



ESG ratings explainability through machine learning techniques

Alessandro Del Vitto¹ · Daniele Marazzina¹ · Davide Stocco¹

Received: 29 November 2022 / Accepted: 5 July 2023
© The Author(s) 2023

Abstract

Environmental, Social, and Governance (ESG) scores are quantitative assessments of companies' commitment to sustainability that have become extremely popular tools in the financial industry. However, transparency in the ESG assessment process is still far from being achieved. In fact there is no full disclosure on how the ratings are computed. As a matter of fact, rating agencies determine ESG ratings (as a function of the E, S and G scores) through proprietary models which public knowledge is limited to what the data provider effectively chooses to disclose, that, in many cases, is restricted only to the main ideas and essential principles of the procedure. The goal of this work is to exploit machine learning techniques to shed light on the ESG ratings issuance process. In particular, we focus on the Refinitiv data provider, widely used both from practitioners and from academics, and we consider white-box and black-box mathematical models to reconstruct the E, S, and G ratings' assessment model. The results show that it is possible to replicate the underlying assessment process with a satisfying level of accuracy, shedding light on the proprietary models employed by the data provider. However, there is evidence of persisting unlearnable noise that even more complex models cannot eliminate. Finally, we consider some interpretability instruments to identify the most important factors explaining the ESG ratings.

Keywords ESG ratings · Corporate social responsibility · Machine learning · Model explainability

1 Introduction

Environmental, Social, and Governance (ESG) scores are the response of rating agencies to the increase in demand for quantitative information capable of assessing a company's sustainable profile. ESG scores measure non-financial complementary information and aim to increase the accuracy of performance forecasts and risks assessment for public and private companies. In the last few years, sustainability ratings have gained more and more interest, and they now play a crucial role in the world of Socially Responsible Investments (SRI). Sustainability (2020) documents that 65% of investors surveyed use ESG ratings at least

✉ Daniele Marazzina
daniele.marazzina@polimi.it

¹ Department of Mathematics, Politecnico di Milano, Via Bonardi 9, 20133 Milan, Italy

once a week. The main reason for this popularity is that they represent one of the few instruments allowing investors to consider sustainable data in a simple and direct way. Even though there are several possible strategies through which investors can integrate ESG pieces of information in their sustainable portfolios, they all require great amounts of deep analysis and specific data (Billio et al., 2021) which these ratings overcome.

ESG, Corporate Social Responsibility (CSR), SRI and sustainable investment are concepts related to the integration of sustainable goals into business and investment decision-making. While these terms are often used interchangeably, they have distinct meanings and implications. ESG refers to the three broad categories of criteria that are used to evaluate a company's performance, policy and risk profile. The term ESG was officially employed for the first time in UN Global Compact (2004). The study, endorsed by many popular institutions, develops guidelines and makes recommendations on how to better integrate Environmental (E), Social (S), and Governance (G) issues in all financial industry fields, from asset management to regulation. The report aims to raise awareness among all market participants. The underlying belief is that, in the globalized, interconnected, and competitive framework where we operate, the management of ESG risks is an essential component for successful businesses. Environmental pillar includes a company's impact on the environment, such as its carbon emissions, water usage or waste production. Social factors refer to the company's impact on society, including its treatment of employees, suppliers and local communities. Governance pillar relates to the company's management, e.g., its leadership structure and board diversity. Therefore, ESG ratings are primarily used by investors to evaluate the sustainability of a company's business model and assess its long-term objectives.

CSR refers to a company's actions to improve its impact on the society and the environment, beyond its legal obligations. CSR includes a broad range of activities, such as philanthropy, community engagement, and environmental initiatives. SRI and sustainable investments are investment approaches that consider sustainability-related factors alongside more traditional financial metrics (Billio et al., 2021). SRI and sustainable investment seek to identify companies that are making positive contributions to the society and the environment (positive screening), while avoiding those that have negative impacts (negative screening). Sustainable investment strategies can also include screening companies based on specific ESG criteria, for instance investing in companies that have strong ESG performance relative to their peers. Sustainable investors may use a variety of metrics to evaluate investments, including ESG factors and the United Nations' Sustainable Development Goals.

In summary, ESG is a set of criteria used to evaluate a company's performance and risk profile, while CSR refers to a company's voluntary actions to improve its impact on society and the environment. SRI and sustainable investment are investment approaches that consider ESG and other factors to identify companies that are making positive contributions to the society and the environment, and to generate financial returns while doing so.

Similar to how credit ratings enable investors to screen firms based on their creditworthiness, ESG rating agencies provide investors a tool to evaluate companies' ESG performance. However, these two are fundamentally different. Firstly, ESG performance is still a nebulous notion, whereas creditworthiness is very clearly defined as the risk of default of a company. Secondly, ESG reporting is still in its infancy, while financial reporting standards have already been developed and harmonized over the previous century. Despite most relevant companies providing some form of ESG disclosure, there are still competing reporting standards, and none is still considered as a benchmark globally. Among other projects, the Global Reporting Initiative¹ promotes sustainability reporting by defining a comprehensive view on

¹ <https://www.globalreporting.org/standards/>.

a company's material issues, impacts and management; the United Nations Global Compact² encourages businesses to adopt environmentally and socially responsible policies and to report on their implementation; the OECD Principles for Corporate Governance³ are a set of standards for evaluating and improving the corporate governance. Finally, with a more circumscribed influence, the EU taxonomy⁴ represents one of the most advanced classification systems which establishes a list of environmentally sustainable economic activities.

The great rise in ESG interest gained momentum during the 2008–2009 financial crisis, during which firms with higher social capital exhibited stock returns that were several percentage points higher than firms with low corporate social responsibility intensity (Lins et al., 2017). The rising ESG consciousness observed worldwide quickly brought a wave of research studies analyzing the impact of ESG-related risks on financial markets. On the theoretical side, in Pástor et al. (2021) and Pedersen et al. (2021) an extension of the traditional Capital Asset Pricing Model (CAPM) is introduced. Investors act in a traditional efficient-frontier framework and they include ESG-related information in their portfolio composition stage. Empirically, several studies find evidence of an impact (still weak) of ESG information on the financial market. Pelizzon et al. (2021) employ a quasi-natural experiment to prove the presence of a transitory price pressure on stocks lead from an incorrect assessment of some investors about the meaning of the change in ESG ratings. Berg et al. (2022a) study the impact of ESG rating changes in the ownership of mutual funds with dedicated ESG investment strategies. The authors find a correlation between rating upgrades and downgrades and stocks' long-term returns. They also show that companies correct their behaviors, after being downgraded, only for what concerns the Governance pillar score, concluding that, while ESG ratings seem to have an impact on financial markets, their effect on the real economy is still limited.

Lin et al. (2019) discuss the effect of CSR on Corporate Financial Performance (CFP), measured by three different economic metrics: Return on Equity (ROE), Return on Assets (ROA), and Return on Invested capital (ROI). The results show a negative link between CSR and CFP, supporting the idea of a trade-off between committing to the realization of ESG objectives and optimizing the firm from the economic point of view in favor of shareholders.

The lack of clarity and common definitions characterizing the world of sustainable investments, which causes weak or even heterogeneous results in unveling the financial influence of ESG ratings, reflects also in the methodology exploited in the ESG issuance. Each data provider has developed its own ESG rating system with its specific characteristics, which translates into different sustainability assessments for the same companies. Berg et al. (2022b) document significant divergences in the ESG ratings of five major data providers and they identify three specific sources of divergence: scope, measurement, and weight. Scope divergence is due to what each rater deems to be the sustainable themes to include in the screening process. Measurement divergence concerns the set of indicators the data provider selects to assess the firm's quality for each pillar. Lastly, weight divergence regards discrepancies in the importance attribution mechanisms of the different agencies, meaning that rating agencies take different views on the relative importance of attributes. They find that measurement divergence is responsible for more than half of the overall discrepancy and that scope and weight are moderately less relevant, yet definitely non negligible. Similar results are discussed in (Billio et al., 2021). The authors show that the disagreement on characteristics,

² <https://unglobalcompact.org/>.

³ <https://www.oecd.org/corporate/principles-corporate-governance/>.

⁴ https://finance.ec.europa.eu/sustainable-finance/tools-and-standards/eu-taxonomy-sustainable-activities_en.

attributes and standards in defining E, S, and G components leads to heterogeneous ESG assessments across different rating agencies and, additionally, this lack of common results disperses the effects of ESG investors preferences on asset prices.

Considering ratings coming from different data providers, Berg et al. (2021b) analyze how ESG performance affects firms' stock returns, considering and adjusting for the noise and the confusion presented in the ESG information published by rating agencies. They find a positive correlation between a firm sustainable performance and expected stock returns, and they conclude that, in the investment decision process, it is better to consider several ESG ratings simultaneously, since, despite the different levels of noise, the data that different agencies provide is still informative. Bams and van der Kroft (2022) address the problem of ESG information asymmetry, highlighting consequent incentives for rating inflation and showing that ESG-rating-based portfolios are less sustainable than the market portfolio. They underline the tendency of sustainable investors to rely on third-party ESG ratings for their investment and divestment decisions, arguing that this phenomenon, together with the significant information asymmetry that comes from the unstandardized nature of ESG disclosures, incentivizes firms to artificially inflate their sustainability ratings obtaining a lower cost of capital. The authors explain that this can be achieved by reporting very optimistic future performance promises without really committing to realizing them. In support of this thesis, they show that ESG ratings are extremely sensitive to improvements in firms' future performance estimates, typically ambitious and hardly achievable, to the point that they are negatively correlated with companies' actual sustainable performances.

The lack of a common framework severely questions the reliability and effectiveness of ESG ratings as measures of corporates' sustainable performances and as tools for socially responsible investment decision-making, addressing a call for action to policy makers to build a standardized regulatory framework for sustainability disclosures and assessments.

In addition to the aforementioned controversies about the ESG framework as a whole, the main problem preventing investors and policymakers from having an effective way to precisely assess the reliability of the aggregation process exploited to determine ESG scores is the lack of transparency concerning the rating system. Rating agencies issue ESG scores through proprietary models which public knowledge is limited to what the data provider effectively chooses to disclose, that, in many cases, is restricted to the main ideas and essential principles followed in the rater's methodology, whose structure is specific to the considered agency. Due to this reason, the rater's algorithm can be considered a black-box model which inner working is unclear.

