Matti Viherkoski

Investigating Financial Drivers of ESG Scores

An Interpretable Machine Learning Approach

School of Finance and Accounting
Master's degree
Finance

Vaasa 2025

UNIVERSITY OF VAASA
School of Finance and Accounting
Author: Matti Viherkoski
Title of the thesis: Investigating Financial Drivers of ESG Scores: An Interpretable Machine Learning Approach
Degree: Master of Science in Economics and Business Administration
Degree Programme: Master's Programme in Finance
Supervisor: Timo Rothovius
Year: 2025 **Pages:** 115

ABSTRACT:
Growing interest in sustainable finance has increased the demand for transparent and reproducible assessments of corporate sustainability. The study has been grounded in the idea that ESG ratings reflect both sustainability-related practices and potentially the economic capacity to disclose and implement them. This thesis examined the extent to which financial information explains variation in ESG ratings. The objective has been to assess how far ESG outcomes are predictable from firm-level characteristics, to identify the financial factors most consistently associated with them and to analyse the underlying structure of these relationships.

The empirical analysis was based on data from firms included in the STOXX Europe 600 index for the period 2014–2023, obtained from the London Stock Exchange Group database. The dataset contained firm-level annual observations, and the dependent variables consisted of the overall ESG score and its environmental, social, and governance pillars. The independent variables comprised profitability, leverage, liquidity, efficiency, and valuation ratios, together with firm size, industry, and year identifiers.

Methods combined supervised machine learning with model-agnostic interpretability. Model evaluation relied on standard regression metrics, and explainable artificial intelligence methods.

The results indicated that firm-level financial characteristics explain a substantial portion of cross-sectional variation in ESG assessments. Nonlinear models outperformed linear alternatives, demonstrating that relationships between financial and sustainability indicators are complex and potentially interactive. The analysis highlighted firm size, operational efficiency, and capital structure as key predictors of higher ESG scores, whereas high profitability margins and liquidity were not systematically associated with higher assessed sustainability.

The findings suggested that financial structure had influence over the measured sustainability, implying that ESG scores can partly reflect underlying economic fundamentals in addition to non-financial performance. The study showed that interpretable machine learning offered a practical framework for understanding these linkages, but also that financial data alone could not fully account for the multidimensional nature of sustainability. Future research was encouraged to integrate non-financial and textual data, apply longitudinal designs, and examine regulatory developments to better capture the dynamic relationship between corporate finance and sustainability outcomes.

KEYWORDS: machine learning, modelling, responsible investing, sustainability reporting, corporate responsibility

VAASAN YLIOPISTO
Laskentatoimen ja rahoituksen yksikkö
Tekijä: Matti Viherkoski
Tutkielman nimi: Investigating Financial Drivers of ESG Scores: An Interpretable Machine Learning Approach
Tutkinto: Kauppatieteiden maisteri
Oppiaine: Rahoitus
Työn ohjaaja: Timo Rothovius
Vuosi: 2025 **Pages:** 115

TIIVISTELMÄ:
Kestävän rahoituksen kasvava merkitys on lisännyt tarvetta läpinäkyville ja toistettaville yritysvastuuta kuvaaville arviointitavoille. Tutkimus on perustunut oletukseen, että ESG-arvosanat heijastavat sekä vastuullisuuskäytäntöjä, sekä mahdollisesti niiden raportointiin ja toimeenpanoon tarvittavaa taloudellista kapasiteettia. Tutkimuksen tavoitteena on ollut arvioida, kuinka pitkälle ESG-tuloksia voidaan ennustaa yritystason taloudellisista tunnusluvuista, tunnistaa mitkä taloudelliset mittarit linkittyvät näihin vahvimmin, ja analysoida näiden suhteiden rakenteellista luonnetta.

Empiirinen analyysi koostui STOXX Europe 600 -indeksin yrityksistä vuosilta 2014–2023 koottuun aineistoon, joka on peräisin London Stock Exchange Groupin tietokannasta. Aineisto koostui yritystason havainnoista vuosittain; riippuvina muuttujina olivat yritysten vastuullisuuspisteytykset, ja selittävinä muuttujina olivat kannattavuutta, velkaantuneisuutta, maksuvalmiutta, tehokkuutta ja arvostustasoa kuvaavat suhdeluvut, sekä yrityksen koko, toimialaluokitus ja vuositunniste.

