UNIVERSITÀ
DI PAVIA

Department of Economics and Business Sciences

Master's Degree Course in Finance

ESG and Financial Performance: a Machine Learning Analysis on European Companies

Supervisor:

Prof. Paolo Giudici

Candidate:

Guglielmo Zatti

Student ID: 543611

A. Y. 2024/2025

Abstract

In recent years, ESG (Environmental, Social, Governance) themes have seen their role increasing with international and European entities releasing several initiatives to encourage the adoption of sustainability practices by companies and the disclosure of their performance regarding environmental, social and governance aspects of their business activities. In this new scenario, hugely centered around the concept of sustainable finance, ESG metrics have assumed a key role as they provide investors with a benchmark to evaluate the risks and opportunities linked to their different investment opportunities. However, there is still uncertainty revolving around the ESG scores. Indeed, it is unclear which are the determinants of such ESG metrics. In particular, the extent to which traditional financial variables or sectoral affiliation can explain firms' sustainability performance is still subject to debate. This thesis investigates the predictability and determinants of the Bloomberg ESG score for a sample of 600 European firms included in the STOXX Europe 600 index. The analysis applies a range of machine learning techniques, including linear regression, logistic regression, regression trees, random forests, gradient boosting, and neural networks, to predict the ESG score and its three individual components (E, S, and G), using financial variables and sectoral dummies as explanatory features. The study pursues two main objectives: first, to evaluate the predictive performance of different models; and second, to assess their interpretability to identify the key drivers of ESG performance. Model performance is evaluated through out-of-sample metrics while interpretability is examined using both model-specific approaches and SHAP values. Moreover, a bootstrap out-of-bag procedure is applied to verify the robustness of the results of both predictive performance and interpretability. The results show that the overall ESG score, as well as the social and governance components, exhibit limited predictability based on the selected variables. In contrast, the environmental score demonstrates higher predictive accuracy and a more stable interpretability structure. Tree-based models, particularly gradient boosting, consistently outperform alternative approaches, highlighting the importance of non-linear relationships. By the way, the general findings suggest that ESG scores, especially the social and governance dimensions, are difficult to explain through firms' economic characteristics and their sectoral affiliation while they also outline the importance of the methodology's choice.

Introduction

In the most recent years, ESG (Environment, Social, Governance) practices have gained an increasingly crucial role in the economic and financial scenario worldwide. In this sense, the attention towards the sustainability of companies' business activities has grown significantly both from the part of investors and public institutions with the goal of promoting more sustainable and responsible development models.

This increasing relevancy of these themes linked to sustainability has been fostered globally by a series of initiatives and international organizations. For instance, the United Nations has issued Sustainable Development Goals (SDG). The SDGs involve organization and people around the world and are expected to be accomplished effectively by 2030 (Junius et al., 2020). Moreover, the increasing number of ESG practices led to the birth of several global reporting initiatives, such as Global Reporting Initiatives and the United Nations Global Compact. Lastly, the Principles for Responsible Investment (PRI), supported again by the United Nations, represent another initiative aimed at encouraging the integration of environmental, social and governance (ESG) elements into investment decisions (Kim & Yoon, 2022).

Sustainability initiatives have begun taking place also in the European territory, where the EU has moved towards the sustainable practices by proposing in 2014 the European Non-Financial Reporting Directive, known also as Directive 2014/95/EU, which requires all listed companies, along with other non-listed but above certain levels of capital, sales and employees, to include into their financial reports also a sustainability statements. The Directive has established that the reports should contain essential information regarding the company's pledges and achievements on various issues about the environmental, social and governance topics that are key elements for the business of each company (Cicchiello et al., 2022). Moreover, the NFRD has introduced for the very first time a standardization of how reports must be done according to specific frameworks. This is an essential point to make them comparable and guarantee transparency. Further development of this Directive, the Corporate Sustainability Reporting Directive (CSRD),

released on 21 April 2021, has updated the reporting requirements of Directive 2014/95/EU, supported by the increasing need for sustainability disclosure (Pulino et al., 2022). Furthermore, European Union Sustainable Finance Disclosure Regulation (SFDR), proposed in 2021, has pushed investors to be more aware of the risks deriving from investing in the less sustainable firms, thus favoring high-rated ESG firms (Gonçalves et al., 2023)

In this context, sustainable finance has experienced a significant growth, both globally and in Europe, as now an increasing number of investors have begun integrating the ESG criteria in their investment decisions since they recognize the fact that certain environmental, social and governance factors could really affect the risks of their portfolios as well the long term outcomes of the firms performance. In this sense, ESG metrics serve as an essential instrument to measure firms' nonfinancial performance in terms of sustainability and provide a clear benchmark for investors.

However, there are still some doubts regarding the ESG scores, mainly revolving around the way through which they are measured by rating agencies. It is also unclear which are the determinants of such ESG metrics. In particular, the extent to which traditional financial variables can explain firms' sustainability performance is still subject to debate. In this context, machine learning techniques could provide a solid solution since they represent a more flexible approach in this kind of research with respect to traditional models, making their employment highly convenient.

However, while prior literature has extensively investigated the relationship between ESG and financial performance, fewer studies have focused on predicting ESG scores themselves and identifying their key drivers. For this reason, this work aims at connecting with the second type of literature mentioned by utilizing a set of machine learning algorithms in the attempt to predict the Bloomberg ESG score and its three pillars, environmental (E), social (S) and governance (G) scores. More in detail, the analysis is based on a sample of 600 European firms included in the STOXX Europe 600 index and uses their financial data for the year 2023, such as total assets, return on assets, EBITDA, total equity, asset turnover and debt-to-equity ratio, plus ten different sectoral dummies, representing the sector to which each firm belongs, as predictors of the underlying scores.

In other words, the work tries to discover if European companies' sustainability performance is affected by their economic characteristics or their affiliation to a specific sector.

In this sense, the study will rely mainly on two aspects of machine learning techniques: their predictive performance, which specifically refers to the capacity of each model to correctly predict the ESG score and its three pillars, and their explainability or interpretability, which refers instead to identifying the features that contribute the most to the predictions of the various scores in each model and common patterns among different methods. The robustness of both elements is then tested through a specific procedure.

Hence, the work is organized as follows: the first chapter focuses on the ESG metrics, trying to define the ESG score and its pillars and the reasons for which they are becoming increasingly important for firms and investors through the main theoretical frameworks. It also gives more information regarding the uncertainties of the ESG metrics adopted and lastly provides a brief review of some works from the literature to which this work relates.

The second chapter instead talks about the methodology adopted in this work: first it provides a definition of the machine learning techniques employed and then it briefly summarizes the tools used to access the accuracy of the predictive performance and the interpretability of the utilized methods as well as the procedure to test the robustness of these two elements.

Moving on, the third chapter contains the description of the dataset used for the analysis, explaining all the variables employed and the way data have been managed before starting with the proper analysis. Lastly, the chapter provides a brief early exploratory analysis by computing a correlation matrix between the numerical variables in the dataset to observe their relationships.

After that, the fourth and fifth chapter present and describe the results regarding respectively the predictive performance and the explainability of the ML methods first on the ESG and then on its three components E, S and G scores, focusing on the comparison across different models and scores. The robustness of these findings is also included in the two chapters.

Finally, the conclusion section summarizes the work, mainly the results of the last two chapters.

Chapter 1: ESG metrics

1.1 Definition and importance of ESG metrics

The concept of ESG score refers to a set of sustainability indicators used to assess the extent to which a company operates in an environmentally and socially responsible manner while maintaining sound governance practices (Tahmid et al., 2022). More specifically, the ESG framework evaluates corporate performance across three main dimensions: environmental (E), social (S), and governance (G).

The environmental dimension concerns the impact of firms' activities on natural systems and includes aspects such as climate change, energy use, pollution, waste management, and biodiversity preservation. The social pillar focuses on the relationships between firms and their stakeholders, including issues related to human rights, working conditions, health and safety, equality, and the firm's role within the broader community. The governance component, instead, relates to the internal structures and processes that guide corporate decision-making. It includes elements such as board composition and independence, ownership structure, transparency, disclosure practices, internal controls, and the prevention of unethical behaviors such as corruption or tax avoidance (Tahmid et al., 2022; Demiraj et al., 2025). More broadly, ESG factors can be interpreted as those environmental, social, and governance elements that may influence the financial performance, risk profile, and long-term sustainability of firms and other economic entities (European Banking Authority, 2021).

Although ESG factors are not strictly financial in nature, they can significantly influence firms' risk profiles through their impact on sustainability and ethical conduct, ultimately affecting financial and market performance. For instance, the environmental dimension can affect firms both directly and indirectly. On the one hand, physical events such as extreme weather conditions may disrupt business operations and supply chains. On the other hand, the transition toward more sustainable economic systems may require firms to adapt their business models, potentially affecting costs and profitability.

The social component also plays a key role, as it relates to firms' interactions with employees, customers, and communities. Compliance with standards concerning labor rights, health and safety, and social responsibility can have significant implications for operational continuity. Stricter regulations or inadequate practices may lead to increased costs or reputational damage, potentially affecting firm performance.

Finally, governance factors concern the internal structures and processes that ensure accountability, transparency, and ethical behavior. Weak governance practices may undermine investor confidence and damage corporate reputation, with negative consequences for both market performance and long-term sustainability (Demiraj et al., 2025; European Banking Authority, 2021).

Hence, firms should always take into account the ESG metrics while running their business activities because that could really lead to overall better performance and valuation over time.

1.2 Theoretical frameworks

There are several theoretical frameworks that may help to understand the potential effects of EGS initiatives and disclosure on firms' performance and value. The most used and considered are stakeholder theory, agency theory, legitimacy theory and signaling theory

Stakeholder theory, originally developed by Edward Freeman, emphasizes that firms should not focus exclusively on shareholders, but rather consider the interests of a broader set of stakeholders. These include, among others, employees, customers, suppliers, investors, local communities, and public institutions, all of whom can influence or be affected by corporate activities (Junius et al., 2020; Garrido-Ruso et al., 2024). From this perspective, firms are encouraged to adopt management practices that go beyond pure financial objectives and incorporate social and environmental considerations. In this context, ESG initiatives and disclosure practices play a crucial role, as they allow companies to address stakeholders' expectations and strengthen their relationships with them. The growing awareness of sustainability issues has increased stakeholders' sensitivity toward ESG performance. For example, employees may prefer to work for

companies that adopt environmentally responsible practices, while customers increasingly value transparency in sustainability reporting. At the same time, regulatory authorities are playing a more active role in enforcing ESG-related standards, potentially imposing restrictions on firms that fail to comply (Shalhoob et al., 2022). Consequently, stakeholder theory suggests that firms are increasingly incentivized to adopt and disclose ESG practices to improve their competitiveness and build solid relationships with their stakeholders.

Agency theory focuses on the relationship between principals and agents, as well as on the separation between ownership and control within firms. In this framework, the principal, typically the shareholders, delegates the decisions to the agent, namely the manager, who is expected to act in the principal's interest. However, due to differing incentives, managers may pursue their own objectives, potentially at the expense of shareholders (Pulino et al., 2022). This divergence of interests can give rise to agency problems, often caused by information asymmetry between the two parties (Demiraj et al., 2025). In the context of ESG, agency conflicts may manifest when managers allocate firm resources to sustainability initiatives for private gain. This can result in overinvestment in ESG projects or in the strategic use of such initiatives to mask weak financial performance. In this regard, transparent ESG disclosure plays a crucial role in mitigating information asymmetry and reducing agency costs. By providing reliable information on both financial and non-financial performance, firms enable shareholders to better monitor managerial actions. Moreover, improved ESG disclosure can help reduce exposure to various risks, including environmental liabilities, legal issues, and reputational damage. Overall, a higher level of ESG disclosure enhances transparency and accountability, thereby aligning the interests of managers and shareholders and contributing to a reduction in agency-related inefficiencies (Helfaya et al., 2023).

Thirdly, signaling theory is focused on the role of information in business transactions that can be shared to reduce information asymmetry (Pulino et al., 2022). It suggests that one of the main goals of companies is to deliver a "signal" to external parties regarding the goodwill of the companies. This is usually achieved by disclosing more information, especially ones related to companies' sustainable activities (Junius et al., 2020). In other

words, managers can reduce information asymmetry by sharing voluntary information with external stakeholders. More in detail, companies are willing to invest their financial resources to disclose important information about their sustainability commitments to provide stakeholders with information that cannot be traced in other ways. According to this theory, ESG disclosure information is hence used as a tool to provide voluntary information on sustainability efforts and disclosure of ESG performance indicators. Therefore, firms use it to show their sustainability achievements, legitimize their existence, and maintain or regain their corporate reputation (Helfaya et al., 2023).

Lastly, legitimacy theory, which is connected with the previous one, supposes the existence of a "social contract" that incentivizes companies to be socially accepted by the community in which they operate or, in other words, "legitimate" their actions (Junius et al., 2020). This social contract can then be seen as something that binds the companies to the society where they are operating. This link, from an ESG perspective, drives them to engage in sustainable development activities because the public interest is rising with respect to the sustainability business. From this point of view, by using ESG reporting, organizations show their compliance with societal norms to the public. Therefore, legitimacy theory is an important motivator for companies to disclose more ESG information to legitimize their existence and achieve sustainable growth through the social acceptance of their communities (Helfaya et al., 2023). Moreover firms, by acting in this way, are also able to improve their financial performance by gaining the approval of the broader society and being perceived as socially responsible entities (Demiraj et al., 2025).

1.3 The uncertainty of ESG metrics

In this new framework of the financial system, heavily centered around sustainability practices, ESG metrics have become a crucial factor because they represent the main tools to measure the sustainability performance of the firms. In this way, they become essential to help investors take the right decisions by making them aware of the risks and opportunities linked to the sustainability of their different investment options.

However, despite their increasing importance, there is still a certain degree of uncertainty revolving around the measurement of such metrics. There are indeed still some barriers,

such as the low availability of reliable and comparable data, as well as the cost of data and the limited access to the necessary resources to conduct this kind of analysis. These problematics make it very difficult to measure and manage the potential environmental, social, and governance risks and opportunities (Belkhiria et al., 2025).

Moreover, ESG data are often characterized by significant heterogeneity and lack of standardization, leading to substantial divergence across rating providers (Berg et al. 2022). Indeed, rating agencies, like Bloomberg or Refinitiv for instance, given their different methodologies adopted for measuring the metrics, can display a great discrepancy of ESG scores for the same companies, with even a low level of correlation between them. These alternative definitions of ESG score lead to the identification of different investment universes and consequently to the creation of different benchmarks, thereby affecting sustainable investments (Billio et al., 2020).

Nevertheless, these indicators remain widely used in both academic research and practice, making it essential to investigate their determinants and predictive drivers. In this sense, a machine learning analysis may help to better understand the hidden contributors and patterns of the ESG metrics that are sometimes difficult or impossible to detect using traditional linear analysis methods (Belkhiria et al., 2025). In this sense, this work attempts to use ML learning methods to discover whether the economic characteristics of the companies or their sectoral affiliation can represent important factors of the ESG score and its three pillars.

1.4 Literature review

This work tries to relate to the literature revolving around the prediction of the ESG metrics, which despite not being as popular and rich as that trying to

For instance, D'Amato et al. (2021) investigate if structural data as balance sheet items and income statements items for traded companies affect ESG scores through machine learning techniques, especially with the random forest. To pursue this goal, they consider the Bloomberg ESG scores and balance sheet data of a subset of companies listed in the STOXX Europe 600 Index. Thei main finding is that balance sheet items have a significant predictive power on the ESG score and that the random forest reaches a better predictive

outcome compared to the classical regression approach based on GLM, demonstrating the ability to capture the non-linear pattern of the predictors.

Secondly, Raza et al. (2022) artificial intelligence based on different machine learning techniques, including random forest, decision tree, artificial neural networks (ANN), K-nearest neighbor (KNN), naive Bayes, support vector machine (linear), support vector machine (radial basis function), and various degrees of polynomial regression, to assess how balance sheet and income statement data impact the Thomson Reuters ESG pillar score for non-financial public companies of USA, UK, and Germany from 2008 to 2020. The main finding is that the balance sheet and income statement are crucial in explaining the ESG score, and the ANN algorithm outperforms the others with minimum RMSE and MAE values.

Lastly, Del Vitto et al. (2023) utilize machine learning algorithms to replicate the Refinitiv ESG ratings in all three pillar scores, with satisfying levels of accuracy, and interpretability instruments to understand which are the most important features for such models. Their analysis underscores the relevance of interpretable models in ESG research by examining how different predictors contribute to ESG outcomes within machine learning frameworks, thereby addressing the complexity and multidimensionality of ESG metrics.

As previously discussed, this study keeps the approach of the works described above while expanding it: indeed, this study adopts different types of machine learning methods to see which of them better performs at predicting the ESG score and its pillars by comparing their performance metrics. Moreover, it relies on financial variables as predictors, but it includes also sectoral dummies to verify if the sectoral affiliation as well could affect the ESG performance of the firms. Variables of different nature are instead excluded. Lastly, even the models' interpretability plays a huge role in this study, with explainability techniques being adopted to understand which features among financial and sectoral variables emerge as most important contributors to the predictions of different models and if there are some common patterns between them.

Chapter 2: Methodology

2.1 Models of Machine Learning

Machine learning (ML) is a set of data analysis techniques through which artificial intelligence systems can identify patterns, support decision-making processes, and improve their performance over time through approaches based on data. ML techniques can be applied to both structured and unstructured data with the purpose of extracting meaningful patterns and insights. As a result, they can support decision-making by reducing uncertainty and providing useful inputs for problem-solving (De Lucia et al., 2020).

ML methodologies are generally divided into two main groups: supervised and unsupervised learning models. In supervised learning, models are trained on datasets in which observations are associated with known output variables, allowing the algorithm to learn the relationship between inputs and outputs (Palynska et al., 2024). Common applications include classification and regression tasks, where the dependent variable represents either a categorical label or a continuous value (Hu et al., 2020). In this context, the objective is to capture the relationships among variables and to derive a general rule that can be used for prediction.

Unsupervised learning, by contrast, does not rely on predefined output variables. Instead, the algorithm seeks to identify patterns, similarities, or structures within the data autonomously. The goal of these models is therefore to uncover underlying relationships among the observed variables and to detect relevant patterns in the dataset (Hu et al., 2020).

This work relies exclusively on supervised learning techniques as it tries to predict the ESG score and its three pillars E, S and G, which represent thus the target variable, through six different representative financial variables and ten sector dummies, intended then as

explanatory variables. A brief description of the specific models employed in the analysis is provided in the next section.

2.1.1 Linear models: Linear Regression and Logistic Regression

Linear regression is a supervised learning model that assumes a linear relationship between a dependent variable and one or more independent (explanatory) variables. In its general form, the model expresses the dependent variable as a linear combination of the explanatory variables and an error term, which captures unobserved factors affecting the outcome (Wang et al., 2025). The effect of each independent variable on the target variable is represented by its corresponding coefficient. These coefficients are commonly estimated using the ordinary least squares (OLS) method, which minimizes the sum of squared residuals between observed and predicted values (Del Vitto et al., 2023). The resulting model is highly interpretable, as each coefficient can be understood as the marginal effect of a given explanatory variable on the dependent variable, holding all other variables constant (Wang et al., 2025).

On the other hand, logistic regression is the only classification model used in this study. It can be seen as an extension of the linear regression since it uses the same function form. However, there are some differences, mainly regarding the estimation of coefficients. In particular, logistic regression uses a sigmoid function, which approximates the prediction output in a range from 0 to 1 (Chase et al., 2022). In the case of binary classifications, like in this study, prediction is basically the probability of either class to happen. It is also important to state that, since the ESG metrics considered in this study are continuous, four different binary variables for the ESG score and its three pillars have been created by using as thresholds the medians of such scores inside the analyzed data sample.

2.1.2 Nonlinear tree models: Regression trees, Random Forest, Gradient Boosting

Regression trees are machine learning models that recursively split the data analyzed into increasingly similar groups based on specified criteria until a stopping rule is reached. In particular, the model starts with all the data gathered in the so-called root node of the tree,

and they are then divided repeatedly according to a certain decision test, with a child node being generated for each block of the partition. Once the tree reaches its stopping condition and quits splitting, a prediction is made based on the values contained in the terminal leaf nodes (Hu et al., 2020; De Lucia et al., 2020).

The random forest, first proposed by Breiman, is essentially an extension of decision trees. It is in fact based on the principle of ensemble learning as it combines multiple decision trees to obtain a more robust ML model (Lin & Hsu, 2023). The regression trees are trained on different random subsets of the available data and random subsets of the input variables from the initial training data set. This means that each tree is given a different subset sampled from the original data set with replacement (Chase et al., 2022; Scornet et al., 2015). Moreover, only a random subset of the data characteristics is chosen for each node division, increasing the variability of the trees. Finally, an overall prediction is achieved as an average of the outputs of all the single trees (Palynska et al., 2024). This procedure makes the random forest a more robust model than many others as it can reduce effectively the risk of overfitting that may occur when using a single regression tree for instance (De Lucia et al., 2020)

Lasty, gradient boosting is another ensemble learning algorithm which, just like random forest, is based on the display of several decision trees. The main difference lies in the way through which such trees are generated: this time, instead of training multiple trees on random subsets, each of them is now trained on the error of the previous ones. In other words, the gradient boosting, rather than minimizing the total error on random trees, builds each new model so that it minimizes the combined error of the former ones (Chase et al.,2022; Zhang & Zhao, 2025). The final prediction is then obtained as a weighted average of the outputs of the single trees, where every of them is weighted according to its capacity to reduce the error. Gradient boosting is particularly effective in managing highly complex data, but it is more exposed to potential overfitting problems compared to the random forest (Palynska et al., 2024).

