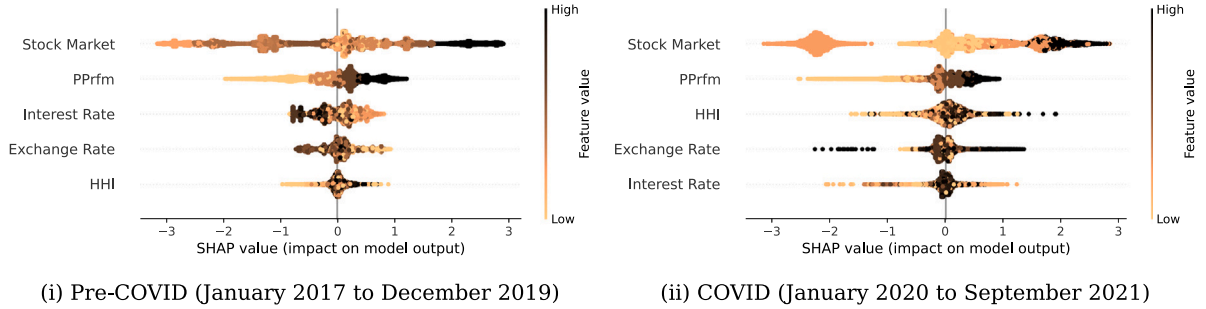


Fig. 6. SHAP summary plot in subperiods. The figure depicts two distinct periods: pre-COVID (on the left) and COVID (on the right). In each figure, the x-axis represents SHAP values. The color gradient on the right y-axis indicates the values of the various features, while the left y-axis lists these features according to their significance.

Table 4 shows that the correlations are consistent across countries, albeit with a few exceptions. For instance, the correlation with past performance in Sweden, Germany, and Japan is found to have the opposite sign to the whole-sample analysis in both the XGBoost and conventional linear regression models. This is likely due to the small samples for these countries.[17]

5.2. Influence of COVID-19

Following the cross-sectional robustness tests, we next investigate the dimension of the time series. Specifically, we focus on the influence of COVID-19 by dividing the data into two subperiods: before and after December 2019.[18] Fig. 6(i) presents the SHAP summary plot prior to the onset of the COVID-19 pandemic, showing a similar pattern to the whole-sample results in Fig. 3. Conversely, the COVID-19 subperiod in Fig. 6(ii) shows a much higher importance, but with an opposite influence, of Interest Rate on fund performance. While this finding of low Interest Rate values associating with poor fund performance is not surprising, it does underline the importance of having a complete sample for the xAI analysis that is balanced between crisis and non-crisis periods.

[17] The corresponding SHAP summary plots are presented in Appendix E.

[18] The performance metrics of XGBoost model before and after the COVID-19 outbreak can be found in Appendix F.

Table 4
Correlation between fund performance and the features. The figure represents Pearson Correlation coefficients and associated *p*-values for several key financial parameters from January 2017 to September 2021. These parameters include the stock market (defined as the logarithmic return), interest rates, exchange rates, PPrfm (a measure summarizing fund performance over the most recent four quarters), and the Herfindahl–Hirschman Index (HHI, an indicator of market concentration and diversification). Significance levels are as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. The superscript 'a' signifies values represented as 0.000 or −0.000, namely minimal amounts that round to zero at the third decimal place.

Country	Sample size	Model accuracy	Features				
			Stock market	Interest rates	Exchange rate	PPrfm	HHI
G10 ex. USA	37,080	0.84	0.140***	−0.090***	−0.240***	0.070***	0.010***
			(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
USA	40,860	0.84	0.150***	−0.140***	−0.200***	0.090***	0.020***
			(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
BEL	270	0.68	0.140**	−0.020	−0.150**	0.010	−0.120**
			(0.022)	(0.713)	(0.015)	(0.882)	(0.043)
CAN	14,328	0.85	0.150***	−0.110***	−0.260***	0.140***	0.010*
			(0.000)	(0.000)	(0.000)	(0.000)	(0.099)
CHE	2,394	0.80	0.140***	−0.040**	−0.180***	0.040*	0.050**
			(0.000)	(0.041)	(0.000)	(0.059)	(0.025)
FRA	4,518	0.84	0.130***	−0.080***	−0.260***	0.020**	−0.010*
			(0.000)	(0.000)	(0.000)	(0.029)	(0.081)
GBR	8,334	0.83	0.150***	−0.070***	−0.230***	0.050***	−0.010**
			(0.000)	(0.000)	(0.000)	(0.000)	(0.040)
DEU	2,484	0.84	0.110***	−0.070***	−0.260***	−0.040**	0.000[a]
			(0.000)	(0.000)	(0.000)	(0.033)	(0.957)
ITA	216	0.73	0.220***	−0.140**	−0.200***	0.040	0.080
			(0.001)	(0.042)	(0.003)	(0.516)	(0.221)
JPN	306	0.74	0.230***	−0.100*	−0.190***	−0.070	−0.050
			(0.000)	(0.075)	(0.001)	(0.244)	(0.359)
NLD	954	0.80	0.110***	−0.140***	−0.220***	0.020	0.030
			(0.000)	(0.000)	(0.000)	(0.617)	(0.351)
SWE	3,276	0.86	0.130***	−0.060***	−0.270***	−0.040***	−0.000[a]
			(0.000)	(0.000)	(0.000)	(0.001)	(0.819)