Menetelmissä yhdistettiin ohjattua koneoppimista malliriippumattomaan tulkittavuuteen. Mallien arviointi perustui standardoituihin regressiometriikoihin ja tuloksia selitettäviin tekoälymenetelmiin.

Tulokset osoittivat, että yritystason taloudelliset ominaisuudet selittävät merkittävän osan ESG-arviointien poikkileikkausvaihtelusta. Epälineaaristen mallien todettiin suoriutuvan lineaarisia vaihtoehtoja paremmin, osoittaen, että taloudellisten ja kestävyysindikaattorien väliset suhteet olivat monimutkaisia ja mahdollisesti vuorovaikutteisia. Analyysi korosti yrityksen kokoa, toiminnan tehokkuutta ja pääomarakennetta ESG-pisteiden keskeisinä ennustajina, kun taas korkeat kannattavuusmarginaalit ja likviditeetti eivät olleet systemaattisesti yhteydessä korkeampaan arvioituun vastuullisuustasoon.

Tulokset osoittivat, että rahoitusrakenne vaikutti mitattuun vastuullisuuteen, mikä viittasi siihen, että ESG-pisteet heijastivat osittain taustalla olevia talouden perustekijöitä ei-taloudellisen informaation lisäksi. Tutkimus osoitti, että tulkittava koneoppiminen tarjosi käytännöllisen viitekehyksen näiden yhteyksien ymmärtämiseen, mutta pelkkä taloudellinen informaatio ei riitä selittämään kestävyyden luonnetta. Jatkotutkimuksiksi suositeltiin ei-taloudellisen ja tekstimuotoisen aineiston integrointia, pitkittäisasetelmien hyödyntämistä sekä sääntelykehitysten tarkastelua, jotta yritysrahoituksen ja kestävyystulosten dynaamista suhdetta voidaan kuvata täsmällisemmin.

AVAINSANAT: koneoppiminen, mallintaminen, vastuullinen sijoittaminen, kestävyysraportointi, yritysvastuu

Contents

Figures

Tables

1 Introduction

The current investment landscape has been undergoing changes propelled by the growing demand and representation for Socially Responsible Investing (SRI) (D'Amato et al., 2022). Alongside financial characteristics, SRI considers companies ethical, social, and environmental values, aiming to generate financial returns while considering positive outcomes for stakeholders from the point of view of sustainability. Arguably the most important characteristics and metrics for socially responsible investors are Environmental, Social and Corporate Governance (ESG) characteristics. Large financial data and rating institutions play pivotal roles in serving market participants benchmarks for guidance in their investment decision making processes (D'Amato et al., 2022). These institutions provide ESG ratings for companies, providing socially responsible investors with a quantifiable tool to compare and analyze sustainability of companies. However, the accuracy and reliability of these ratings remain subjects of scrutiny. While ESG ratings are gaining traction, the accuracy of existing scores continues to be widely questioned, providing a need for further research and refinement in methodologies (Chowdhury et al., 2023).

An obstacle for investors and policymakers is the inability to accurately evaluate the reliability of the aggregation process used to determine ESG scores. This challenge stems from the lack of transparency in the rating system. Rating agencies generate ESG scores using proprietary models, and the information available to the public is often limited to what the agency chooses to disclose. In many cases, this disclosure is restricted to the fundamental principles of the methodology, which varies between agencies. Consequently, from the perspective of outside stakeholders, algorithms used by rating agencies can be considered as black-box models, where the inner workings are obscure (Del Vitto et al., 2023).

Furthermore, although different papers have found nonlinear relationships between sustainability metrics and their constituting indicators, Berg et al. (2022) discovered that six major ESG ratings are developed using linear models. These ratings rely on ad hoc weighted averages, meaning that the model weights assigned by the rater are assumed

to accurately reflect the relative importance of different ESG aspects. However, this approach overlooks the nuanced nature of ESG factors and may not fully capture their actual significance in sustainability assessment. Referring to Berg et al. (2022) findings, Svanberg et al. (2022) argue that because complex concepts such as ESG are not likely to have solely linear relationships with the features constituting ESG indicators, ESG ratings are unlikely to represent the degree of actual sustainability of companies.

Further relating to the issue of ESG ratings accurately representing the degree of sustainability, Billio et al. (2021) find that due to raters' disagreements on characteristics and their significances defining components of ESG leads to varying sustainability assessments among rating agencies and thus disseminates sustainable investors preferences regarding asset prices.