This work wants to contribute to the existing literature and proposes a tool for the solution of the transparency problem, unveiling, understanding, and explaining, through machine learning methods, the model that is exploited by data providers to issue ESG ratings.

To the best of our knowledge, there are no academic research papers relating machine learning methods and ESG ratings to uncover the proprietary algorithm exploited in the score issuance process, even though machine learning has been applied in several ESG-related studies. D'Amato et al. (2022) apply random forests algorithms to predict companies' ESG scores from structural financial data as balance sheet items. De Lucia et al. (2020) move on the opposite direction, predicting companies' ROE and ROA from ESG indicators. Zanin (2022) analyzes with statistical and machine learning methods the impact of firms' ESG scores on credit ratings.

In this work, we consider the Refinitiv data provider, and we construct algorithms able to replicate and predict Refinitiv's ESG scores, considering both fully explainable models (the so-called white-box) as well as machine learning black-box models. We start from the overall set of ESG granular data points collected by the data provider and apply simple regression

models, like Ridge and Lasso regularization, and black-box models, like Random Forest and Artificial Neural Networks (ANN), to predict companies' individual E, S, and G sustainability ratings. Notice that the Refinitiv's proprietary model changes for different industrial sectors for the E and S pillars, and it changes in the G depending on the firm's country. More precisely, for E and S, we analyze the 'Financial and Insurance', 'Manufacturing', and 'Information & Professional, Scientific, and Technical Services' sectors, while for the G pillar we consider the United States of America, European, and Chinese geographical regions. Then, we address the problem of acquiring a thorough understanding of the Refinitiv ESG score assessment model. We apply the Shapley Values model explainability technique to recover a general understanding of the algorithm and to interpret specific companies' ratings. In the end, we show that machine learning models are able to predict Refinitiv ESG scores if trained with a suitable selection of features, and that it is possible to understand and explain, from a general point of view, the behavior of the proprietary model.

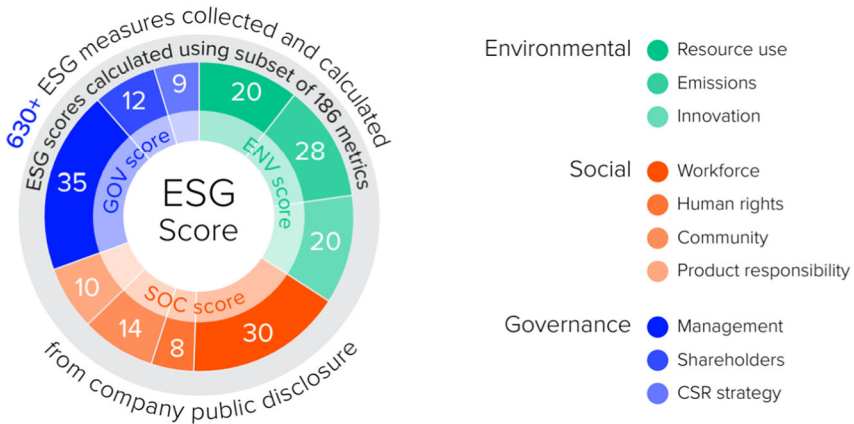
The remainder of the paper is structured as follows. In Sect. 2, we introduce and describe Refinitiv ESG scores, explaining what is known about the assessment process. In Sect. 3, we report in detail our replication methodology, providing a thorough description of the machine learning models we applied, the data, and the preprocessing stage. The replication results are reported in Sect. 4. Section 5 concerns the problem of model explainability. In Sect. 6, we conclude the paper analyzing, interpreting, and explaining our results.

2 Refinitiv ESG scores

Refinitiv is one of the major data providers in the financial world. Its ESG data have been analyzed by more than 1 500 academic research papers (Berg et al., 2021a). Originated from a rebranding of Thomson Reuters' Financial & Risk business in 2018, Refinitiv's ESG dataset, as well as Refinitiv's ESG scores, were initially constructed and defined by ASSET4, a company acquired by Thomson Reuters in 2009, that was one of the firsts to gather and analyze environmental and social pieces of information. Refinitiv ESG data cover over 80% of the global market cap, providing information on over 12,000 public and private companies (Huber et al., 2017; Refinitiv, 2022).

Refinitiv ESG ratings are percentile rank scores, ranging from 0, the worst possible score, to 100, the best possible score. They are designed to objectively measure a company's relative ESG performance, commitment, and effectiveness. Refinitiv argues that its sustainability ratings are completely data-driven, effectively balanced to take into account the most material industry metrics and adjusted to compensate for transparency and market cap biases (Refinitiv, 2022). Refinitiv scores are based only on company-disclosed data available on their websites, reports, media, and news (Billio et al., 2021; Refinitiv, 2022).

Even though ESG scores change primarily once a year, when the company sustainability report is disclosed, they are updated weekly and remain subject to changes for five fiscal years before being marked as *definitive* (Refinitiv, 2022). This data rewriting process is partially due to company restatements and data corrections, but also to changes in the aggregation rule applied by Refinitiv itself. Berg et al. (2021a) provide thorough documentation of this phenomenon, showing that the most significant rewriting source appears to be the second one. They also show that the rewriting process, due to methodological changes in the rating attribution, introduces in the data-set a positive correlation between ESG ratings and firms' stock market performances that cannot be observed in the original information set. Moreover, it is not always clear to which time extent these changes can prolong since the declared cut-off



- (a) Rating methodology: ESG Scores are aggregation of 186 metrics, selected from more than 630 data points.
- (b) ESG categories: the 186 metrics are divided in 10 categories.

Fig. 1 Refinitiv ESG Scores, *source*: Refinitiv (2022)

period was only 3 years before February 2021 and it was changed later to five. They conclude by arguing that this ongoing phenomenon plausibly comes from the agency's incentive to retroactively strengthen the link between its ESG scores and financial returns.

It must be noticed that the overall Refinitiv ESG score is determined by the combination of the three individual pillars (E, S, and G) for which we also have a quantitative percentile rating. These scores are computed relatively to the framework in which the company is operating; they are not and should not be intended as absolute scores. This relative computation is not performed in the same manner across the three factors. The Environmental and Social pillar scores are determined through the comparison with industry peer companies, whereas the Governance pillar score is obtained relatively to the firm's country of incorporation. In our attempt to replicate and understand Refinitiv ESG Scores we will consider these three pillars separately and set up specific models to take into account the differences in treatment. Moreover, as stated in (Refinitiv, 2022), scores are computed with algorithms adjusted for each industrial sector for the Environmental and Social pillar, and for each country for the Governance pillar, see Sect. 3.2.

Refinitiv ESG Scores are the result of the aggregation of 186 selected metrics, selected from more than 630 data points, to power the model with the most comparable and significant information per industry, as summarized in Fig. 1a. The aggregation rule effectively considers, for each sector, only a subset that goes from 70 to 170 of the 186 variables. Therefore, each subset contains the features for our learning models, i.e., the inputs of the white and black-boxes, while the E, S and G scores are the output that the models have to learn. Notice that the values of all the features are given by the data provider, while the formulas which compute the scores starting from the features is not fully publicly disclosed, therefore our interest in considering learning models to shed light on it. These features are grouped into ten categories, which are associated to the three pillars E, S, and G, and represent the scope for which the single data have been considered in the sustainable screening model, as it is represented in Fig. 1b. In all the tests we consider the 2021 E, S and G scores.

3 Methodology

We approach the problem of discovering the model underlying the sustainability ratings assessment by constructing and optimizing multiple machine learning regression algorithms to replicate the scores of the underlying model with the best possible accuracy.

In the modeling phase we initially consider simple white-box algorithms, like linear regression models, k-nearest neighbor algorithm, and decision trees, and then we move to black-box algorithms, like random forests and ANNs. In all the models, the learning process consists of the minimization of the Root Mean Squared Error (RMSE) on a training set, measuring and comparing models' performances onto an independent test set. To be able to work with the great amount of data and to overcome the problems of missing values, zero-variance predictors, and redundant features, we first perform a substantial preprocessing stage.

In the following, in Sect. 3.1 we deal with the considered white- and black-box algorithms, as well as with the interpretability tools, necessary to understand the functioning of the black-box tools, while in Sects. 3.2 and 3.3 we present the data-set we worked with, and how data has been preprocessed, e.g., how we deal with missing values.

3.1 Modeling

We exploit several regression algorithms⁵ among the most used in supervised learning, a subcategory of machine learning and artificial intelligence where data-sets are used to train algorithms to predict outcomes (in our case the E, S, G scores) accurately. We deal with white-box models, that is simpler models such as linear regression and decision trees which usually provide less predictive capacity and are not always capable of modelling the inherent complexity of the data-set, e.g., feature interactions, being however significantly easier to explain and interpret, and with black-box models, such as neural networks, and boosting or ensemble models, that often provide great accuracy. The inner workings of this last class of models are however harder to understand, and it is necessary to use interpretability algorithm to try to explain their functioning, e.g., estimate the importance of each feature on the model predictions. For an introduction to the considered models and the interpretability of black-box models, we refer to Hull (2021).

Among the models, we first of all consider the linear regression, as well as Lasso and Ridge regressions, which can be considered as L^1 and L^2 penalized version of the linear one. Given a data-set with n observations, each one consisting of p features, x_{ij} , $i = 1, \dots, n$, $j = 1, \dots, p$, and an output y_i , $i = 1, \dots, n$, the linear regression models takes the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon_i,$$

ϵ_i being the error term. The computation of the coefficients β_0, \dots, β_p is done through the least-square method, i.e., minimizing

$$\frac{1}{n} \sum_{i=1}^n \epsilon_i^2 = \frac{1}{n} \sum_{i=1}^n (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} - y_i)^2.$$

⁵ We use the Keras' Scikit-learn library for Python.

The Lasso and Ridge regression are L^1 and L^2 penalized version of the linear one due to the introduction of penalizing terms, that is, the function to be minimized is

$$\frac{1}{n} \sum_{i=1}^n (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} - y_i)^2 + \lambda \sum_{i=1}^p |\beta_i|$$

for Lasso,

$$\frac{1}{n} \sum_{i=1}^n (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} - y_i)^2 + \lambda \sum_{i=1}^p (\beta_i)^2$$

for Ridge regression.

