Borsa Istanbul Review 25 (2025) 1403–1419

Contents lists available at ScienceDirect

Borsa Istanbul Review

journal homepage: www.elsevier.com/journals/borsa-istanbul-review/2214-8450

Predicting mutual fund performance with machine learning: Are ESG pillar scores relevant predictors of fund return?☆

Francisco Climent [a], Alexandre Momparler [b], Pedro Carmona [c,*]

[a] Universitat de València, Department of Financial and Actuarial Economics, Spain
[b] Universitat de València, Department of Corporate Finance, Spain
[c] Universitat de València, Department of Accounting, Spain

ARTICLE INFO

Keywords:
Mutual fund
Financial performance
ESG
Sustainability
Machine learning
Fund selection

ABSTRACT

The growing emphasis on sustainability within the business community has led to a progressive integration of Environmental, Social, and Governance values (ESG) into the investment process. This study aims to identify those key fund characteristics that best predict financial performance, with a special focus on the individual impact of ESG Pillar scores. Using the Extreme Gradient Boosting (XGBoost) algorithm, a machine learning technique known for enhancing predictive accuracy, we analyze cross-sectional data on Euro-denominated equity mutual funds with a global scope over a five-year period (2020–2024). In addition, this paper evaluates the advantages of the XGBoost algorithm in predicting mutual fund returns by comparing its performance against two benchmark models: OLS regression and a deep learning architecture. Our findings reveal that ESG Pillar Social score is the second most important predictive factor and that it is positively associated with fund performance, whereas ESG Pillar Environmental score ranks fifth in predictive power and it shows a negative relationship with performance. These insights offer practical value for values-driven investors, financial advisors, and fund managers by supporting investment decisions that align financial performance with sustainability considerations. This study addresses a critical gap in the literature by analyzing the individual effects of each ESG pillar, rather than relying on aggregated ESG scores as commonly done in prior research. The use of cross-sectional data provides a detailed representation of these relationships over a five-year span. Our approach expands the literature by using advanced machine learning to show links between sustainability and fund returns.

1. Introduction

Mutual funds are essential financial instruments that play a key role in gathering savings from both individual and corporate investors. Offered by institutional investors, these investment vehicles allow for significant capital diversification, minimizing risk and maximizing potential returns.

Diversification may be the most important benefit of investing in mutual funds because it allows investors to invest in a portfolio that includes a large number of securities, thereby reducing the risk associated with investing in a single asset. Professional Management is another advantage of investing through mutual funds, as they are managed by experienced professionals who analyze the market and make informed investment decisions. Individual investors can also have access to investment opportunities that may be out of reach for them individually. Despite their numerous advantages, it is crucial for investors to understand the risks associated with mutual funds. Factors such as management fees, market volatility, and their own risk tolerance are essential aspects to consider before making an investment.

Socially responsible investors would like to ensure that their investments align with their values. Environmental, social, and governance (ESG) factors play an essential role in the current business environment and mutual fund ESG scores provide investors with valuable information about how well the fund incorporates sustainability factors into its investment strategy. By examining ESG ratings, investors can get a clearer picture of how a mutual fund aligns with their values

☆ The authors gratefully acknowledge financial support from (1) PID 2023-153128NB-I00 funded by MCIU/AEI/10.13039/501100011033/FEDER,UE; and (2) UV-INV_AE-3663662 funded by University of Valencia.

* Corresponding author.
E-mail addresses: f.jose.climent@uv.es (F. Climent), alexandre.momparler@uv.es (A. Momparler), pedro.carmona@uv.es (P. Carmona).

https://doi.org/10.1016/j.bir.2025.07.019
Received 11 March 2025; Received in revised form 28 July 2025; Accepted 28 July 2025
Available online 15 August 2025

and its commitment to sustainable and responsible investing.

Investing in sustainable mutual funds represents an attractive option for socially responsible individual investors and companies willing to maximize their savings while promoting sustainable and responsible corporate behavior. The overall “ESG score” (see Figs. 1 and 2) is a crucial tool that helps investors compare the sustainability performance of different funds.

Several well-known rating agencies and organizations are currently providing ESG ratings for mutual funds and companies. These agencies vary in their methodologies and focus areas, so investors often look at ratings from multiple sources to get a better picture of a fund’s or a company’s ESG performance. At the present time, the two main agencies offering ESG ratings for mutual funds are Refinitiv Eikon and Morningstar. As all ESG data used in this research work is collected from the Refinitiv Eikon database, we deem it appropriate to provide a short description of how mutual fund ESG metrics are generated.

Refinitiv Eikon ESG scores measure a fund portfolio company’s relative ESG performance based on company-reported data. ESG scores include 630 company-level measures, of which a subset of 186 of the most comparable and relevant per industry power the overall company assessment and scoring process (see Fig. 1). These are grouped into 10 categories that reformulate three Pillar scores (environmental, social, and governance) which comprise the final ESG score (see Fig. 2).

Access to precise, standardized, and comparable ESG data empowers investors to assess the social and ethical implications of their investments, going beyond mere financial metrics. This represents a transformative shift toward an era where investments are not only anticipated to generate financial returns but also to foster a sustainable and equitable future.

According to the Global Sustainable Fund Flows Report (Morningstar, 2023), Europe is the most important ESG market as it represents the largest part of the global market of sustainable funds, which includes both open-end and exchange-traded funds with a clear focus on sustainability. As of December 2023, Europe leads the sustainable funds market with 84 % of total assets invested in sustainable funds and 73 % of registered sustainable funds. The US comes second with 11 % of sustainable fund assets and 9 % of funds (see Table 1). Note that Europe and the US together account for 95 % of global sustainable fund assets and 82 % of sustainable funds.

Investors are motivated by many different factors when choosing sustainable mutual funds, and social responsibility as well as long-term profit are essential for most of them. These diverse motivations reflect a growing recognition that financial performance and social responsibility can go hand-in-hand, making sustainable mutual funds an appealing choice for a broad range of investors.

Companies committed to social responsibility values often exhibit a greater degree of transparency in their operations. This increased transparency can have a ripple effect, potentially enhancing their financial performance.

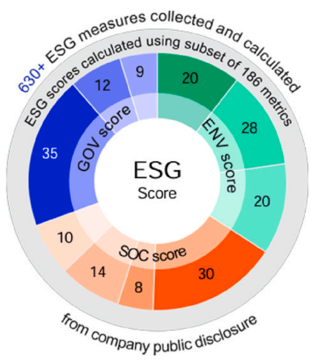

Fig. 1. ESG metrics.
Source: Environmental, social, and governance scores from LSEG, Dec 2023.

The wide range of sustainable funds available makes mutual fund selection a complex task. To facilitate fund selection for socially responsible investors aiming to maximize their savings, we assess the impact of overall ESG scores as well as the individual impact of each of the three ESG Pillars on fund financial performance. To the best of our knowledge, this is the first study that analyses the individual impact on fund performance of ESG Pillar scores.

The results of our study help recognize those fund features with greater predictive capacity for mutual fund financial performance during our analysis time frame (2020–2024). It is noteworthy that both Social and Environmental scores (two factors that comprise the ESG score) are among the five most important fund features for predicting fund financial performance.

Modern portfolio theory (MPT) provides a foundational framework for understanding the benefits of diversification in mutual fund investing, emphasizing the trade-off between risk and return (Markowitz, 1952). According to MPT, investors can optimize their portfolios by combining assets with varying risk profiles, thereby maximizing expected returns for a given level of risk. This theoretical perspective supports the rationale for mutual funds as vehicles for diversification and professional management.

In recent years, ESG-related management theories have further expanded the traditional view of investment by incorporating non-financial factors into decision-making. Stakeholder theory, for example, argues that companies—and by extension, mutual funds—-should consider the interests of all stakeholders, not just shareholders, to achieve long-term value creation (Freeman, 1984). Empirical evidence supports this approach: studies have found that funds with higher ESG scores often demonstrate improved risk-adjusted returns and resilience during market downturns (Broadstock et al., 2021; Whelan et al., 2021).

Moreover, the resource-based view of the firm suggests that integrating ESG practices can serve as a source of competitive advantage, as these practices may enhance reputation, operational efficiency, and stakeholder trust (Barney, 1991; Hart, 1995). Recent empirical research has shown that mutual funds with strong ESG profiles can attract more inflows and deliver superior performance, particularly when social and environmental factors are considered separately (Momparler et al., 2025).

This study integrates theoretical perspectives with empirical evidence to offer a more comprehensive understanding of how ESG factors—analyzed through the lens of modern portfolio theory and management frameworks—impact mutual fund performance.

The remainder of this paper is organized as follows. In section 2, the existing literature on sustainable mutual funds and fund financial performance is reviewed. Section 3 describes the sources of data and the characteristics of the sample. Section 4 describes the methodology, the variables, and model specification. Then, empirical results are reported in section 5. Finally, the main findings, implications, and limitations of our study are discussed in section 6.

2. Literature review

For several decades, researchers and academics have investigated the relationship between ESG factors and financial performance. Most studies report a positive correlation between ESG scores and various indicators of corporate performance, including operational efficiency, stock returns, and reduced capital costs. High ESG ratings are increasingly regarded as key indicators of corporate transparency and effective management, driving improved financial performance and higher returns. Recent research also suggests that ESG investing, under specific conditions, may improve risk management and yield returns comparable to traditional investment strategies.

Several studies indicate that ESG investing may offer effective protection during economic downturns. A recent study examining the impact of ESG performance during the global financial crisis precipitated by the COVID-19 pandemic demonstrated that companies with

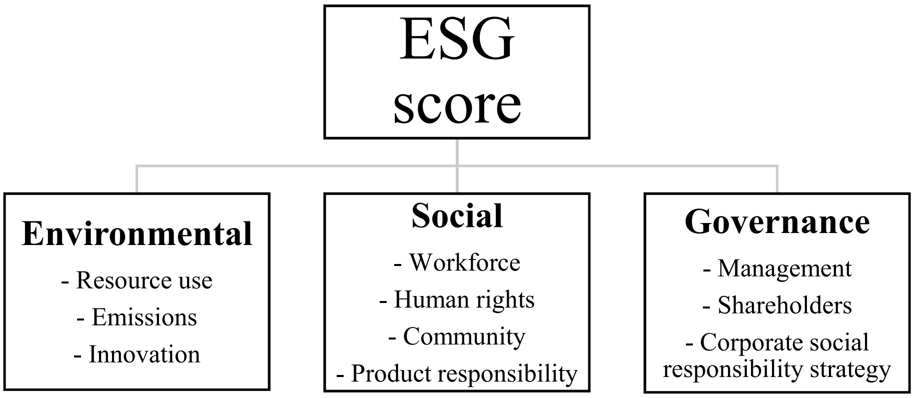

Fig. 2. ESG score three Pillars.
Source: Environmental, social and governance scores from LSEG, Dec 2023.

Table 1
Global Sustainable Funds (2023).

Region	Assets		Funds	
	USD Billion	% Total	#	% Total
Europe	2492	84 %	5433	73 %
United States	324	11 %	647	9 %
Asia ex-Japan	62	2 %	595	8 %
Australia/New Zealand	33	1 %	263	4 %
Japan	25	1 %	235	3 %
Canada	31	1 %	312	4 %
Total	**2967**		**7485**	

Source: Morningstar Direct, Manager Research. Data as of December 2023.

higher ESG scores outperformed those with lower scores (Broadstock et al., 2021).