1.1 Purpose and motivation

Against this backdrop, the purpose of this thesis is to contribute to the evolving literature of sustainable finance and ESG investing by applying machine learning techniques to predict ESG scores using financial statement items, and analyzing the information learned by the model, to investigate the potential relationships between them. The results from investigating these relationships can improve the understanding of how and what financial characteristics affect the ESG scores of companies, and of the inner workings of the "black box models or methods" used at rating companies sustainability scores; the paper investigates whether, and to which extent a company's ESG performance can be predicted using traditional financial statement items, and to examine which financial features are most important in explaining ESG scores using Explainable AI tools, which are methods for interpreting machine learning models.

Considering data and methodology, within previous literature this thesis has most similarities with the work of Chowdhury et al. (2023), and D'Amato et al. (2021 and 2022), but uses combination of different variables that their studies previously recommended

as significant and uses larger dataset with more recent observations than, for example, D'Amato et al. (2021), covering years of Covid-19 pandemic. During recent years the importance of sustainability has also kept rising, and LSEG's ESG score methodology might have quietly changed, as the more specific methodology is not disclosed.

1.2 Research hypotheses

This thesis hypotheses that RandomForest and XGBoost machine learning methods can achieve notable reduction in variability of prediction than standard mean prediction would from simply using financial data on firm-year level observations, and that the information learned by the model can provide further insights on how ESG scores are affected by different financial characteristics. If this hypothesis is true, it brings up questions about the nature of sustainability scores: To what degree do these scores accurately capture sustainability of companies, considering that ESG ratings struggle with the problem of transparency of the models, and real sustainability effects of companies should at least in theory be rather independent from solely financial metrics? To what extent is it appropriate to compare ESG scores of different companies as a proxy for real sustainability on a continuous 0 – 100 scale, rather than on a form of weighted or adjusted scale, if companies with some financially homogenous metrics are consistently placed on different ESG score quantiles than others?

1.3 Structure of the study

In conclusion, in this thesis I aim to use supervised machine learning algorithms to predict ESG scores based on financial statement ratios and analyse the patterns learned by the models. By doing so, the study seeks to examine whether publicly available financial data can approximate ESG ratings and to what extent ESG assessments may be driven by quantifiable financial indicators.

The thesis is organized as follows: Chapter 2 provides a review of the existing literature regarding to sustainability, its impact on companies and how it can be measured, ESG ratings and predictive modelling approaches.

Chapter 3 outlines the data used in the research, data processing and pre-processing steps and imputation methods.

Chapter 4 outlines the methodology, covers the general information concerning the machine learning techniques used, and explains further model specific information considering the specific model data and calibration.

Chapter 5 presents the results of the models, including their performance, and further explainable AI analysis of the best performing models learned patterns from the training dataset.

Chapter 6 discusses the findings considering previous studies and evaluates the implications of model results and interpretability, discusses the limitations of the study, and suggests directions for future research.

Finally, the last chapter concludes the study.

2 Literature review

This chapter introduces key concepts and findings considering the topic from the literature. In this chapter, corporate sustainability is first defined and its importance shortly introduced with common metrics to measure sustainability. The chapter continues by reviewing previous literature considering predictive modelling of ESG scores and concludes with summary table consisting of the key studies referred to in this chapter.

2.1 Corporate Sustainability and ESG Scores

The next subchapters define corporate sustainability, its importance for firms, and outlines ways companies present it.

2.1.1 Defining Corporate Sustainability

Corporate sustainability refers to a company's ability to conduct business in a way that is environmentally sound, socially responsible, maintains transparent governance principles, and sustains long term economic viability (Ahmad et al., 2024). In practice, this means integrating ecological integrity, social welfare, and good governance into corporate strategies while continuing creating value for shareholders. This concept aligns with the "triple bottom line" of people, planet, and profit, emphasizing that sustainable firms balance financial performance with social and environmental stewardship. According to the OECD, corporate sustainability entails embedding environmental and social considerations into core business operations and strategy.

2.1.2 Importance of Sustainability for Companies

Companies are increasingly recognizing that strong sustainability practices can confer significant benefits. One key driver is investor demand: a growing share of global investments now incorporates Environmental, Social, and Governance (ESG) factors. As of the late 2010s, roughly $30 trillion in assets were managed using ESG criteria (D'Amato et

al., 2021). Investors increasingly view sustainability as linked to long-term financial performance and effective risk management, prompting firms to improve ESG performance to attract capital.