These models are white-box algorithms, i.e., fully explainable. Their inner working is very straightforward: the model's outcome is computed through a weighted sum of each sample's features. Therefore, the model's explanation can be obtained directly from the variables' weights β_0, \dots, β_p exploited in the algorithm. However, linear regression is prone to overfitting in case a large number of regressors (the features) is considered. Ridge regression prevents overfitting, increasing bias (but lowering variance). Lasso regression is designed to select features by shrinking coefficient towards zero due to its penalization term; however Lasso selects only one feature from a group of correlated features, the selection being arbitrary. This does not impact on the predictability power of the algorithm, but on its interpretability, if not coupled with ad-hoc methods, see, e.g., Wang et al. (2019). This is the reason why we do not use Lasso for interpretability in our analysis, but we only consider its predictive ability. We refer to (Hull, 2021, Chapter 3) for details.

Among supervised machine learning white-box models, we deal with neighbors regression, that assigned to a firm a score based on the mean of the scores of its nearest neighbors, the distance between the firms being computed according to a predefined norm. The main problem of this method is the curse of dimensionality, since it works well if the number of features is not large. We refer to (Hull, 2021, Chapter 2) for details. We also consider decision trees, that predict the score variable by learning simple decision rules inferred from the data features. Decision trees are simple to understand and to interpret, trees can be visualized, see, e.g., Azzone et al. (2022), however this method is prone to overfitting, creating over-complex trees that do not generalize the data well, see (Hull, 2021, Chapter 4) for details.

To improve the ability of algorithms to predict scores, ensemble learning models are considered, where multiple models (often called "weak learners") are trained to solve the same problem and combined to get better results, at the cost of losing the explainability of the predictions (that is why they are referred as black-box algorithms). Among ensemble models we deal with random forests, based on the bagging ensemble, that is the score is computed as a mean of the scoring of several small decision trees, each one trained exploiting a random subset of firms and a random subset of features, and the boosting Ada regression, where subsequent weak learners are tweaked in favor of those instances (in our case the firms) having a large prediction error by previous learners. We refer to Dietterich (2000) for a review of ensemble algorithms.

To conclude, we consider ANNs, inspired by the biological neural networks that constitute human brains. An ANN is based on a collection of connected units or nodes called artificial neurons. Each connection can transmit a signal to other neurons. An artificial neuron receives signals, processes them, and sends the processed signal, that is, its output, to neurons connected to it. The signal is a real number, and the output of each neuron is computed by some non-linear function of the sum of its inputs. The connections are called edges. Neurons and edges typically have a weight adjusted as learning proceeds. The weight increases or

Table 1 Regression models and hyperparameters values

Model	Hyperparameters
LinearRegression	None
LassoRegression	$\lambda \in \{0.00001, 0.0001, 0.001, 0.01, 0.1, 0.2, 0.4, 0.6, 0.8, 1\}$
RidgeRegression	$\lambda \in \{0.01, 0.1, 0.5, 1, 2, 10, 20, 30, 100\}$
KNeighborsRegressor	$n_neighbors \in \{5, 10, 15, 20\}$ and $p \in \{1, 2, 4, 8, 10\}$
DecisionTreeRegressor	$max_depth \in \{5, 10, 15, 20\}$, $min_samples_leaf \in \{1, 2, 5\}$, and $min_samples_split \in \{1, 2\}$
RandomForestRegressor	$n_estimators \in \{200, 400, 500\}$
AdaBoostRegressor	$n_estimators \in \{200, 400, 500\}$ and $learning_rate \in \{0.001\}$

For details on the hyperparameters, we refer to Appendix C in the supplementary materials

decreases the strength of the signal at a connection. ANN can exhibit very complex structures, it is not possible to fully explain the prediction results, however their ability to predict scores even in complex scenarios is fully recognized. We refer to (Hull, 2021, Chapter 6) for details.

We document in Table 1 the models and the hyperparameter combinations we have considered in the optimization procedure, with the exception of the ANNs that we discuss below. A hyperparameter is a parameter whose value is used to control the learning process: we train the learning algorithms on different parameters on the training set, choosing the ones which achieve the minimum RMSE in the test set.

The optimal Ridge L^2 penalization and Lasso L^1 penalization coefficients are $\lambda=10$ and $\lambda=1e-4$, respectively, with the exception of the Social score for the Manufacturing sector companies, in which the parameters achieving the best results are $\lambda=0.5$ and $\lambda=1e-5$, respectively, and the three Governance pillar score scenarios for which the best Ridge parameter is $\lambda=1$.

For what concerns the KNeighborsRegressor algorithm there is no general rule, obtaining different hyperparameters for each specific case.

The optimal parameter choices for the simple decision tree model are $max_depth=10$, $min_samples_leaf=5$, and $min_samples_split=2$ in all the cases except for Chinese companies G score in which $max_depth=5$ works better. The lowest RMSE is generally reached by RandomForests and AdaBoostRegressors with $n_estimators=500$, but we also have that, in the Inf sector E score replication, $n_estimators=200$ are enough for both algorithms.

Considering ANNs, we deal with three models:

- *ANN1*—a straightforward three-layer ANN with two layers of 64 and 32 hidden units completed with a single-neuron output layer with linear activation.
- *ANN2*—a deeper architecture considering a five-layer ANN with Dense layers of 128, 64, 32, and 16 hidden units before the necessary single-neuron output layer with linear activation.
- *ANN3*—a five-layer ANN with Dense layers of 256, 128, 64, and 32 hidden units completed with the necessary single-neuron output layer with linear activation.

We use hyperbolic tangent (\tanh) activation functions in the hidden layers for all the models.

3.1.1 Interpretability

Model explainability, or global interpretability, techniques are very useful tools that allow us to understand, to some extent, how a black-box model works. In our paper we consider the SHAP (SHapley Additive exPlanations) technique, see Lundberg and Lee (2017). It originates from a method of cooperative game theory that allows one to measure a player's strength, considering the weighted marginal contributions on all the possible coalitions that he can join. It is translated to the world of machine learning imagining that the predicted value is determined through a coalition game played by the different feature values. This technique estimates each feature's relevance by measuring its impact on the model's output value. The Shapley Value of a feature can be defined as follows: given the current values of all the features, the impact of the value of a particular feature on the difference between the considered prediction and the average prediction (obtained computing the average between all the predictions of the training set) is its estimated Shapley Value (Molnar, 2022).

The advantages of Shapley Values are numerous. Thanks to the solid theory behind cooperative games we have a reasonable foundation that guarantees the four properties of efficiency, symmetry, dummy player (variable), and additivity. These characteristics imply that the difference from the average prediction is fairly distributed among all the model's features values, that symmetric players get the same value (symmetric features, in the sense that they influence the output in the same way, are attributed the same relevance), that a player that does not contribute to any coalition does not get anything (a feature that never changes the output is irrelevant), and, lastly, that estimated values are additive. On the other hand, the main disadvantage of Shapley Values is the fact that they are very expensive from the computational point of view, which is particularly relevant in our case given the large number of features considered.

Shapley Values can be used for both global and local interpretability. Global interpretability focuses on the general prediction model decisions, while local interpretation focuses on the specifics of a single observation. In our analysis, the global feature importance estimation technique is based on the magnitude of variables' attributions. It estimates the average Shapley Value of each feature on the whole data-set, and represents the results, ordered according to the estimated mean relevance, through a straightforward 'beeswarm' plot, designed to display an information-dense summary of how the top features impact the model's output.

Dealing with the single prediction interpretations (local interpretability), Shapley Value shows the contribution of each feature to push the model output of the considered observation, e.g., the firm, from the base value (the average model output over the training data-set) to the model output associated with the observation. Given a single observation, a set of SHAP values, one for each feature, is calculated. We recall that SHAP values are additive: their sum is equal to the difference between the considered prediction and the average prediction.

3.2 Data

We deal with the 2021 ESG score,⁶ and we collect the most possible granular pieces of information available on Refinitiv's Eikon software from the former Asset4 data-set. We divide under the three pillars E, S, and G and we consider only companies with a market

⁶ All data were downloaded in September 2022. We also consider 2019 and 2020 ESG scores, obtaining comparable results both in ability to predict the scores and in interpretability.

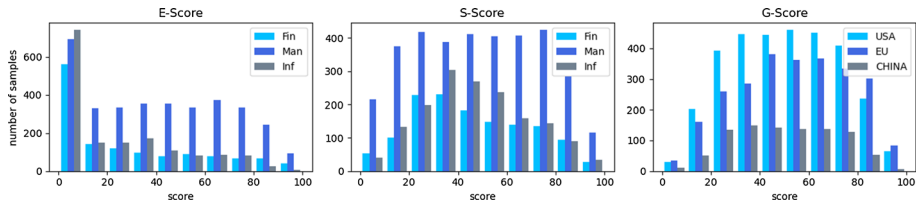


Fig. 2 Refinitiv 2021 ESG Scores by pillar distributions: E, S, and G scores histograms by industry sector and by geographical region

cap of over one million USD. We repeat this procedure for three different industries of the NAICS Sector Classification⁷:

1. *Finance and Insurance* (**Fin**)
2. *Manufacturing* (**Man**)
3. *Information & Professional, Scientific and Technical Services* (**Inf**)

and three geographical regions:

1. *The United States of America* (**USA**)
2. *European Countries*⁸ (**EU**)
3. *China* (**CHINA**)

We choose these three sectors/regions since they have the largest number of firms. Notice that the Man sector is of particular interest for the high waste production, which has a large impact on the environment,⁹ and therefore with a problem related to the E pillar. The Fin and Inf sectors are also of interest for the S pillar, as shown in Tamimi and Sebastianelli (2017); for the Fin sector, the ESG rating is also important when green bonds are issued, see Grishunin et al. (2023). For the regions, our choice considers three of the most important economies. The fact that the G score strongly depends on the law of the firm's country is another important driver to deal with these regions.

For each sector, we consider all the firms worldwide. Notice that with the threshold on market capitalization we delete only 9% of all the firms, this percentage being constant for any pillar, avoiding considering small firms, which could provide less (or less reliable) information to Refinitiv. In the end, we collect 180 attributes concerning the Environmental pillar, 281 associated to the Social one and, lastly, 152 regarding the Governance. We end up with a total of 1340, 3446, and 1606 firms for each one of the three industry sectors and 3137, 2568, and 955 companies for each geographical region, respectively. Following Refinitiv's Eikon process, we consider the industry split for the E and S pillar, while the geographical partition for the G pillar. In Fig. 2, we report the E, S, and G scores histograms by industry sector and by geographical region of our data-set. Notice that the empirical distribution of the Environmental score is very different from the other two; in particular, several scores are zero or close-to-zero E-rated firms, while this does not happen in the Social and Governance cases.