A substantial body of empirical research has explored the link between ESG factors and financial performance. For example, Ruf et al. (2019), Dolvin et al. (2019), Steen et al. (2020), Abate et al. (2021), Raghunandan and Rajgopal (2022), Del Vitto et al. (2023), Momparler et al. (2025), Shin et al. (2024), and Vilas et al. (2024) have all contributed to our understanding of how ESG integration can influence mutual fund returns. These studies, grounded in finance and management theories such as stakeholder theory and the resource-based view, generally find that higher ESG ratings are associated with improved risk-adjusted returns, greater resilience during periods of market stress, and, in some cases, enhanced long-term value creation. For instance, Ruf et al. (2019) and Steen et al. (2020) highlight that funds with high ESG ratings tend to perform better during economic downturns, supporting the notion that ESG can serve as a buffer against market volatility.

The literature has also evolved methodologically, particularly with the adoption of advanced machine learning techniques to analyze the ESG–performance relationship. Li and Rossi (2021) pioneered the use of machine learning to capture the non-linear and interactive effects of fund characteristics, demonstrating that models such as Boosted Regression Trees outperform traditional linear approaches. Kanade et al. (2022) extended this work by applying artificial neural networks to mutual fund selection in the Indian market, finding that these models offer superior predictive accuracy. Kaniel et al. (2023) further advanced the field by incorporating a wide array of variables—including stock and fund characteristics as well as macroeconomic indicators—into neural network models, revealing that fund-specific features often outweigh stock-level data in predicting alpha. De Miguel et al. (2023) emphasized the importance of including macroeconomic variables in machine learning models to improve robustness and out-of-sample performance. Del Vitto et al. (2023) focused on the explainability of ESG ratings using both white-box and black-box machine learning models, providing insights into the factors driving ESG assessments. Momparler et al. (2025) applied Extreme Gradient Boosting (XGBoost) to US-registered global equity funds, confirming that ESG scores are among the top predictors of fund returns. Shin et al. (2024) used Categorical Boosting and SHAP values to analyze ESG and financial performance in project-driven sectors, while Vilas et al. (2024) combined logistic regression, neural networks, random forest, and gradient boosting to model the influence of ESG on fund flows and returns, finding that while ESG matters, traditional financial variables often play a larger role.

Building on these methodological advances, our study contributes to the literature by applying XGBoost to a cross-sectional sample of Euro-denominated global equity funds, with a particular focus on the individual impact of each ESG pillar. Unlike many previous studies that rely on aggregate ESG scores or traditional linear models, our approach disaggregates ESG into its Environmental, Social, and Governance components and leverages advanced machine learning to capture complex, non-linear relationships. This enables a deeper understanding of how specific sustainability factors influence mutual fund performance and helps identify which ESG pillars are most relevant for investors aiming to achieve both financial returns and value alignment.

In summary, the integration of empirical evidence with theoretical frameworks from finance and management underscores the significant role ESG factors can play in shaping mutual fund performance. By synthesizing insights from prior research with advanced analytical techniques, this study seeks to deepen our understanding of the mechanisms through which ESG considerations influence financial outcomes—particularly within the context of modern portfolio theory and stakeholder-oriented management.

The objective of this research is to assist socially responsible investors in the fund selection process. We conduct an empirical analysis utilizing machine learning techniques, with a particular emphasis on the XGBoost algorithm, to identify the fund characteristics most predictive of fund performance.

3. Data

The data was collected in 2024 from the Refinitiv Eikon Database (https://www.refinitiv.com). We selected European mutual equity funds with a global geographic scope, focusing on actively managed share-based assets in euros, with continuous five-year performance records, and a minimum investment requirement of up to €10,000, in order to concentrate on retail funds and exclude institutional funds.[1] To ensure the consistency of performance ratios, the sample comprises only capitalization funds; therefore, distribution funds are excluded.

[1] Our research aligns with prior work by Momparler et al. (2025), employing a model that uses the five-year annualized return as the dependent variable alongside annual explanatory variables.

Consequently, the resulting sample comprises 235 funds.

The dataset spans the period from January 1, 2020, to December 31, 2024, capturing a recent and highly relevant phase of the mutual fund market. All data were sourced directly from the Refinitiv Eikon platform, ensuring a high level of consistency and reliability.

The objective of this paper is to identify the fund characteristics that best predict mutual fund performance. The explanatory variables are related to various fund attributes: first, traditional variables based on size, costs, and investment style (FUNDTNA, TER, and STYLEMATRIX); second, variables focused on performance, volatility, manager skill, and market risk (TER, ANNUAL RETURN, ANNUAL STANDARD DEVIATION, ALPHA, and BETA); finally, variables related to mutual fund sustainability (ESGSCORE, ENVSCORE, SOCSCORE, and GOVSCORE).

The variables used in the analysis are defined as follows:

- Fund Return (ANNUAL RETURN): The annual total return of the fund, including both capital appreciation and dividends, calculated on a calendar-year basis.
- ESG Score (ESGSCORE): A composite measure reflecting the fund's overall performance on Environmental, Social, and Governance criteria, based on peer-relative rankings.
- Environmental Score (ENVSCORE): The fund's score on environmental factors, such as emissions, resource use, and environmental management practices.
- Social Score (SOCSCORE): The fund's score on social factors, including community impact, human rights, product responsibility, and workforce practices.
- Governance Score (GOVSCORE): The fund's score on governance factors, such as board structure, transparency, shareholder rights, and executive compensation.
- Total Expense Ratio (TER): The annual cost of managing the fund, expressed as a percentage of assets under management, covering management fees and other operating expenses.
- Fund Total Net Assets (FUNDTNA): The total net value of the fund's assets, calculated as total assets minus total liabilities, measured in thousands of euros.
- Five-Year Annualized Standard Deviation of Returns (ANNUALSD5Y): A measure of the fund's return volatility over the past five years, indicating the degree of variability in performance.
- Alpha (ALPHA): A risk-adjusted performance metric showing the difference between the fund's actual returns and its expected returns, given its level of market risk (beta).
- Beta (BETA): A measure of the fund's sensitivity to market movements, indicating how much the fund's returns move in relation to the overall market.
- Style Matrix (STYLEMATRIX): A classification of the fund's investment style and size, based on the Lipper Style Box, which categorizes funds by market capitalization and investment approach (e.g., value, growth, or core).

These variables were selected to capture a comprehensive view of each fund's financial characteristics, risk profile, and sustainability orientation. By defining and including these variables, the analysis aims to provide clear insights into the factors that most strongly predict mutual fund performance in the current market environment.

4. Methodology

We aim to develop a robust regression prediction model using financial variables, a common approach in quantitative finance research. This model should be able to pinpoint the most crucial features that can forecast fund performance, a goal often pursued in portfolio management and investment analysis (Harvey et al., 2016). Our primary goal is to identify features that contribute to a fund's high performance, a key objective in performance attribution, while ensuring our model doesn't overfit and can generalize its predictions effectively, critical aspects of model validation and robustness (Hastie et al., 2009). To accomplish this, we will apply the Extreme Gradient Boosting (XGBoost) algorithm, a powerful technique within the gradient boosting family and a popular choice in machine learning applications for its efficiency and performance (Chen & Guestrin, 2016). Gradient boosting is especially well-suited for regression tasks and is known to produce highly predictive models. It achieves this by adaptively combining numerous simple tree models, ultimately enhancing predictive performance, a strategy rooted in ensemble learning principles. Boosting methods, like XGBoost, are highly effective in addressing regression problems and can significantly improve predictions for various regression techniques while simultaneously reducing the variance of the statistical learning method. Essentially, boosting is a sequential, ensemble-based approach that averages many tree models, leading to highly accurate predictions, a characteristic that has led to its wide adoption in various domains (Natekin & Knoll, 2013).

Boosting's strength lies in its ability to create multiple trees and combine their predictions into a single, enhanced prediction that significantly improves regression accuracy. Chen and Guestrin (2016) highlight XGBoost as a powerful and efficient implementation of the gradient boosting framework, building upon the foundational work of Friedman (2002) and Friedman (2001).

The process of building these trees is sequential. Each new tree is trained on a modified version of the original dataset, learning from the errors of its predecessors. This implies that the construction of each tree is heavily dependent on the trees that came before (James et al., 2017). Fundamentally, XGBoost iteratively adds new trees to the model. Each tree is designed to correct the residuals (the errors) left by the previous trees, as clarified by James et al. (2017). This process continues until no further improvement is possible. The power of XGBoost lies in its ability to sequentially refine the model, with each new tree building upon the knowledge of its predecessors to create a highly accurate and robust predictive model.

The core principle of gradient boosting, as explained by Friedman (2002), is to create and combine multiple "weak learners"—typically decision trees—to form a "strong rule" that achieves superior predictive accuracy. Essentially, gradient boosting leverages the collective strength of many simple models to mitigate individual weaknesses and improve overall performance, a strategy based on ensemble learning principles (Zhou, 2012). Through iterative fine-tuning, these weak learners are integrated to create a robust model capable of delivering more precise predictions.

XGBoost, a highly effective implementation of gradient boosting (Chen & Guestrin, 2016), employs an ensemble of classification or regression trees. By aggregating the predictions of multiple trees, it overcomes the limitations of single trees that often suffer from lower predictive power. While previous research has utilized XGBoost with classification trees to forecast bank failures (Carmona et al., 2019; Climent et al., 2019), this study distinguishes itself by employing XGBoost based on regression trees, showing its versatility in addressing diverse predictive tasks.

A distinguishing feature of XGBoost models is their incorporation of regularization, which mitigates overfitting compared to other boosting techniques and enhances the model's ability to generalize to new data. Regularization, also referred to as feature penalization, helps control variable weights or penalizes complexity, effectively performing variable selection and reducing the challenges associated with high-dimensional data (Bühlmann & Van De Geer, 2011). XGBoost integrates regularization directly into its learning objective, a key distinction from simpler gradient boosting and random forests (Chen & Guestrin, 2016).

Furthermore, XGBoost operates by sequentially adding predictors to an ensemble, with each new predictor aiming to correct the errors of its predecessor (Friedman, 2001). However, unlike AdaBoost, which adjusts instance weights at each iteration, XGBoost focuses on fitting the new predictor to the residual errors made by the previous one. This approach contributes to the algorithm's efficiency and predictive power.

To achieve the best possible XGBoost model for a given dataset, it is essential to identify the optimal values for its hyperparameters, including the ideal number of decision trees (Probst et al., 2019). This requires an iterative process of exploring various combinations of hyperparameter values to determine the configuration that yields the highest predictive performance (Bergstra & Bengio, 2012).

In machine learning, hyperparameters are adjustable settings that control the learning process of an algorithm and influence the performance of the resulting model (Claesen & De Moor, 2015). Unlike model parameters, which are learned directly from the data during training, hyperparameters are set prior to training and remain fixed throughout the learning process. They can significantly impact the model's ability to find patterns, generalize to new data, and ultimately achieve its objective (Feurer & Hutter, 2019, pp. 3–33). Examples of hyperparameters in XGBoost include the learning rate in gradient descent, the number of trees, or the regularization.[2] Selecting appropriate hyperparameters is crucial for optimizing model performance and avoiding issues like overfitting or underfitting.

Estimating the optimal values for hyperparameters directly from the data is not feasible. These values must be determined either through the practitioner's expertise or by utilizing tuning techniques like cross-validation. Hyperparameter tuning involves identifying the combination of hyperparameters that results in a model with the highest performance on unseen data (Feurer & Hutter, 2019, pp. 3–33). A common approach to discovering this optimal combination is to create a grid, which is essentially a list or range of suitable values for each hyperparameter. Subsequently, all possible models resulting from these parameter combinations are trained. However, when the number of parameter configurations is extensive, the training process can become computationally expensive and time-consuming. In such cases, practitioners may choose to terminate the process once a satisfactory level of performance is achieved (Yang & Shami, 2020).