Research further suggests that companies with strong sustainability profiles may be more resilient during periods of crisis. For example, firms with high ESG ratings experienced better stock return performance during the 2008 financial crisis (Lins et al. 2017), (D'Amato et al., 2021).

Beyond investor considerations, Pilz (2024) suggests that sustainability also offers reputational and operational advantages. Embracing ESG can enhance a company's brand and consumer trust, while also helping identify risks and opportunities within operations (Pilz, 2024).

Pilz (2024) further suggests that sustainability initiatives can lead to cost savings like improved energy efficiency and can spur innovation. Integrating ESG considerations also enables more informed decision-making and strengthens relationships with stakeholders. Surveys of executives consistently show that sustainability is no longer seen as a niche issue, but as essential for long-term success and risk mitigation (Pilz, 2024).

In summary, companies should care about sustainability not only for ethical reasons but also because it aligns with financial prudence, stakeholder expectations, and evolving regulatory trends in today’s business environment.

2.1.3 Measuring Sustainability: ESG Scores and Other Metrics

Corporate sustainability is commonly assessed using standardized measurement frameworks, with ESG scores standing out as among the most widely adopted. ESG—an acronym for Environmental, Social, and Governance—represents three dimensions used to evaluate a firm’s sustainability-related performance (Del Vitto et al., 2023). An ESG score serves as a summary indicator that reflects how effectively a company manages its risks

and externalities across these three domains. The environmental dimension typically includes metrics such as greenhouse gas emissions, resource consumption, and waste management; the social dimension covers issues such as labour practices, community relations, and product safety; and the governance dimension focuses on board structure, ethical conduct, and transparency (Del Vitto et al., 2023).

These scores are generally produced by independent ESG rating agencies or data providers, which assess company disclosures, news sources, and other information to benchmark sustainability performance relative to industry peers. Leading providers include MSCI ESG Ratings, Sustainalytics, S&P Global (CSA/DJSI), and Refinitiv (formerly Thomson Reuters/Asset4), each applying distinct methodologies. For example, Refinitiv's ESG framework evaluates over 12,000 companies and assigns percentile-based scores ranging from 0 (lowest) to 100 (highest), based on industry-relative performance (Del Vitto et al., 2023). The expansion of ESG scoring systems reflects growing demand for quantifiable sustainability metrics, and they have become a key tool for investors to quickly evaluate a company's sustainable profiles (Del Vitto et al., 2023).

It is important to recognize that ESG scores can vary substantially across rating providers due to differences in data sources, weighting schemes, and evaluation methodologies. Berg et al. (2022) documented significant divergence among ESG scores assigned by six leading rating agencies, underscoring the lack of standardization in sustainability assessment practices*. Nevertheless, ESG ratings remain widely used as a proxy for corporate sustainability performance in academic research and investment practice (D'Amato et al. 2021).* While these composite scores offer a convenient summary measure, firms often supplement them with more granular sustainability metrics for internal tracking and disclosure purposes.

For further information considering largest ESG report providers and their rating methods, Huber et al. (2017) have constructed comprehensive summary table of the topic in

their paper *"ESG Reports and Ratings: What They Are, Why They Matter",* that can also be found in this papers Appendix 1.

In addition to ESG scores, several other frameworks and metrics are used to assess corporate sustainability. Many companies publish sustainability reports in accordance with sustainability directives such as Corporate Sustainability Reporting Directive (CSRD) or European Sustainability Reporting Standards (ESRS), that follow established sustainability standards such as the Global Reporting Initiative (GRI), which mandate detailed qualitative and quantitative disclosures. Other benchmarks include sustainability indices, such as the Dow Jones Sustainability Index (DJSI) and FTSE4Good, which rank companies based on structured questionnaires and performance criteria. Organizations may also pursue third-party certifications or ratings, such as B Corp certification or Carbon Disclosure Project (CDP) scores, particularly for environmental performance.

Furthermore, concepts like Corporate Social Responsibility (CSR) and alignment with the UN Sustainable Development Goals (SDGs) are used to qualitatively gauge a company's contributions to sustainable development. are often used to qualitatively evaluate a firm's contributions to sustainable development. These diverse measurement approaches complement ESG ratings. For example, a company may receive a high ESG score from MSCI, be included in the DJSI, and disclose its sustainability efforts in line with GRI standards—together offering a more holistic view of corporate sustainability. In this thesis, however, the primary focus is on ESG scores as a quantifiable measure of sustainability performance, given their widespread use in financial markets and research.