6. Conclusion

Explainable artificial intelligence (xAI) techniques currently show great potential to enrich information content in financial and economic studies. On one hand, our findings supplement Gu et al. (2020) by providing further information on the importance and directional influence specific to each explanatory feature. We showed that a state-of-the-art machine learning model (specifically, XGBoost) together with xAI techniques can produce reliable and interpretable results in financial studies. On the other, we leveraged the advantage of machine learning models in analyzing highly nonlinear problems with large datasets as in Li et al. (2023). We further examined the diversification implications for portfolio holdings across the G10 countries, finding good performance of equity funds to be associated with a moderate degree of diversification. Moreover, our results as implied by xAI were able to replicate statistical characteristics such as signs and significance of the benchmark linear regression model at both aggregate and country levels. We therefore advocate the benefits of applying xAI to complex questions that remain open in the finance literature. Additionally, adaptations of the xAI approach that cater to the endogenous or exogenous nature[19] of input variables form a prospective direction for future research with the enhancement of interpretability.

Data availability

The authors do not have permission to share data.

Acknowledgment

We thank the participants of the XTAI 2022 workshop in Swansea, Wales for helpful comments and Francel Dessou for his excellent research assistance. V.R.R. Kovvuri thanks Mitacs Inc. for the generous scholarship of the Globalink Research Award. H. Fu gratefully acknowledges financial support from the Social Sciences and Humanities Research Council of the Government of Canada and the AMF–GIRIF Fund at Laval University. H. Fu also thanks the grant from the Program of financial support to faculties for knowledge mobilization activities and the dissemination and promotion of research results, from Presses de l'Université Laval Development Fund. We all gratefully acknowledge financial support from the cooperation programme between the governments of Quebec and Wales, and honor the financial support we received for knowledge exchange in the Horizon 2020 project CID. Finally, we thank the two anonymous referees for their insightful feedback and constructive suggestions.

[19] Controllable versus uncontrollable features in the context of Kovvuri et al. (2022).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.frl.2023.104419.

References

Adadi, A., Berrada, M., 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access 6, 52138–52160.
Babaei, G., Giudici, P., Raffinetti, E., 2022. Explainable artificial intelligence for crypto asset allocation. Finance Res. Lett. 47, 102941.
Berger, T., 2023. Explainable artificial intelligence and economic panel data: A study on volatility spillover along the supply chains. Finance Res. Lett. 54, 103757.
Bussmann, N., Giudici, P., Marinelli, D., Papenbrock, J., 2021. Explainable machine learning in credit risk management. Comput. Econ. 57, 203–216.
Camanho, N., Hau, H., Rey, H., 2022. Global portfolio rebalancing and exchange rates. Rev. Financ. Stud. 35 (11), 5228–5274.
Carhart, M.M., 1997. On persistence in mutual fund performance. J. Finance 52 (1), 57–82.
Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 785–794.
Duell, J., Fan, X., Burnett, B., Aarts, G., Zhou, S.M., 2021. A comparison of explanations given by explainable artificial intelligence methods on analysing electronic health records. In: BHI 2021.
Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. J. Financ. Econ. 33 (1), 3–56.
Ferson, W., Chen, Y., 2021. How many good and bad funds are there, really? In: Handbook of Financial Econometrics, Mathematics, Statistics, and Machine Learning. World Scientific, pp. 3753–3827.
Gil-Bazo, J., Ruiz-Verdú, P., 2009. The relation between price and performance in the mutual fund industry. J. Finance 64 (5), 2153–2183.
Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. Rev. Financ. Stud. 33 (5), 2223–2273.
Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H., 2009. The elements of statistical learning: Data mining, inference, and prediction, second ed. In: Springer Series in Statistics.
Jung, J., Lee, H., Jung, H., Kim, H., 2023. Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. Heliyon 9.
Kothari, S., Warner, J.B., 2001. Evaluating mutual fund performance. J. Finance 56 (5), 1985–2010.
Kovvuri, V.R.R., Liu, S., Seisenberger, M., Fan, X., Muller, B., Fu, H., 2022. On understanding the influence of controllable factors with a feature attribution algorithm: a medical case study. In: INISTA 2022.
Li, A., Liu, M., Sheather, S., 2023. Predicting stock splits using ensemble machine learning and SMOTE oversampling. Pac.-Basin Finance J. 78, 101948.
Lin, B., Bai, R., 2022. Machine learning approaches for explaining determinants of the debt financing in heavy-polluting enterprises. Finance Res. Lett. 44, 102094.
Lipton, A.F., Buetow, G.W., 2000. Interest rate sensitivity of equity mutual funds. J. Wealth Manag. 2 (4), 61–71.
Lundberg, S., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30, 4765–4774.
McLemore, P., Sias, R., Wan, C., Yüksel, H.Z., 2022. Active technological similarity and mutual fund performance. J. Financ. Quant. Anal. 57 (5), 1862–1884.
Molnar, C., 2023. Interpretable machine learning, second ed. Online version available at https://christophm.github.io/interpretable-ml-book/. (Accessed 5 July 2023).
Moncada-Torres, A., van Maaren, M.C., Hendriks, M.P., Siesling, S., Geleijnse, G., 2021. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. Sci. Rep. 11 (1), 6968.
Oh, N.Y., Parwada, J.T., 2007. Relations between mutual fund flows and stock market returns in Korea. J. Int. Financ. Mark. Inst. Money 17 (2), 140–151.
Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144.
Shapley, L.S., 1953. A value for n-person games. Contrib. Theory Games 2 (28), 307–317.