To fine-tune an XGBoost regression model with an optimal set of hyperparameters, we employed the machine learning technique known as cross-validation (Probst et al., 2019). Specifically, for each considered hyperparameter combination, we opted for 10-fold cross-validation ($k = 10$). This entailed training the XGBoost algorithm on ten different models, each built on nine distinct subsets of the data (comprising 90 % of the observations). The final goodness-of-fit estimation was then derived by averaging the results obtained on the remaining 10 % of each of the ten subsets.

Upon completing this cross-validation process for every hyperparameter combination, an additional model was trained using 100 % of the observations, serving as the final model for that specific combination. Consequently, for each combination, a total of 11 models ($k+1 = 10 + 1$) were estimated. Internal goodness-of-fit metrics were then calculated for both the final model and the cross-validation models.

These 11 models were then multiplied by the number of hyperparameter combinations under consideration. Given that approximately 20 combinations were explored, the total number of models estimated in the pursuit of the most predictive model reached 220 (11*20).

All models were fitted in *R* (R Core Team, 2024) version 4.4.1, and for XGBoost using *h2o* package version 3.44.0.3 (LeDell et al., 2024) and *tidyverse* package version 2.0.0 (Wickham, 2023).

[2] For a comprehensive understanding of the hyperparameters that influence the learning process of the XGBoost algorithm, readers are encouraged to refer to the following URL (accessed in October 2024): https://xgboost.readthedocs.io/en/stable/parameter.html.

5. Results

5.1. Correlational analysis

Fig. 3 shows a correlogram plot of all variables, where the small dot sizes and lack of intense color suggest the absence of strong correlations. A correlogram plot is a graphical representation of the correlation matrix, with each dot's color and size corresponding to the magnitude of the respective correlation coefficient. Preliminary analysis reveals the presence of correlations among the ESG-related variables. However, this does not denote a concern for the XGBoost algorithm, as it is inherently robust to multicollinearity (Chen & Guestrin, 2016), unlike traditional linear regression models (Gujarati & Porter, 2009).

Additionally, in Fig. 4, we present the correlations (and 95 % confidence intervals) between all the predictors or variables and the return of investment funds (ANNUAL5Y) within the dataset considered. This analysis suggests that ANNUAL5Y has a relatively positive correlation with some variables (ALPHA5Y, FUNDTNA, and SOCSCORE). On the other hand, there is a certain negative correlation with the variable TER.

The positive correlations with ALPHA5Y, FUNDTNA, and SOCSCORE indicate that as these variables increase, the five-year annualized return of the investment funds (ANNUAL5Y) also tends to increase:

- Higher values of ALPHA5Y (a measure of risk-adjusted performance) might be associated with higher investment fund returns. This could imply that funds with better risk-adjusted performance tend to generate higher returns.
- Larger FUNDTNA (fund total net assets) could be linked to higher returns. This might suggest that larger funds have advantages in generating returns, possibly due to economies of scale or access to a wider range of investment opportunities.
- Higher SOCSCORE (a measure of social responsibility) could be positively related to returns. This might indicate that investors are increasingly valuing social responsibility and that funds with strong social responsibility practices may attract more investment and generate higher returns.

The negative correlation with TER implies that as the value of TER increases, ANNUAL5Y tends to decrease. The cost of managing a fund has a negative relation with its returns. The TER represents the annual operating expenses of a fund, including management fees,

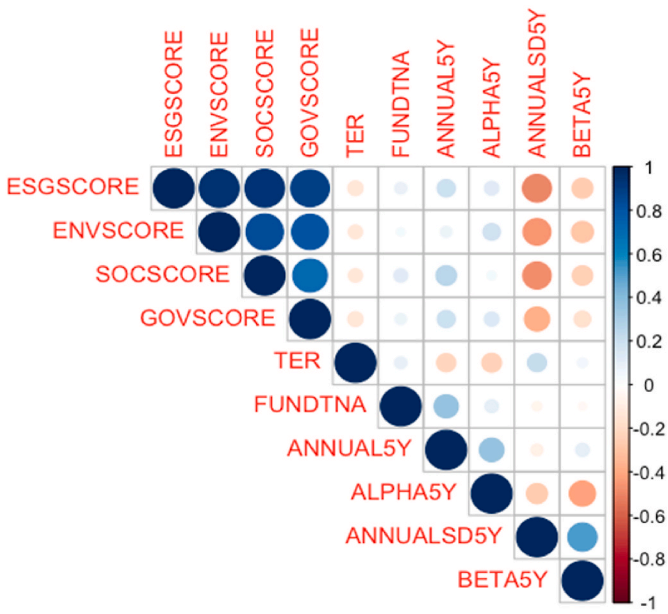

Fig. 3. Correlogram plot of all predictors.

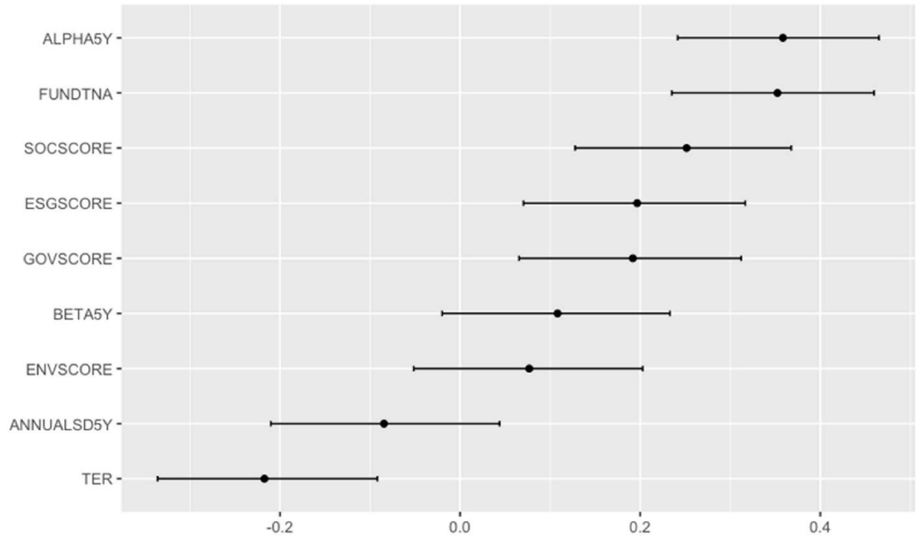

Fig. 4. Correlations (and 95 % confidence intervals) between all independent variables and return of investment funds.

administrative costs, and other expenses, and these expenses directly reduce the returns generated by the fund.

These results are also reflected in Fig. 5, which shows the ten independent variables with the highest correlation to the dependent variable, return of the investment funds—red indicates negative correlations and blue indicates positive ones.

5.2. XGBoost model and performance

We tune and train a model incorporating all variables specified in section 3. The aim is to identify the most influential variables that offer the strongest predictive power in the context of investment funds in the European context. This process entails determining optimal values for the aforementioned hyperparameters to generate an effective XGBoost model. To assess the model's performance on an independent sample, we employ a widely accepted machine learning technique, cross-validation, specifically 10-fold cross-validation (James et al., 2017). Consequently, we will report performance results based on the outcomes obtained during this cross-validation process.

The training sample contains all observations from the database, and the error is calculated based on this data. While these results tend to be overly optimistic and should not be taken as definitive, they are essential for initial model development. The cross-validation samples, on the other hand, provide a more realistic estimate of how the model might perform on unseen data. As such, the cross-validation results are more conservative and should be prioritized when evaluating the model's overall performance and generalizability.

To identify a good combination of hyperparameters that provide a very robust model with high predictive capacity, we have applied what is known as random search within the space of all possible combinations of hyperparameters; that is, among all these possible arrangements of the search space, 20 parameter combinations have been considered to

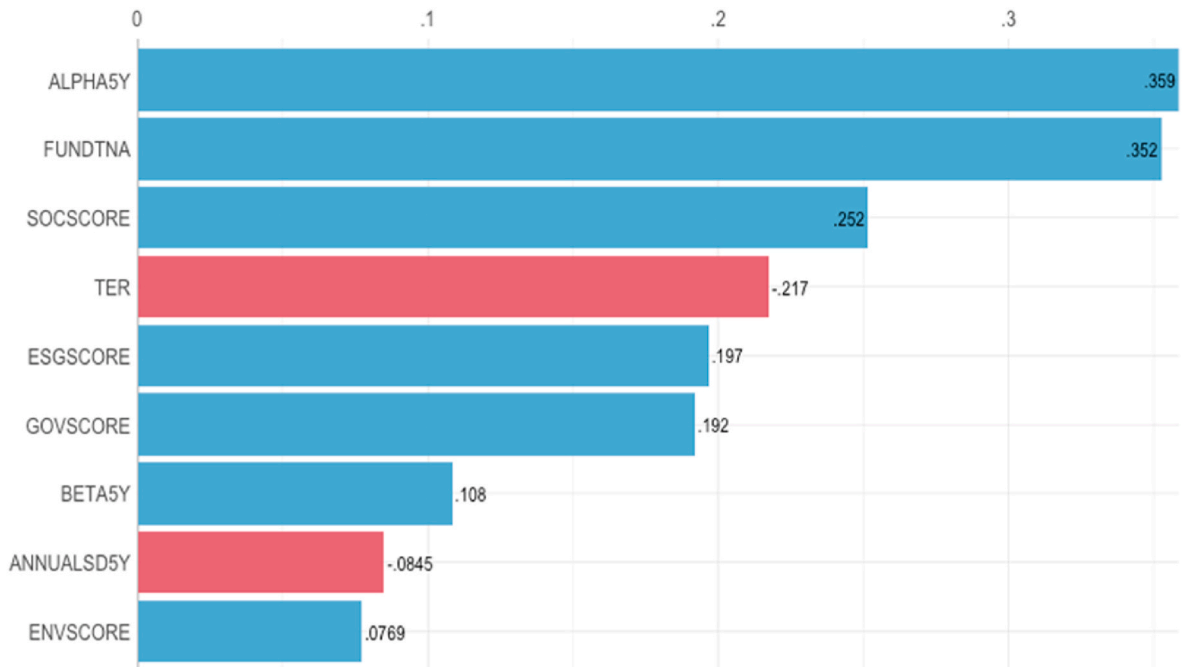

Fig. 5. Identification of the ten independent variables with the highest correlation to return of investment funds.

build 20 primary XGBoost models, along with their respective cross-validation models.

The final model selected was the one with the best RMSE indicator throughout the hyperparameter optimization process. Its value on cross-validation with k = 10 was 2.57, and on the total training test sample, it was 0.24. These low values prove that the model is well fitted.

Fig. 6 visually depicts the model fitting procedure, where it is observed that the optimal value for the XGBoost model, in terms of achieving the lowest possible RMSE (root mean squared error) in the cross-validation samples (orange line), is obtained with approximately 42 decision trees. Beyond this number of decision trees, a slight decrease of this indicator is observed, although it improves in the training sample. Furthermore, as expected, the cross-validation results are worse than those of the training sample. Consequently, the ideal number for model fitting is about 44 decision trees, and in accordance with the observations, there is no apparent problem of "over-fitting," and the model could be applied without issues to independent samples.