2.1.3 Neural Networks

Neural networks are a class of ML models inspired by the structure and the working of the neurons in human brains. A generic neural network is characterized by three main elements: neurons, hidden layers, and output layers.

Neurons are put together in various layers, with outputs of one layer becoming the inputs of the next one. In particular, each neuron performs a simple learnable calculation, typically a weighted sum of its inputs, to obtain the output and then applies a non-linear activation function, allowing the network to capture complex and non-linear relationships between variables (De Lucia et al., 2020). Lastly, the output layer takes the former layers' output as its input to derive the final prediction of the network, usually through a linear regression function (Wang & Ji, 2025; Hu et al., 2020).

However, neural networks are often characterized by a lower interpretability compared to more traditional models such as linear regression or decision trees. It is also important to state that this work will deal with neural networks with just one hidden layer, which can be seen as an extension of linear basis model.

2.2 Evaluation methods of the models' performance

The first main objective of this work is to investigate the predictive performance of the adopted models. In this sense, the goal is to understand if the considered ML methods generally can effectively predict the ESG score and its three pillars trough financial data and sectoral dummies. Secondly, the purpose is also to establish which of the different techniques offers the best performance compared to the others. In order to achieve these tasks, some specific evaluation metrics were adopted such as the root mean square error (RMSE), the mean absolute error (MAE), and the coefficient of determination, also referred to as R squared score (R^2 score).

The RMSE measures the average magnitude of the prediction errors. In other words, it captures the average discrepancy between observed and predicted values. Hence, the lower the value of the RMSE of a model is, the better the accuracy of that model will be. On the other hand, the MAE measures the average absolute difference between the actual and predicted values of the model. This means that it focuses on the average length of

prediction errors regardless of whether they are positive or negative. Compared to the RMSE, MAE results are less sensitive to large outliers (Patel et al., 2026: Lin & Hsu, 2023).

Lastly, the R^2 score indicates instead the proportion of the variance explained by the model, or better the extent to which the predicted values explain the actual values of the target variable. It is usually included in a range that goes from 0 to 1 and so, the closer its value gets to 1, the more the formers are closer to the second ones (Lin & Hsu, 2023).

The only exception is given by the logistic regression, which, being a classification model, relies on different tools to measure its predictive performance. The most commonly used are accuracy and secondly the AUC (area under the curve) and ROC (receiver operating characteristic) curve. The first one refers to the proportion of all correctly predicted samples with respect to the total sample. ROC represents instead graphically the trade-off between the true positive rate and the false positive rate while the AUC quantifies the area under the ROC curve and provides a scalar measure of the model's predictive capability (Zhang & Zhao, 2026).

2.3 Models' interpretability

In addition to predictive performance, a second objective of this study is to guarantee the interpretability of the employed models. This implies identifying the variables that contribute most to their predictions. In other words, the gal is to understand which financial characteristics or sectors mostly affect the sustainability performance of the companies

To be fair, most ML models already have mechanisms the determine the features' importance: for example, linear models like linear and logistic regression are explainable through their own parameters, more specifically their coefficients, while nonlinear tree models like regression trees, random forest and gradient boosting can rely on the feature importance plot, which attributes to each variable a value according to the number of splits that they contribute to generate. Lastly, neural networks can be made explainable by permutating and reshuffling the values of the independent variables (Calzarossa et al., 2025).

However, another important aspect in this field is to ensure comparability between the models in terms of features' importance through a common measure. To address this issue, SHapley Additive exPlanations (SHAP), a model agnostic tool, are employed. This technique, inspired by the Shapley values, derives from cooperative game theory, where the players represent the explanatory variables while rewards are the contributions that each of them gives to the prediction of the model. In this context, Shapley values of each variable are computed as the weighted impact that they produce on the model's output (Del Vitto et al., 2023; Palynska et al., 2024; Zhang & Zhao, 2026). Indeed, after Shapley values are calculated locally for every observation, a global contribution for each variable to the model's prediction can be obtained by averaging them over all observations (Calzarossa et al., 2025).

Additionally, to investigate the consistency in terms of explainability across different models and scores, correlation matrices based on normalized SHAP values are computed. In particular, the Spearman rank correlation coefficient is used to measure the degree of association between the feature importance rankings obtained from each model.

The choice of Spearman correlation is motivated by its non-parametric nature, as it evaluates the strength and direction of a monotonic relationship between two variables based on their ranks rather than their absolute values (Spearman, 1904). This makes it particularly suitable in this context, where the objective is to compare the relative importance of features across models rather than their exact numerical contributions.

The resulting correlation matrices allow for a direct comparison of the ranking structures produced by different modeling approaches, highlighting whether models identify similar or divergent patterns in terms of feature importance. High correlation values indicate a strong agreement between models in the ranking of predictors, while low or negative values suggest substantial differences in how models interpret the underlying relationships.

Overall, this analysis provides additional insight into the stability and robustness of the interpretability results, complementing the evidence obtained from the SHAP values and allowing for a more comprehensive comparison across methodologies.

2.4 Robustness of the results

Finally, the study also attempts to verify the robustness and consistency of the results regarding the predictive performance and the interpretability of the models involved. In order to pursue this goal, a bootstrap out-of-bag (OOB) procedure is implemented.

Bootstrapping consists of repeatedly generating training samples from the original dataset through sampling with replacement, so that each sample excludes a subset of observations (Breiman, 1996). These excluded observations, referred to as out-of-bag samples, can be used as a validation set to evaluate model performance.

This approach allows for a reliable out-of-sample assessment of predictive performance, while also providing a framework to examine the stability of model interpretability measures, such as SHAP-based feature importance. In particular, in the case of this work, a bootstrap OOB procedure of 100 iterations is implemented to verify the robustness of the results regarding the models' predictive performance through the computation of the mean values of the RMSE and the R^2 along with their respective standard deviations. On the other hand, a bootstrap OOB of 50 iterations is executed to assess the robustness of the results in terms of models' interpretability by calculating the average normalized SHAP values which are used again to construct the correlation matrices. Further information about the procedures will be given in the corresponding sections.

Chapter 3: Data and sample

3.1 Dataset description

To pursue its aims, this analysis considers 600 European firms included in the STOXX Europe 600 index. In particular, the STOXX Europe 600 index, which represents one of the main benchmarks for European equity markets, contains large, medium, and small capitalized companies from 17 countries that belong to the European region. Indeed, besides the countries from the Eurozone, the index also refers to Great Britain, Switzerland, and Scandinavian countries (D'Amato et al., 2021).

Secondly, the study has been conducted by using cross-sectional data from the year 2023, including firm-level information on ESG performance and financial characteristics. The choice of a single-year cross-section allows for a consistent comparison across firms, focusing on contemporaneous relationships between ESG performance and financial variables.

The result is a data set made by 600 observations of 13 initial variables, all taken from a Bloomberg database, referred to as "ESG_600_EUR". Out of those 13 variables, the first three, called "Ticker", "Name" and "Sector" in the sample, are nominal categorical variables representing respectively ticker, name and the sector to which each firm belongs.

The next four, named "ESG_SC", "E_SC", "S_SC" and are instead numerical continuous variables indicating the ESG score and its three pillars environment, social and governance scores obtained through Bloomberg's measurement.

The dataset includes also five other numerical continuous financial indicators, referred to as "TOT_ASS", "ROA", "EBITDA", "TOT_EQ", "ASS_TUR" and "D/E_RATIO", representing respectively total assets, return on assets, EBITDA, total equity, asset turnover and debt to equity ratio of the companies in the data set. They are used as proxies for some

economic characteristics of the firms. Both ESG and financial variables are discussed more deeply in the following sections.

It is also important to underline that not all variables in the data set have 600 observations each since some of them present some missing values.

3.2 Target variables: Bloomberg ESG scores

As previously said, the dependent variables considered in this study are the ESG score and its three individual components, environmental (E), social (S), and governance (G) scores, provided by Bloomberg. Such indicators from Bloomberg, included in a range from 0 to 10, are continuous numerical variables which measure specifically a company's exposure and management of financially relevant, industry specific environmental and social issues and opportunities and even governance policies and practices (Bloomberg, 2025). Moreover, the procedure through which they are estimated is extremely transparent: the scores are based on publicly available information disclosed by companies. They consider also the disclosure of quantitative data as a parameter of performance. They reflect thus not only firms' ESG performance but also their level of transparency in reporting sustainability-related information.

These indicators were used as target variables for all models employed except for the logistic regression, for which, as previously said in Chapter 2, four binary variables, "ESG_bin", "E_bin", S_bin" and "G_bin", have been appositively created using as thresholds the median values of the four respective scores.

3.3 Explanatory variables: financial indicators and sectoral dummies

Looking instead at the features of the models, they are basically divided into two main groups: the first one consists of five financial indicators, each of them reflecting a different key economic characteristic of the considered firms.

These variables include total assets, return on assets (ROA), EBITDA (Earnings Before Interest, Taxes, Depreciation and Amortization), total equity, asset turnover, and debt over equity ratio (D/E ratio). Total assets measure the total amount of assets owned by a firm

and serve as proxy for companies' size and market power. Secondly, return on assets measures a firm's capacity to generate profit from its asset base and is used as a proxy for companies' profitability. Thirdly, EBITDA reflects earnings before the impact of financing and accounting decisions and serves as a proxy for operational efficiency by excluding non-cash charges. Total equity instead represents the firm's capital base and provides information on the shareholders' stake in the company. It is used as a proxy for the net worth of the considered companies. In addition, D/E ratio is used to measure firms' financial leverage, indicating the extent to which a firm relies on debt relative to equity financing. Finally, asset turnover captures the ability of a firm to create revenues from its assets and serves as a proxy for companies' productivity. Taken together, these variables provide a comprehensive representation of firm size, profitability, financial structure, and efficiency, which are key factors potentially influencing ESG performance.

On the other hand, to consider also the sector to which each firm belongs, a set of sectoral dummy variables has been created and included among the explanatory variables. This choice has been made with the goal of understanding if the affiliation of firms among certain industries can influence firms' ESG performance. Firms are classified according to the Global Industry Classification Standard (GICS), which groups companies based on their primary business activities. In this way, the 600 companies in the data set have been assigned to one of eleven different sectors: financials, real estate, materials, industrials, health care, communication services, information technology, consumer staples, consumer discretionary, energy, and utilities. The precise distribution of the companies among the 11 different sectors is shown in Table 1.

Sector	Number of companies	Share of total
Financials	125	20,83%
Real estate	30	5,00%
Materials	48	8,00%
Industrials	131	21,83%
Health care	52	8,67%
Communication service	29	4,83%
Information technology	26	4,33%
Consumer staples	45	7,50%
Consumer discretionary	62	10,33%

Energy	21	3,50%
Utilities	31	5,17%
Total	600	100,00%

Table 1: sectoral distribution of the companies included in dataset

Through this procedure, for each sector, a binary variable has been constructed taking the value of 1 if the firm belongs to the corresponding sector and 0 otherwise. However, to avoid multicollinearity in linear models, one sectoral dummy representing the communication service sector is omitted. Even though it is not required for tree models and neural networks, this restriction is made also in those cases to ensure a more coherent evaluation in terms of predictive performance and explainability across different models.

3.4 Data processing

Prior to model estimation, several data preprocessing steps are implemented to ensure the quality and correctness of the analysis. The first one refers to the management of the missing values of the numerical variables in the data set, which are handled by substituting the missing observations with the median value of the corresponding variable. This approach is preferred over mean imputation as it is less sensitive to extreme values and outliers.

Subsequently, the dataset is divided into a training set and a test set using a 70–30 split. In this way, a training and a test sample are created, including 70% and 30% of the available observations. The models are then trained on the training data, while their predictive performance (as well as their interpretability) is evaluated on the test set in order to obtain an out-of-sample assessment of their generalization ability.

Lastly, in the case of the neural network models, feature scaling is applied. In particular, the numerical variables are standardized to ensure efficient training and convergence of the neural network. This procedure is not done instead with linear and tree-based models since their evaluation in terms of predictive performance and interpretability remains unchanged with respect to the scale of the input variables.

3.5 Exploratory Analysis

Before beginning with the machine learning analysis, a preliminary exploratory analysis is conducted to examine the main characteristics of the dataset and identify potential relationships among variables. More specifically, a correlation matrix is computed for all numerical variables included in the analysis (Figure 1). The results of this preliminary analysis serve as a first indication of the underlying structure of the data.

Looking at the matrix, it is possible to notice some relevant findings: first of all, the overall ESG score is strongly correlated with its environmental and social components, while the correlation with the governance pillar is comparatively lower. This suggests that the overall Bloomberg ESG measure is more closely aligned with environmental and social dimensions.

The correlations between ESG variables and financial indicators are generally weak, indicating the absence of strong direct relationships. This suggests that the link between firms' economic characteristics and ESG performance may be more complex and potentially non-linear.

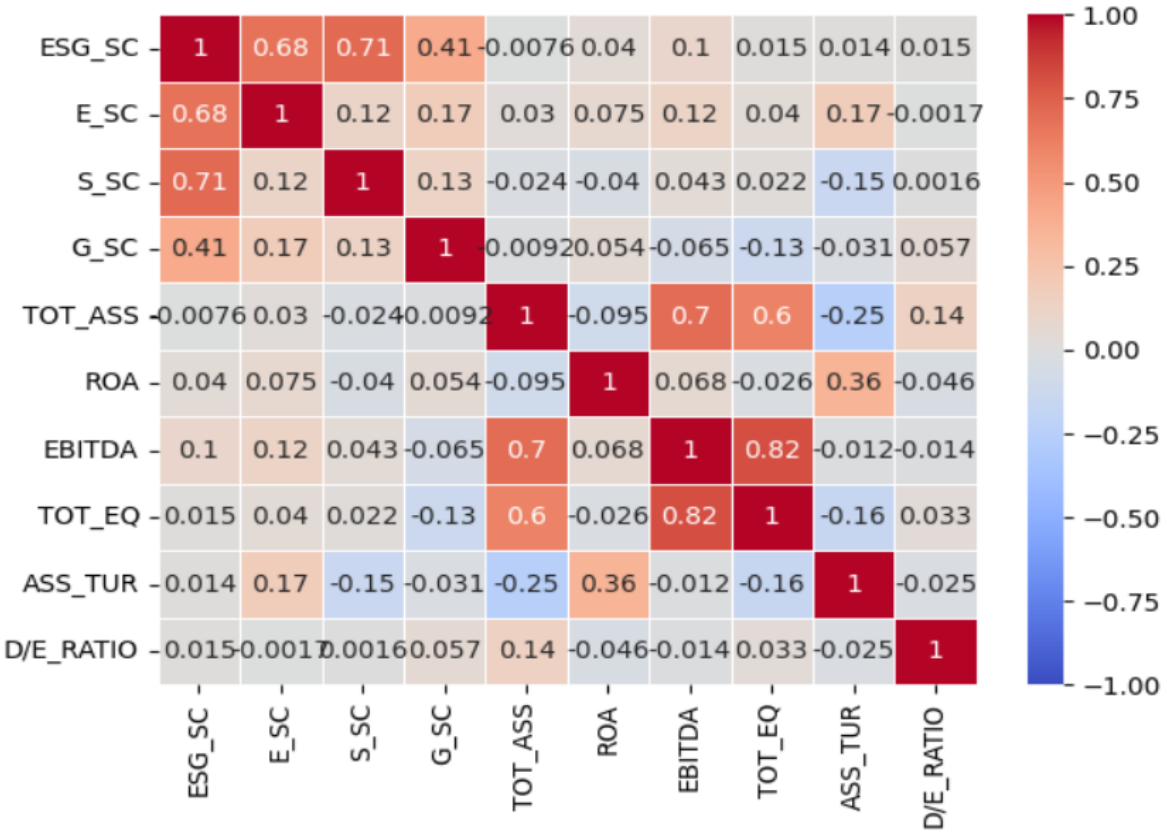

Fig.1: correlation matrix of the numerical variables of the dataset

Regarding instead the three components of the ESG score, they do not show strong correlations with each other. This confirms how they capture different aspects of a firm's sustainability performance and the fact that a strong performance in one pillar does not automatically translate into a similar one in the others.

Moreover, they generally have weak correlations with financial indicators too. Overall, these findings suggest that even the three ESG dimensions are influenced by different factors and that their relationship with financial variables is limited when assessed through simple linear correlations.

Chapter 4: Predictive performance of the models on the ESG score

4.1 Evaluation framework

As previously stated, the analysis is conducted by applying the models described in Chapter 2 to four different target variables, namely the overall ESG score and its three components: environmental (E), social (S), and governance (G) scores. This translates into four separate model specifications for each methodology considered.

This chapter presents and discusses the results related to the predictive performance of the models. The analysis first aims to compare their predictive outcomes for every different dependent variable considered and then establish which are the ones performing the best. In addition, a bootstrap out-of-bag procedure with 100 iterations is implemented on each model to assess the robustness and consistency of the results.

This is done by looking at the out-of-sample metrics, like R^2, RMSE, and MAE, of the different methods obtained on the test set of the data. In the context of the bootstrap, average performance metrics and their standard deviations are computed across iterations. However, only average R^2 and RMSE are reported, together with their standard deviations. The MAE is not included, as it provides information similar to RMSE and would therefore be largely redundant for the purpose of assessing robustness and consistency.

The only exception is given the logistic regression, which, being a classification method, relies instead on different metrics such as accuracy and ROC AUC score as predictive performance indicators. For these reasons, it is not considered too much in this chapter and just as an additional support to the main findings since it cannot be compared directly with the other models. Moreover, it is excluded from the bootstrap procedure. However, logistic regression will be further considered in the next chapter, where the focus will shift to model interpretability.

The next chapter will in fact focus on model interpretability, with particular attention to the contribution of the explanatory variables to the models' predictions, as well as on the robustness of these findings again through the bootstrap approach.

4.2 Results on ESG score

Starting from the predictive performance of the models having the ESG score as target variable, all of them exhibit relatively limited outcomes (Table 2).

In particular, the R^2 remains constantly close to zero through all specifications, suggesting that the selected features, the financial variables and sectoral dummies, explain only a very small portion of the variability in the ESG score. Hence, the differences among firms in terms of economic characteristics and sectoral membership are not sufficient to account for the observed variations in the ESG score. This suggests that such variations may be determined by other firm-specific factors that are not captured by the available financial data or the sector to which companies belong. This is consistent with the weak correlations observed during the exploratory analysis.

Model	R squared	RMSE	MAE
Linear Regression	0.0299	1.103	0.861
Regression Tree	0.0546	1.089	0.858
Random Forest	0.0683	1.081	0.829
Gradient Boosting	0.0207	1.108	0.849
Neral Network	0.0199	1.109	0.860

Model	Accuracy	ROC AUC score
Logistic regression	0.578	0.598

Table 2: predictive performance of models with the ESG score as target variable

Despite the overall low predictive performance of models, there are still some interesting observations that can be made: among them, random forest achieves the best performance, having the highest R^2 and the lowest error metrics (RMSE and MAE). The regression tree model also performs relatively well compared to the others, while linear regression shows weaker results, highlighting the limitations of linear approaches in capturing the underlying relationships. Furthermore, more complex models such as gradient boosting and neural network do not lead to performance improvements and, in fact, exhibit lower predictive accuracy than the random forest and the regression tree.

Even the logistic regression model exhibits limited predictive performance, with an accuracy of 0.578 and a ROC AUC score of 0.598. Although these values are slightly above the level of a random classifier, they indicate a weak ability to discriminate between firms with high and low ESG scores, confirming the struggles to identify potential relationships between the ESG score and the financial indicators or the sector affiliation of the firms.

Overall, these findings show that, even though the predictive power of the considered features remains limited across all cases, there are still some models, in particular the tree-based ones like regression trees and random forest, which can lead to a slight improvement, leading to think that the relationships between ESG score and the indicated features may be more complex and non-linear.

4.3 Results on Environmental score

Interestingly, as outlined by the corresponding metrics (Table 2.1), the predictive performance of the employed models improves significantly with the environmental score as target variable compared to the overall ESG score.

Model	R squared	RMSE	MAE
Linear Regression	0.142	1.944	1.593
Regression Trees	0.181	1.899	1.506
Random Forest	0.198	1.879	1.472
Gradient Boosting	0.218	1.856	1.465
Neral Network	0.177	1.903	1.557

Model	Accuracy	ROC AUC score

Logistic regression	0.617	0.684

Table 2.1: predictive performance of the models with the E score as target variable

Indeed, R^2 values are considerably higher across all methodologies, suggesting that the financial variables and sectoral dummies are quite more informative in explaining the environmental dimension of ESG performance. These results consistently imply that the E score is more predictable than the aggregate ESG score through the considered features. This may be due to a stronger and more direct relationship between the environmental component and the observed characteristics of the firms.

Among the considered models, gradient boosting achieves the best performance in terms of both R^2 and error metrics, followed closely by random forest, while the regression tree also produces relatively good outcomes. On the other hand, linear regression and neural networks, although still providing reasonable estimates, exhibit a lower predictive capacity. Thereby, the stronger predictions provided by tree-based ensemble methods suggest again the presence of non-linear relationships and interaction effects between the E score and the independent variables, which are instead not fully captured by linear models.