In the following we always work on each pillar and on each sector/country, separately. As stated in Refinitiv (2022), Refinitiv exploits algorithms adjusted for each sector/country.

⁷ The "North American Industry Classification System", a general overview of this sector categorization can be found at <https://www.census.gov/naics/>.

⁸ We have to consider the European region as a whole, since, otherwise, focusing on the individual countries, we would have a too meager data-set to optimize and train machine learning models.

⁹ <https://www.epa.gov/trinationalanalysis/manufacturing-sectors>.

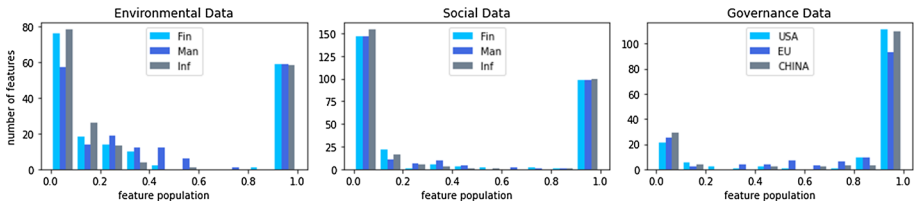


Fig. 3 Missing data problem: distribution of the feature population among the different pillars, different sectors/countries (1 = 100% refers to the features without missing values, while 0 refers to the empty ones.)

Therefore, given that our goal is to spread light on the different algorithms, we construct white and black-boxes algorithms for the different sectors/countries, and not for the sample as a whole, e.g., predicting the E score considering all the firms of the three sectors together.

As shown in Fig. 3, the data at our disposal is highly unpopulated. As an example, in the E pillar, there are no firms without at least one missing value. The problem is particularly severe in the case of the Environmental and Social scores where a large quantity of the sustainability attributes are completely, or nearly, empty. For what concerns the S score, we see that, in each industry, more than half of the variables have values for less than 5% of the data-set samples and shall be removed from the data-set. It is also interesting to notice that the majority of the unpopulated features are numerical; on the contrary, categorical variables are typically nearly complete. This difference comes from the fact that almost all of the categorical variables are Boolean and Refinitiv's analysts fill in with default zero values the data points for which no relevant information is found in the firm's public disclosure (Refinitiv, 2022).

3.3 Preprocessing

We cannot directly use all granular ESG pieces of information because of the large quantity of variables (the predictors, or features, of our learning process). In machine learning, feature selection techniques are employed to reduce the number of input variables by eliminating redundant or irrelevant features. Feature selection has mainly three benefits: (i) decreases overfitting, since fewer redundant data means fewer chances of making decisions based on noise, (ii) improves accuracy, due to the decrease of misleading data, and (iii) reduces the training time. Moreover, some data are missing. Therefore the needs of a preprocessing of the data-set.

More precisely, in our case, some predictors are completely unpopulated, some present very low variability, and others are overlapping in their definition, showing high correlation with each other. We preprocess our data as follows: first, we prepare the data-set with a robust feature selection (see Sect. 3.3.1); second, we fill in the missing values of the remaining variables (see Sect. 3.3.2); then, we split the data-set into a training and a test set, see Table 2; and finally, we apply a MinMaxScaling procedure to scale every predictor and the target score to the unitary interval, that is, given x the value of a feature for one firm, and m (M) the minimum (maximum) value of this feature over all the firms of the sector/country, we rescale x as

$$\frac{x - m}{M - m}.$$

We repeat this preprocessing independently for each distinct industry sector and for each geographical region considered, since a variable that is not contemplated by Refinitiv for a specific industry or geographical region could be relevant for another one, e.g., a variable

can be highly correlated with others in the E pillar, and therefore removed, but not in the S one, where it is considered in the learning process.

3.3.1 Feature selection

Our feature selection process consists of three sequential steps: we firstly eliminate highly unpopulated features, then low variance predictors, and, finally, redundant attributes, i.e., those highly correlated with other features. For details on the thresholding methodology considered—e.g., we consider a feature as populated if the number of non-empty samples is larger than a fixed threshold, otherwise we remove it—we refer to Appendix B in the supplementary material.

As highlighted in Sect. 3.2, the problem of unpopulated variables is particularly severe. Even though the empty attributes vary across different sectors, there are a few that are completely void across all industries and are consequently removed, e.g., in the E score model *Percentage of Green Products*, and several very specific metrics concerning EPA Upstream and Downstream Scope 3 Emissions.¹⁰

In addition to the unpopulated features of the data-set, several predictors with very low variability across companies are removed. For example, in the E score model for financial companies, we remove *Agrochemical 5% Revenue* and *Animal Testing Reduction*, since the former is false for all the firms and the latter is true only for one sample.

In Refinitiv's original data-set there are numerous cases of perfectly positively correlated variables. We just mention an example: in the E pillar score, *Total CO2 Equivalent Emissions To Million Revenues USD* and *CO2 Equivalent Emission Total* are perfectly collinear, therefore the latter is drop out.

We summarize in Table 2 the number of predictors we consider at the end of each feature selection process for each pillar, by industry for E and S and by geographical region for G. Detailed information about the selected variables for the E, S, and G models can be found in Appendix A in the supplementary materials, where we report complete lists of the considered variables.

3.3.2 Missing values

We operate under the assumption that every incomplete data is due to non-disclosed information by the company and not to a failure in the data collection procedure by the rating agency. Following this hypothesis, missing values cannot be treated neutrally, as it is done in many other applications. The lack of data must be weighted by our algorithm to take into account and compensate for transparency biases, as it is done by Refinitiv itself. Due to this reason, we complete the data-set as follows: first of all, for each feature with missing values, we compute its correlation with the target score. If the correlation is positive (negative), a high (low) value of the feature will probably be related to a high score. Therefore, in order to penalize missing values, we replace them with the minimum value of the feature on the whole data-set in case of positive correlation, whereas, if it is negative, we use the maximum registered one. In this way, we are confident that we penalize the missing values, and not treat them impartially (as in the case we replace them with the mean or median value).

¹⁰ EPA is the acronym for the United States Environmental Protection Agency. Scope 3 emissions are the results of activities from assets not owned or controlled by the reporting organization. Details can be found at <https://www.epa.gov/climateleadership>.

Table 2 Number of selected features (Feat) and samples (Samples), and how the samples are split in each training (Train) and test (Test) set

	E score				S score				G score				
	Feat		Test		Feat		Test		Feat		Test		
	Samples	Train	Samples	Train	Samples	Train	Samples	Train	Samples	Train	Samples	Train	Test
Fin	56	1340	1206	134	84	1340	1206	134	USA	86	3137	2980	157
Man	77	3446	3273	173	82	3446	3273	173	EU	85	2568	2388	180
Inf	41	1606	1477	129	64	1606	1477	129	CHINA	70	955	859	96

4 Results

In Table 3 we report the Environmental and Social scores results under the RMSE metric¹¹ for all the models and for all the considered industry sectors, underlining the models capable of replicating Refinitiv's ratings with the best accuracy. The best result is reached predicting the Environmental score for companies in the Inf sector with the regularized linear regression models (both Ridge and Lasso). On the other hand, the lowest accuracy is observed for the Social pillar for Manufacturing companies: in this case the best performing models are ANN1 and ANN3. The worst performing model from a general point of view is clearly the decision tree, which presents the worst results in four out of the six scenarios. Also the K-Neighbors has similarly poor performances. On the contrary, the models that seem to obtain the best results are the Ridge and Lasso regressions, which consistently present low average prediction errors. We also notice a large overfitting for random forest and Ada Boost methods, since the errors on the training set are largely smaller than the corresponding on the test set. Moreover, the results show that the more complex models ANN2 and ANN3 are able to provide better predictions with respect to the more straightforward ANN1 only in half of the cases.

This first analysis suggests that Lasso, Ridge, and ANN methods are the best ones, due to the high accuracy in predicting on the test set, and the low (or missing) overfitting. Moreover, the fact that ANN1 provides an accuracy similar to ANN3 in both the training and the test sets, and the good results for regularized linear regressions, suggest that the proprietary formula should be not that complex.

We now move to the Governance pillar, Table 4. The best accuracy in replication of the G score is reached by the ANN2 in the specific case of American companies. European companies are the ones for which we observe the poorest accuracy level: in this case, the Lasso linear regression is the best model. As before, a large overfitting is present for decision trees and random forests.

Given the preliminary results described above, in the following we will focus on Lasso and Ridge regressions, as well as ANN. Considering Lasso and Ridge regressions, in Fig. 4 we report, for each pillar and for each industry/region, the Cross Validation RMSE results for both the Lasso and Ridge regressions. More precisely, we randomly split the data-set into training set and test set 10 times, computing each time the error, and reporting in the figures the 10 errors as well as their average. In line with the previous results, for both models the best average RMSE is achieved in the case of the E score for firms belonging to the Inf industry sectors, whereas the worst average RMSE score is obtained in the European companies case. We also see that these scenarios are the ones with the most variability in the results of the model from 10 different training and test sets choices.

We now move to ANN: in Figs. 5, 6 and 7 we plot the ANN1 predictions in comparison to Refinitiv's scores for the Environmental, Social, and Governance pillars, respectively, on a randomly selected subset of a hundred firms from the test sets. Firms are ordered accordingly to their E, S, and G ratings, scaled in the unit interval. From the graphs we see that the model accurately predicts low rated firms, especially in the E score case for which we have numerous samples close-to-zero employed in the learning process. Conversely, this model seems to moderately underestimate higher scores. Nevertheless, the errors are well-distributed and moderate in value, as shown in Fig. 8, where the absolute residuals histograms on the whole test sets are reported.

¹¹ We report in the Appendix D in the supplementary materials the results under the Mean Absolut Error (MAE) and R^2 measures.