Fig. 7 suggests a lack of correlation between the cross-validation residuals and fitted or predicted values of the best identified model—-which is a desirable characteristic in a regression model assuming normally distributed errors. The zero slope of the fitted line further reinforces this observation, indicating no dependency between these two components. This evidence supports the adequacy of the final XGBoost regression model and the absence of major concerns regarding its fit.

Additionally, Fig. 7 reveals that most of the residuals fall within a narrow range of −2.50 to 2.50, suggesting a close alignment between the model's predictions and the actual observed values. This observation strongly supports the conclusion that the final fitted XGBoost model demonstrates a high degree of accuracy and predictive capability.

Furthermore, Fig. 8 displays the cross-validation predictions against the actual returns of the investment funds. The resulting fitted line closely resembles a diagonal, indicating a high degree of agreement between predicted and actual values. This visual further supports the model's strong predictive performance.

To gain further insight into the model's residuals, Fig. 9 presents a box plot illustrating the distribution of cross-validation residuals. Box plots provide a graphical summary of data distribution, highlighting the maximum and minimum values while emphasizing the median to reveal central tendencies. The red dot indicates the mean of the square root of the sum of all residuals (RMSE) from the fitted model. As shown in the diagram, the RMSE is approximately 2.5, while the median is about 1.

Finally, Fig. 10, depicting the reverse cumulative distribution of cross-validation residuals, provides detailed insight into the magnitude of errors in the XGBoost model.

5.3. XGBoost variable importance and model interpretability

To better understand how the model's predictions are influenced by various factors, it's helpful to analyze the overall relationships between the predictor variables and the predicted outcome. This global interpretation can offer valuable insights into the model's behavior and the factors driving its predictions.

To identify which factors are most influential in the model's predictions, we utilize a permutation-based feature importance method (Fisher et al., 2019). This technique involves randomly shuffling the values of each predictor variable and measuring the resulting decrease in model accuracy. By disrupting the relationship between a predictor and the outcome, we can assess its importance: a larger drop in accuracy indicates a more influential predictor. This approach allows us to identify the most important features driving the model's predictions.

Fig. 10 illustrates the relative importance of each independent variable in predicting the dependent variable (investment fund returns).

This figure ranks the relative importance of different variables in predicting annual returns of mutual funds within the XGBoost model:

- ALPHA5Y: This metric, essentially a measure of a fund manager's skill in exceeding expected returns given the risk taken, is the most important predictor. A higher Alpha indicates the fund has outperformed expectations, suggesting that investor confidence in skilled management significantly drives fund value.
- SOCSCORE: As the second most important factor, this highlights the growing relevance of social responsibility in investment decisions. A higher SOCSCORE, reflecting positive performance across various social factors, such as product responsibility and human rights, appears to contribute substantially to a fund's perceived value. This could indicate that investors are increasingly seeking funds that align with their ethical values.
- ANNUAL5DSY: This variable, which measures the volatility of a fund's returns over a five-year period, ranks third. Its prominence suggests that investors may be sensitive to consistency and stability in returns. A higher standard deviation implies greater variability; this might suggest that funds with less stable and predictable returns are likely to yield higher returns.
- FUNDTNA: Representing the fund's net asset value, this factor's importance is likely tied to the concept of scale and resources. Larger funds may have access to a wider range of investment opportunities

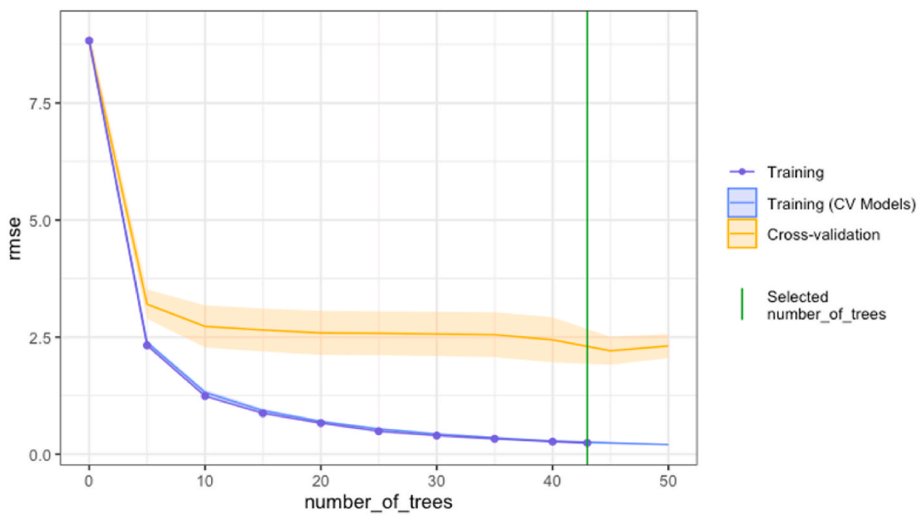

Fig. 6. Learning curve of XGBoost model.

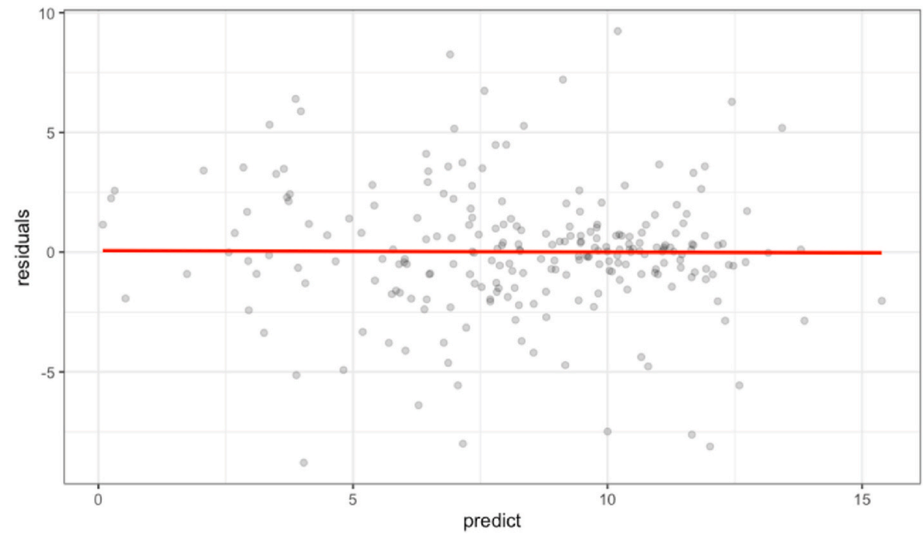

Fig. 7. Residual analysis of final XGBoost model on cross-validation data.

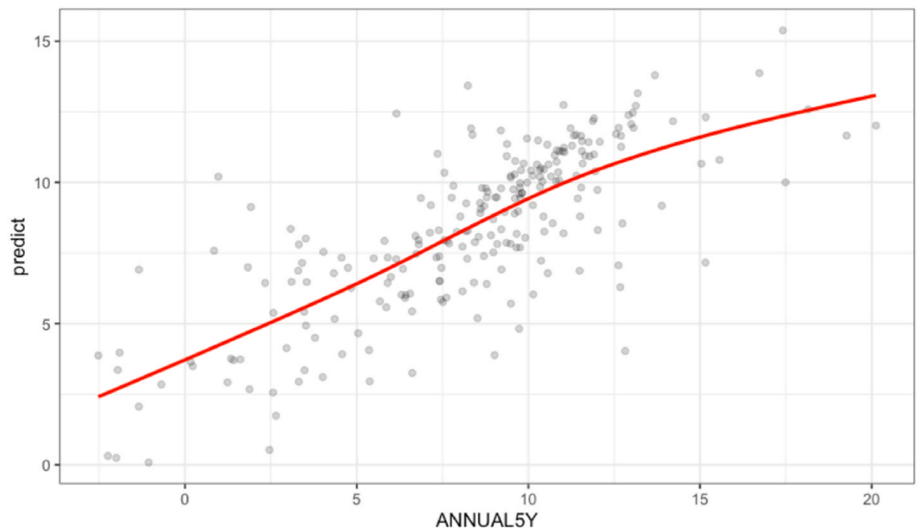

Fig. 8. Cross-validation predictions vs actual data (XGBoost).

or be able to negotiate better terms, contributing to their overall value.

- ENVSCORE: The position of this variable suggests that while it's not the most dominant predictor, it still plays a noticeable role in the model's ability to predict mutual fund performance. It's interesting to note that ENVSCORE is less important than SOCSCORE (Social score) and slightly more important than GOVSCORE (Governance score). This might indicate that, for the fitted model, social factors are considered to be slightly more important than environmental factors when assessing the overall value of investment funds.

It is noteworthy that both SOCSCORE (Pillar Social score) and ENVSCORE (Pillar Environmental score) are among the top five most important variables in predicting mutual fund returns, while ESGSCORE (the overall ESG score) falls outside this top tier. This observation suggests a potential divergence between the perceived importance of individual ESG components and the overall ESG score. Investors may be prioritizing specific aspects of ESG performance, particularly social and environmental factors, rather than relying solely on the aggregate ESG score. This could reflect a shift in investor focus towards specific elements of sustainability, where they carefully consider individual social and environmental impacts instead of relying on a general ESG assessment.

To further explain the inner workings of our model, we apply partial dependence plots, a powerful visualization technique that reveals how specific factors influence the model's predictions (Molnar, 2020). These plots illustrate the relationship between a particular predictor variable and the model's predicted outcome, while accounting for the average effects of all other variables. While not a perfect representation, partial dependence plots effectively highlight overall trends and provide a valuable foundation for interpretation (Greenwell, 2017). By isolating the marginal effect of each predictor, these plots help us understand how different factors contribute to the model's predictions.

Therefore, we will utilize partial dependence plots derived from the XGBoost model to present and analyze the relationship between investment fund performance and each of the most relevant variables. The x-axis of these plots represents the values of the respective variables, while the y-axis shows the corresponding investment fund performance.

Fig. 11 shows the relationship between the ALPHA5Y and the predicted investment fund performance according to the XGBoost model. There is a generally positive relationship between ALPHA5Y and fund performance. As ALPHA5Y increases, the predicted performance tends

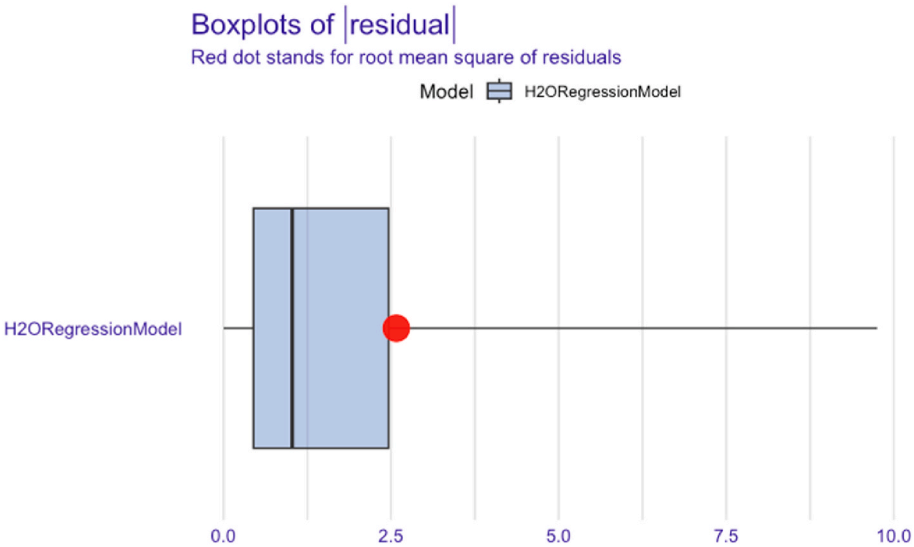

Fig. 9. Distribution of cross-validation residuals (XGBoost).