The better explanatory capacity of the underlying features on the environmental dimension of ESG score seems further supported by the logistic regression on the E score. In fact, with the E score turned into a binary variable, the model achieves an accuracy of 0.617 and a ROC AUC score of about 0.684, indicating a moderate ability to distinguish between firms with high and low levels of environmental score. These results are quite better than those obtained in the case of the logistic regression on the ESG score or in the following ones.

4.4 Results on Social score

When the companies' social score is used as target variable, the predictive performance of the employed models goes back to levels witnessed with the ESG score or gets even worse in some cases. In fact, the level of the R^2 across models remains significantly low and, in the case of the regression tree and random forest, becomes even negative (Table 2.3). This is the sign of a really limited ability from the adopted features to correctly predict

the value of the S score. This implies that the variation of the considered metric, just like in the case of the aggregate ESG score, is probably explained by other factors that are not strictly included in the sphere of the economic characteristics and sectoral membership of European companies.

Model	R squared	RMSE	MAE
Linear Regression	0.0444	1.707	1.355
Regression Trees	-0.0119	1.756	1.411
Random Forest	-0.0551	1.793	1.433
Gradient Boosting	0.0176	1.731	1.368
Neral Network	0.0113	1.736	1.369

Model	Accuracy	ROC AUC score
Logistic regression	0.539	0.577

Table 2.3: predictive performance of the models with the S score as target variable

Interestingly enough, the linear regression model, although its predictive performance remains considerably low, slightly outperforms the non-linear ones, even gradient boosting and neural network which exhibit low but still positive values. This pattern indicates that the limited signal present in the S score may follow a predominantly linear pattern with respect to the financial indicators and sectoral dummies. On the contrary, non-linear models, which are designed to capture complex interactions, do not provide improvements in this context, likely due to the weak and sparse information available in the data.

The logistic regression also shows almost null predictive outcomes as it achieves an accuracy of 0.539 and a ROC AUC score of 0.577 when the S score is turned into a binary variable. Although slightly above the level of a random classifier, this result indicates a minimal ability to discriminate between firms with high and low social scores.

Hence, the overall conclusion is that the S score is the least predictable out of the three scores considered so far through the selected explanatory variables. This likely reflects the presence of factors driving social performance that are not captured by the included features, as well as more complex or weaker relationships between the S score and the observed firm characteristics.

4.5 Results on Governance score

Lastly, with the G score adopted as the independent variable of the models, their predictive outcomes are weak (Table 2.4). Indeed, the R^2 is once more consistently low across all models involved and even negative for some of them, highlighting a poor capacity from the features to correctly predict the G score. This evidence, just like for the cases of ESG and S scores, is probably due to the fact that the variability of the governance performance of firms mostly depends on elements that are not directly related to companies' financial characteristics or sectoral affiliation.

Model	R squared	RMSE	MAE
Linear Regression	-0.0350	1.059	0.841
Regression Tree	0.0069	1.038	0.816
Random Forest	0.0391	1.021	0.794
Gradient Boosting	0.0099	1.036	0.812
Neral Network	-0.0914	1.088	0.867

Model	Accuracy	ROC AUC score
Logistic regression	0.583	0.576

Table 2.4: predictive performance of the models with the G score as target variable

Despite the overall weak results of the models' predictions, a clear pattern emerges: the random forest model is the one performing the best since it produces the highest R^2 and the lowest levels of RMSE and MAE. Moreover, it is immediately followed by gradient boosting and the regression tree, which display low but still positive R^2 values and slightly lower error metrics. On the other hand, linear regression and neural network perform

worse, exhibiting negative R^2 values, which indicates a limited ability to capture the underlying relationships, and higher RMSEs and MAEs.

The relative predominance of tree-based ensemble methods suggests that, although the overall predictive power remains limited, some weak non-linear relationships may exist between the G score and the explanatory variables considered.

Finally, the metrics of the logistic regression model seem to confirm the previous considerations. Indeed, the accuracy of the model is 0.583 while the ROC AUC score is 0.576. This inevitably highlights a weak capacity to discriminate between firms with high or low G scores.

4.6 Bootstrap's results

As stated in the first section of this chapter, to further assess the robustness and stability of the results, a bootstrap out-of-bag procedure with 100 iterations is implemented for the ESG score. The average performance metrics and their corresponding standard deviations can provide indeed additional insight into the variability of the models and allow for a more reliable evaluation of their predictive performance.

4.6.1 Bootstrap on ESG score

Starting again from the models with ESG score as target variable, the results from the bootstrap (Table 3) seem to confirm the former conclusions: the predictive performance remains weak across all models. Furthermore, compared to the results obtained from the single train-test split, the bootstrap analysis provides a more conservative assessment of models' performance.

Model	Average R squared	Std. Dev. Of R squared	Average RMSE	Std. Dev. Of RMSE
Linear regression	-0.0380	0.157	1.203	0.0853
Regression trees	-0.0423	0.0670	1.209	0.0560
Random Forest	-0.0188	0.0583	1.190	0.0543
Gradient Boosting	0.00846	0.0517	1.180	0.0476
Neural Networks	-0.554	0.696	1.454	0.273

Table 3: results of the bootstrap out-of-bag on models with the ESG score as target variable

Indeed, while some models previously exhibited slightly positive R^2 values, these new findings show that, on average, the general explanatory power of the features is almost null. In particular, mean R^2 values are close to zero or negative for most specifications, with only gradient boosting achieving a slightly positive one, with also pretty low level of variability, highlighting the consistency of such results.

Despite the overall low performance, tree-based ensemble models, especially gradient boosting and random forest, continue to exhibit relatively better outcomes, achieving slightly higher R^2 values and lower RMSE values while also showing smaller standard deviations, which indicates a satisfactory degree of stability compared to the other models. In contrast, the neural network performs substantially worse, with a heavily negative average R^2 and very high variability, suggesting strong sensitivity to sample variation and limited reliability in this context.

4.6.2 Bootstrap on E score

Moving on, the results of the OOB bootstrap for the models on the E score also confirm and strengthen those observed in the baseline analysis (Table 3.1). In this sense, the bootstrap analysis provides a more reliable assessment of model performance, confirming that the environmental score is consistently more predictable. In fact, although the overall explanatory power remains moderate, the fact that average R^2 values remain moderately positive across all specifications, combined with a generally low variability, except for the neural network, indicates that the selected independent variables contain meaningful information for this specific ESG dimension.

Model	Average R squared	Std. Dev. Of R squared	Average RMSE	Std. Dev. Of RMSE
Linear regression	0.0806	0.146	2.008	0.150
Regression trees	0.0762	0.0781	2.024	0.0944
Random Forest	0.0907	0.0649	2.001	0.0879
Gradient Boosting	0.1291	0.0551	1.965	0.0774
Neural Networks	-0.754	2.224	2.626	0.847

Table 3.1: results of the bootstrap out-of-bag on models with the E score as target variable

Moreover, in line with the previous findings, models like gradient boosting and random forest outperform others with better metrics and lower standard deviations, emerging as the most consistent and reliable ones. This suggests again the presence of non-linear relationships and interaction effects between the environmental score and the explanatory variables. In contrast, the neural network performs substantially worse, with a significantly negative average R^2 and very high variability, indicating instability and limited reliability in this context.

4.6.3 Bootstrap on S score

Thirdly, the bootstrap results for the social score confirm the very limited predictive performance observed in the baseline analysis (Table 3.2). Indeed, average R^2 values remain close to zero across all models, with some specifications even exhibiting negative values, indicating that the explanatory variables provide little to no useful information for predicting the S score. This reinforces the idea that the social score is particularly difficult to predict using financial indicators and sectoral dummies, likely reflecting the importance of unobserved and qualitative factors driving this dimension.

Model	Average R squared	Std. Dev. Of R squared	Average RMSE	Std. Dev. Of RMSE
Linear regression	0.0110	0.135	1.862	0.138
Regression trees	-0.0763	0.0751	1.939	0.0777
Random Forest	-0.0168	0.0562	1.878	0.0837
Gradient Boosting	0.0233	0.0493	1.836	0.0641
Neural Network	-0.595	0.717	2.336	0.438

Table 3.2: results of the bootstrap out-of-bag on models with the S score as target variable

Moreover, compared to the results obtained from the single train-test split, gradient boosting emerges as the best-performing model. However, it is followed by linear regression, which, despite slightly higher variability, still outperforms the other non-linear

approaches. In contrast, the neural network once again shows substantially worse performance and higher variability, confirming its limited reliability in this context. By the way, this suggests that the superior performance of linear regression in the baseline analysis may be partly driven by sample-specific effects, whereas gradient boosting captures more stable patterns in the data. This evidence supports the relative predominance of this last model in capturing the variability of the S score, although no clear and consistent predictive structure emerges across models.

4.6.4 Bootstrap on G score

Lastly, looking at the results regarding the bootstrap out-of-bag from the models having the G score as target variable (Table 3.3), it is possible to see that, just like in the baseline analysis, the general predictive performance remains low. Indeed, all models have negative average R^2 values, indicating that the explanatory variables fail to capture meaningful variations in the G score.

However, just like in the case of the single train-test split, a pattern could be identified: tree-based ensemble methods, especially gradient boosting and random forest, perform better by showing higher R^2 values and lower error metrics with also smaller standard deviations with respect to other models considered. This evidence suggests once again that, despite still showing a drastically weak explanatory capacity, random forest and gradient boosting emerge as the most reliable and robust predictors of the G score. The other models instead, especially neural networks, are the ones characterized by the worst predictive accuracy and consistency.

Model	Average R squared	Std. Dev. Of R squared	Average RMSE	Std. Dev. Of RMSE
Linear regression	-0.103	0.129	1.155	0.0674
Regression trees	-0.0944	0.0651	1.1515	0.0489
Random Forest	-0.0738	0.0591	1.1327	0.0483
Gradient Boosting	-0.0603	0.0515	1.1290	0.0515
Neural Networks	-1.116	1.404	1.5449	0.4368

Table 3.3: results of the bootstrap out-of-bag on models with the G score as target variable

4.7 Results review

In summary, the predictive performance of the different models on the ESG score and its three components, E, S and G, both on the single train-test split and on the 100 iterations of bootstrap out-of-bag, reveal some interesting and consistent conclusions.

First of all, for the models on the ESG score, the analysis highlights that, given the constantly very low R^2 values, the explanatory power of the considered features, representing firms' economic characteristics and sectoral affiliation, is drastically limited. However, this situation moderately varies when considering the specific dimensions of the ESG score: although the overall results remain limited, some differences can be identified.

In particular, models on S and G score basically show the equal or even worse levels of R^2 of those observed on the models on the ESG score, indicating that even the social and governance dimension are likely not explainable trough sectoral membership or firm-specific characteristics. However, the specifications on E score exhibit better results with higher levels of R^2 in the single train-test split and in the bootstrap OOB process, reflecting a higher capacity from the features to explain the variability of this specific score.

This implies that industry-specific and economic characteristics of firms are more informative only for a specific component of the ESG score, specifically the environmental dimension, while they are not for the other two, social and governance dimensions. This is probably due to the fact that, while the first one is more connected to those elements, the others depend more on factors that are not closely related to the economic measures or the sector to which companies belong.

Regarding instead the comparison across different methodologies, tree-based models clearly appear as the best in terms of performance and reliability. Indeed, looking especially at the results from the bootstrap procedure, such methods, particularly gradient boosting, show the best values for both R^2 and error metrics. Additionally, they display the lowest levels of variability since they have smaller standard deviations compared to the other techniques employed. This suggests that the weak but existing

relations between the ESG, E, S and G scores and the financials indicators as well as the sectoral affiliation might probably happen non-linearly.

Linear regression exhibits generally worse values of R^2 and error metrics in the single train-test split scenario, except for the case of the S score, and also in the bootstrap OOB procedure with even higher variability. However, the bootstrap method clearly identifies the neural network as the worst performing model: its R^2 are significantly negative across all scores and way lower than those observed with the other methods while the RMSE values are far higher. All of this, combined with higher standard deviations of the two metrics mentioned above, contributes to label the neural network as the worst performing technique out of those adopted while also being the most inconsistent, making the results obtained on the single train-test split unreliable.

Chapter 5: Interpretability results

5.1 Chapter overview

This chapter focuses on the interpretability of the adopted models, with the aim of identifying which of the considered features contribute most to their predictions. In this context, the first part of the chapter is devoted to the analysis of individual models, using their respective explainability techniques to determine the most relevant features for each model and for each ESG-related score. The analysis is limited to linear and tree-based models, as their structure allows for more direct and intuitive interpretation, whereas neural networks are excluded due to their higher complexity and "black-box" nature.

The second part of the chapter adopts a comparative perspective across models through the use of SHAP values, which, as previously discussed, provide a consistent and model-agnostic framework for assessing and comparing feature importance across different methodologies. Within this framework, neural networks are also included, allowing them to be evaluated alongside the other models despite their lower interpretability in model-specific terms. In particular, a 6x6 Spearman correlation matrix, based on the normalized SHAP values of the six models employed, is constructed for each target variable in order to assess whether, and to what extent, the different methods rank the explanatory variables in a similar way in terms of their contribution to the predictions. The aim of this process is to identify possible interpretability patterns across different methodologies. Finally, the normalized SHAP values, and consequently the four correlation matrices, are also computed within a bootstrap out-of-bag framework with 50 iterations, in order to assess the robustness and consistency of the interpretability results.

5.2 Interpretability of the linear models: coefficients of linear and logistic regression

Linear models, including linear and logistic regression, provide a transparent and interpretable framework for understanding the relationships between financial features, sectoral dummies, and ESG performance. As previously stated, their coefficients allow for

an immediate assessment of both the direction and magnitude of each feature's effect. In this section, these models are analyzed first, focusing on the overall ESG score and the individual E, S, and G pillars, providing a clear benchmark for the more complex methods discussed later.

5.2.1. Linear models on ESG score

Starting with the linear models on the ESG score, the outcomes (Table 4) highlight some interpretable relationships. Among the financial variables, the only three that are really significant are total assets, EBITDA, and total equity, while asset turnover exhibits just a marginal significancy on the ESG score.

Explanatory variable	Coefficient	Std. Dev.	t	p-value	Signif.
const	4.604587e+00	2.273750e-01	20.251073	1.897105e-69	***
TOT_ASS	4.364419e-07	1.766420e-07	2.470771	1.376709e-02	*
ROA	5.221970e-03	4.759162e-03	1.097246	2.729870e-01	
EBITDA	2.314930e-05	7.019747e-06	3.297741	1.033993e-03	**
TOT_EQ	-5.625750e-06	2.152497e-06	-2.613592	9.190952e-03	**
ASS_TUR	-2.371691e-01	1.339285e-01	-1.770864	7.710578e-02	.
D/E_RATIO	4.501907e-05	1.013222e-04	0.444316	6.569790e-01	
Sector_Consumer Discretionary	-3.168905e-02	2.642983e-01	-0.119899	9.046046e-01	
Sector_Consumer Staples	4.131557e-02	2.836044e-01	0.145680	8.842241e-01	
Sector_Energy	1.005730e+00	3.330933e-01	3.019364	2.643859e-03	**
Sector_Financials	-2.923764e-01	2.486475e-01	-1.175867	2.401280e-01	
Sector_Health Care	-1.065502e-01	2.685740e-01	-0.396726	6.917149e-01	
Sector_Industrials	4.103972e-01	2.437992e-01	1.683341	9.284445e-02	.
Sector_Information Technology	3.181329e-01	3.141568e-01	1.012656	3.116445e-01	

Sector_Materials	6.223266e-01	2.771545e-01	2.245414	2.511549e-02	*
Sector_Real Estate	5.528768e-01	3.080034e-01	1.795035	7.316599e-02	.
Sector_Utilities	3.932599e-01	2.993894e-01	1.313540	1.895178e-01	

Table 4: coefficients of the linear regression on the ESG score

In particular, the first two have a significant positive association with the ESG score, suggesting that larger firms or those with a higher operational efficiency may accomplish higher ESG scores. On the other hand, total equity has a significant negative impact on the target variable, indicating that companies with higher net worths could have lower ESG scores.

Regarding the sectoral dummies, those referring to energy and materials industries are the only two clearly significant variables while industrials and real estate sectors display just marginal effects on the dependent variable.

Moving on to the results of logistic regression on the ESG score (Table 5), unlike in the case of the linear regression, there is no significance across the financial indicators apart from EBITDA, which shows again a significant positive effect on the ESG score. Return on assets and asset turnover exhibit only some marginal significance.

Explanatory Variables	Coefficient	Std. Err.	z	p-value	Signif.
Intercept	-6.858596e-01	4.279281e-01	-1.602745	0.108991	
TOT_ASS	1.707195e-07	3.950948e-07	0.432098	0.665671	
ROA	1.528234e-02	9.052180e-03	1.688250	0.091363	.
EBITDA	4.149214e-05	1.798353e-05	2.307230	0.021042	*
TOT_EQ	-7.104212e-06	4.720998e-06	-1.504812	0.132373	
ASS_TUR	-4.392391e-01	2.490496e-01	-1.763661	0.077789	.
D/E_RATIO	6.037855e-04	7.315278e-04	0.825376	0.409158	
Sector_Consumer Discretionary	3.079699e-01	4.841396e-01	0.636118	0.524700	
Sector_Consumer Staples	5.648889e-01	5.164751e-01	1.093739	0.274070	

Sector_Energy	1.726401e+00	6.584485e-01	2.621922	0.008744	**
Sector_Financials	9.253145e-02	4.594515e-01	0.201395	0.840389	
Sector_Health Care	4.190202e-01	4.935791e-01	0.848942	0.395913	
Sector_Industrials	1.083511e+00	4.490049e-01	2.413138	0.015816	*
Sector_Information Technology	8.846005e-01	5.698766e-01	1.552267	0.120598	
Sector_Materials	1.709174e+00	5.215299e-01	3.277231	0.001048	**
Sector_Real Estate	1.756715e+00	5.958153e-01	2.948422	0.003194	**
Sector_Utilities	8.868893e-01	5.439535e-01	1.630450	0.103006	

Table 5: Coefficients of logistic regression on the ESG score.

On the contrary, there are four different sector dummies which are really significant: those representing the energy and materials sectors , like in the case of the linear regression, but also Sector_Industrials and Secor_Real Estate, which show this time a clear and not just marginal level of significance. In particular, all these features tend to have a significant negative impact on the prediction of the logistic regression model. These effects appear generally more pronounced than in the linear regression.

Overall, considering the linear methods included in the analysis, some sectoral dummies, especially those representing the energy and materials sectors, as well as industrials and real estate ones, seem to be the most consistently significant variables of these specifications. This may imply that belonging to some specific industries could play a key role in shaping the ESG performance of the companies.

On the other hand, EBITDA appears to be the only financial variable that is consistently significant across both models, indicating that a firm's operational capacity may be an important driver of ESG performance. Other financial variables such as total assets, total equity asset turnover, or ROA are significant in just one of the two models, suggesting that their relevance may depend mostly on the specific modeling approach.

5.2.2 Linear Models on E score

With the environmental score as target variable, the coefficients of the linear models reveal both similarities and differences compared to the aggregate ESG score: in the case

of the linear regression, total assets, EBITDA and total equity are again the only strongly significant variables between the financial indicators, with the first two still exhibiting a significant positive relation with the considered score while the third a negative one. Other variables, including ROA, asset turnover, and the debt-to-equity ratio, do not show statistically significant effects (Table 4.1).

A heavily different result emerges instead when considering sectoral dummies: the variables representing different sectors, such as consumer discretionary, financials, health care, and real estate, display significant negative coefficients, indicating that firms operating in these industries tend to achieve lower environmental scores relative to the reference category. At the same time, sectors that were previously significant in the aggregate ESG model, such as energy and materials, do not appear to play a relevant role in explaining environmental performance.

Explanatory variable	Coefficient	Std. Err	t	p-value	Signif.
Intercept	5.356828	3.784311e-01	14.155356	2.589447e-39	***
TOT_ASS	0.000002	2.939938e-07	5.170562	3.212571e-07	***
ROA	0.003999	7.920902e-03	0.504891	6.138262e-01	
EBITDA	0.000042	1.168330e-05	3.569805	3.866891e-04	***
TOT_EQ	-0.000010	3.582505e-06	-2.888124	4.019330e-03	**
ASS_TUR	-0.209262	2.229036e-01	-0.938800	3.482224e-01	
D/E_RATIO	-0.000030	1.686354e-04	-0.175755	8.605472e-01	
Sector_Consumer Discretionary	-1.333021	4.398844e-01	-3.030390	2.550665e-03	**
Sector_Consumer Staples	-0.710066	4.720165e-01	-1.504324	1.330395e-01	
Sector_Energy	-0.093521	5.543832e-01	-0.168694	8.660959e-01	
Sector_Financials	-2.752759	4.138360e-01	-6.651812	6.676153e-11	***
Sector_Health Care	-1.328564	4.470006e-01	-2.972174	3.079001e-03	**

Sector_Industrials	-0.244640	4.057668e-01	-0.602909	5.468038e-01	
Sector_Information Technology	-0.138010	5.228664e-01	-0.263949	7.919124e-01	
Sector_Materials	-0.156446	4.612817e-01	-0.339155	7.346148e-01	
Sector_Real Estate	-1.495677	5.126249e-01	-2.917682	3.662334e-03	**
Sector_Utilities	-0.869138	4.982883e-01	-1.744248	8.164284e-02	.