Table 3 Results on train and test set: Root Mean Square Error (RMSE) metric

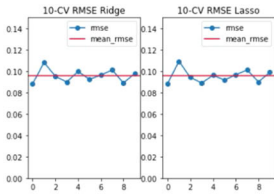
	Training set						Test set					
	Fin		Man		Inf		Fin		Man		Inf	
	E	S	E	S	E	S	E	S	E	S	E	S
LinearRegression	0.089	0.085	0.086	0.082	0.071	0.093	0.084	0.093	0.084	0.098	0.072	0.095
RidgeRegression	0.091	0.087	0.086	0.082	0.072	0.092	0.078	0.092	0.084	0.097	0.071	0.094
LassoRegression	0.089	0.086	0.086	0.082	0.072	0.094	0.079	0.092	0.084	0.097	0.071	0.094
KNeighborsRegressor	0.097	0.092	0.099	0.094	0.074	0.101	0.106	0.107	0.109	0.110	0.090	0.111
DecisionTreeRegressor	0.072	0.074	0.079	0.074	0.065	0.076	0.104	0.117	0.124	0.120	0.087	0.122
RandomForestRegressor	0.041	0.034	0.038	0.035	0.042	0.037	0.096	0.089	0.096	0.101	0.081	0.103
AdaBoostRegressor	0.050	0.053	0.070	0.069	0.049	0.065	0.096	0.093	0.100	0.108	0.081	0.107
ArtificialNeuralNetwork 1	0.089	0.090	0.082	0.080	0.072	0.090	0.083	0.102	0.085	0.096	0.075	0.093
ArtificialNeuralNetwork 2	0.096	0.088	0.084	0.082	0.074	0.093	0.081	0.100	0.089	0.097	0.079	0.095
ArtificialNeuralNetwork 3	0.092	0.087	0.083	0.081	0.072	0.093	0.081	0.098	0.088	0.096	0.077	0.090

We bold the smallest RMSEs for each sector; for each pillar, to highlight the models providing the best replication of Refinitiv scores

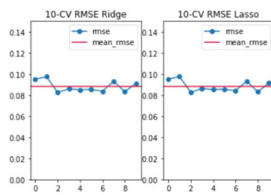
Table 4 Results on train and test set: Root Mean Square Error (RMSE) metric

	Training Set			Test Set		
	USA	EU	CHINA	USA	EU	CHINA
LinearRegression	0.085	0.100	0.085	0.088	0.258	0.092
RidgeRegression	0.085	0.100	0.086	0.087	0.167	0.092
Lasso Regression	0.086	0.100	0.086	0.086	0.103	0.092
KNeighborsRegressor	0.123	0.121	0.154	0.135	0.131	0.163
DecisionTreeRegressor	0.089	0.088	0.145	0.143	0.169	0.206
RandomForestRegressor	0.039	0.046	0.052	0.105	0.110	0.140
AdaBoostRegressor	0.076	0.064	0.038	0.119	0.111	0.141
ArtificialNeuralNetwork 1	0.068	0.093	0.080	0.078	0.112	0.094
ArtificialNeuralNetwork 2	0.071	0.097	0.082	0.075	0.112	0.092
ArtificialNeuralNetwork 3	0.067	0.095	0.082	0.077	0.112	0.092

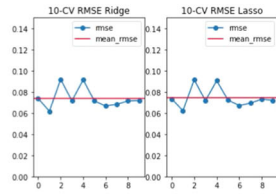
We bold the smallest RMSEs for each country to highlight the models providing the best replication of Refinitiv scores



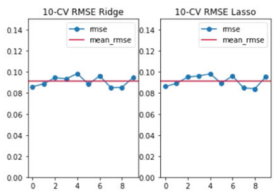
(a) E score, Fin sector.



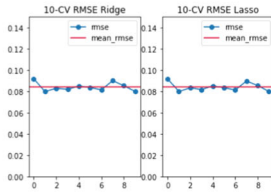
(b) E score, Man sector.



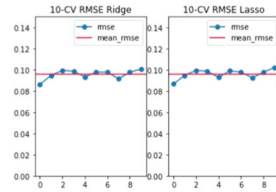
(c) E score, Inf sector.



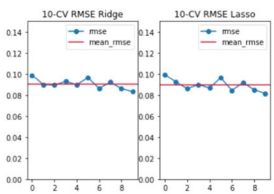
(d) S score, Fin sector.



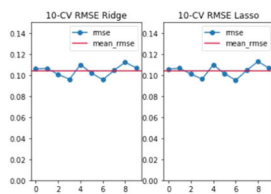
(e) S score, Man sector.



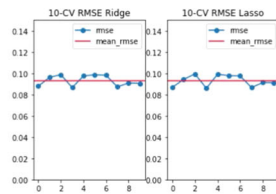
(f) S score, Inf sector.



(g) G score, USA.

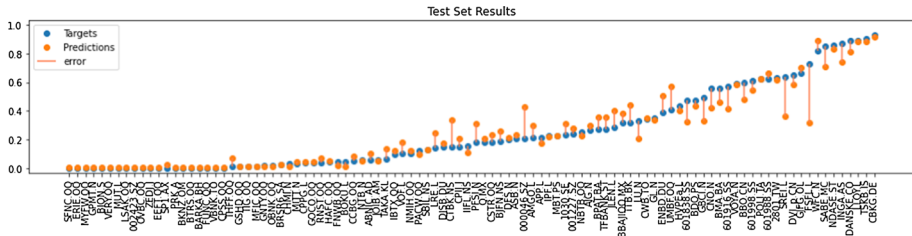


(h) G score, EU.

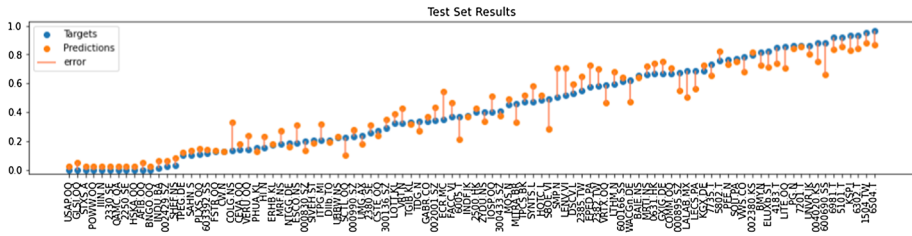


(i) G score, CHINA.

Fig. 4 Ridge and Lasso 10-Fold CrossValidation results. The data-set is split in training and testing set randomly ten times, and we report the 10 Root Mean Square Errors (RMSE) computed on the test set, as well as their mean values



(a) Fin sector targets and predictions.



(b) Man sector targets and predictions.



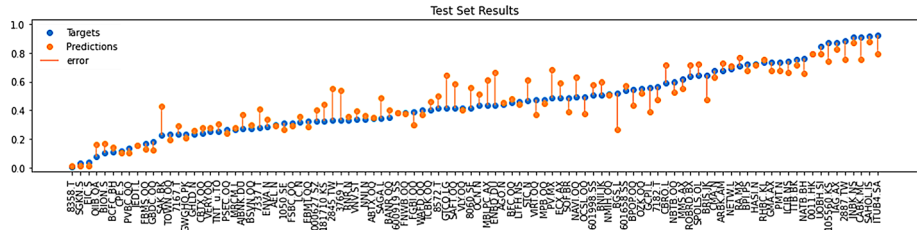
(c) Inf sector targets and predictions.

Fig. 5 ANNI E score test set predictions: companies' target ratings (predictions) are reported in blue (orange)

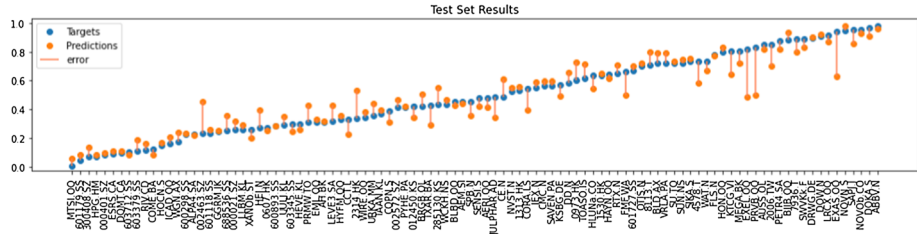
5 Interpretability

In Sect. 4, we sought an approximation of Refinitiv's ESG score model by applying machine learning techniques. In this section we focus on understanding, as thoroughly as possible, how the proprietary ESG score issuance algorithm works, exploiting what our replication models learned.

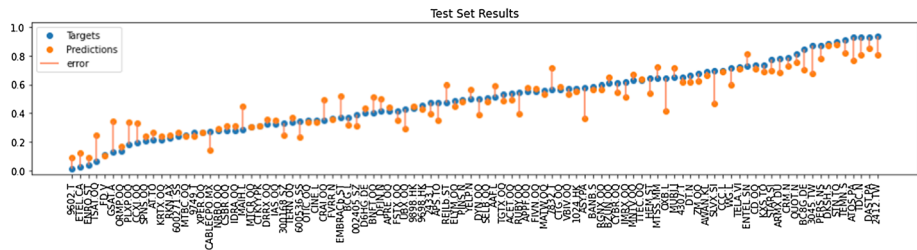
First, we apply a machine learning global interpretability method to explain the inner workings of our ANNI trained model, in such a way that we can make general assumptions about Refinitiv's proprietary assessment algorithm. We also compare the results with the ones of the Ridge regression. Then, in Sect. 5.2, we apply a local interpretability technique to understand single predictions, and therefore to motivate and justify single companies' sustainability scores.



(a) Fin sector targets and predictions.



(b) Man sector targets and predictions.



(c) Inf sector targets and predictions.

Fig. 6 ANNI S score test set predictions: companies’ target ratings (predictions) are reported in blue (orange)

5.1 Global interpretability

Figure 9 contains the ‘beeswarm’ plots, introduced in Sect. 3.1.1 as a global interpretability tools. This representation method draws a dot for each single sample feature value’s Shapley Value on the corresponding level. The dots are colored in blue or red depending on whether the value for that feature is respectively low or high relative to the variable range. In this way, we have information about the ordering of the features accordingly to their average impact, but also on their relationship with the output value since, from the dots’ colors distributions, we can understand if higher values imply higher outcomes or, conversely, if they lead to lower ratings. The plot style is called ‘beeswarm’ since dots with similar values are stacked vertically and this typically leads to plots that recall warms of bees. In Fig. 9 we display only the five most important predictors in each subgroup because, due to the large number of variables, the Shapley Values are very similar after the first two or three features. This trend is evident in the specific values documented in Figs. 13–14 in Appendix E in the supplementary materials, where we report the mean Shapley Values for the most important ten features in the SHAP bar plot.