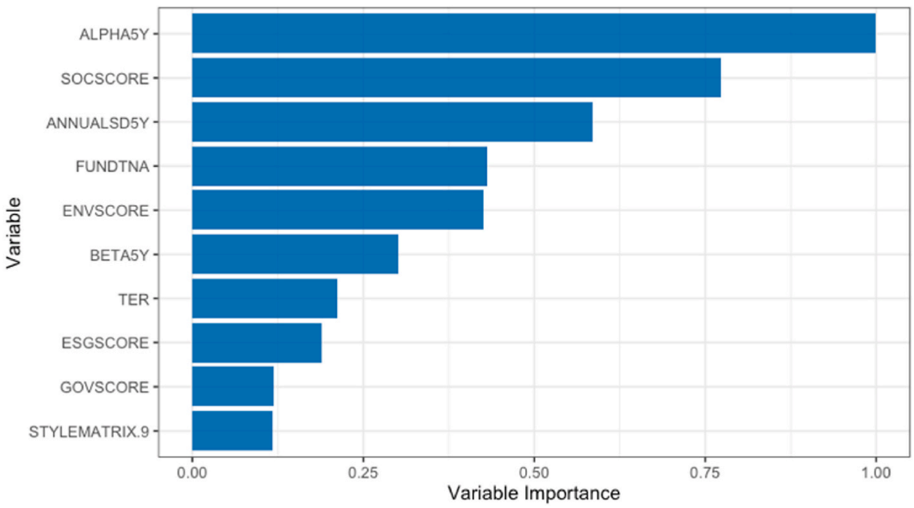

Fig. 10. Variable importance according to the XGBoost model.

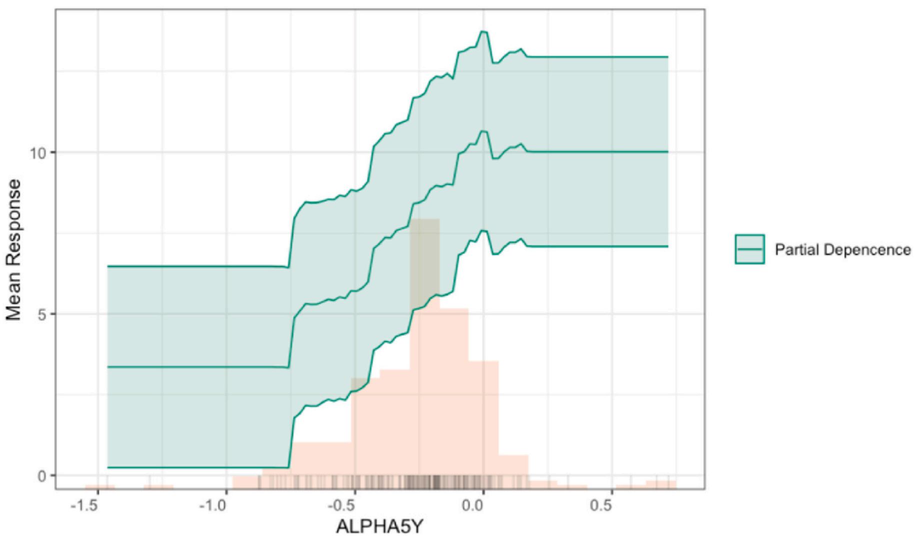

Fig. 11. Partial dependence plot of variable ALPHA5Y.

to increase as well. This aligns with the expectation that funds with better risk-adjusted returns (higher alpha) are likely to have better overall performance. The histogram at the bottom shows the distribution of ALPHA5Y values in the dataset. Most of the data is concentrated around the −0.5 to 0 range, indicating that the model has more information and is likely more reliable in that region. Improving risk-adjusted returns has a particularly strong impact on performance in the mid-range of ALPHA5Y values.

The partial dependence plot of Fig. 12 shows the relationship between SOCSCORE (ESG Pillar Social score) and the predicted investment fund performance. The plot demonstrates a positive relationship between SOCSCORE and fund performance. As the social responsibility score increases, the predicted performance of the fund tends to improve. This suggests that funds with better social responsibility practices are generally associated with higher performance. The relationship is not strictly linear. There's a noticeable "step" around a SOCSCORE of 60, where the impact of further increases in the score appears to become more pronounced. This implies that there might be a threshold beyond which social responsibility has a stronger influence on performance; it seems investors are particularly sensitive to social responsibility factors when the score is already relatively high, potentially viewing it as a sign of strong commitment and positive impact. The histogram at the bottom shows the distribution of SOCSCORE values in our data. The data appears to be relatively evenly distributed across the range, with a slight concentration around the middle.

Therefore, this relationship highlights the growing importance of social responsibility factors in investment decisions, suggesting that investors are increasingly incorporating these considerations into their evaluations of fund performance. This analysis provides further evidence that social responsibility is not just an ethical consideration but also a potential driver of financial performance in the investment fund landscape.

The partial dependence plot of Fig. 13 explains the relationship between ANNUAL5DSY (five-year annualized standard deviation of returns) and the predicted investment fund performance, as determined by the XGBoost model. The plot reveals a generally positive relationship between ANNUAL5DSY and fund performance. As the volatility of returns increases, the predicted performance tends to increase as well. This suggests that funds with less stable and predictable returns are likely to yield higher returns. There's a noticeable "step" around an ANNUAL5DSY value of 20, where the positive impact of volatility appears to become more pronounced. This could indicate a threshold beyond which volatility has a stronger positive effect on performance. The histogram at the bottom shows the distribution of ANNUAL5DSY values in the data. The data appears to be concentrated mostly between 15 and 25, with a few outliers beyond that range.

Therefore, this plot confirms that ANNUAL5DSY is a significant predictor of fund performance, with higher volatility generally associated with higher performance. The positive relationship highlights a degree of risk propensity among investors, who tend to favor funds with less stable and predictable returns. This analysis underscores the importance of considering volatility when evaluating investment fund performance.

The partial dependence plot of Fig. 14 shows the relationship between FUNDTNA (fund total net assets) and the predicted investment fund performance. The plot indicates a generally positive relationship between FUNDTNA and fund performance. As the fund's total net assets increase, the predicted performance tends to improve. This suggests that larger funds, in terms of assets, are generally associated with better performance (figures in thousands of euros):

- Small Funds (FUNDTNA <1000): In this range, increasing the fund's net assets has a relatively small positive impact on predicted performance.
- Mid-Sized Funds (1000 < FUNDTNA <5000): This region shows a steeper increase in performance as FUNDTNA increases, suggesting that growth in assets has a more pronounced positive effect for funds in this size range.
- Large Funds (FUNDTNA >5000): Beyond 5000, the impact of further increases in FUNDTNA on performance levels off, indicating diminishing returns to scale.

The histogram at the bottom shows the distribution of FUNDTNA values in the data. The data appears to be heavily skewed to the right, with most funds having relatively low net assets and a few large funds with significantly higher assets.

Hence, this plot confirms that FUNDTNA is a significant predictor of fund performance, with larger funds generally exhibiting better performance. However, this effect is not unlimited, and there appear to be diminishing returns to scale beyond a certain size. Larger funds may have advantages such as greater diversification, access to a wider range of investment opportunities, and lower operating costs, which could contribute to their better performance.

The partial dependence plot of Fig. 15 illustrates the relationship between ENVSCORE (ESG Pillar Environmental score) and the predicted investment fund performance:

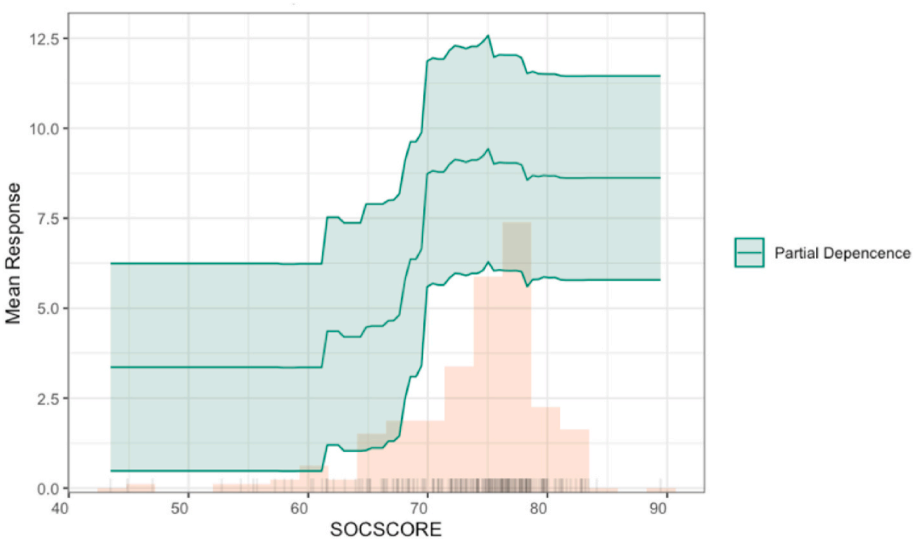

Fig. 12. Partial dependence plot of variable SOCSCORE.

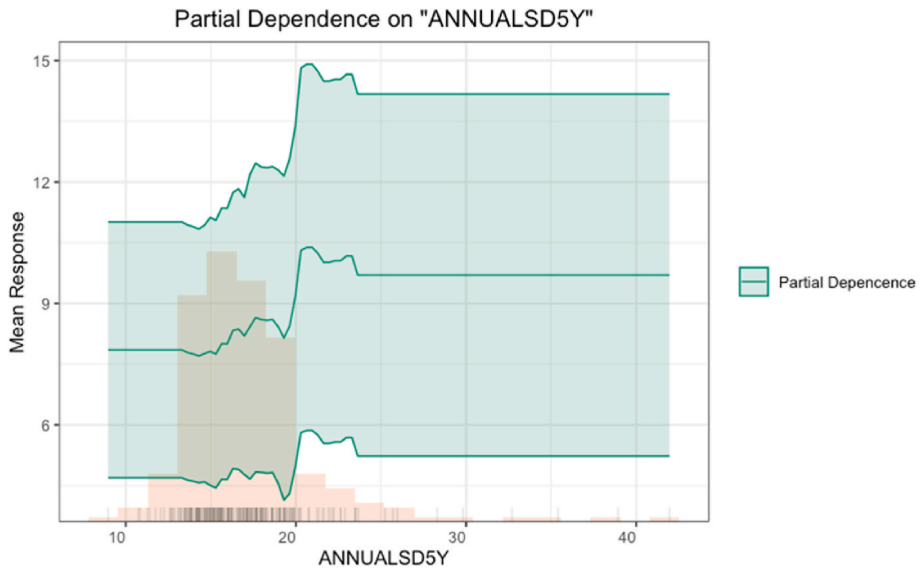

Fig. 13. Partial dependence plot of variable ANNUALSD5Y.