Table 4.1: coefficients of the linear regression on E score

Moving on, the results of the logistic regression on the E score in Table 5.1 tend to confirm the patterns observed in the linear specification: among financial variables, total assets and EBITDA remain positive and statistically significant, reinforcing the role of firm size and operational capacity in increasing the probability of achieving higher environmental performance. Total equity continues to exhibit a negative and significant effect, while asset turnover is only marginally significant.

With respect to sectoral variables, the results highlight a more limited but still relevant set of effects. In particular, the dummies representing the financials and real estate sector still have significant negative impacts on the target variable while the others don't.

The general conclusion is that, when the environmental component is used as dependent variable, linear models highlight that, in contrast with the ESG score, multiple financial indicators, especially total assets, EBITDA and total equity, appear as robust and significant factors of environmental performance of firms. On the other hand, some sectoral dummies, different from those observed in the case of ESG score, such as the financials and real estate sectors, seem to be the most consistent and significant ones across the different specifications.

Explanatory Variable	Coefficient	Std. Err	z	p-value	Signif
Intercept	0.377011	4.225319e-01	0.892267	0.372250	
TOT_ASS	0.000002	4.250092e-07	3.548105	0.000388	***
ROA	0.009307	8.879950e-03	1.048082	0.294601	
EBITDA	0.000076	2.365025e-05	3.202394	0.001363	**
TOT_EQ	-0.000015	5.662800e-06	-2.646205	0.008140	**
ASS_TUR	-0.471782	2.503112e-01	-1.884782	0.059459	.
D/E_RATIO	-0.000296	3.676571e-04	-0.806312	0.420063	
Sector_Consumer Discretionary	-0.760742	4.884192e-01	-1.557560	0.119338	
Sector_Consumer Staples	-0.185818	5.165350e-01	-0.359740	0.719042	
Sector_Energy	0.353279	6.464251e-01	0.546512	0.584714	
Sector_Financials	-1.308104	4.650796e-01	-2.812645	0.004914	**
Sector_Health Care	-0.431554	4.926985e-01	-0.875898	0.381085	
Sector_Industrials	0.366167	4.515622e-01	0.810889	0.417429	
Sector_Information Technology	0.418916	5.796475e-01	0.722708	0.469859	
Sector_Materials	0.726321	5.212182e-01	1.393506	0.163467	
Sector_Real Estate	-2.081435	6.840762e-01	-3.042695	0.002345	**
Sector_Utilities	-0.212652	5.430061e-01	-0.391619	0.695340	

Table 5.1: coefficients of the logistic regression on E score

5.2.3 Linear models on S score

Thirdly, with the social score employed as dependent variable, the results of the linear regression model (Table 4.2) underline again some differences compared to those of the previous scores. In particular, the table shows that none of the financial variables appear to be statistically significant, suggesting that firm-level economic characteristics do not play a relevant role in explaining social performance.

In contrast some sectoral variables, more specifically those referred to the energy and real estate industries, have significant positive effects on the social components while the financials dummy is just marginally significant for the prediction of the outcome.

Explanatory Variable	Coefficient	Std. Err.	t	p-value	Signif.
Intercept	3.411303e+00	3.522992e-01	9.682970	1.159629e-20	***
TOT_ASS	-2.375863e-07	2.736926e-07	-0.868077	3.857092e-01	
ROA	1.274637e-03	7.373938e-03	0.172857	8.628238e-01	
EBITDA	1.156586e-05	1.087653e-05	1.063377	2.880511e-01	
TOT_EQ	-9.837507e-07	3.335121e-06	-0.294967	7.681241e-01	
ASS_TUR	-2.831924e-01	2.075114e-01	-1.364707	1.728715e-01	
D/E_RATIO	4.108475e-05	1.569906e-04	0.261702	7.936436e-01	
Sector_Consumer Discretionary	6.592427e-01	4.095089e-01	1.609837	1.079747e-01	
Sector_Consumer Staples	-1.420913e-02	4.394222e-01	-0.032336	9.742152e-01	
Sector_Energy	1.812698e+00	5.161013e-01	3.512291	4.786553e-04	***
Sector_Financials	7.250789e-01	3.852593e-01	1.882054	6.032650e-02	.
Sector_Health Care	1.617574e-01	4.161338e-01	0.388715	6.976289e-01	
Sector_Industrials	4.837382e-01	3.777473e-01	1.280587	2.008481e-01	
Sector_Information Technology	1.702262e-01	4.867608e-01	0.349712	7.266809e-01	
Sector_Materials	5.525866e-01	4.294287e-01	1.286795	1.986765e-01	
Sector_Real Estate	2.773362e+00	4.772265e-01	5.811416	1.020108e-08	***
Sector_Utilities	5.060565e-01	4.638799e-01	1.090922	2.757581e-01	

Table 4.2: coefficients of the linear regression on S score

Moreover, looking at the results of the logistic regression on the social component (Table 5.2), financial variables appear again as not relevant in terms of significance with only total assets having a significant negative impact on the output's prediction and total equity a marginal positive one.

Meanwhile, among sectoral dummies, those representing the real Estate and energy industries are still the only significant ones.

Explanatory variable	Coefficient	Std. Err.	z	p-value	Signif.
Intercept	-5.660447e-01	4.166086e-01	-1.358697	0.174243	
TOT_ASS	-9.129257e-07	4.485395e-07	-2.035330	0.041818	*
ROA	6.106658e-03	8.773329e-03	0.696048	0.486399	
EBITDA	-7.162323e-06	1.330436e-05	-0.538344	0.590339	
TOT_EQ	1.124829e-05	5.979334e-06	1.881195	0.059945	.
ASS_TUR	-1.044630e-02	2.428334e-01	-0.043018	0.965687	
D/E_RATIO	5.354789e-04	6.764838e-04	0.791562	0.428616	
Sector_Consumer Discretionary	1.642571e-01	4.707080e-01	0.348958	0.727121	
Sector_Consumer Staples	-3.988850e-01	5.172799e-01	-0.771120	0.440636	
Sector_Energy	1.457758e+00	6.472320e-01	2.252295	0.024304	*
Sector_Financials	6.568288e-01	4.456988e-01	1.473705	0.140561	
Sector_Health Care	2.701608e-01	4.769343e-01	0.566453	0.571086	
Sector_Industrials	2.220615e-01	4.342617e-01	0.511354	0.609103	
Sector_Information Technology	-5.754316e-02	5.642125e-01	-0.101988	0.918766	
Sector_Materials	2.806972e-01	4.934601e-01	0.568835	0.569468	
Sector_Real Estate	2.265166e+00	6.722499e-01	3.369530	0.000753	***
Sector_Utilities	4.447507e-01	5.305774e-01	0.838239	0.401896	

Table 5.2: coefficients of the logistic regression on S score

In conclusion, according to linear models, sectoral variables seem to be the real drivers of the companies' social performance with the dummies representing real estate and financials being the only significant ones across the two models. On the other hand, economic characteristics of the firms seem to not have so much importance in determining the social dimension of the ESG score.

5.2.4 Linear models on G score

Lastly, the coefficients of the linear models with the governance score as target variable define a pattern oriented towards the firms' economic characteristics (Table 4.3). Starting as always with the linear regression, total assets, EBITDA, and total equity are one more time the only statistically significant financial variables. On the contrary, sectoral dummies do not show any sign of significance, indicating a null impact on the G score.

Explanatory variable	Coefficient	Std. Err.	t	p-value	Signif.
Intercept	6.699991e+00	2.161056e-01	31.003324	2.024022e-125	***
TOT_ASS	3.717055e-07	1.678871e-07	2.214021	2.721349e-02	*
ROA	6.486180e-03	4.523283e-03	1.433954	1.521215e-01	
EBITDA	1.383421e-05	6.671826e-06	2.073526	3.856187e-02	*
TOT_EQ	-7.960496e-06	2.045813e-06	-3.891117	1.112948e-04	***
ASS_TUR	-9.794072e-03	1.272906e-01	-0.076943	9.386956e-01	
D/E_RATIO	1.244880e-04	9.630035e-05	1.292705	1.966251e-01	
Sector_Consumer Discretionary	-7.606904e-02	2.511988e-01	-0.302824	7.621320e-01	
Sector_Consumer Staples	-1.704269e-01	2.695481e-01	-0.632269	5.274589e-01	
Sector_Energy	-1.963904e-01	3.165842e-01	-0.620342	5.352752e-01	
Sector_Financials	1.086485e-01	2.363237e-01	0.459744	6.458712e-01	
Sector_Health Care	6.015518e-02	2.552626e-01	0.235660	8.137793e-01	
Sector_Industrials	-1.615345e-01	2.317158e-01	-0.697123	4.860034e-01	
Sector_Information Technology	-8.769809e-02	2.985863e-01	-0.293711	7.690832e-01	
Sector_Materials	-1.155754e-02	2.634179e-01	-0.043875	9.650188e-01	
Sector_Real Estate	2.061824e-01	2.927378e-01	0.704324	4.815120e-01	
Sector_Utilities	2.856561e-01	2.845508e-01	1.003884	3.158508e-01	

Table 4.3: coefficients of the linear regression on G score

Furthermore, the coefficients of the logistic regression (Table 5.3) seem to quite confirm this trend: among financial variables, total assets and total equity still have significant effects on the predictions of the models while EBITDA has them but just marginally. However, this time, considering sectoral dummies, real estate sectoral dummy appears as statistically significant while the utilities sectoral dummy does it marginally.

Explanatory Variable	Coefficient	Std. Err.	z	p-value	Signif.
const	-0.518289	4.163917e-01	-1.244714	0.213237	
TOT_ASS	0.000001	4.632066e-07	2.534092	0.011274	*
ROA	0.017224	1.131164e-02	1.522711	0.127831	
EBITDA	0.000027	1.597772e-05	1.713751	0.086574	.
TOT_EQ	-0.000021	6.725682e-06	-3.122193	0.001795	**
ASS_TUR	0.159434	2.410797e-01	0.661333	0.508399	
D/E_RATIO	0.000277	3.575204e-04	0.774327	0.438738	
Sector_Consumer Discretionary	0.181069	4.794778e-01	0.377638	0.705700	
Sector_Consumer Staples	0.185613	5.114485e-01	0.362917	0.716667	
Sector_Energy	0.805851	6.094866e-01	1.322179	0.186108	
Sector_Financials	0.528297	4.510956e-01	1.171142	0.241542	
Sector_Health Care	0.620174	4.844540e-01	1.280149	0.200493	
Sector_Industrials	0.035592	4.424920e-01	0.080436	0.935890	
Sector_Information Technology	0.182586	5.681566e-01	0.321365	0.747933	
Sector_Materials	0.631911	5.015192e-01	1.259993	0.207672	
Sector_Real Estate	1.284451	5.668535e-01	2.265931	0.023456	*
Sector_Utilities	0.970327	5.438866e-01	1.784062	0.074414	.

Table 5.3: coefficients of the logistic regression on G score

Overall, based on the coefficients of the two linear models, financial variables, especially total assets and total equity but also EBITDA, seem to be the most important in determining the level of the firms' governance performance. On the other hand, sectoral

affiliation doesn't seem to have much significance since none of the industry dummies own meaningful coefficients in both models.

5.3 Interpretability of tree-based models: graphs and importance plots of regression trees, random forest and gradient boosting

The following section of this chapter focuses instead on the interpretability of the tree-based models employed in the analysis, which include regression trees, random forest, and gradient boosting. As previously outlined in Chapter 2, their interpretability revolves around the importance of plots, which allow them to rank the variables that contribute the most to predictions of the considered model. Moreover, in the case of regression trees, their explainability can be accessed also through their apposite representation graphs. Variables appearing in the top splits of the tree are generally among the most important, as they contribute the most to the reduction of prediction error.

5.3.1 Tree models on ESG score

Starting again with the models with the ESG score as target variable, in particular from the regression tree, it is possible to notice how financial indicators are clearly the most important features. Indeed, they generate all the splits of the underlying regression tree while sectoral dummies never appear in it (Figure 2). This implies that there may be some potential non-linear relationships between financial data and the ESG score.

These impressions are consistently confirmed by the importance plots of the random forest and gradient boosting (Figures 3&4), which clearly show that financial variables are by the most relevant contributors to the models' predictions. On the contrary, sectoral dummies, even though they appear at least in the importance plots, emerge as the least significant variables for both random forest and gradient boosting.

In conclusion, financial data are clearly the most crucial in contributing to the predictions of non-linear tree models. In this sense, the debt-to-equity ratio represents the main predictor in all three models, suggesting that the companies' level of financial leverage could really affect their ESG performance through some non-linear relationships. In

contrast, the affiliation to some specific sectors doesn't seem to influence the firms' overall ESG score.

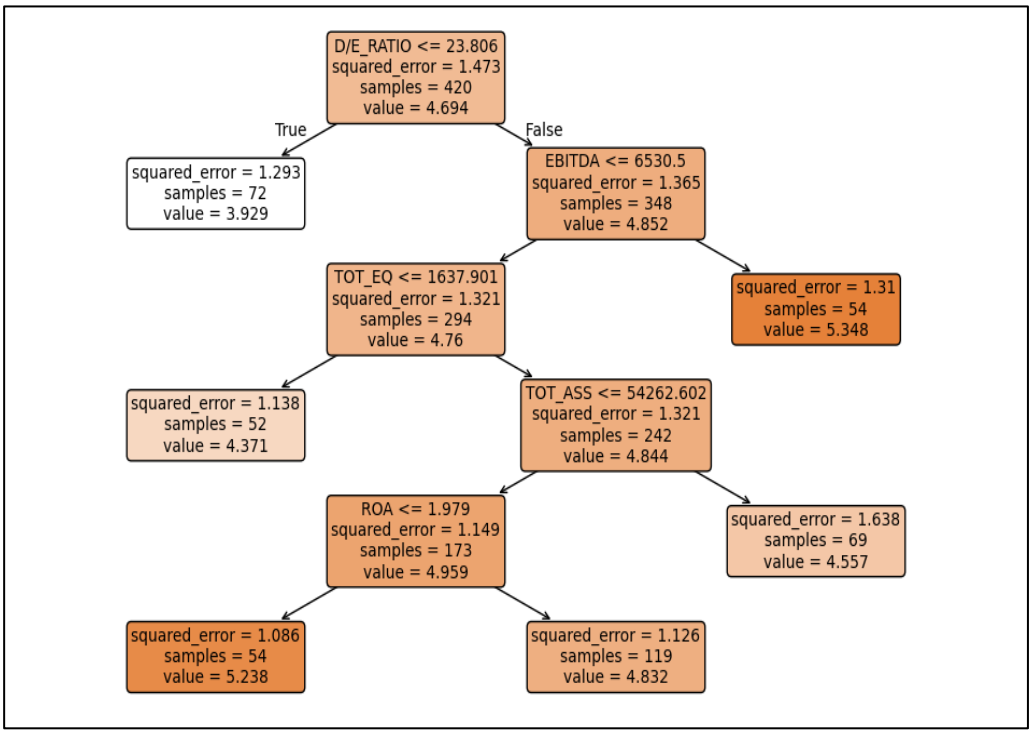

Figure 2: graphical representation of the regression on ESG score

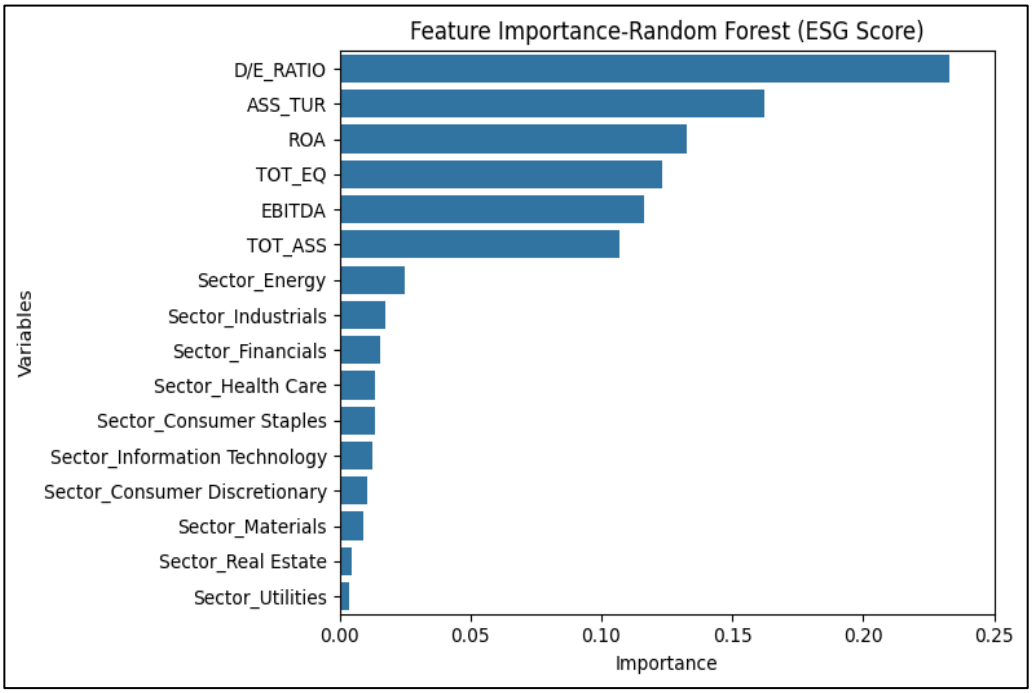

Figure 3: features' importance plot of the random forest on ESG score

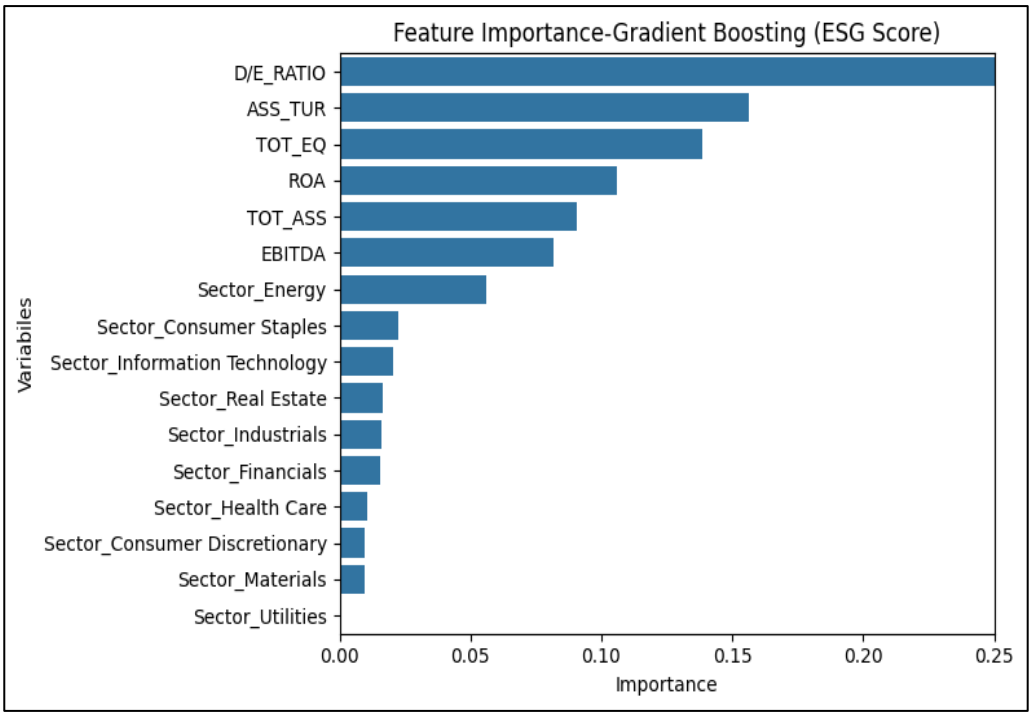

Figure 4: features' importance plot of the gradient boosting on ESG score.

5.3.2 Tree models on E score

Secondly, tree-based models using the environmental score as dependent variable exhibit some similarities and differences with respect to the results of those on the aggregate ESG score.

First of all, looking at the regression tree in Figure 2.1, financial variables, especially EBITDA, are still the most important variables in the model since they create most of the nodes. However, sectoral dummies also play a relevant role in this case. Indeed, the dummy representing the financials industry emerges as the most relevant variable of the tree while the one representing the industrials sector also contributes to the prediction by creating a split.

Most of these results can also be acknowledged through the importance plots (Figures 3.1&4.1): in the plot of random forest, financial variables are still the most meaningful for the prediction but also the financials dummy shows a discrete degree of importance. Lastly, regarding the gradient boosting, Sector_Financials is the second most relevant variable for the model's prediction behind EBITDA. It is then followed by other financial indicators. The remaining sectoral variables are instead irrelevant for both random forest and gradient boosting.

In conclusion, financial data, based on tree models, seem to be again the key determinants of the predictions of the tree-based models. In particular, EBITDA is the most important across all of them, indicating that the firms' operational efficiency might have an impact on their environmental score through non-linear relationships. However, unlike in the case of the general ESG score, the affiliation to certain sectors, more specifically the financials industry, seems to affect substantially the environmental performance of the companies. This could imply that the E score is more sector-driven with respect to the other components of the ESG score.