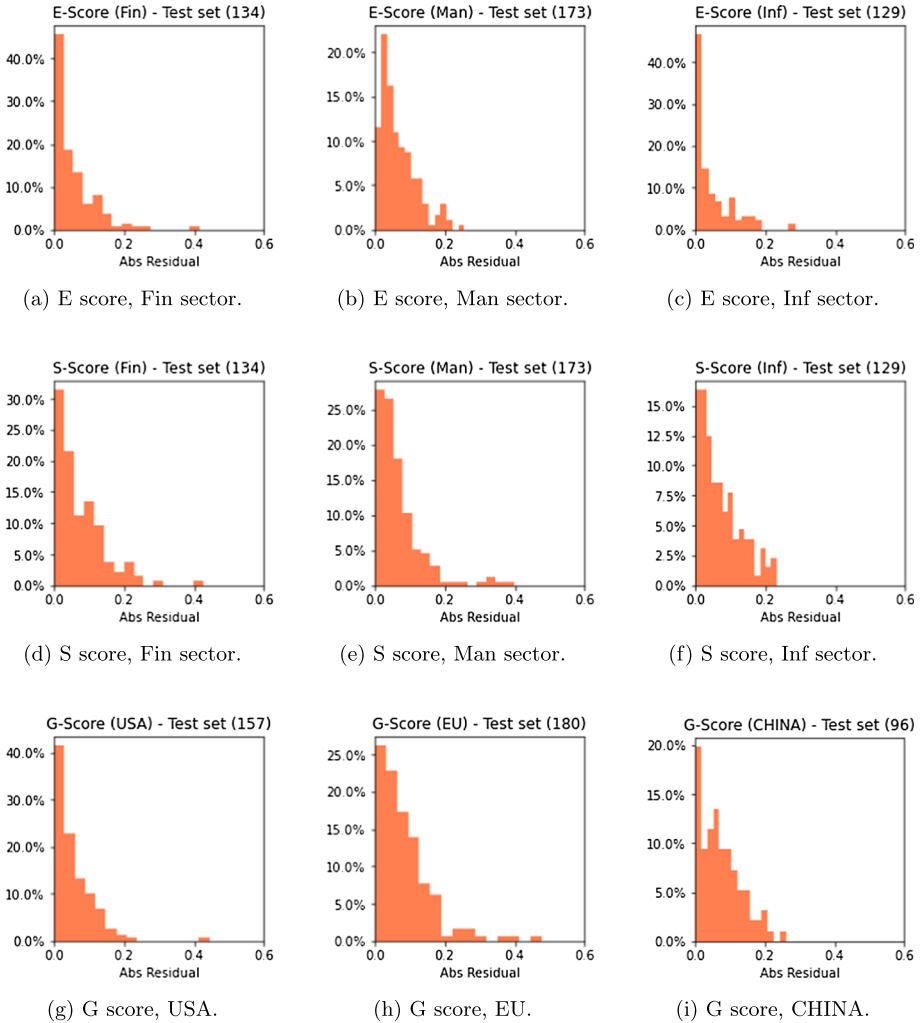


Fig. 8 ANNI test set absolute residuals histograms. We show the distribution of the absolute value of the difference between the Refinitiv score and the predicted score, considering all the firms of the test set

score is ‘Total Energy Use To Revenues USD in Million’, which is reasonable since a higher energy consumption implies lower environmental sustainability.

5.1.2 S score model interpretation

For what concerns the Social score, Fig. 9d–f show that the most relevant feature overall seems to be whether the company has a policy for the freedom of association of its employees, being the predictor with the highest average significance for the Fin and Man sectors. It is also in the top five for the Inf industry. We also notice that the presence of external awards for social, ethical, or community activities appears to have a significant impact. In the Financial sector a low-price line of products or services offered and designed for lower-

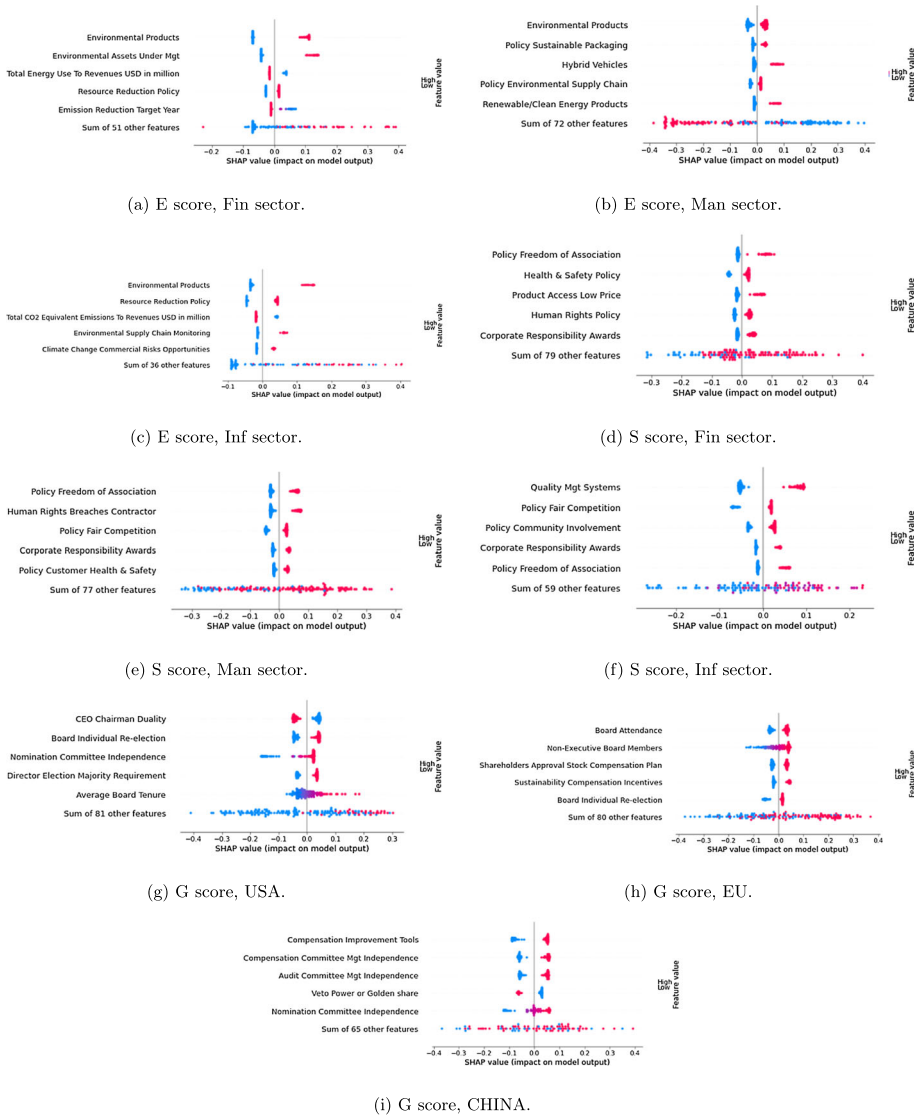


Fig. 9 ANNI SHAP beeswarm plots. We display the five most important predictors in each subgroup according to the Shapley Value analysis, and their impact on the scoring

income categories has the third greatest average impact, which is reasonable. The results show that, in the Manufacturing and Information industries, it is also important to have a policy for fair competition.

5.1.3 G score model interpretation

In the Governance score, the features that have the greatest impact, accordingly to Fig. 9g–i, vary among the different geographical regions. The only two variables shared between two regions are ‘*Nomination Committee Independence*’ and ‘*Board Individual Re-election*’.

It is interesting that, for USA, if the chairman of the board is simultaneously the CEO of the firm, then the company is penalized. We have that this is the factor with the greatest average influence. In the European region, companies that report data concerning board members’ attendance rates and firms with higher percentages of non-executive board members are rewarded. In other words, this means that the model values transparency and the presence of individuals without executive responsibilities on the company’s board. Lastly, for Chinese companies, the results show that the model rewards the presence of improvement tools for the employees’ compensation and independence elements in the compensation process of management and the audit committee definition. Additionally, from the analysis, it also emerges that the presence of veto owners lowers the company’s G score.

Figure 9 also shows that it is not possible to focus on the first five features only, since the sum of the remaining features has a non-negligible impact. Below we better analyze what happens if we only consider the most important features.

5.1.4 Features’ number reduction

Given the ordering of the ANN1 model’s features accordingly to their estimated relevance, we now want to understand whether it is possible to reduce the number of features considered by the replication model, still obtaining satisfying ESG score predictions. This could be an important information for companies, since it would result in defining a small number of features having the largest impact in the ESG score issuance, and therefore to be improved to increase their ESG scores. To answer this question, we train and evaluate a Ridge regularized linear regression algorithm to predict E, S, and G scores, considering only the ten and five most important features for the ANN1. The model’s training is performed consistently with what is presented in Sect. 3.1. In Table 5 we report the MAE, RMSE, and R^2 metrics. We also highlight the difference between the RMSE achieved considering only the ten most important attributes and the one obtained exploiting the whole set of variables. We recall that an increase in RMSE implies a loss in the prediction accuracy considering only the ten/five features. In Table 5 we also report the minimum number of features necessary to reach a R^2 (coefficient of determination) greater or equal to 0.7, that is the variation in the scores that is predictable from the features is greater or equal than 70%. For the list of the features, see Appendix E.3. The R^2 provides a measure of how well observed outcomes are replicated by the model. We notice that the E score seems to depend mainly on few features - in the Fin sector, 2 features gives a R^2 of 71%, increasing to 87.3% with 5 features (and almost no gain in adding further 5 features). On the opposite, the G pillar exhibits a strong dependence on a large number of features, especially for China, since we need 22 features to get a R^2 of 70%, while with 10 features we only get 46.7%. Results on MAE and RMSE show the same trend.

Therefore, our analysis shows that in some cases the 10 most relevant features (sometimes 5 or less than 5) present a significant predictive power. This is the case, for instance, of the S pillar considering the Fin and Inf industry sectors. Therefore, for the firms belonging to this sectors, an improvement on only few features, like ‘*Policy Freedom of Association*’, or ‘*Corporate Responsibility Awards*’, see Appendix E.3, could result in an overall improvement of the S score. On the opposite, the G pillar requires a larger number of features in order to

have a good enough prediction, and therefore, to get an improvement of the G score, firms should act on a large set of features.