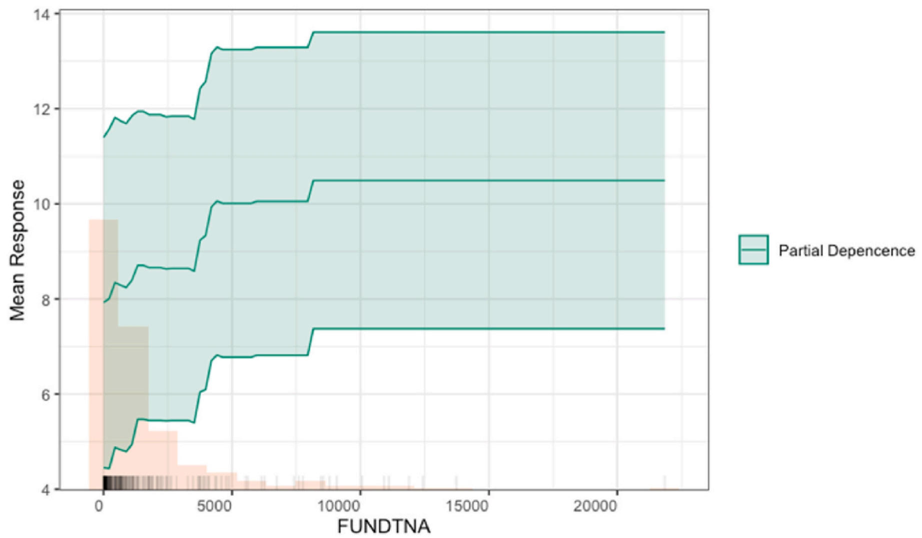

Fig. 14. Partial dependence plot of variable FUNDTNA.

- The plot shows a generally negative relationship between ENVSCORE and fund performance, particularly in the mid-range of ENVSCORE values. As the environmental score improves, the predicted performance tends to decrease, suggesting that funds with better environmental practices are, counterintuitively, associated with lower performance.
- The relationship is non-linear, with a noticeable “step” around an ENVSCORE of 65. This indicates that the impact of environmental performance on predicted performance isn’t uniform across its entire range.
- Lower ENVSCORE (below 55): In this range, improvements in the environmental score have a relatively small negative impact on predicted performance.
- Mid-range ENVSCORE (55–70): This region shows the steepest decrease in performance as ENVSCORE increases, suggesting that environmental performance is a more significant driver of lower predicted performance in this range.
- Higher ENVSCORE (above 70): The impact of further improvements in the environmental score levels off, indicating diminishing negative returns to performance beyond a certain point.
- The histogram at the bottom shows the distribution of ENVSCORE values in the data. The data appears to be concentrated mostly between 50 and 80, with a few outliers on either side.

The dataset might include a proportion of funds invested in industries with inherently high environmental impacts, such as fossil fuels, mining, or heavy manufacturing. In these sectors, companies with lower ENVSCOREs might actually be more profitable due to lower environmental compliance costs or less stringent regulations. Some funds might be specifically focused on prioritizing fossil fuel energy sources or less environmentally friendly practices. These funds might exhibit higher returns in the short term, even though their long-term sustainability is questionable.

While understanding the overall behavior of a model is important, it’s equally decisive to explain how the model arrives at predictions for individual instances. This is known as local interpretability. Instead of focusing on average effects, local explanations explore the specific factors driving individual predictions (Molnar, 2020). SHAP (SHapley Additive exPlanations) plots provide this granular level of interpretation. By showing the contribution of each feature to a specific prediction,

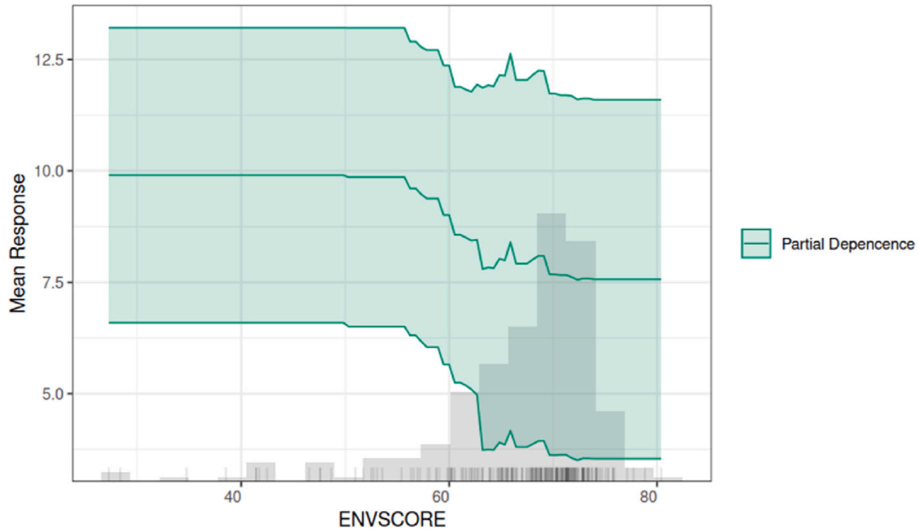

Fig. 15. Partial dependence plot of variable ENVSCORE.

SHAP plots offer a clear and intuitive explanation of the model's decision-making process for that instance.

SHAP plots offer a powerful approach to interpreting machine learning models. Based on game theory, SHAP values fairly distribute the contribution of each feature to a prediction, providing both local and global interpretability (Lundberg & Lee, 2017). These plots help visualize how individual features influence a specific prediction by showing the magnitude and direction of their impact. By examining SHAP plots, we can gain valuable insights into the model's decision-making process and understand which features are driving predictions.

To demonstrate the versatility of SHAP plots, we examine their application in understanding two contrasting predictions from our dataset: one with a low predicted value (Fig. 16) and another with a high predicted value (Fig. 17). The first figure shows a prediction of 0.20, while the second shows a much higher prediction of 17.67. The range of the predicted returns of the mutual funds extends from −2.31 to 20.50. This analysis demonstrates how SHAP plots can reveal the distinct combinations of factors driving different outcomes. These two SHAP plots illustrate how the XGBoost model arrives at different predictions for two distinct investment funds (number 22 and number 1 in our dataset):

- Although ALPHA5Y is the most influential factor in both figures, its impact is substantially negative in the first one. The fund in the second SHAP plot has a considerably higher ALPHA5Y value, which strongly contributes to its higher predicted value. This underscores the consistent importance of this factor in driving higher predicted values across different funds.
- In the first SHAP plot, SOCSCORE has a negative impact on the prediction. The SOCSCORE value in this plot is lower than in the second one, where this variable has an important positive impact.
- ANNUALSD5Y (volatility) has a higher value in the prediction of the observation that shows high fund performance.
- FUNDTNA (fund size) has a more substantial and positive impact in the second SHAP plot. Its value is considerably higher than in the first plot.

In summary, the predictions are different because the two funds have distinct characteristics. The fund with the higher prediction has a much higher ALPHA5Y value, and higher values of the variables SOCSCORE, ANNUALSD5Y, and FUNDTNA. These differences in input features lead to different SHAP values and ultimately different predictions.

These SHAP plots highlight the model's ability to capture the nuances of individual funds. Even though ALPHA5Y is a dominant factor in both cases, the model considers the unique combination of features for each fund to arrive at distinct predictions. This demonstrates the model's capacity for personalized assessment and its ability to provide insights into the specific factors driving the value of each investment

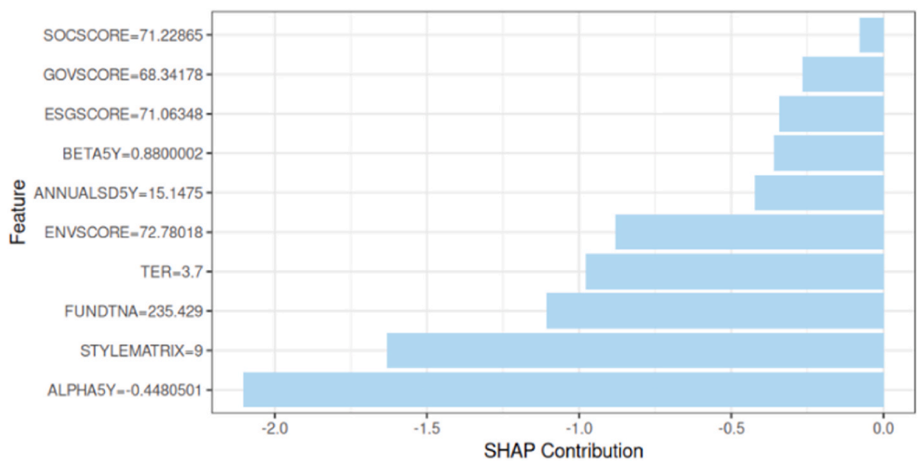

Fig. 16. SHAP explanation for XGBoost model for a fund with low return prediction (number 22 of our dataset, prediction 0.20).

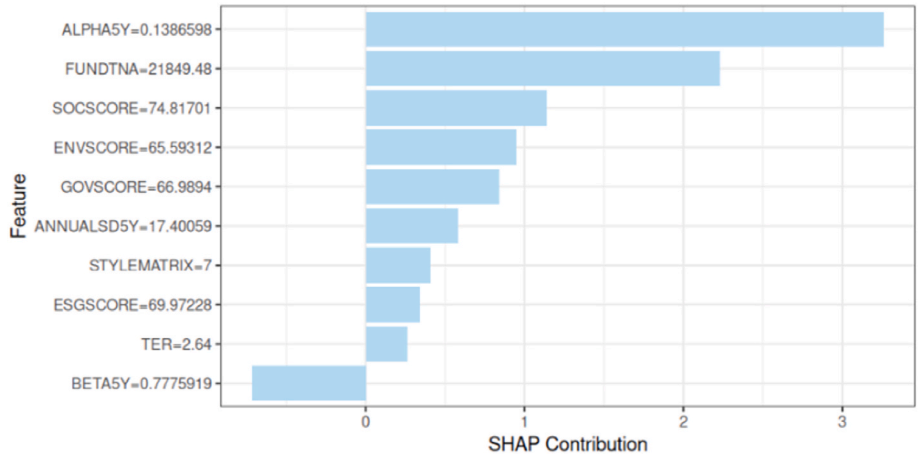

Fig. 17. SHAP explanation for XGBoost model for a fund with high return prediction (number 1 of our dataset, prediction 17.67).

fund.

5.4. Comparing XGBoost with two other regression models

Following the development of a predictive model using the XGBoost algorithm, with robust performance predicting mutual fund returns, we now turn to a comparative evaluation. This evaluation involves two distinct methodologies: a classical OLS regression model and a modern deep learning model. The former is a well-established statistical technique commonly applied to examine linear relationships, whereas the latter is an advanced machine learning approach known for its ability to learn complex data patterns. By comparing these models, we seek to obtain insights into the relative strengths and limitations of each method in the context of mutual fund returns.

OLS regression, also known as linear regression, is a traditional statistical method that estimates the linear relationship between one or more independent variables and a dependent variable. Its primary objective is to minimize the sum of squared residuals—that is, the differences between observed and predicted values. The OLS model is expressed mathematically as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

where:

- Y is the dependent variable
- β_0 is the intercept
- $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients associated with the independent variables $X_1, X_2, \ldots, X_p$
- ε represents the error term.

The estimation process involves determining the values of the coefficients (β) that minimize the sum of the squared residuals. Once estimated, the model can be used to predict the outcome for new observations based on their independent variable values.

In contrast, deep learning—a subset of machine learning—employs artificial neural networks with multiple layers to identify complex data patterns and hierarchical representations (Heaton et al., 2018). Unlike conventional models, deep learning automates feature extraction, minimizing reliance on explicit feature engineering (LeCun et al., 2015).