2.2 Predictive Modelling of ESG Scores

In recent years, a growing body of research at the intersection of sustainable finance and machine learning has focused on predicting ESG scores using a variety of data sources. The motivation for this work is twofold: first, to identify the factors that influence ESG ratings, thereby offering insight into the rating process and the relationship between

financial and sustainability performance; and second, to develop predictive models capable of estimating ESG scores where data are missing or of forecasting future ESG outcomes, with potential applications for investors and corporate decision-makers. Leveraging the increasing availability of ESG ratings and firm-level financial data, researchers have employed a wide range of machine learning (ML) methods, ranging from linear regressions to advanced deep learning architectures, to model ESG scores. The following review surveys recent literature on ESG score prediction, organizing studies by model type, from linear and regularized models to ensemble and deep learning approaches.

2.2.1 Linear and Regularized Regression Models

Since sustainability has become more topical in recent years, many studies have tried to make ESG scoring more transparent by building predictive models. A study by Licari, J. et al. (2021) used traditional linear regression to predict ESG scores across a large and global dataset consisting of 19,000+ companies in 96 countries between years 2004–2020. The paper found that traditional models struggle to handle the complexity and inconsistency of ESG score construction, highlighted the limitations of traditional statistical methods in modelling ESG ratings. In the paper, predicting ESG scores using liner regression achieved weak prediction performance – an R^2 of 31.13%, suggesting it captured only a small portion of the variation in ESG scores (31.13% of ESG score variation was attributable to the independent variables in the model). The paper presents multiple potential reasons for poor performance of traditional models, including complex nature of ESG rating methodologies, varying data sources, subjective weighting of ESG attributes, direct company engagement, and coverage caps considering smaller firms, and varying regulation between sectors and emerging markets.

Del Vitto, Marazzina, and Stocco (2023) investigate the transparency of proprietary ESG ratings by attempting to replicate the ESG scoring methodology used by LSEG (formerly Refinitiv). Using a combination of machine learning methods—including regularized linear models (Ridge and Lasso regressions), Random Forest, and Artificial Neural Networks—they model the Environmental, Social, and Governance (ESG) pillar scores based

on Refinitiv's full set of sustainability indicators and financial variables. A key contribution of their study is the demonstration that interpretable models such as Lasso and Ridge, often referred to as “white-box” methods, can achieve predictive performance comparable to more complex black-box models like neural networks. These linear models also offered the advantage of minimal overfitting and strong generalizability across sectors. The authors report high predictive accuracy for the Environmental pillar and moderate accuracy for the Social and Governance scores. The reduced accuracy for the social pillar is attributed to its broader and less quantifiable scope, while regional variation in Governance scores reflects differing institutional contexts and data availability—prompting caution when making cross-country comparisons (e.g., between the U.S. and China). Their analysis also reveals that feature importance varies across industries and geographies, underscoring the contextual nature of ESG rating mechanisms. Overall, the findings suggest that a well-specified linear model using relevant financial and ESG indicators can approximate Refinitiv’s ESG ratings with surprising accuracy,

In a study of Taiwanese companies, *Lin and Hsu (2023)* included a multiple linear regression as a benchmark for ESG score prediction. The authors emphasized the importance of establishing interpretable baseline models, particularly in the context of Taiwan’s unique market characteristics, including technology-driven economy, limited stock circulation, and heightened information asymmetry. Although they found that the linear models were consistently outperformed by more advanced machine learning techniques, they still demonstrated moderate predictive accuracy and served as a transparent reference point for evaluating more complex approaches. The authors noted that linear models struggled to capture the nonlinear relationships inherent in ESG ratings, especially in the presence of multicollinearity among financial and governance-related variables. Nonetheless, the inclusion of linear regression highlighted the trade-off between model simplicity and predictive power, underscoring its value in contexts where interpretability and transparency are prioritized.