The different behaviour among the sectors of the same pillar, e.g., Man sector in the E pillar with respect to the Fin and Inf ones, China in the G pillar with respect to USA and EU, suggests once again that Refinitiv scoring algorithm deeply changes moving among the sectors/countries.

5.1.5 Comparison with the ridge model

We now compare our estimation of the most important features for ANN1 with the Ridge regression's weights. We are interested in understanding whether the two models agree on the most important ESG features. For the accuracy of the two models, we refer to Tables 3 and 4. Results reported in Appendix E in the supplementary materials confirm what we have previously found from the ANN1 model's interpretation, strengthening our understanding of the underlying proprietary model. The features associated with the most significant weights are often the ones with high relevance in the ANN1's explanation results. In fact we found that, in all the considered industries and geographical regions, at least half of the top ten most impactful features for our ANN1 are also in the Ridge model's top ten.

Ridge's results confirm that '*Environmental Products*' is the most important feature in Refinitiv's Environmental pillar model. Indeed, the feature has the largest absolute weight in all three industry sectors, confirming the ANN1 interpretability results. The Ridge and ANN models' comparison is particularly interesting in the Inf sector's case for the Environmental and Social scores. Eight out of the top ten features are the same in these scenarios. Moreover, in the Inf sector for the E pillar, the top three attributes coincide exactly, confirming that the presence of a 'Resource Reduction Policy' and the total CO2 equivalent emissions variable are two major factors. Concerning the Social pillar, we notice that in the Financial and Manufacturing sectors the two models agree on seven out of the top ten attributes, confirming that a policy for the freedom of association is crucial in determining a company's Social score. Concerning the Governance factor score, the results are less consistent than the other pillars but still satisfying: in the American and European cases, the models agree on five out of the top ten attributes, six in the Chinese geographical region. The only factor present in all the feature interpretability results (except for the EU case of the ANN model) is related to diversity in the board.

We also notice that the Ridge model's weights confirm the differences between industry sectors' most relevant features we found in ANN1's interpretations. Attributes like 'Environmental Assets Under Mgt' in the financial sector, 'Hybrid Vehicles' and 'Noise Reduction' in the manufacturing sector, are deemed to be relevant only in the specific framework where it is reasonable and not in all the sectors or geographical regions. This trend highlights the differences in treatment reserved to non-industry-peer companies.

In conclusion, our analysis shows that simple linear regression algorithm, like Ridge regularization, are able to provide, in most cases, the same level of accuracy reached by more complex and capable models like ANN, see Tables 3 and 4. Moreover, the main drivers of Ridge and ANN predictions are the same. This suggests that Refinitiv ESG Scores do not require overly complex algorithms and can be replicated with simple regression models. Nevertheless, the results in Tables 3 and 4 highlight the presence of noise in the score attribution mechanism of the rating agency, which cannot be deleted exploiting more complex models.

Table 5 Ridge test-set performance measures (Root Mean Square Error—RMSE, Mean Absolute Error—MAE, and coefficient of determination— R^2) considering only the ten/five most relevant features

	E			S			G		
	Fin	Man	Inf	Fin	Man	Inf	USA	EU	CHINA
10 most relevant features									
MAE	0.066	0.112	0.053	0.076	0.085	0.077	0.105	0.114	0.127
RMSE	0.102	0.141	0.080	0.100	0.109	0.098	0.134	0.140	0.153
R^2	0.874	0.794	0.906	0.812	0.840	0.777	0.665	0.618	0.467
RMSE variation	+0.024	+0.057	+0.009	+0.011	+0.015	+0.008	+0.059	+0.037	+0.061
5 most relevant features									
MAE	0.076	0.134	0.061	0.100	0.114	0.100	0.136	0.139	0.137
RMSE	0.104	0.166	0.091	0.124	0.143	0.128	0.166	0.170	0.174
R^2	0.873	0.678	0.888	0.710	0.708	0.676	0.539	0.411	0.325
RMSE variation	+0.026	+0.082	+0.020	+0.035	+0.047	+0.038	+0.091	+0.067	+0.082
Minimum number of features necessary to reach a $R^2 \geq 0.7$									
	2	6	2	5	5	6	11	16	22

In the last lines of each subtable we report the difference between the RMSE reached by the models considering only ten features and the one obtained with the complete set of variables (RMSE Variation, that is $RMSE_{10F} - RMSE_{ESG}$ and $RMSE_{5F} - RMSE_{ESG}$). In the last line of the table we report the minimum number of features necessary to reach a $R^2 \geq 0.7$

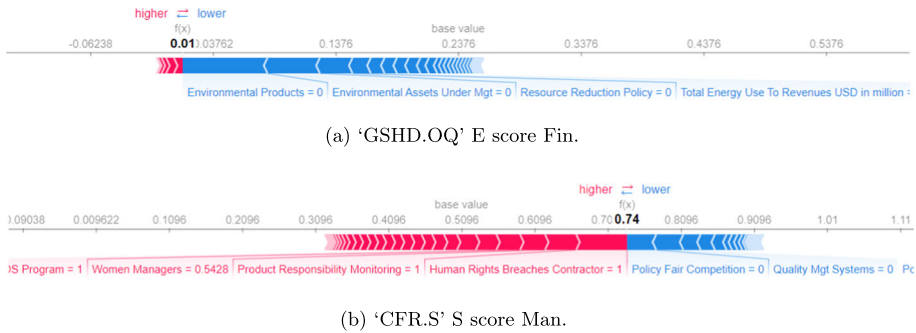


Fig. 10 Two ANNI's predictions interpretations

5.2 Local interpretability

In Fig. 10 we show the interpretation of two test set samples' predictions: we choose the samples 'GSHD.OQ', and 'CFR.S' from the samples with the best replication accuracy for the E and S score, respectively. 'GSHD.OQ' (Fig. 10a) stands for 'Goosehead Insurance Inc', an American insurance company (Fin sector). For this firm both Refinitiv's Environmental score and our model's predictions are very close to zero, 1.31% and 1.30% respectively. The model's prediction is driven from the fact that the firm does not deal environmental products ('*Environmental Products*'=0), has no environmental assets under management ('*Environmental Assets Under Mgt*'=0), and lacks a resource reduction policy ('*Resource Reduction Policy*'=0).

'CFR.S' (Fig. 10b) stands for 'Compagnie Financiere Richemont SA', a Swiss jewelry Maison and specialist watchmaker company. Refinitiv's Social score for this firm is equal to 73.71, very close to our model's result, 73.96%. This rating is mainly motivated by the fact that, despite lacking fair competition and quality of management systems policies ('*Policy Fair Competition*'='*Quality Mgt Systems*'=0), the company has a high number of women in managing roles, monitors its product from the responsibility point of view, and is ready to end a partnership with a sourcing company if human rights criteria are not met ('*Women Managers*'=0.5428, and '*Product Responsibility Monitoring*'='*Human Rights Breaches Contractor*'=1).

6 Discussions and conclusions

The purpose of this work is to replicate, understand, and explain the proprietary black-box model that rating agencies exploit to issue ESG ratings in an attempt to contrast the problem of transparency in the world of socially responsible investments and to provide investors and policymakers with a useful and effective tool to assess the reliability of sustainability scores.

In the first part of the work, we show that through the employment of machine learning algorithms, we are able to replicate Refinitiv ESG ratings with satisfying levels of accuracy in all three pillar scores. In the second part, we show that with the application of machine learning interpretability techniques we are able to explain the choices of replication models, make suppositions on which are the most relevant features in the rater's proprietary algorithm, and interpret single companies scores even when complex black-box replication algorithms are employed.

We obtain remarkable results in replicating the Environmental pillar scores, with accuracy, measured in terms of RMSE, of 0.078, 0.084, and 0.071 in the three scenarios, respectively. On the contrary, in the regression of the Social factor score, we are able to achieve an accuracy of 0.089, 0.096, and 0.090, as reported in Sect. 4. Because this difference in replication accuracy is consistent in all three considered sectors with the same conditions, we deem it reasonable to imply that the model replication task of the Social pillar is generally more complex than the Environmental one. In our opinion, the difference is likely to be rooted directly in the definition of the two pillars, with the environmental point of view being more clearly outlined and easier to measure effectively. In support of this idea, we also have that the amount of social attributes collected and employed by the data provider is significantly greater than the environmental ones, suggesting a more nebulous scope.

The Governance factor results vary on the geographical regions we consider. While we obtain a low RMSE for USA companies, equal to 0.075, we have higher outcomes for European and Chinese firms, equal to 0.103 and 0.092, respectively. We believe that the EU results can be motivated by the fact that the European region comprises many individual countries that, ideally, should have been treated separately to correctly mimic Refinitiv's methodology. We stress that we choose to consider the region as a whole since we would have too few samples to train the machine learning models with the single European countries. Similarly, the reason behind the less-than-optimal regression accuracy for Chinese companies might be the restricted number of firms we have at our disposal. However, we notice that, in the case of the Financial sector, we have a similar amount of samples and the results we obtain are significantly better. This phenomenon suggests that the model underneath the Governance score of Chinese companies requires more information to be understood, so it presents a higher level of noise.

It is also interesting to notice that, on the one hand, the Governance factor is the one for which the transparency problem is less severe since the regulatory framework is more evolved and the data are generally more accessible. The Governance attributes are strictly intertwined with information like corporate management structures, members, and decision systems, and they are consequently easier to collect and verify from the outside. This aspect emerges in particular from the lower percentage of non-populated features and the smaller amount of Governance-related attributes considered by the data provider and collected in the data-set. On the other hand, we have that, first, despite this advantage, the Governance score results do not show better levels of accuracy, and second, the number of selected features in the Governance pillar models, starting from a lower amount of available attributes, is generally higher with respect to what happens in the other factors, implying that the feature selection procedure is less effective in determining which are the relevant variables.

Two additional interesting behaviors emerge from what we have obtained. First, we see that straightforward models like the simple linear regression algorithm and the Lasso and Ridge regularization are able to provide, in most cases, the same level of accuracy reached by more complex and capable models like ANN and random forests. Second, we notice that increasing the number of parameters and the generalization power of the ANNs model by considering deeper and broader architectures does not yield better results consistently or remarkably. These two trends, in our opinion, are evidence that what we can learn from Refinitiv ESG Scores does not require overly complex algorithms and can be replicated with simple regression models.