The core component of deep learning is the artificial neuron, which processes inputs by applying learnable weights and biases and generates output via an activation function. These neurons are structured into interconnected layers, creating a layered architecture (Fig. 18). The input layer ingests raw data, hidden layers iteratively refine the information through nonlinear transformations, and the output layer delivers the final prediction (e.g., mutual fund returns). This layered design enables the model to capture intricate relationships within financial datasets, such as nonlinear interactions between different independent variables.

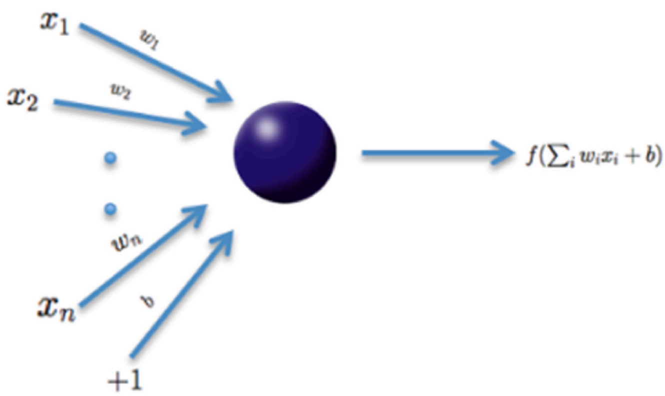

Fig. 18. Neuron from a hidden layer.
Source: Candel and Ledell (2018).

Deep learning models, including the multi-layer feedforward neural networks illustrated in Fig. 19, rely on extensive training data and backpropagation algorithms to iteratively update network parameters (weights and biases). These adjustments aim to optimize model accuracy by reducing discrepancies between predicted and observed outcomes (Heaton et al., 2018). In such architectures, each input connection to a neuron is assigned a learnable weight (denoted as w1,w2, …,wn), which is calibrated during training to enhance predictive performance.

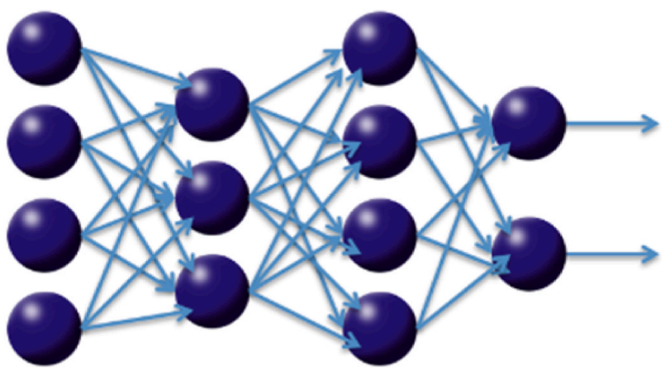

Fig. 19. Multi-layer feedforward neural networks.
Source: Candel and Ledell (2018)

Signal transmission between layers is governed by activation functions (Candel & Ledell, 2018), which apply nonlinear transformations to inputs, enabling the network to model complex financial relationships—such as interactions between market volatility, leverage ratios, and operational efficiency. The input layer consumes raw financial data, hidden layers progressively refine feature representations, and the output layer generates predictions (e.g., mutual fund returns).

This iterative optimization process allows deep learning models to autonomously identify nonlinear associations and hierarchical patterns within datasets, surpassing the capabilities of linear methods like OLS regression. For detailed examinations of neural network architectures, training methodologies, and practical applications in financial analytics, see Romero Martínez et al. (2021).

Deep learning models offer distinct advantages in predicting mutual fund returns due to their ability to process high-dimensional datasets and model nonlinear interactions among variables (Barboza et al., 2017). These capabilities position them as viable alternatives to conventional methods such as OLS regression, which assumes linearity and may overlook complex financial patterns.

For this study, a deep neural network tailored for regression tasks will be implemented. The architecture includes multiple hidden layers with nonlinear activation functions (e.g., ReLU or sigmoid), allowing the model to hierarchically learn intricate relationships within mutual fund data—such as the interplay between the different independent variables. By iteratively refining these representations through backpropagation, the network adapts to capture factors influencing fund performance, thereby improving prediction accuracy compared to linear approaches.

This study evaluates the advantages of the XGBoost algorithm in predicting mutual fund returns by comparing its performance against two benchmark models: OLS regression and a deep learning architecture. Table 2 presents key performance metrics for all three models on the cross-validation data, including the baseline XGBoost results. The results indicate that the XGBoost model outperforms both the OLS regression and deep learning models across all metrics, such as MSE, RMSE, MAE, mean residual deviance, and R2.

Fig. 20 displays the reverse cumulative distribution of absolute prediction errors for the XGBoost, logistic regression, and deep learning models on cross-validation data, which provides detailed insight into the magnitude of errors in the different models. In this visualization, a lower area under the curve indicates better predictive accuracy. The XGBoost model demonstrates the smallest area, reflecting its superior performance in minimizing prediction errors. This result implies that XGBoost captures the relationships between financial predictors and mutual fund returns more effectively than the other models. Notably, XGBoost residuals exceeding 5 account for less than 10 % of the total, and this indicates that the model exhibits a low frequency of large prediction errors.

To sum up, the comparative evaluation of XGBoost, OLS regression, and deep learning models highlights XGBoost's supremacy in mutual fund returns prediction. It achieves better performance metrics values than both traditional and deep learning approaches. Furthermore, its residual distribution profile confirms stronger alignment with observed outcomes, indicating a robust capacity to model factors driving mutual fund returns. These findings demonstrate the practical value of ensemble learning frameworks like XGBoost for financial risk assessment, particularly in contexts requiring high-dimensional data analysis and nonlinear pattern recognition.

Table 2
Performance metrics of the three fitted regression models on cross-validation data.

Model	MSE	RMSE	MAE	Mean Residual Deviance	R2
XGBoost	6.648	2.578	1.769	6.648	0.594
OLS Regression	8.094	2.845	2.081	8.094	0.506
Deep Learning	9.459	3.075	2.080	9.458	0.423

5.5. *Comparative assessment of model performance*

In the "Comparing XGBoost with two other regression models" subsection, our findings both support and extend those obtained by previous researchers using classical OLS regression models. Consistent with the literature, our results confirm that machine learning techniques—particularly XGBoost—offer superior predictive accuracy compared to traditional linear regression approaches when forecasting mutual fund returns (Li & Rossi, 2021; Momparler et al., 2025). Specifically, our XGBoost model achieved a lower RMSE (2.578) and higher R2 (0.594) on cross-validation data than both the OLS regression (RMSE: 2.845, R2: 0.506) and deep learning models (RMSE: 3.075, R2: 0.423), indicating a more robust capacity to capture the complex, nonlinear relationships among fund characteristics and performance.

These findings are congruent with those of Li and Rossi (2021), who demonstrated that boosting algorithms outperform OLS regression in predicting mutual fund performance due to their ability to model intricate interactions among variables. Similarly, Momparler et al. (2025) found that XGBoost not only improved predictive accuracy but also identified ESG scores as significant predictors of fund returns, a result echoed in our analysis. Our results also align with broader evidence from financial machine learning research, which consistently shows that ensemble methods like XGBoost surpass OLS regression in both accuracy and adaptability (see also Kaniel et al., 2023 Nti and Somanathan, 2022).

However, our methodology provides new insights beyond those obtained with classical OLS regression. Notably, while OLS regression assumes linear relationships and may overlook complex interactions, our XGBoost model shows nonlinear and threshold effects in the impact of ESG Pillar scores on fund performance. For example, we find that the Social Pillar score (SOCSCORE) is the second most important predictor and is positively associated with returns, while the Environmental Pillar score (ENVSCORE) ranks fifth and shows a negative, nonlinear relationship with performance. These subtle findings—such as the nonlinear effects of SOCSCORE and ENVSCORE—are not readily captured by linear models and highlight the added value of advanced machine learning approaches in uncovering subtle patterns within the data.

In summary, our results both corroborate and extend the conclusions of prior research using OLS regression. The proposed methodology not only supports the established predictive relevance of ESG factors but also uncovers new, nonlinear dynamics in their relationship with mutual fund performance, demonstrating the practical and analytical advantages of XGBoost over classical regression models.

6. Conclusions

This study examines the predictive power of ESG Pillar scores in forecasting mutual fund performance using Extreme Gradient Boosting (XGBoost) on a sample of Euro-denominated global equity funds (2020–2024). The primary objective is to identify key fund characteristics influencing returns, with a focus on the individual effects of Environmental (ENVSCORE), Social (SOCSCORE), and Governance (GOVSCORE) scores. Results reveal that SOCSCORE emerges as the second most influential predictor of fund returns, showing a positive correlation, while ENVSCORE ranks fifth but with a negative relationship with performance. These findings directly address the research question of whether ESG Pillar scores independently impact financial outcomes, demonstrating that specific sustainability factors hold distinct predictive value beyond aggregate ESG metrics.

The results challenge assumptions about the uniform financial relevance of ESG components. While prior studies often emphasize composite ESG scores, this research highlights the divergent roles of individual Pillars. The positive association of SOCSCORE aligns with growing investor emphasis on social responsibility, whereas the

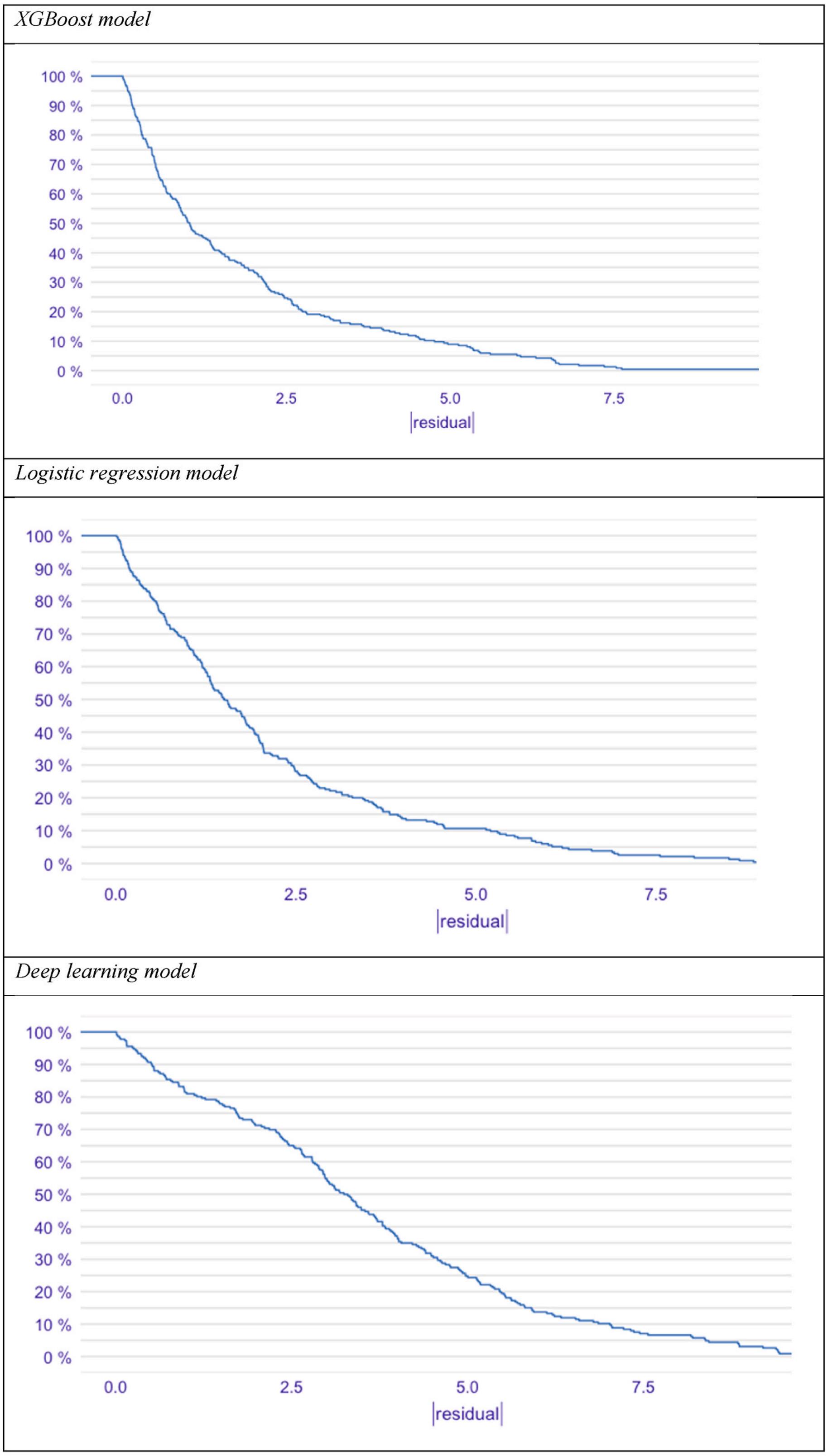

Fig. 20. Reverse cumulative distribution of residuals on cross-validation data.