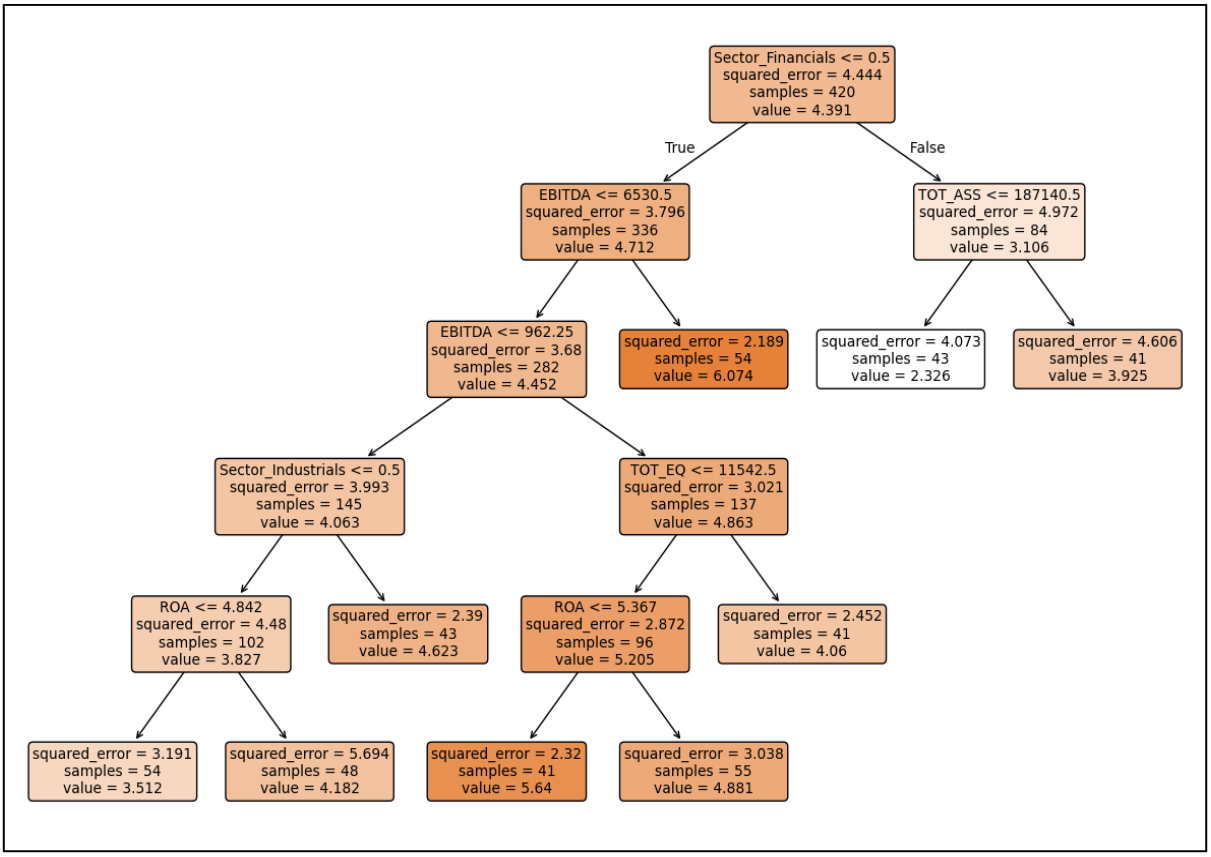

Figure 2.1: graphical representation of the regression tree on E score

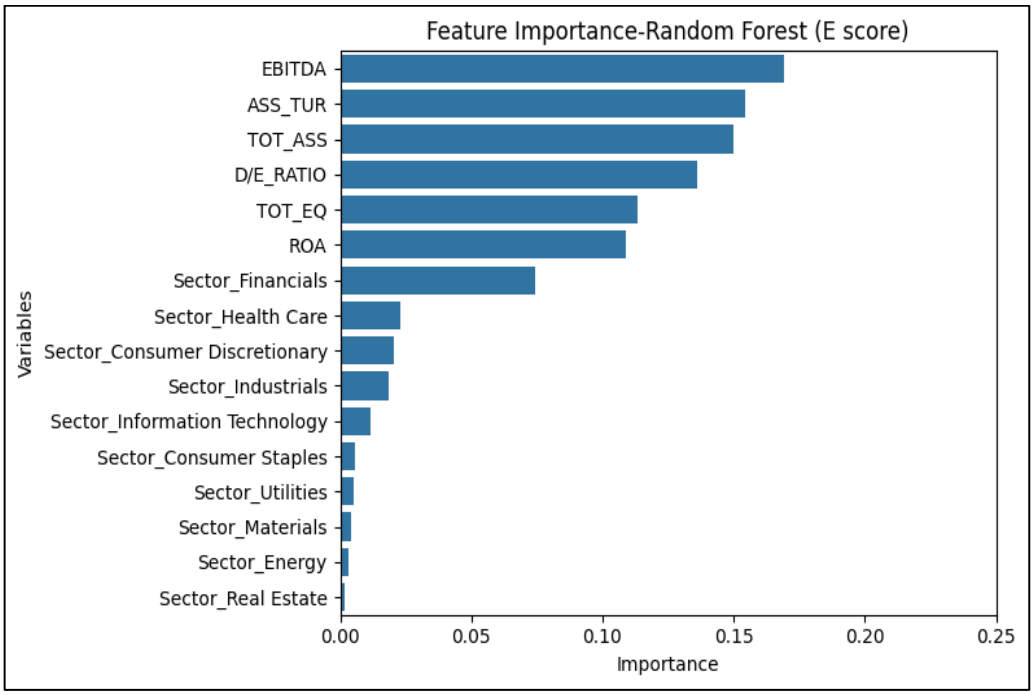

Figure 3.1: features' importance plot of the random forest on E score

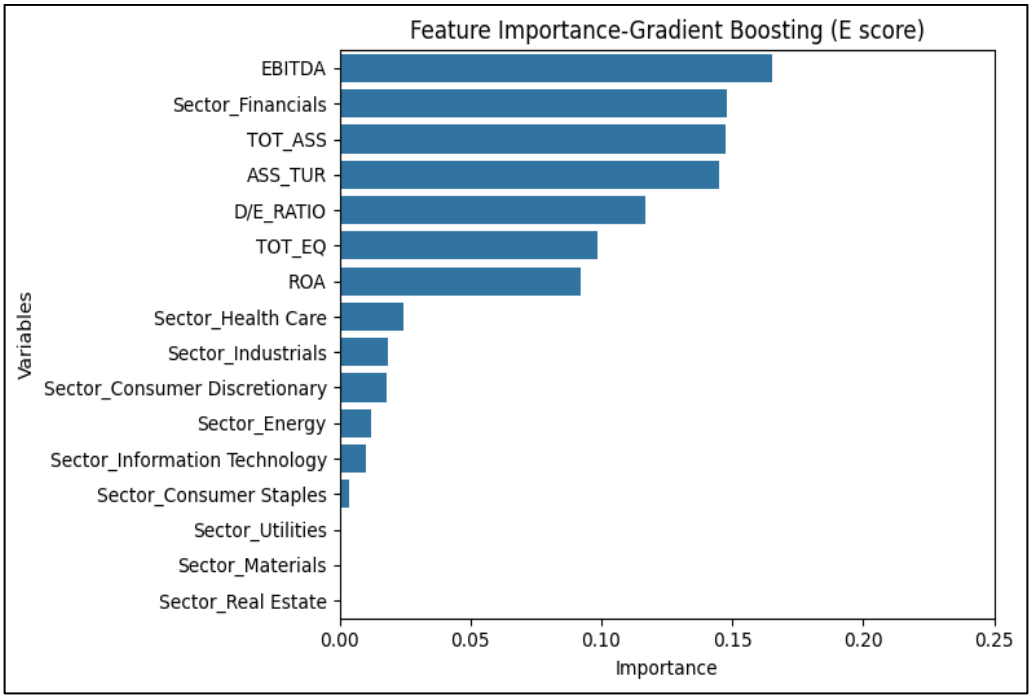

Figure 4.1: features' importance plot of the gradient boosting on E score.

5.3.3 Tree models on S score

Regarding the tree-based models on the social component, some similarities arise with both models on ESG and E score: looking at the regression tree in Figure 2.2, financial variables are the only important features of the model as they generate all of its nodes just like in the case of the ESG score. In this context, total equity emerges as the most meaningful of them for the prediction, being in the top split, while EBITDA and total assets give noticeable contributions.

In the importance plot random forest (Figure 3.2), instead, financial indicators are clearly still the most important contributors to the model's prediction, with debt-to-equity ratio being this time the best one followed by the total equity. However, like in the case of the E score, also sectoral affiliation seems to matter: some industry dummies, especially that referring to real estate sector but even that of the energy sector, exhibit levels of importance above those of the other sectors considered.

This evidence is further supported by the importance plot of the gradient boosting (Figure 3.2), where the dummy representing the real estate sector is even the most significant variable of the model while the energy industry one still has a discrete level of relevance. By the way, most of the importance is still concentrated in the financial variables as they occupy the majority of the top positions in the plot, with EBITDA being the best among them.

Overall, when the S score is used as target variable, the non-linear tree models show that, just like for the ESG score, financial indicators, especially EBITDA, total equity and debt to equity ratio, represent again the main factors of the social performance of the companies. This suggests the presence of non-linear relationships between the S score and the economic characteristics of the firms.

However, like the E score, also sectoral membership could be relevant. In particular, the affiliation to some specific industries, like those of real estate and energy, seems to really affect the companies' levels of social score.

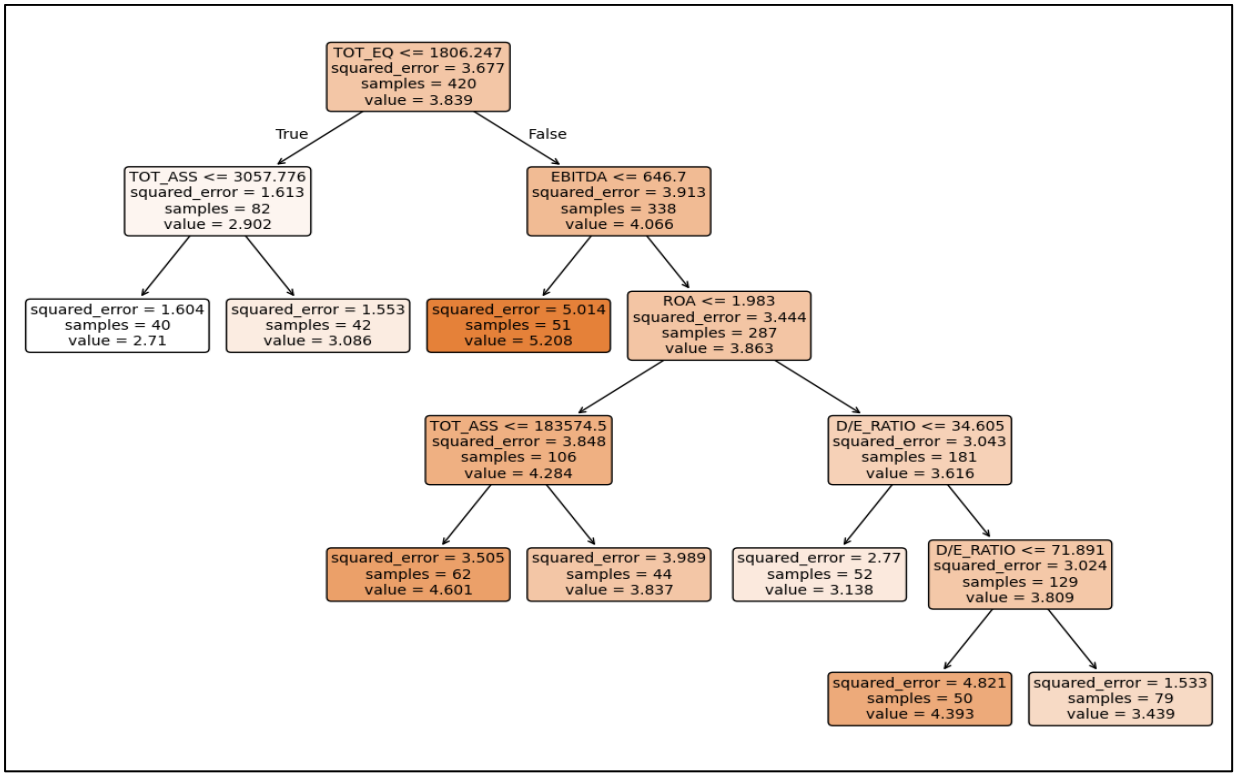

Figure 2.2: graphical representation of the regression tree on the S score

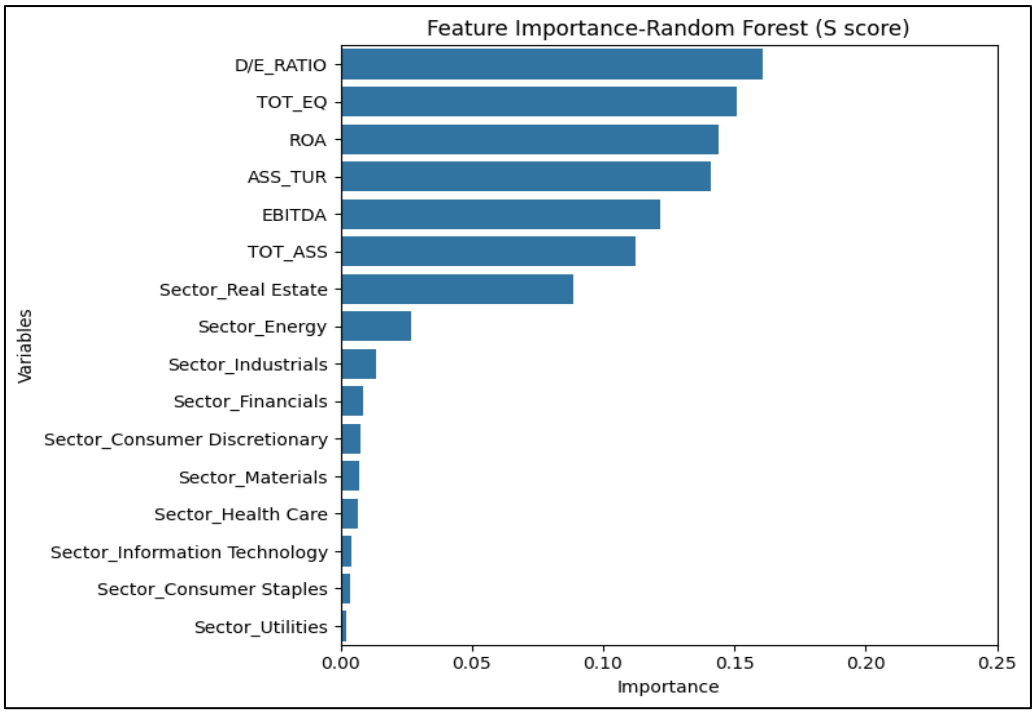

Figure 3.2: features' importance plot of the random forest on the S score

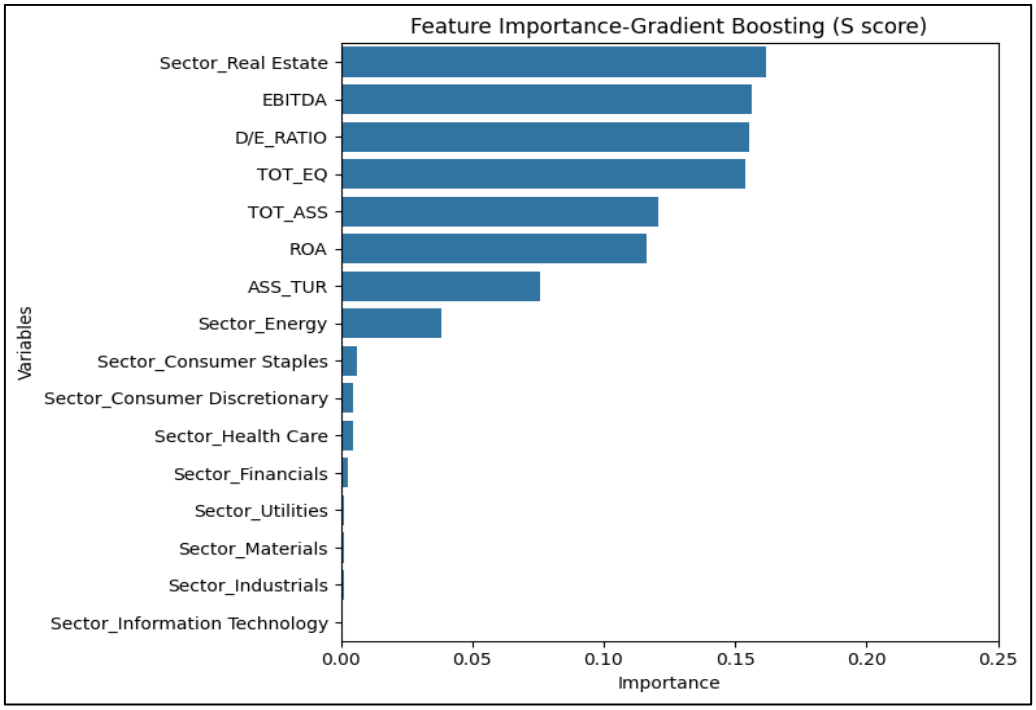

Figure 4.2: features' importance plot of the gradient boosting on the S score

5.3.4 Tree models on the G score

Finally, the tree models with the governance score as target variable reveal a similar pattern to that observed with the aggregate ESG score: first, in the regression tree (Figure 2..3), financial indicators are the only variables creating the splits, meaning that they are the only meaningful ones in the model. In particular, the debt-to-equity ratio is again the most important variable as it is selected for the top split and even last one. Also, asset turnover seems to have an important role in the model though.

The importance plots further suggest this conclusion, as the random forest plot (Figure 3.3) clearly highlights how financial data, especially to debt-to-equity ratio and asset turnover, are still the most relevant features by far while sectoral dummies show minimal levels of contribution to the prediction of the model.

Lastly, in the importance plot of gradient boosting (Figure 4.3), although the dummy representing the energy industry has a discrete level of significance compared to the others, financial indicators are still predominant in the aggregate contribution to the

prediction . In particular, asset turnover is the most important this time followed by total equity.

Overall, tree and ensemble models on the G score lead to a similar conclusion of those on ESG score: financial variables are the only significant factors for the companies' governance performance. More specifically, firms' economic characteristics such as financial leverage, net worth, or productivity seem to have a relevant impact on the G score through some non-linear relationships.