In the end, considering the great number of granular attributes employed by the data provider and the severe problems of missing information and variable redundancy characterizing the E, S, and G data-sets, we believe that we are able to approximate Refinitiv's proprietary model with a satisfying level of accuracy. Nevertheless, the results highlight the

presence of noise in the score attribution mechanism of the rating agency. One source of noise could be due to the way the data provider solves the missing value problem, which is not fully disclosed.

In the paper, we show that it is possible to draw conclusions on the inner workings of the proprietary black-box model by applying machine learning global interpretability techniques to explain what the replication models have learned. The interpretation is straightforward when considering simple linear regression models. However, through the Shapley Values technique, we show that we can estimate each ESG attribute's relevance even when complex black-box machine learning algorithms are exploited. In particular, we find very interesting results by comparing the Ridge linear regression and our ANN1 replication model. The outcomes of the two algorithms are very consistent, agreeing on the majority of the most relevant features in each scenario. The interpretability outcomes allow us to conclude, for instance, that having an environment-dedicated line of products and services is a crucial attribute in Refinitiv's Environmental score model. This result is in line with the literature: eco-friendly products have shown a considerable impact on customers' decisions (D'souza et al., 2006). Eco-labeled products are also more likely to be developed in companies with environmental targets strategies, which can be related also to management remuneration scheme (Ullah & Nasim, 2021), proving effectiveness of sustainability-related corporate policies. Moreover, in the Inf sector for the E pillar, the presence of a '*Resource Reduction Policy*' and the total CO2 equivalent emissions variable are two major factors. The latter, in particular, has become a traditionally employed information for carbon risk studies in sustainable finance literature (Bolton and Kacperczyk, 2021; Ehlers et al., 2022; Huij et al., 2022). Concerning the S pillar, our analysis shows that a policy for the freedom of association is crucial in determining a company's Social score, confirming Anner (2012). Concerning the Governance factor score, the results are less consistent, the only factor present in almost all the feature interpretability results is related to diversity in the board, which relation with CSR and corporate performance has been widely analyzed in the literature (Carter et al., 2003; García-Meca et al., 2015; Harjoto et al., 2015). Due to the large diversity of the main drivers of the G score, our interpretability analysis suggests that comparing the G scores of firm based in different countries, e.g., China and USA, is not meaningful, since the main figures providing the scores are too different.

Thanks to this methodology, we can study how the ESG models' relevant features and weights vary across the different industry sectors and geographical regions, analyzing the specific treatment reserved for non-industry-peer companies. The Ridge model's weights confirm the differences between industry sectors' most relevant features we found in ANN1's interpretations. Attributes like 'Environmental Assets Under Mgt', in the financial sector, 'Hybrid Vehicles', and 'Noise Reduction', in the manufacturing sector, are deemed to be relevant only in the specific framework where it is reasonable and not in all the sectors or geographical regions.

In conclusion, we think that this interpretability method allows one to achieve a deeper understanding of the rating agency's issuance system and to better integrate the information provided by the sustainability performance indicator in the decision process. We also show that through the local interpretability application of the Shapley Values method, we can meaningfully motivate and explain the ratings associated with single companies.

What we find has important implications for companies, investors and market makers. Our results are specific to the Refinitiv data provider. However, our methodology can be applied in any framework where it is possible to gain access to the ESG metrics exploited in the issuance process. This technique allows firms to understand and motivate the ratings they receive from specific data providers. Moreover, it enables companies to compare the different

methodologies exploited by separate rating agencies, highlighting the most relevant features in each issuance process. This approach can be extremely useful when there is disagreement on the company's ESG score, explaining and underlying the motivations behind each rater's assessment. Furthermore, we believe this method can be a functional tool to better integrate ESG Ratings' information even for investors and, in general, for any market maker wanting to precisely assess companies' sustainability ratings.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10479-023-05514-z>.

Acknowledgements The present research is part of the activities of "Dipartimento di Eccellenza 2023-2027" and of the European Union's Horizon 2020 COST Action "FinAI: Fintech and Artificial Intelligence in Finance - Towards a transparent financial industry" (CA19130).

Funding Open access funding provided by Politecnico di Milano within the CRUI-CARE Agreement. The authors did not receive any funding support from these projects for this article.

Declarations

Conflict of interest All the Authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Anner, M. (2012). Corporate social responsibility and freedom of association rights: The precarious quest for legitimacy and control in global supply chains. *Politics & Society*, 40(4), 609–644.
- Azzone, M., Barucci, E., Moncayo, G. G., & Marazzina, D. (2022). A machine learning model for lapse prediction in life insurance contracts. *Expert Systems with Applications*, 191, 116261.
- Bams, D., & van der Kroft, B. (2022). Divestment, information asymmetries, and inflated ESG ratings. SSRN: 4126986.
- Berg, F., Fabisik, K., & Sautner, Z. (2021a). Is history repeating itself? The (un)predictable past of ESG ratings. *European Corporate Governance Institute: Finance Working Paper 708/2020*, SSRN: 3722087.
- Berg, F., Kölbel, J., Pavlova, A., & Rigobon, R. (2021b). ESG confusion and stock returns: Tackling the problem of noise. SSRN: 3941514.
- Berg, F., Heeb, F., & Kölbel, J. (2022a). The economic impact of ESG ratings. SSRN: 4088545.
- Berg, F., Kölbel, J., & Rigobon, R. (2022). Aggregate confusion: The divergence of ESG ratings. *Review of Finance*, 26(6), 1315–1344.
- Billio, M., Costola, M., Hristova, I., Latino, C., & Pelizzon, L. (2021). Inside the ESG ratings: (Dis)agreement and performance. *Corporate Social Responsibility and Environmental Management*, 28(5), 1426–1445.
- Bolton, P., & Kacperczyk, M. (2021). Do investors care about carbon risk? *Journal of Financial Economics*, 142(2), 517–549.
- Carter, D. A., Simkins, B. J., & Simpson, W. G. (2003). Corporate governance, board diversity, and firm value. *Financial Review*, 38(1), 33–53.

- De Lucia, C., Paziienza, P., & Bartlett, M. (2020). Does good ESG lead to better financial performances by firms? Machine learning and logistic regression models of public enterprises in Europe. *Sustainability*, *12*(13), 5317.
- Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple classifier systems: First international workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings I*, pp. 1–15. Springer.
- D'souza, C., Taghian, M., & Lamb, P. (2006). An empirical study on the influence of environmental labels on consumers. *Corporate Communications: An International Journal*, *11*(2), 162–173.
- D'Amato, V., D'Ecclesia, R., & Levantesi, S. (2022). ESG score prediction through random forest algorithm. *Computational Management Science*, *19*(2), 347–373.
- Ehlers, T., Packer, F., & de Greiff, K. (2022). The pricing of carbon risk in syndicated loans: Which risks are priced and why? *Journal of Banking & Finance*, *136*, 106180.
- García-Meca, E., García-Sánchez, I.-M., & Martínez-Ferrero, J. (2015). Board diversity and its effects on bank performance: An international analysis. *Journal of Banking & Finance*, *53*, 202–214.
- Grishunin, S., Bukreeva, A., Suloeva, S., & Burova, E. (2023). Analysis of yields and their determinants in the European corporate green bond market. *Risks*, *11*(1), 14.
- Harjoto, M., Laksmana, I., & Lee, R. (2015). Board diversity and corporate social responsibility. *Journal of Business Ethics*, *132*, 641–660.
- Huber, B., Comstock, M., Polk, D., & LLP, W. (2017). ESG reports and ratings: What they are, why they matter. Harvard Law School, <https://corpgov.law.harvard.edu/2017/07/27/ESG-reports-and-ratings-what-they-are-why-they-matter/>.
- Huij, J., Laurs, D., Stork, P. A., & Zwinkels, R. C. (2022). Carbon beta: A market-based measure of climate risk. SSRN: 3957900.
- Hull, J. (2021). *Machine learning in business: An introduction to the world of data science*. Amazon Fulfillment Poland Sp. z oo,
- Lin, W. L., Law, S. H., Ho, J. A., & Sambasivan, M. (2019). The causality direction of the corporate social responsibility: Corporate financial performance nexus—Application of panel vector autoregression approach. *The North American Journal of Economics and Finance*, *48*, 401–418.
- Lins, K. V., Servaes, H., & Tamayo, A. M. (2017). Social capital, trust, and firm performance: The value of corporate social responsibility during the financial crisis. *The Journal of Finance*, *72*(4), 1785–1824.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.) *Advances in Neural information processing systems*, vol. 30, pp. 4765–4774.
- Molnar, C. (2022). *Interpretable machine learning. A guide for making black box models explainable*. 2nd edition, <https://christophm.github.io/interpretable-ml-book>.
- Pástor, L., Stambaugh, R. F., & Taylor, L. A. (2021). Sustainable investing in equilibrium. *Journal of Financial Economics*, *142*(2), 550–571.
- Pedersen, L. H., Fitzgibbons, S., & Pomorski, L. (2021). Responsible investing: The ESG-efficient frontier. *Journal of Financial Economics*, *142*(2), 572–597.
- Pelizzon, L., Rzeznik, A., & Hanley, K. W. (2021). The salience of ESG ratings for stock pricing: Evidence from (potentially) confused investors. CEPR Discussion Paper DP16334.
- Refinitiv. Environmental, social and governance scores from Refinitiv, (2022). https://www.refinitiv.com/content/dam/marketing/en_us/documents/methodology/refinitiv-ESG-scores-methodology.pdf.
- Sustainability. Rate the raters 2020: Investor survey and interview results, (2020). <https://www.sustainability.com/globalassets/sustainability.com/thinking/pdfs/sustainability-ratetheraters2020-report.pdf>.
- Tamimi, N., & Sebastianelli, R. (2017). Transparency among S&P500 companies: An analysis of ESG disclosure scores. *Management Decision*, *55*(8), 1660–1680.
- Ullah, S., & Nasim, A. (2021). Do firm-level sustainability targets drive environmental innovation? insights from brics economies. *Journal of Environmental Management*, *294*, 112754.
- UN Global Compact. Who cares wins: Connecting financial markets to a changing world. Technical report, (2004).
- Wang, H., Lengerich, B. J., Aragam, B., & Xing, E. P. (2019). Precision lasso: Accounting for correlations and linear dependencies in high-dimensional genomic data. *Bioinformatics*, *35*(7), 1181–1187.
- Zanin, L. (2022). Estimating the effects of ESG scores on corporate credit ratings using multivariate ordinal logit regression. *Empirical Economics*, *62*(6), 3087–3118.