negative link to ENVSCORE suggests potential trade-offs between environmental commitments and short-term returns, possibly reflecting higher compliance costs or sector-specific dynamics. Methodologically, the study contributes by demonstrating XGBoost's superiority over traditional regression and deep learning models in financial prediction, achieving lower prediction errors (RMSE: 2.57) and higher explanatory power (R^2: 0.594).

Practically, these insights enable values-driven investors to prioritize social criteria without sacrificing returns while cautioning against simplistic reliance on overall ESG ratings. Fund managers can leverage these findings to align portfolio strategies with performance-sensitive sustainability metrics.

Several limitations should be acknowledged, as they may affect the generalizability and interpretation of our findings. First, the use of cross-sectional data offers only an account of the relationships between ESG factors and fund performance within a specific time frame.

Second, our sample is limited to Euro-denominated global equity funds, which may not be representative of other fund categories or geographic regions. Replicating the analysis across different asset classes (e.g., fixed income) and markets (e.g., North America, Asia) would help to determine the robustness of our findings.

Third, our reliance on a single data source (Refinitiv Eikon) may introduce potential biases or limitations related to the specific ESG ratings and financial data available on that platform. Future research could benefit from incorporating data from multiple sources to enhance the reliability and validity of the results.

Fourth, the complexity of the XGBoost model, while providing superior predictive performance, can make it challenging to fully understand the underlying mechanisms driving the observed relationships. While we have used partial dependence plots and SHAP values to interpret the model, further research is needed to explore the features linking ESG factors to fund performance.

Finally, our analysis may be subject to survival bias, as we only included funds with continuous five-year performance records. This may exclude funds that performed poorly and were subsequently liquidated, potentially skewing our results towards better-performing funds.

Future research should expand to longitudinal analyses to assess ESG effects over market cycles and examine sector-specific variations in environmental score impacts. Replication across geographic regions and fund categories (e.g., fixed-income) would test generalizability. Additionally, qualitative investigations into the mechanisms driving the negative ENVSCORperformance relationship could inform ESG integration strategies. Addressing survival bias through alternative sampling methods and exploring nonlinear interactions among ESG Pillars could further refine predictive frameworks.

This work advances sustainable finance literature by dissecting ESG components' heterogeneous impacts and establishes machine learning as a critical tool for modern investment analysis.

Author contribution form

Alexandre Momparler: Conception, Design, Literaturereview, Writerwriter, Criticalreview, Francisco Climent: Conception, Design, Supervision, Criticalreview, Pedro Carmona: Design, Data collection and/or processing, Analysis and/or interpretation, Writer, Criticalreview.

References

Abate, G., Basile, I., & Ferrari, P. (2021). The level of sustainability and mutual fund performance in Europe: An empirical analysis using ESG ratings. *Corporate Social Responsibility and Environmental Management, 28*(5), 1446–1455. https://doi.org/10.1002/csr.2175

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications, 83*, 405–417.

Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of Management, 17*(1), 99–120. https://doi.org/10.1177/014920639101700108

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research, 13*(2).

Broadstock, D., Chan, K., Cheng, L., & Wang, X. (2021). The role of ESG performance during times of financial crisis: Evidence from COVID-19 in China. *Finance Research Letters, 38*, Article 101716. https://doi.org/10.1016/j.frl.2020.101716

Bühlmann, P., & Van De Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Springer Science & Business Media.

Candel, A., & Ledell, E. (2018). Deep learning with H2O. https://h2o-release.s3.amazonaws.com/h2o/rel-wheeler/4/docs-website/h2o-docs/booklets/DeepLearningBooklet.pdf Accessed June 2024.

Carmona, P., Climent, F., & Momparler, A. (2019). Predicting failure in the US banking sector: An extreme gradient boosting approach. *International Review of Economics & Finance, 61*, 304–323.

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

Claesen, M., & De Moor, B. (2015). *Hyperparameter search in machine learning. arXiv preprint arXiv:1502.02127*.

Climent, F., Momparler, A., & Carmona, P. (2019). Anticipating bank distress in the Eurozone: An extreme gradient boosting approach. *Journal of Business Research, 101*, 885–896.

De Miguel, V., Gil-Bazo, J., Nogales, F., & Santos, A. (2023). Machine learning and fund characteristics help to select mutual funds with positive alpha. *Journal of Financial Economics, 150*(3), Article 102737. https://doi.org/10.1016/j.jfineco.2023.103737

Del Vitto, A., Marazzina, D., & Stocco, D. (2023). ESG ratings explainability through machine learning techniques. *Annals of Operations Research*. https://doi.org/10.1007/s10479-023-05514-z

Dolvin, S., Fulkerson, J., & Krukover, A. (2019). Do "Good Guys" finish last? The relationship between morningstar sustainability ratings and mutual fund performance. *Journal of Investing, 28*(2), 77–91. https://doi.org/10.3905/joi.2019.28.2.077

Feurer, M., & Hutter, F. (2019). *Hyperparameter optimization. Automated machine learning: Methods, systems, challenges*.

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research, 20*(177), 1–81.

Freeman, R. E. (1984). *Strategic management: A stakeholder approach*. Pitman.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Journal of Computational Statistics & Data Analysis, 38*(4), 367–378.

Greenwell, B. M. (2017). Pdp: An R package for constructing partial dependence plots. *R J, 9*(1), 421.

Gujarati, D. N., & Porter, D. C. (2009). *Basic econometrics* (5th ed.). New York: McGraw Hill Inc.

Hart, S. L. (1995). A natural-resource-based view of the firm. *Academy of Management Review, 20*(4), 986–1014. https://doi.org/10.5465/amr.1995.9512280033

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

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer.

Heaton, J., Goodfellow, I., Bengio, Y., & Courville, A. (2018). Deep learning. *Genetic Programming and Evolvable Machines, 19*(4), 305–307. https://doi.org/10.1007/s10710-017-9314-z

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning* (2nd ed.). Springer.

Kanade, A., Sherdiwala, M., & Herekar, S. (2022). Performance of machine learning algorithms to predict right mutual funds of Indian companies. In V. K. Gunjan, & J. M. Zurada (Eds.), *Studies in computational intelligence: 1027. Modern approaches in machine learning & cognitive science: A walkthrough*. Springer. https://doi.org/10.1007/978-3-030-96634-8_3.

Kaniel, R., Lin, Z., Pelger, M., & Van Nieuwerburgh, S. (2023). Machine-learning the skill of mutual fund managers. *Journal of Financial Economics, 150*(1), 94–138. https://doi.org/10.1016/j.jfineco.2023.07.004

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

LeDell, E., Gill, N., Aiello, S., et al. (2024). *h2o: R Interface for the 'H2O' Scalable Machine Learning Platform. R package*. Retrieved October 2024, from https://CRAN.R-project.org/package=h2o. version 3.44.0.3.

Li, B., & Rossi, A. G. (2021). *Selecting mutual funds from the stocks they hold: A machine learning approach*. Available at: SSRN 3737667 papers.ssrn.com.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance, 7*(1), 77–91. https://doi.org/10.2307/2975974

Molnar, C. (2020). *Interpretable machine learning*. Lulu.com.

Momparler, A., Carmona, P., & Climent, F. (2025). Catalyzing sustainable investment: Revealing ESG power in predicting fund performance with machine learning. *Computational Economics, 65*, 1617–1642. https://doi.org/10.1007/s10614-024-10618-0

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics, 7*, 21. https://doi.org/10.3389/fnbot.2013.00021

Nti, I. K., & Somanathan, A. R. (2022). A scalable RF-XGBoost framework for financial fraud mitigation. *IEEE Transactions on Computational Social Systems, 11*(2), 1556–1563.

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9*(3), Article e1301.

R Core Team. (2024). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved October 2024, from http://www.R-project.org/.

Raghunandan, A., & Rajgopal, S. (2022). Do ESG funds make stakeholder-friendly investments? *Review of Accounting Studies, 27*, 822–863. https://doi.org/10.1007/s11142-022-09693-1

Romero Martínez, M., Carmona Ibáñez, P., & Pozuelo Campillo, J. (2021). Utilidad del Deep Learning en la predicción del fracaso empresarial en el ámbito europeo. *Revista de Métodos Cuantitativos para la Economía y la Empresa, 32*, 392–414. https://doi.org/10.46661/revmetodoscuanteconempresa.5172

Ruf, B., Das, N., Chatterjee, S., & Sunder, A. (2019). Investments in ESG-rated mutual funds: Is good better than great? *The Journal of Wealth Management, 22*(1), 56–61. https://doi.org/10.3905/jwm.2019.1.070
Shin, J. J., Kwak, Y. H., & Park, C. (2024). Using machine learning to analyze ESG performance in Project-driven sectors. *Procedia Computer Science, 239*, 840–845. https://doi.org/10.1016/j.procs.2024.06.242
Steen, M., Taghawi, J., & Gjolberg, O. (2020). Is there a relationship between Morningstar's ESG ratings and mutual fund performance? *Journal of Sustainable Finance & Investment, 10*(4), 349–370. https://doi.org/10.1080/20430795.2019.1700065
Vilas, P., Andreu, L., & C. Serrano-Cinca, C. (2024). The limited role of sustainability in mutual fund investor decisions: A machine learning approach. *Expert Systems with Applications, 247*, Article 123288. https://doi.org/10.1016/j.eswa.2024.123288
Whelan, T., Atz, U., Van Holt, T., & Clark, C. (2021). ESG and financial performance: Uncovering the relationship by aggregating evidence from 1,000 plus studies published between 2015–2020. Center for Sustainable Business (NYU – Stern). Retrieved June 22, from https://www.stern.nyu.edu/sites/default/files/assets/documents/ESG%20Paper%20Aug%202021.pdf.
Wickham, H. (2023). Tidyverse. *R package*. Retrieved October 2024, from version 2.0.0. https://tidyverse.tidyverse.org https://github.com/tidyverse/tidyverse.
Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing, 415*, 295–316.
Zhou, Z. H. (2012). *Ensemble methods: Foundations and algorithms*. CRC Press.