Notably, linear models allow researchers to identify which financial ratios and indicators have the most explanatory power for ESG scores, albeit assuming a linear relationship. Commonly influential variables include profitability metrics, leverage, firm size, and industry-specific factors, which are consistent with broader empirical findings on the determinants of ESG performance. Regularization techniques such as Lasso regression further enhance model parsimony by shrinking the coefficients of less relevant predictors toward zero, thereby highlighting a core subset of explanatory features (Del Vitto et al., 2023). While linear models generally exhibit lower predictive accuracy than nonlinear approaches in more complex environments, they nevertheless provide a transparent and reasonably effective baseline for ESG score modelling, particularly when interpretability and variable selection are of primary importance.

2.2.2 Tree-Based Ensemble Models (Random Forests and Boosting)

A significant portion of recent ESG prediction research employs tree-based ensemble models, including Random Forests (RF) and gradient boosting frameworks such as XGBoost and LightGBM. These models are suited to capture nonlinear relationships and complex interactions among predictors, making them effective in financial modelling contexts. They have likewise shown strong performance in predicting ESG scores across various studies. As they can handle high-dimensional data and model heterogeneity, they have become popular choice in studies aiming to replicate ESG scores or forecast sustainability performance. Moreover, ensemble models such as RF offer built-in mechanisms for estimating feature importance, which can provide insights into the relative contribution of predictors to ESG outcomes—albeit still with less transparency than linear models.

Random Forests model was used by D'Amato, D'Ecclesia, & Levantesi (2021) in one of the pioneering works to link financial fundamentals with ESG ratings. Using data from 109 STOXX Europe 600 index companies during the 2010s, the authors trained the model on balance sheet and income statement ratios to predict Bloomberg's ESG disclosure scores. The study aimed to assess the predictive power of conventional financial

variables in explaining variation in sustainability ratings. Among the models tested, Random Forest delivered the highest predictive performance, achieving an R^2 of approximately 0.62, which outperformed linear regression and other baseline models. Key predictors identified included firm size, profitability, and leverage. The authors concluded that financial statement items constitute a robust explanatory basis for ESG scores, providing empirical support for the notion that sustainability assessments, although seemingly non-financial in nature, are linked to a firm's financial characteristics.

Complementing the findings of D'Amato et al. (2021), Lin et al. (2019) study had already found negative link between corporate social responsibility and corporate financial performance measured by ROE, ROA and ROI, which supports the theory that a trade-off exists between optimising financial performance metrics and carrying out sustainability objectives. However, a few other studies have later pointed out that the trade-off only negatively affect companies which financial performance is below optimal to begin with.

D'Amato et al. (2022) expand on their earlier study, aiming to assess structural data and balance sheet items effect on ESG scores of regularly traded stocks. In this study, they instead use Refinitiv (LSEG) ESG scores, with larger sample of companies across 2009 – 2019 and find that balance sheet items present a significant predictive power on ESG score. Based on their findings the Random Forest algorithm performs best at predicting ESG scores compared to classical regression approach, as it can capture nonlinear relationships between ESG scores and predictive variables, which their study shows to occur consistently.

Cini and Ferrari (2025) took this approach a step further by introducing a time dimension: they trained an RF classification model to predict a firm's next-year ESG rating class using current financial ratios and risk indicators. Using panel data from 2016 to 2021 for European companies, their model categorized firms into ESG performance tiers (e.g., high, medium, or low) with high out-of-sample accuracy. This is notable as it demonstrates forward-looking predictive power – essentially showing that there is informational

content in financial fundamentals that anticipates improvements or declines in ESG performance. The authors described their model's accuracy as "unprecedented," suggesting practical applications in estimating ESG ratings for firms that lack current evaluations, such as small-cap or privately held companies.

Beyond Random Forests, boosting algorithms have also gained traction in ESG score prediction due to their high predictive accuracy and ability to model complex nonlinear relationships. Gradient boosting machines such as XGBoost have been applied in ESG studies with promising results. A study by Choi, Chen, & Lee (2024), compared multiple ML models on a dataset of Korean companies' financial ratios over three years, aiming to predict the companies' ESG ratings. They evaluated linear models, tree ensembles, and neural networks, and while applying SHAP (Shapley Additive Explanations) to interpret the variable importance. In their results, XGBoost was found to be the most effective model, achieving an F1-score of 85.1% in classifying ESG ratings.

Similarly, Lin and Hsu (2023) included XGBoost in their evaluation of ESG prediction models for Taiwanese firms and found it to perform competitively, although an alternative model—Extreme Learning Machine (ELM)—slightly outperformed it in their dataset. However, the literature also cautions that particularly in ESG applications where datasets can consist of relatively small panels, boosting algorithms require careful hyperparameter tuning to prevent overfitting.