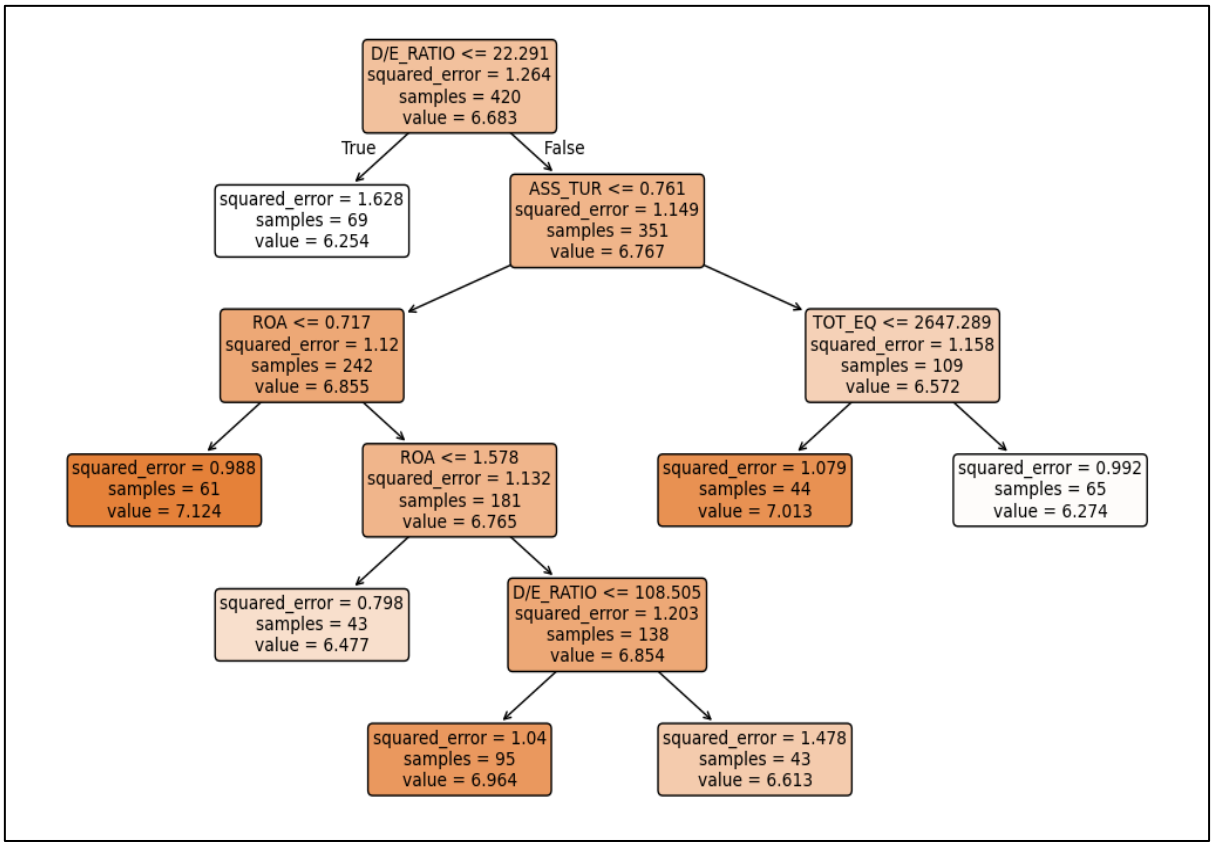

Figure 2.3: graphical representation of the regression tree on the G score

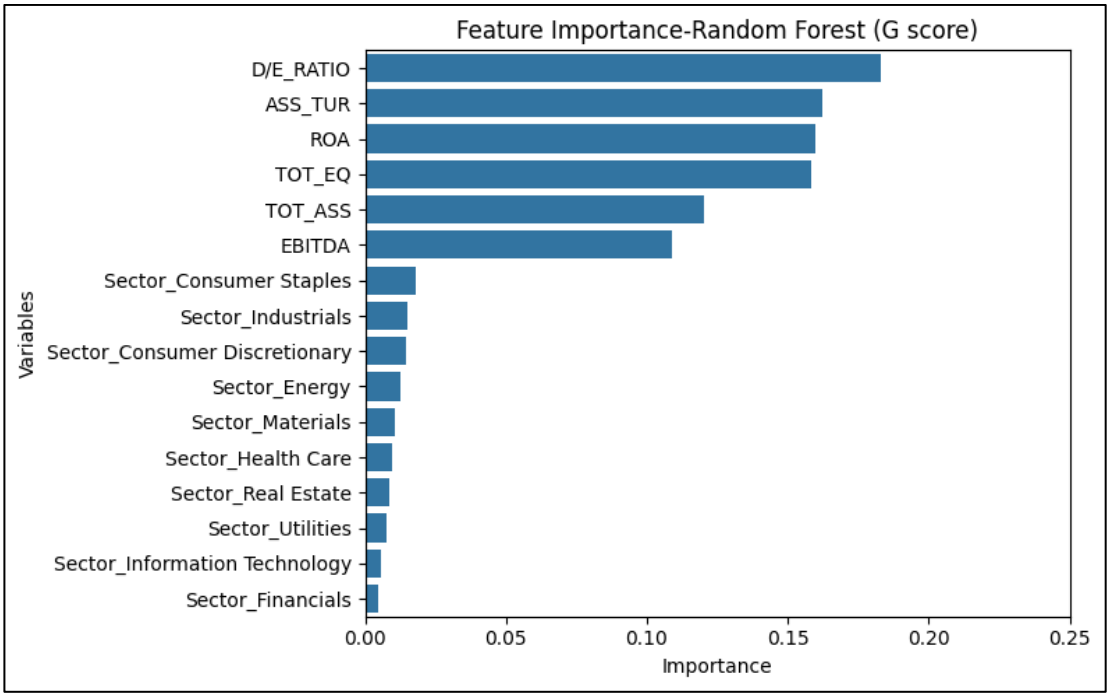

Figure 3.3: features' importance plot of the random forest on the G score

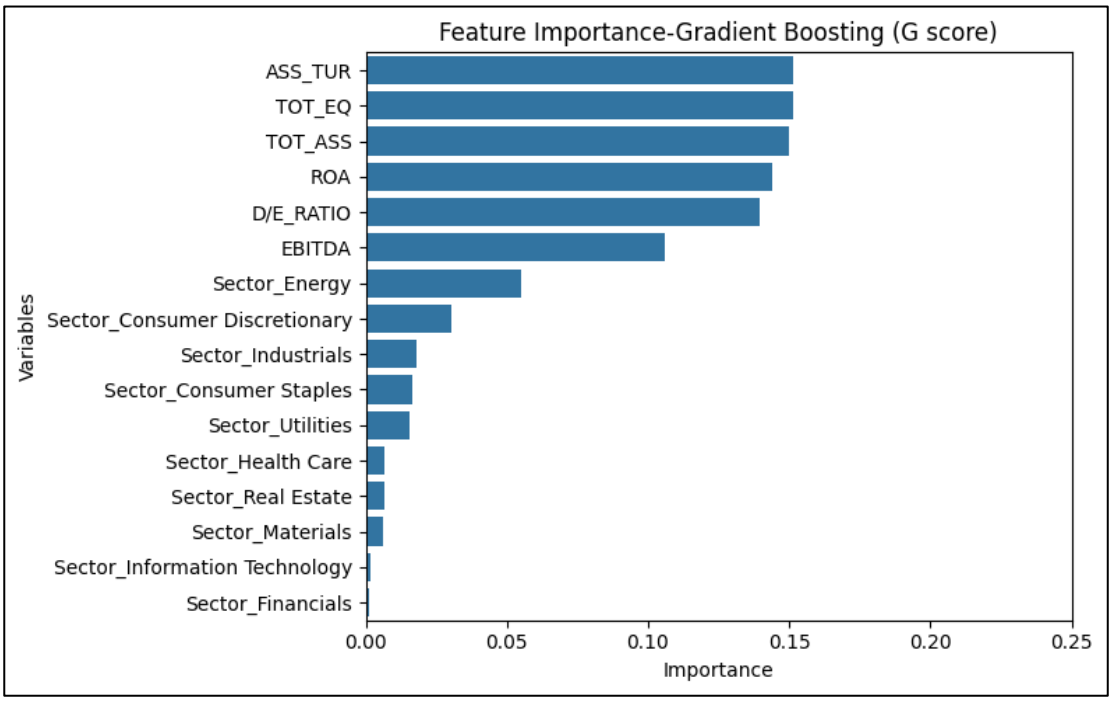

Figure 4.3: features' importance plot of the gradient boosting on the G score

5.4 Model agnostic metrics for interpretability: normalized SHAP values

5.4.1 SHAP values of models on ESG score

The comparison of the normalized mean SHAP values across models (Figure 5) reveals both consistent patterns and relevant differences in the drivers of the ESG score. In general, the financial variables emerge as the most relevant across all models.

In particular, EBITDA and total equity are consistently ranked among the most important variables in almost all specifications, highlighting the key role of firms' levels of net worth and operational efficiency in affecting their ESG performance. Moreover, the debt-to-equity ratio appears to be a dominant contributor to the predictions of tree-based models, since it clearly represents the most influential variable in them.

Regarding the sectoral dummies, they seem to be particularly more meaningful in just the linear modes and the neural network, in which sectors like financials, health care, materials or industrials constantly occupy the top positions of the rankings. This indicates that the sectoral affiliation could also play a crucial role in explaining the differences in terms of ESG score across firms.

By the way, the relative importance of financial and sectoral variables is considerably different across different types of models: on one hand, tree-based models tend to attribute basically all the aggregate importance to a limited number of financial indicators, which occupy are all indeed ranked as the most relevant variables in any of them. At the same time, linear methods as well as the neural network display more balanced approaches by distributing it also across several sectoral dummies while keeping some economic measures very important for their predictions.

In conclusion, companies' ESG levels are likely determined by a combination of economic and sectoral factors, whose contributions widely differ depending on the modeling approach adopted.

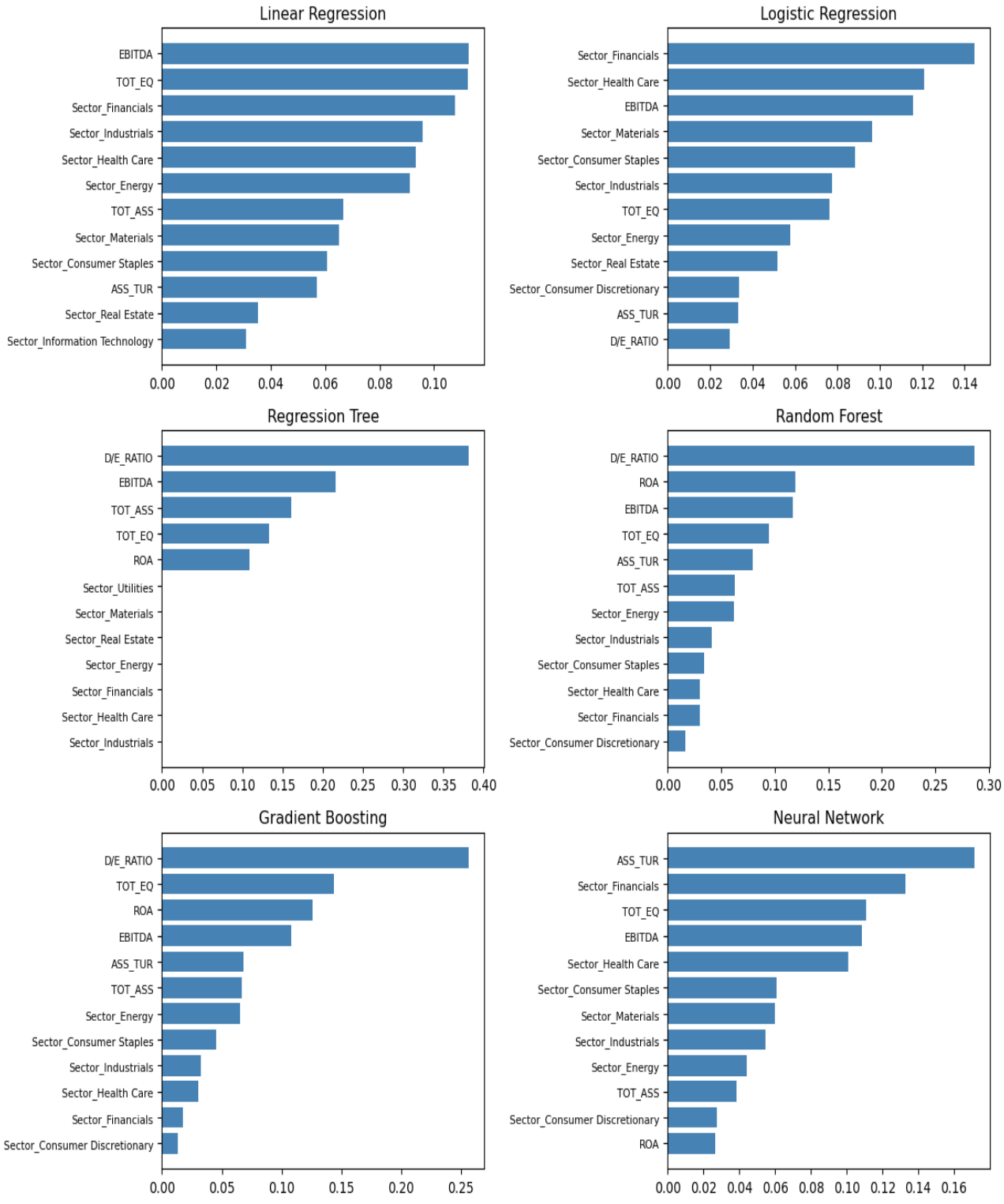

Figure 5: The ranking of the ten most important variables across different models with the ESG score as the target variable according to the normalized mean SHAP values.

5.4.2 Normalized SHAP values of models on E score

The analysis of the normalized SHAP values on the environmental score (Figure 5.1) reveals some differences with the previous ones relative to the ESG score: indeed, sectoral variables emerge as main contributors across all the considered models, with the dummy referred to the financials industry consistently ranking as the top feature in almost every specification.

Furthermore, dummies representing other sectors, such as health care, consumer discretionary, consumer staples or industrials, tend to show a considerable level of importance across the various models.

At the same time, financial data still contribute substantially to the predictions, although to a lesser extent compared to the case of the ESG score. Among them, EBITDA, total assets, and total equity appear as the most relevant indicators, especially in tree-based models but also in others like the logistic regression or the neural network.

Hence, environmental performance of the companies seem to be guided mainly by factors related to the sectors to which they belong but with their economic characteristics still playing a quite important role.

In this sense, the real difference with the models on the ESG score is the fact that the normalized SHAP values for the E score suggest that the environmental dimension is characterized by a more homogeneous interpretability across the models. Indeed, the ranking of feature importance is largely consistent with some sectoral variables, in particular the financials sector, and financial indicators emerging as predominant predictors in almost all methodologies.

This indicates that the factors determining the companies' environmental performance are less affected by the choice of modeling approach, leading to a more consistent and interpretable framework.

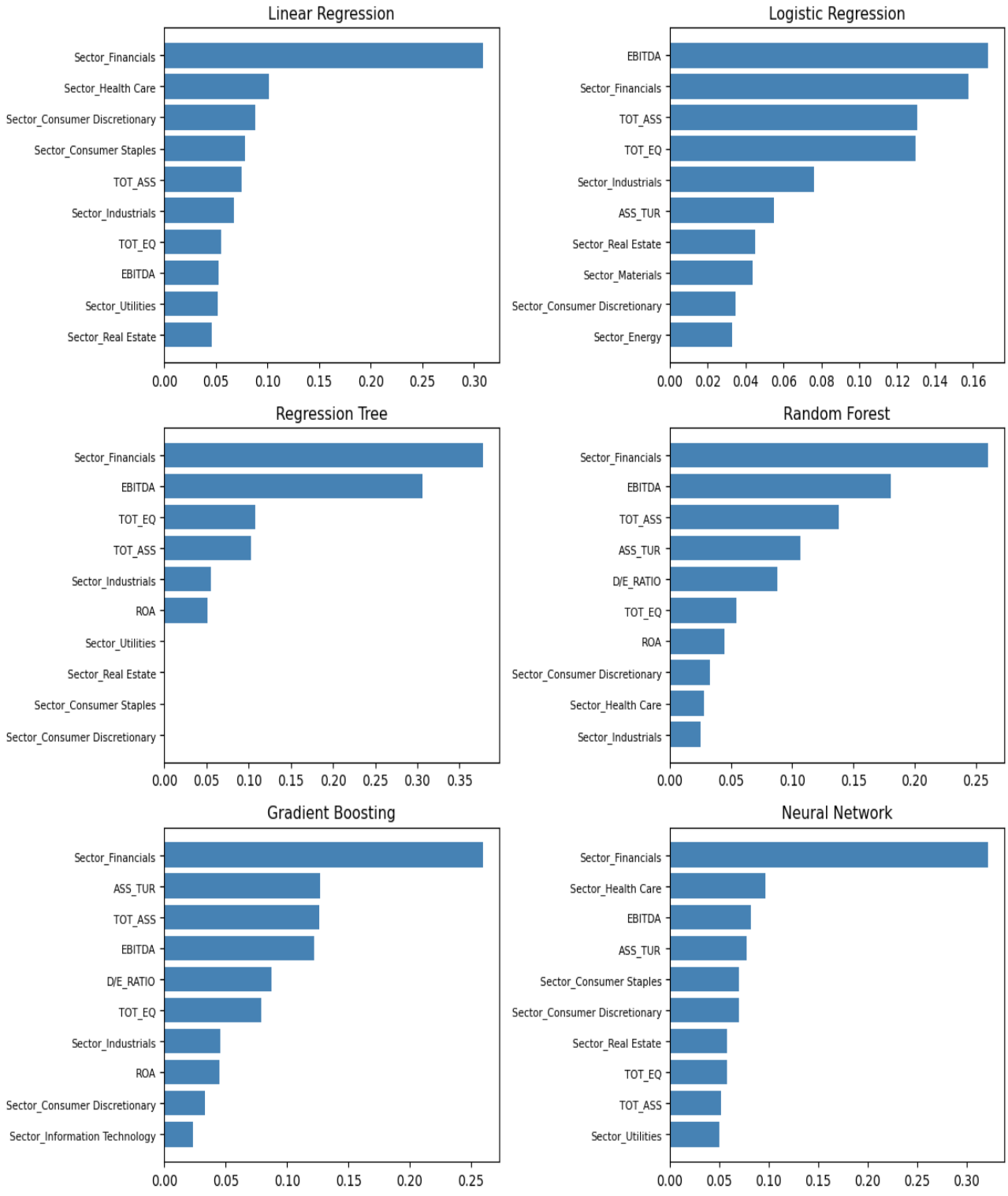

Figure 5.1: The ranking of the ten most important variables across different models with the E score as the target variable according to the normalized mean SHAP values.

5.4.3 Normalized SHAP values on S score

Moving forward with the analysis of the normalized SHAP values on the S score (Figure 5.2), the results highlight a more heterogenous scenario in terms of interpretability with respect to the E score and more like that observed with the ESG score. Indeed, the importance of financial and sectoral variables, which generally both contribute to the predictions, seems to vary a lot according to the type of model considered.

In particular, sectoral dummies, especially those representing the real estate, energy or financials sectors, emerge as the most relevant predictors of the linear regression model. On the other hand, logistic regression and the neural network tend to adopt a more balanced approach by including also financial variables like total equity, total assets, and asset turnover among their top contributors.

In contrast, tree-based models assign greater importance to financial indicators such as total equity, return on assets, and the debt-to-equity ratio, suggesting the presence of non-linear relationships between firm characteristics and social performance. By the way, the dummies representing real estate and energy industries provide important contributions to the predictions of both random forest and gradient boosting, indicating that also industry-specific factors may be relevant for them.

Overall, the normalized SHAP values outline how the social component is probably characterized by a more complex structure in which both firm-specific characteristics and industry affiliation play a key role. In fact, even though some variables like total equity or the real estate sector dummy are important in almost every model, many others differ depending on the model employed. As a result, the identification of the main drivers of social performance appears to be more sensitive to the modeling approach, leading to lower consistency in feature importance rankings compared to the environmental score.

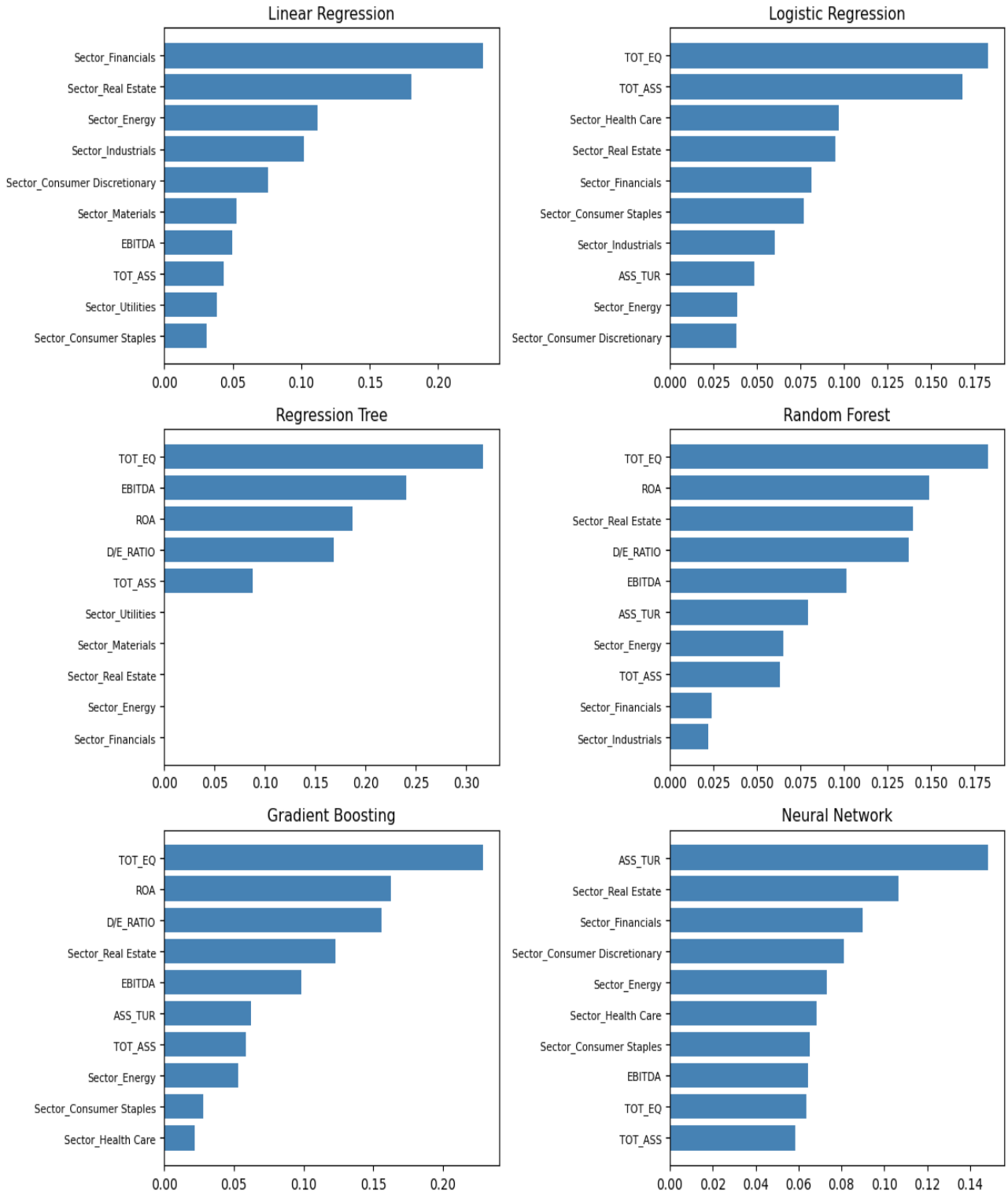

Figure 5.2: The ranking of the ten most important variables across different models with the S score as the target variable according to the normalized mean SHAP values.

5.4.4 Normalized SHAP values on the G score

Lastly, regarding the normalized SHAP values of the models on the governance score, the graphs (Figure 5.3) show a relatively consistent pattern, where financial variables are without doubt the primary contributors of the predictions. This suggests that governance performance is mainly driven by firm-specific characteristics rather than industry affiliation.

This evidence is first supported by non-linear tree models, in which all the financial indicators, especially the debt-to-equity ratio, asset turnover or total equity, are clearly the most important features in these types of models while sectoral dummies are largely ignored.

Moreover, economic measures keep being the most relevant variables in the linear models as well, where EBITDA, total equity, ROA, and total assets emerge as the top ranked features. However, unlike tree models, some sectoral variables, like industrials and utilities sectors, hold a moderate level of importance.