In summary, ensemble tree-based models have demonstrated strong predictive performance in ESG score modelling. By capturing nonlinear relationships and complex feature interactions, methods such as Random Forest and XGBoost often outperform linear regression models, which assume constant marginal effects. For instance, the impact of profitability on ESG scores may vary nonlinearly, strengthening or diminishing beyond certain thresholds. The collective evidence from recent studies suggests that these models can effectively learn the functional mapping between financial ratios and ESG ratings, with reported R^2 values and classification metrics substantially exceeding baseline

accuracy (e.g., Choi et al., 2024; D'Amato et al., 2021). For more recent example, Alsayyad and Fadel (2025) findings demonstrated high predictive performance in their comprehensive machine learning study on ESG scores and employed panel data with best R^2 scores reaching over 0.9.

The findings generally indicate that a considerable portion of the variance in ESG ratings can be explained by financial data. However, it should be noted that each study's results depend on the specific dataset and ESG rating agency used, as each have unique methodologies. Additionally, several studies point to diminishing returns: once a robust tree-based model is in place, even more complex approaches may not dramatically improve accuracy, as we discuss next.

2.2.3 Deep Learning Models (Neural Networks)

Given the success of machine learning in predicting ESG scores, researchers have investigated deep learning approaches, such as multilayer artificial neural networks (ANNs), to see if they can further improve prediction performance. Neural networks can, in theory, capture very complex nonlinear interactions in data. However, in the context of ESG score prediction, deep learning has been explored less in comparison to tree-based models, and the empirical results are mixed.

Del Vitto et al. (2023) evaluated multiple ANN architectures in their effort to replicate Refinitiv's ESG scoring methodology. The authors tested both shallow and deep networks with varying the number of layers and hidden units, and benchmarked their performance against simpler models, including Lasso regression and Random Forest. They found that increasing the depth and complexity of the neural networks did not consistently improve prediction accuracy. In some cases, a simpler ANN with fewer hidden layers performed comparably to, or better than, more complex architectures. Furthermore, the highest overall performance was achieved by the regularized linear models and the simpler ANN, rather than by the deeper ANNs or ensemble methods. These findings suggest that while ESG–financial relationships are nonlinear, they may not require deep

architectures to model effectively. This could be due to the moderate size of structured ESG datasets and the risk of overfitting when models include too many parameters relative to the data available (Del Vitto et al. (2023).

Other studies reinforce the view that deep learning should be applied with caution in the context of ESG score prediction. Choi et al. (2024) included a neural network in their model comparison when classifying ESG ratings for Korean firms but ultimately found the tree-based XGBoost model superior.

Lin and Hsu (2023) studied ESG score prediction of Taiwanese non-financial companies using 27 financial metrics with corporate governance indicators. The used an Extreme Learning Machine (ELM), a form of single-layer neural network with random weights and reported that ELM achieved excellent performance (R^2 of over 0.9 for multiple models), slightly outperforming Random Forest and XGBoost in predicting ESG scores for their dataset. While ELM is technically a neural approach, it is not a deep learning model. It rather offers an efficient architecture for capturing nonlinearities in relatively small datasets.

These findings suggest that neural models—especially shallow or lightweight variants like ELM—can perform competitively in ESG prediction. However, evidence from recent studies indicates that deep neural networks have not consistently outperformed boosting or ensemble tree methods when using structured financial data alone. Although hybrid deep learning approaches incorporating unstructured data such as ESG reports or news sentiment are gaining attention, they fall outside the scope of predictions based on structured data and are rather new topic in the research. As such, the incremental benefit of deep learning over more interpretable machine learning models remains limited in this domain, particularly given concerns around overfitting, data volume, and model transparency. An advantage of neural networks is their flexibility in integrating heterogeneous data sources, such as combining structured financial indicators with unstructured information like textual disclosures or ESG news. However, this flexibility

comes at the cost of reduced model interpretability, which presents a limitation for sustainability assessment. To address this concern, recent studies have increasingly employed explainable AI (XAI) techniques to interpret the internal logic of complex models. For example, both Del Vitto et al. (2023) and Choi et al. (2024) applied SHAP (Shapley Additive Explanations) to their ESG prediction models, enabling them to identify which input features, such as the debt-to-equity ratio, return on assets, or carbon emissions, had the greatest influence on predicting ESG scores.