The neural network further supports this mixed but predominantly financial-driven structure, where both financial and sectoral variables contribute, but with firm-level characteristics remaining slightly more dominant.

All these results combined suggest that the governance component is primarily determined by the firms' economic characteristics while sectoral affiliation still plays a role but highly marginal. Regarding the consistency of interpretability across models, the level seems lower compared to the case of environmental score, but still higher than that observed for the social and ESG scores, suggesting an intermediate degree of stability in feature importance rankings. Indeed, even though some financial variables like EBITDA, ROA or total assets seem relevant in most specifications, others like debt-to-equity ratio or asset turnover tend to have great importance only in the non-linear models.

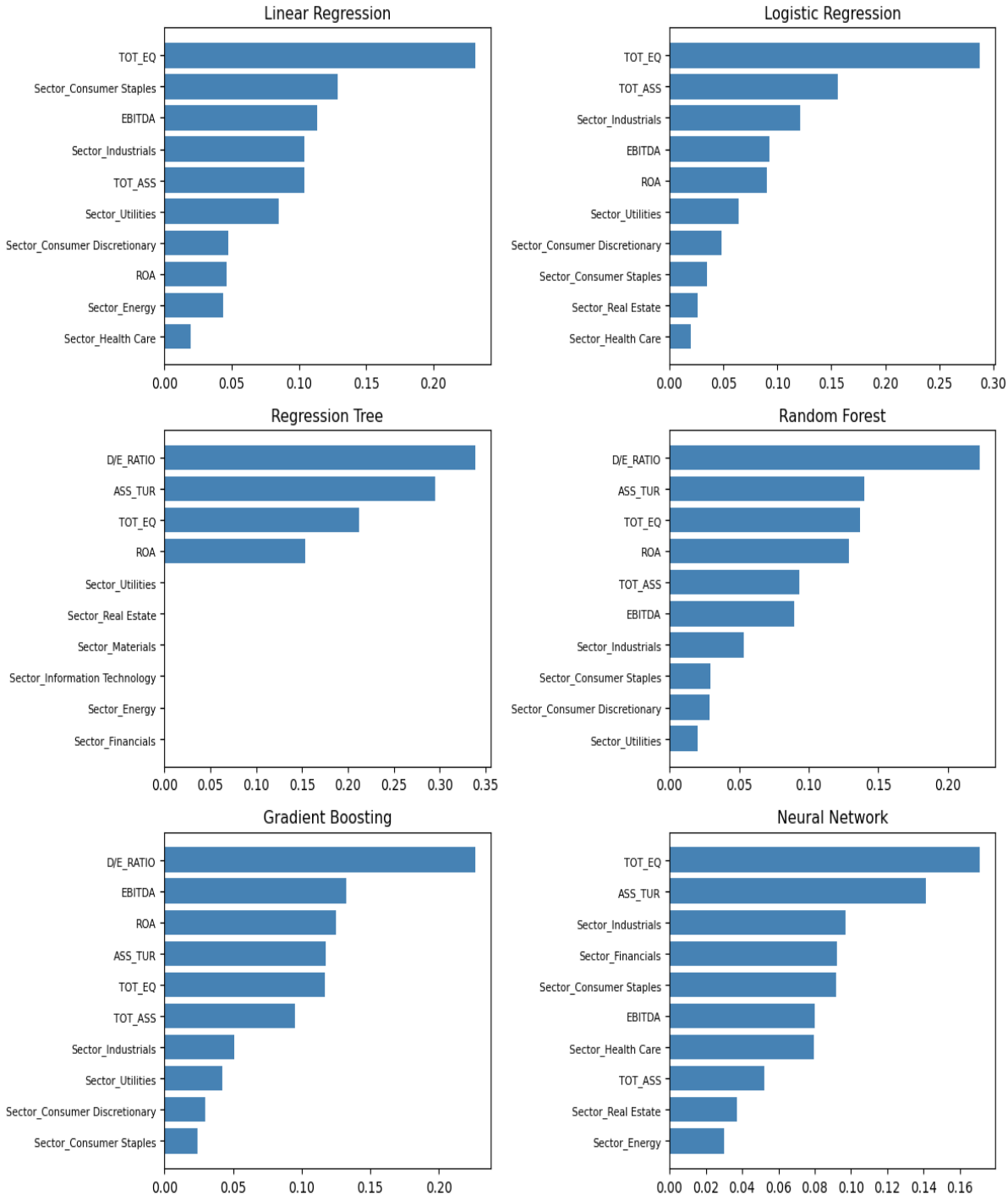

Figure 5.3: The ranking of the ten most important variables across different models with the G score as the target variable according to the normalized mean SHAP values.

5.5 Consistency and robustness of interpretability across models: correlation matrices and bootstrap procedure

5.5.1 Matrices on ESG score

The correlation matrix based on the normalized SHAP values (Figure 6) provides strong evidence regarding the consistency of the interpretability across the different models with the ESG score as the target variable. In fact, it reveals a clear clustering structure with two distinct group of models: on one hand, linear regression, logistic regression and the neural network exhibit high pairwise correlations, meaning that they are similar in the terms of features' ranking and they capture a common and mainly linear structure in the data. On the other hand, the regression tree, random forest and boosting also manifest high levels of correlation with each other, especially the last two. This again means that these models tend to rank the same features as the most important for their predictions.

In contrast, the correlations between these two groups are consistently low, and in some cases even negative, highlighting a marked divergence in how linear and non-linear approaches interpret the relationship between firms' economic characteristics and sectoral affiliation with the ESG performance.

Moreover, the comparison between the original correlation bootstrap and the one generated through the bootstrap out-of-bag with 50 different iterations shows that the identified structure remains stable under resampling as the latter (Figure 7) still generates the two groups of models described before. This proves the robustness of the findings on the ESG score.

Lastly, these findings are perfectly coherent with those observed by features' rankings in the former section, which have marked a distinction between linear methods, driven largely by sectoral effects and few selected financial indicators, and the non-linear tree-based ones, where financial variables play a largely more dominant role.

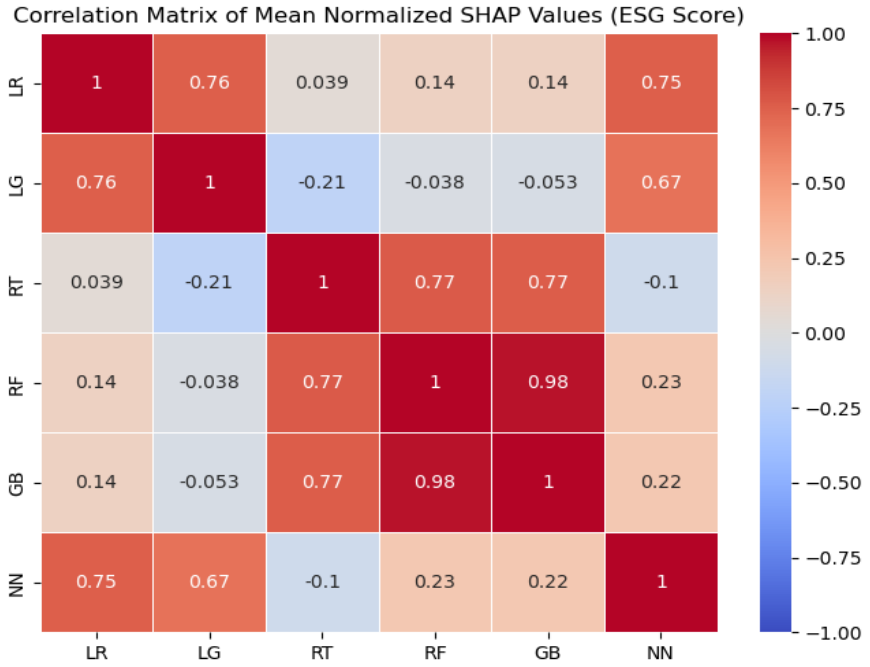

Figure 6: correlation matrix based on the normalized mean SHAP values of the different models on the ESG score obtained on the single train test split.

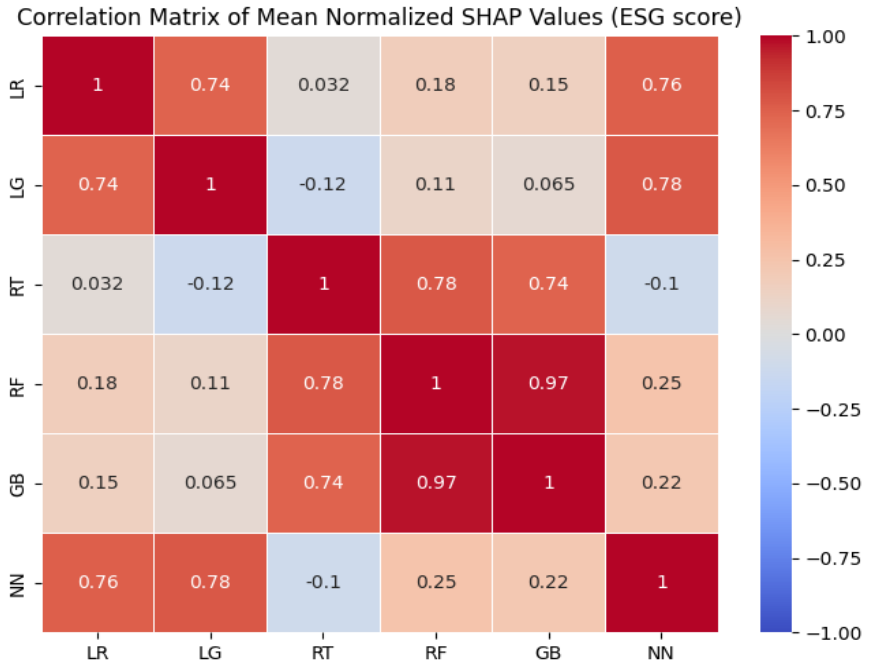

Figure 7: correlation matrix based on the normalized mean SHAP values of the different models on the ESG score obtained through the 50 bootstrap iterations

5.5.2 Correlation matrices of E score

The correlation matrix for the E score (Figure 6.1) reveals a less polarized structure compared to the ESG case. While tree-based models still exhibit strong internal consistency, particularly between random forest and gradient boosting, the distinction between linear and non-linear approaches appears less pronounced.

In particular, logistic regression shows relatively high correlations with tree-based methods, especially with the regression tree, as well as with the neural network, suggesting that different types of specifications can capture similar patterns in the data. The neural network and linear regression, in turn, maintain an intermediate position, being strongly correlated with each other while also displaying moderate associations with the other models.

Overall, the absence of a clear separation between linear and non-linear methods points to a more homogeneous interpretability structure compared to the ESG score, with feature importance rankings appearing broadly similar across models. This evidence is coherent with the results obtained from the analysis of normalized SHAP values, where both sectoral variables, mainly that referring to the Financials sector, and specific financial indicators, such as EBITDA and total equity, consistently emerged as key contributors across most specifications.

As a result, the determinants of the E score appear to be less sensitive to the choice of modeling approach, leading to a more stable, robust, and interpretable framework. This conclusion is further supported by the bootstrap out-of-bag correlation matrix (Figure 7.1), which exhibits basically identical correlation patterns and consistency of interpretability across different models. Thereby, this reinforces the reliability and consistency of the findings.

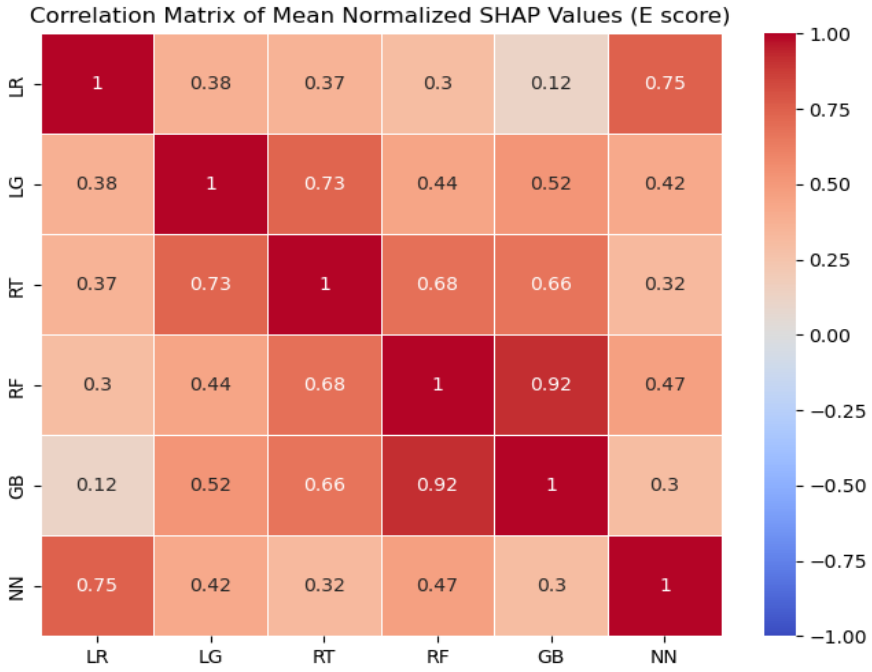

Figure 6.1: correlation matrix based on the normalized mean SHAP values of the different models on the E score obtained on the single train test split.

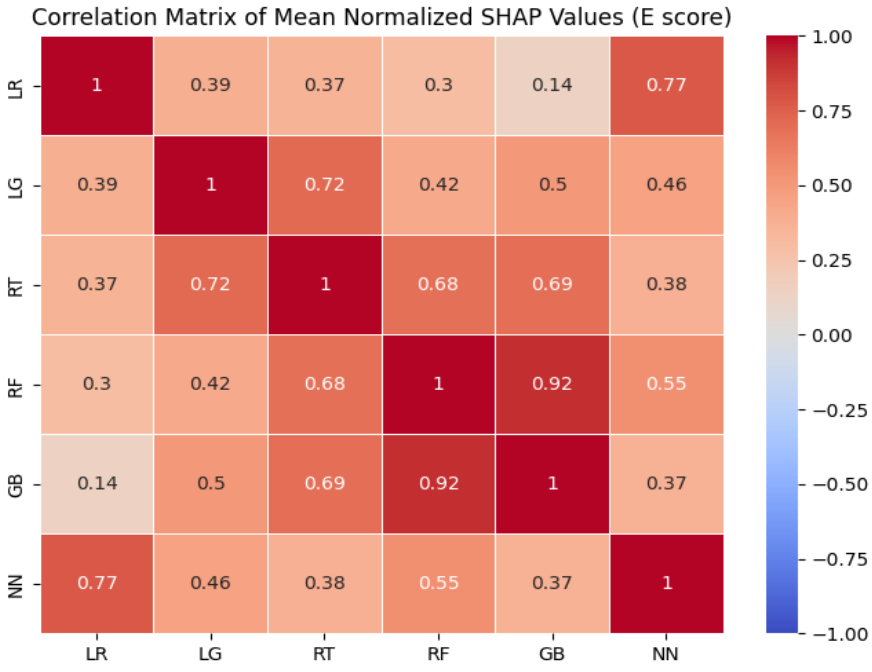

Figure 7.1: correlation matrix based on the normalized mean SHAP values of the different models on the E score obtained through the 50 bootstrap iterations

5.5.3 Correlation matrices of the S score

Moving forward, the correlation matrix based on the normalized SHAP values of the models on the S score (Figure 6.2) reveals a more fragmented and less stable interpretability structure with respect to that observed in the case of the E score.

Indeed, even though the tree models still have strong pairwise correlations, especially random forest and gradient boosting, their correlation with the other methods becomes weaker. In this sense, linear regression shows a very low and even negative, proving that it generates a completely different ranking of the features with respect to such models. Logistic regression occupies an intermediate position, displaying slightly positive correlations with gradient boosting and random forest while a higher one with the neural network. The latter acts as a partial bridge between linear and non-linear approaches, showing relatively strong associations with linear models and weaker correlations with ensemble methods.

These results indicate that, unlike the E score, the determinants of the social component are more sensitive to the choice of modeling approach, leading to a more heterogeneous interpretability structure more similar to that of the ESG score. This evidence is consistent with the findings from the analysis of the normalized SHAP values, where both financial and sectoral variables have emerged as relevant have contributors, but with rankings that vary considerably across models except for variables like total equity or the real estate sector dummy.

This pattern is further confirmed by the correlation matrix obtained through the bootstrap out-of-bag procedure (Figure 7.2), which presents very similar values and preserves the same structural relationships among models. The consistency between the two matrices suggests that the observed results are robust and not driven by sample-specific effects, reinforcing the conclusion that the identification of the main drivers of the S score depends significantly on the modeling technique adopted.

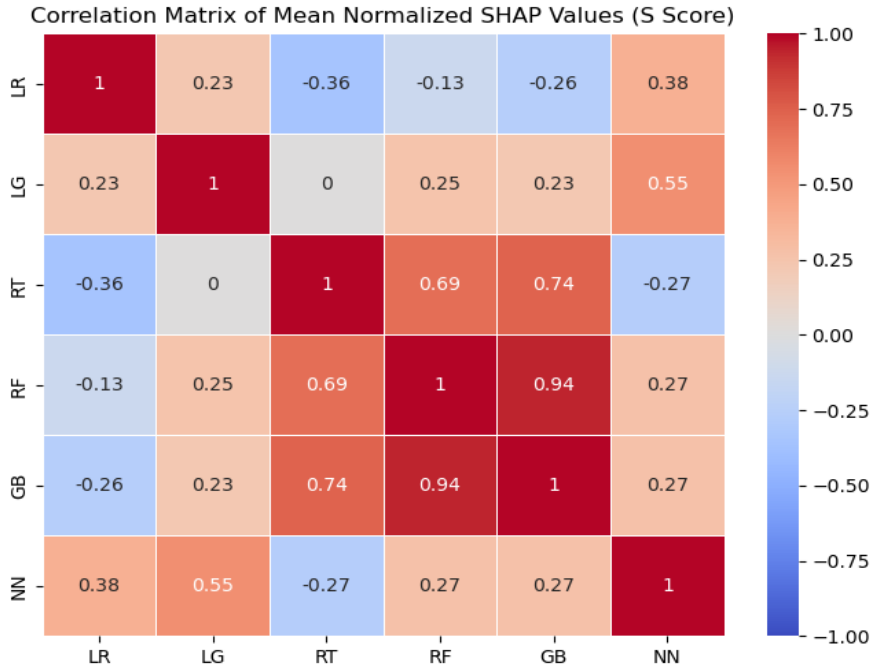

Figure 6.2: correlation matrix based on the normalized mean SHAP values of the different models on the S score obtained on the single train test split.

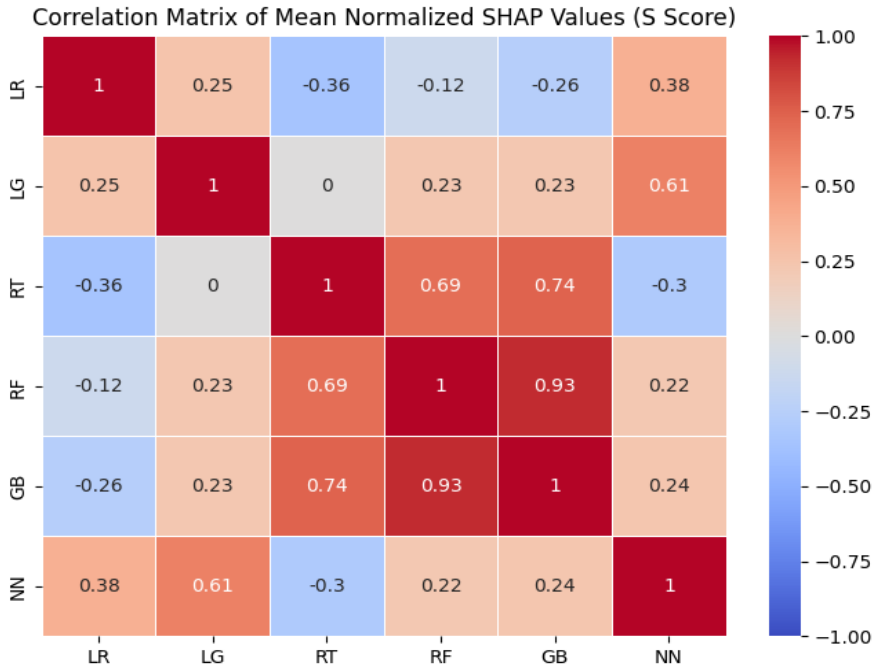

Figure 7.2: correlation matrix based on the normalized mean SHAP values of the different models on the S score obtained through the 50 bootstrap iterations

5.5.4 Correlation matrices of the G score

Finally, looking at the correlation matrix of the governance score (Figure 6.3), a more coherent and less polarized explainability framework emerges compared to that relative to the S score.

In particular, linear regression and logistic regression have a strong correlation between themselves. On the other hand, tree-based models keep a high level of internal coherence, confirming that these methods capture almost identical patterns. In this regard, the correlation matrix for S score first seems to identify a similar structure to that created by the correlation matrix relative to the aggregate ESG score, with two clearly distinct groups of models in terms of explainability.

However, unlike the case of ESG score, the relationship between linear and non-linear tree models appears less conflicting, as most correlations are positive and of moderate magnitude. This suggests that, despite methodological differences, the various models tend to identify broadly similar drivers of the governance component. The only exceptions are the regression tree, which doesn't show any alignment with linear models, and neural network, which keeps a more isolated position, displaying relatively low correlations with the other approaches, although without generating strong divergences.

In conclusion, these findings seem to confirm those obtained from the analysis of the normalized SHAP values, where financial variables, such as total equity, total assets, and EBITDA, consistently emerged as key contributors across most models, leading to a relatively stable and reliable identification of the main drivers of corporate governance performance.

Lastly, the comparison with the correlation matrix created through the bootstrap out-of-bag procedure (Figure 7.3), which presents very similar values and preserves the same structural relationships among models, highlights the robustness of the results and suggests that the observed interpretability structure is not driven by sample-specific effects.

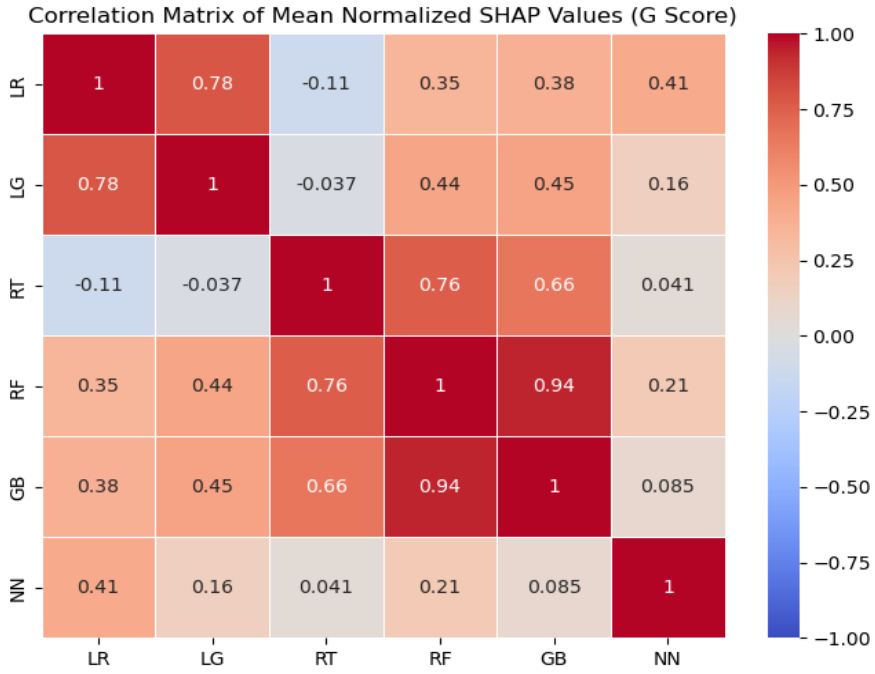

Figure 6.3: correlation matrix based on the normalized mean SHAP values of the different models on the G score obtained on the single train test split.

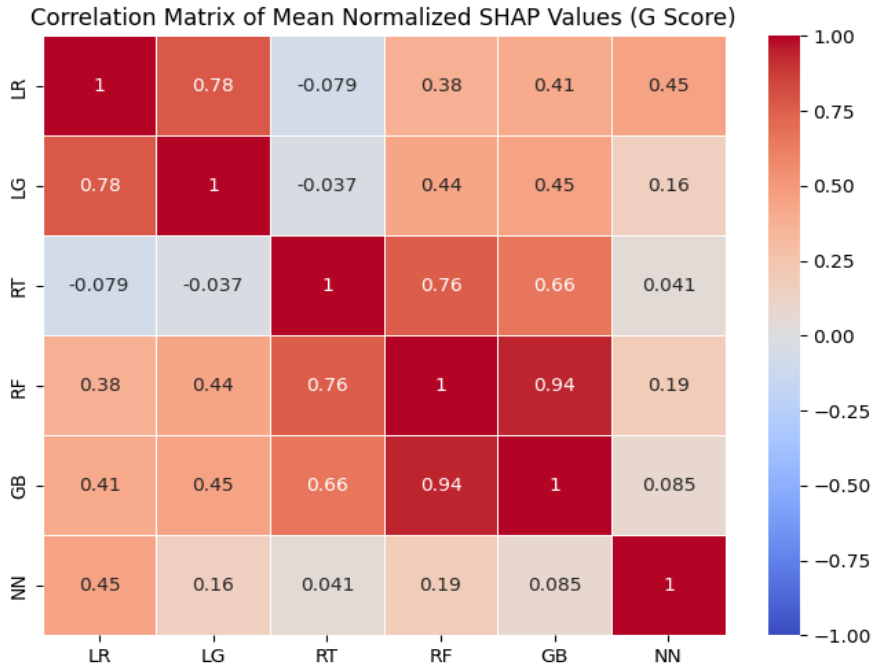

Figure 7.3: correlation matrix based on the normalized mean SHAP values of the different models on the G score obtained through the 50 bootstrap iterations

5.6 Results review

In summary, the analysis conducted in this chapter, based on both model-specific and model-agnostic techniques to access explainability, highlights substantial differences in interpretability across the ESG score and its three components, as well as across the various modeling approaches.

For the E score, sectoral affiliation, particularly to the financials sector, emerges as the main determinant of firms' environmental performance, while financial variables such as EBITDA, total assets, and total equity still play a significant role. Moreover, these findings are remarkably consistent across different models and interpretability techniques. Indeed, the same variables metioned above systematically appear among the most important predictors, resulting in a robust and stable interpretability framework that is only marginally affected by the choice of methodology.

A relatively homogeneous pattern is also observed for the G score, especially when comparing linear models and tree-based ensemble methods. In this case, the interpretability structure is strongly centered on financial indicators, which represent then the primary drivers of governance performance. Variables such as total equity, EBITDA, and total assets consistently rank among the most relevant features across models. However, the degree of homogeneity is slightly lower than in the E score, mainly due to the more isolated behavior of specific models, such as the neural network and the regression tree. Another reason behind the smaller degree of consistency may be that linear models tend to rank certain financial variables higher while non-linear models give more importance to other financial data (D/E ratio for instance).

In contrast, the S score exhibits a much higher degree of heterogeneity in terms of interpretability. Tree-based models assign substantial importance to financial indicators, while still recognizing some sectoral variables, particularly those related to the real estate and energy sectors, as relevant contributors. Linear regression, on the other hand, tends to prioritize sectoral dummies almost exclusively, whereas logistic regression and the neural network adopt more balanced approaches, distributing importance across both financial and sectoral features. As a result, the interpretability structure appears

fragmented, with tree-based models representing only clearly consistent group. This suggests that the identification of the main drivers of social performance is highly sensitive to the modeling approach.

Finally, the aggregate ESG score, resulting as a combination of its three different dimensions, also displays a heterogeneous interpretability framework but characterized by a clear distinction between model families. Linear models, together with the neural network, tend to emphasize sectoral variables as the dominant predictors, while still assigning some relevance to financial indicators. In contrast, tree-based models predominantly rank financial variables as the most important features, largely disregarding sectoral effects. This divergence further confirms that, for the ESG score, interpretability is strongly influenced by the choice of methodology.

Chapter 6: Conclusions

To summarize, this study has conducted an analysis of 600 European firms belonging to the STOXX Europe 600 index. In this sense, the work has tried to predict the Bloomberg ESG score and its three pillars, the environmental, social and governance scores, through six different ML techniques, namely linear regressions, logistic regressions, regression trees, random forest, gradient boosting and neural networks. The predictors or explanatory variables employed are six financial indicators, specifically total assets, EBITDA, ROA, total equity, asset turnover, D/E ratio, but also ten different sectoral dummies, representing the sectors to which the various firms belong. The main purpose was in fact to understand whether firms' economic characteristics and affiliation to some specific sectors could affect their ESG performance.

In this sense, the analysis has been conducted on two main lines: predictive performance of the ML models and their interpretability. Regarding the first aspect, metrics such as R^2 score, RMSE, and MAE have been adopted for all models on all four different scores. For the second, the analysis was based on the observation of the model-specific explainability tools, except for the neural network, as well as of the model-agnostic normalized SHAP values. Furthermore, the second ones have been then used to create a 6X6 Spearman correlation matrix for all four scores to identify possible interpretability patterns across different models in terms of variables' ranking. Lastly, a bootstrap out-of-bag approach was also implemented to assess the robustness and consistency of the findings both in terms of predictive performance and interpretability.

The study has highlighted different outcomes for the ESG score and its three components. Starting from the predictive performance, the analysis revealed that the overall explanatory power of the considered financial and sectoral variables is heavily limited: consistently low R^2 values for the ESG score, as well its components, have been found across all models. This indicates that firms' economic characteristics and sectoral affiliation play a minimal role in determining their sustainability performance. However, the models on the E score always display higher levels of R^2 compared to their

counterparts on the other scores, suggesting a higher degree of predictability with respect to the S and G scores through the considered features. This indicates that the environmental dimension's variability is more closely linked to measurable firm-level and sector-specific factors.

The interpretability analysis confirmes these patterns. For the E score, feature importance was consistent across models, with sectoral affiliation, especially the financials sector, and key financial indicators such as EBITDA, total assets, and total equity emerging as the primary drivers. The G score also displayed a relatively coherent structure, centered on financial metrics, although with slightly less consistency. In contrast, the S score showed a fragmented interpretability pattern, with significant variability across models, indicating that its main drivers are highly sensitive to the methodological approach. The aggregate ESG score also displayed heterogeneity, with linear models and neural networks emphasizing sectoral variables, while tree-based models focused on financial indicators.

In conclusion, the environmental component of the ESG score emerges as the best metric both in terms of predictability and interpretability being the most homogenous. This can be explained by its direct link to tangible, measurable firm-level activities and sector-specific regulatory obligations, such as energy efficiency, emissions reduction, and compliance with environmental standards. These factors are moderately captured by financial indicators and sector affiliation, and their effects are relatively consistent across firms and industries. In contrast, the S and G scores are largely driven by qualitative or cultural factors that are almost unobservable in financial data and vary more across firms, making them more difficult to predict and leading to more fragmented interpretability across different modeling approaches.

Regarding instead the comparison across different models, tree-based methods, and gradient boosting in particular, demonstrated better performance despite still being very low. Their ability to capture non-linear relationships and interactions between financial variables and sectoral affiliation, combined with robustness across bootstrap iterations, makes them more reliable and accurate than linear models and neural networks. Neural networks, although flexible, exhibited high instability and lower predictive accuracy in this context.

Furthermore, tree-based models show the highest consistency in feature importance rankings, identifying the same variables as most relevant. These are predominantly financial indicators, with a few sector-specific exceptions, such as the financials sector for the E score and the real estate sector for the S score, highlighting both the primary role of financial characteristics in explaining ESG performance and the robustness of interpretability provided by tree-based approaches. On the other hand, linear regression always tends to consider sectoral variables as its main contributors while logistic regression and neural network adopt different approaches depending on the score they are trying to predict. This leads to different interpretability patterns between them and the tree-based models across the four different scores.

In conclusion, the findings highlight both the strengths and weaknesses of using machine learning to analyze ESG performance. While certain dimensions, particularly the environmental one, are partially predictable and interpretable, ESG scores as a whole remain complex and influenced by factors beyond standard financial or sectoral membership. The study also demonstrates the importance of model choice, showing that both predictive performance and interpretability can vary substantially depending on the methodology adopted. Lastly, the work proves the importance of the bootstrap OOB procedure, which, in both cases, has confirmed the results obtained on the single train-test splits, guaranteeing their consistency and robustness across different samples.

References

-Ahelegbey, Daniel Felix, and Paolo Giudici. “Multidimensional Inequality Metrics for Sustainable Business Development.” *Mathematics*, vol. 12, no. 22, 20 Nov. 2024, p. 3633, https://doi.org/10.3390/math12223633.

-Alessandro Del Vitto, et al. “ESG Ratings Explainability through Machine Learning Techniques.” *Annals of Operations Research*, 19 July 2023, https://doi.org/10.1007/s10479-023-05514-z.

-Belkhiria, Sina, et al. “Predicting Environmental Social and Governance Scores: Applying Machine Learning Models to French Companies.” *Journal of Risk and Financial Management*, vol. 18, no. 8, 26 July 2025, pp. 413–413, https://doi.org/10.3390/jrfm18080413.

-Berg, Florian, et al. “Aggregate Confusion: The Divergence of ESG Rating.” *Review of Finance*, vol. 26, no. 6, 23 May 2022, pp. 1315–1344, https://doi.org/10.1093/rof/rfac033.

-Billio, Monica, et al. “Inside the ESG Ratings: (Dis)Agreement and Performance.” *Corporate Social Responsibility and Environmental Management*, vol. 28, no. 5, 1 Sept. 2021, pp. 1426–1445, https://doi.org/10.1002/csr.2177.

-Breiman, Leo. “Bagging Predictors.” *Machine Learning*, vol. 24, no. 2, Aug. 1996, pp. 123–140, link.springer.com/article/10.1007/BF00058655.

-Calzarossa, Maria Carla, et al. “An Assessment Framework for Explainable AI with Applications to Cybersecurity.” *Artificial Intelligence Review*, vol. 58, no. 5, 27 Feb. 2025, https://doi.org/10.1007/s10462-025-11141-w.

-Chase, Randy J., et al. “A Machine Learning Tutorial for Operational Meteorology. Part I: Traditional Machine Learning.” *Weather and Forecasting*, vol. 37, no. 8, Aug. 2022, pp. 1509–1529, https://doi.org/10.1175/waf-d-22-0070.1.

-Cicchiello, Antonella Francesca, et al. “Non-Financial Disclosure Regulation and Environmental, Social, and Governance (ESG) Performance: The Case of EU and US Firms.” *Corporate Social Responsibility and Environmental Management*, vol. 30, no. 3, 28 Oct. 2022, https://doi.org/10.1002/csr.2408.

-De Lucia, Caterina, et al. “Does Good ESG Lead to Better Financial Performances by Firms? Machine Learning and Logistic Regression Models of Public Enterprises in Europe.” *Sustainability*, vol. 12, no. 13, 1 July 2020, p. 5317, https://doi.org/10.3390/su12135317.

-Demiraj, Rezart, et al. “ESG Scores Relationship with Firm Performance: Panel Data Evidence from the European Tourism Industry.” *Pressacademia*, 31 Jan. 2023, https://doi.org/10.17261/pressacademia.2023.1674. Accessed 16 Feb. 2023.

-Gonçalves, Tiago Cruz, et al. “Environmental, Social and Governance Scores in Europe: What Drives Financial Performance for Larger Firms?” *Economics and Business Letters*, vol. 12, no. 2, 13 July 2023, pp. 121–131, reunido.uniovi.es/index.php/EBL/article/view/18659, https://doi.org/10.17811/ebl.12.2.2023.121-131.

-Helfaya, Akrum, et al. “Investigating the Factors That Determine the ESG Disclosure Practices in Europe.” *Sustainability*, vol. 15, no. 6, 21 Mar. 2023, p. 5508, www.mdpi.com/2071-1050/15/6/5508, https://doi.org/10.3390/su15065508.

-Hu, Linwei, et al. “Supervised Machine Learning Techniques: An Overview with Applications to Banking.” *International Statistical Review*, vol. 89, no. 3, 4 May 2021, https://doi.org/10.1111/insr.12448.

-Junius, David, et al. “THE IMPACT of ESG PERFORMANCE to FIRM PERFORMANCE and MARKET VALUE.” *Jurnal Aplikasi Akuntansi*, vol. 5, no. 1, 31 Oct. 2020, pp. 21–41, https://doi.org/10.29303/jaa.v5i1.84.

-Koundouri, Phoebe, et al. “The Impact of ESG Performance on the Financial Performance of European Area Companies: An Empirical Examination.” *ICSD 2021*, vol. 15, no. 1, 20 Sept. 2021, https://doi.org/10.3390/environsciproc2022015013.

-Lee, Kahyun. “Bridging Financial Disclosures and ESG Ratings: A Data-Driven Predictive Framework.” *Quantitative Finance and Economics*, vol. 10, no. 1, 2026, pp. 86–107, https://doi.org/10.3934/qfe.2026005. Accessed 24 Feb. 2026.

-Lin, Hsio-Yi, and Bryant Hsu. “Empirical Study of ESG Score Prediction through Machine Learning—a Case of Non-Financial Companies in Taiwan.” *Sustainability*, vol. 15, no. 19, 23 Sept. 2023, pp. 14106–14106, https://doi.org/10.3390/su151914106.

-María Garrido-Ruso, et al. “Does ESG Implementation Influence Performance and Risk in Smes?” *Corporate Social-Responsibility and Environmental Management*, vol. 31, no. 5, 4 Apr. 2024, https://doi.org/10.1002/csr.2783.

-Palynska, Marta, et al. “The Impact of the ESG Factor on Industrial Performance an Analysis Using Machine Learning Techniques .” *Ssrn.com*, 17 June 2024, papers.ssrn.com/sol3/papers.cfm?abstract_id=4910370.

-Patel, Sanskruti, et al. “Predicting ESG Scores Using Machine Learning for Data-Driven Sustainable Investment.” *Analytics*, vol. 5, no. 1, 9 Jan. 2026, pp. 7–7, https://doi.org/10.3390/analytics5010007.

-Pulino, Silvia Carnini , et al. “Does ESG Disclosure Influence Firm Performance?” *Sustainability*, vol. 14, no. 13, 22 June 2022, p. 7595, https://doi.org/10.3390/su14137595.

-Raza, Hassan, et al. “Applying Artificial Intelligence Techniques for Predicting the Environment, Social, and Governance (ESG) Pillar Score Based on Balance Sheet and Income Statement Data: A Case of Non-Financial Companies of USA, UK, and Germany.” *Frontiers in Environmental Science*, vol. 10, 4 Oct. 2022, https://doi.org/10.3389/fenvs.2022.975487.

-Rezart Demiraj, et al. “The Moderating Role of Worldwide Governance Indicators on ESG–Firm Performance Relationship: Evidence from Europe.” *Journal of Risk and Financial Management*, vol. 18, no. 4, 14 Apr. 2025, pp. 213–213, https://doi.org/10.3390/jrfm18040213.

-Sassen, Remmer, et al. “Impact of ESG Factors on Firm Risk in Europe.” *Journal of Business Economics*, vol. 86, no. 8, 23 Apr. 2016, pp. 867–904, https://doi.org/10.1007/s11573-016-0819-3.

-Scornet, Erwan, et al. “Consistency of Random Forests.” *The Annals of Statistics*, vol. 43, no. 4, Aug. 2015, pp. 1716–1741, projecteuclid.org/euclid.aos/1434546220, https://doi.org/10.1214/15-aos1321.

-Shalhoob, Hebah, and Khaled Hussainey. “Environmental, Social and Governance (ESG) Disclosure and the Small and Medium Enterprises (SMEs) Sustainability Performance.” *Sustainability*, vol. 15, no. 1, 23 Dec. 2022, p. 200. *mdpi*, www.mdpi.com/2071-1050/15/1/200, https://doi.org/10.3390/su15010200.

-Tahmid, Tahani, et al. “Does ESG Initiatives Yield Greater Firm Value and Performance? New Evidence from European Firms.” *Cogent Business & Management*, vol. 9, no. 1, 20 Nov. 2022, https://doi.org/10.1080/23311975.2022.2144098.

-Wang, Jingyuan, and Jiahao Ji. “A Tutorial on Regression Analysis: From Linear Models to Deep Learning -- Lecture Notes on Artificial Intelligence.” *ArXiv.org*, 2025, arxiv.org/abs/2512.04747. Accessed 5 Apr. 2026.

-Zahid, R.M. Ammar, et al. “The Role of Audit Quality in the ESG-Corporate Financial Performance Nexus: Empirical Evidence from Western European Companies.” *Borsa Istanbul Review*, vol. 22, no. 2, Sept. 2022, https://doi.org/10.1016/j.bir.2022.08.011.

-Zhang, Jianfeng, and Zexin Zhao. “Corporate ESG Rating Prediction Based on XGBoost-SHAP Interpretable Machine Learning Model.” *Expert Systems with Applications*, vol. 295, 7 July 2025, p. 128809, www.sciencedirect.com/science/article/pii/S0957417425024273?ssrnid=5246044&dgcid=SSRN_redirect_SD, https://doi.org/10.1016/j.eswa.2025.128809.

-European Banking Authority. *EBA REPORT on MANAGEMENT and SUPERVISION of ESG RISKS for CREDIT INSTITUTIONS and INVESTMENT FIRMS EBA/REP/2021/18 EBA REPORT on MANAGEMENT and SUPERVISION of ESG RISKS for CREDIT INSTITUTIONS and INVESTMENT FIRMS 2*. 2021,

https://www.eba.europa.eu/sites/default/files/document_library/Publications/Reports/2021/1015656/EBA%20Report%20on%20ESG%20risks%20management%20and%20supervision.pdf

-Spearman, C. “The Proof and Measurement of Association between Two Things.” *The American Journal of Psychology*, vol. 15, no. 1, Jan. 1904, pp. 72–101, https://doi.org/10.2307/1412159.

-Zar, Jerrold H. “Spearman Rank Correlation.” *Encyclopedia of Biostatistics*, 15 July 2005, https://doi.org/10.1002/0470011815.b2a15150.

-Kim, Soohun, and Aaron Yoon. “Analyzing Active Fund Managers’ Commitment to ESG: Evidence from the United Nations Principles for Responsible Investment.” *Management Science*, vol. 69, no. 2, 18 Apr. 2022, https://doi.org/10.1287/mnsc.2022.4394.

Bloomberg. *Bloomberg ESG Scores Overview & FAQ*. Apr. 2025.