A related line of research explores the integration of natural language processing (NLP) and model robustness techniques. For example, Lee et al. (2022) proposed an AI framework for predicting firm-specific ESG ratings by analysing governance and social-related datasets using a combination of machine learning and NLP algorithms. In addition to evaluating multiple models for prediction accuracy, their study addressed the vulnerability of ESG systems to adversarial attacks, which they describe as malicious manipulations of input data that can distort rating outcomes. They introduced a method for detecting such attacks, contributing to the growing emphasis on data reliability and security in ESG analytics. While such hybrid approaches extend beyond structured financial data and remain relatively novel, they showcase how new AI technology can be applied to expand the scope and resilience of ESG prediction models.

The use of XAI contributes to making obscure models more transparent, which is important in the ESG domain where stakeholders seek to understand the drivers behind sustainability ratings. Insights from these interpretability tools also reinforce broader findings in the literature across both complex and simpler models, a relatively consistent set of financial variables frequently emerges as key predictors of ESG performance; profitability, firm size, leverage, and industry-specific environmental or social factors are among the most cited drivers, hinting that certain financial fundamentals hold robust explanatory power across models and contexts.

In summary, while model performance varies, recent literature confirms that ESG scores can be predicted with reasonable accuracy using firm-level financial data and machine learning techniques. The table below summarizes key studies from this literature review, highlighting their methods, data sources, and key findings.

Table 1. Summary of key studies.

Study	Data	Methods	Key Findings	Add. Insights
D'Amato et al. (2021)	Euro Stoxx 600 firms, Bloomberg ESG scores	Random Forest vs. linear models	RF achieved R^2 ~0.62, indicating financial metrics explain ESG scores	Highlighted the importance of structural financial data in ESG ratings.
Del Vitto et al. (2023)	Refinitiv ESG scores, global firms by sector	Lasso, Ridge, Decision Tree, Random Forest, ANN	Lasso and a shallow ANN were best predictors of ESG; deeper ANNs didn't improve much	ESG ratings can be largely replicated with a selected feature set and ML models
Lin & Hsu (2023)	Taiwan companies, ESG index scores (2018–2021)	SVM, Random Forest, XGBoost, Extreme Learning Machine (ELM)	High accuracy (~0.9+) R^2 with different models	Integrating financial and governance indicators is effective
Chowdhury et al. (2023)	6171 firms from 2005 to 2019	Six machine learning classification models	The RFC model was superior with 78.50% accuracy	findings highlight the relationship between firm size, liquidity, and ESG investing.
Choi et al. (2024)	Korean firms, ESG ratings from a local agency	Multiple (linear, RF, XGBoost, deep NN) + XAI	XGBoost was best (F1 ~85%), beating deep neural nets	Financial factors (leverage, profitability) were significant
Cini & Ferrari (2025)	Euro Stoxx 600 (2016–2021), MSCI (or similar) ESG rating classes	Random Forest classification to predict next-year ESG rating class from current financial ratios + a systemic risk metric	High out-of-sample accuracy in classifying ESG ratings one year ahead	Investors can forecast sustainability improvements or deteriorations using financial data

3 Data and Data Processing

Data for this study is gathered from LSEG database. LSEG is one of the largest and most important financial data and ESG score providers, covering more than 80% of the global market capitalization (LSEG.com). LSEG ESG ratings are percentile rank scores, ranging from 0 (lowest) to 100 (highest). These ratings aim to objectively assess a company's relative ESG performance. LSEG states that their ESG ratings are data-driven and consider the most crucial industry metrics and are adjusted to consider biases related to transparency and market capitalization. However, their scores are not exempt from some of the main problems with ESG scores.

The original dataset consists of six hundred European companies from StoxxEurope600 index from years 2014 to 2023. Original variables are selected based on suggestions and previous findings within the literature, based on Chowdhury et al. (2023), D'Amato et al. (2021) and D'Amato et al. (2022). D'Amato et al. (2021) convincingly argue that using ratios that consider the overall financial statements of the companies representing profitability, liquidity and solvency ratios is more informative and improves the characterization of companies, in contrast to using absolute financial statement values when aiming to explain ESG scores, which is why in this paper, the step of testing model characteristics importances using solely financial statement values is skipped, and financial statement ratios are used instead.

After testing for variable correlations and initial model performances some potentially influential variables originally recommended in the literature, such as NI/NS (Net income / Net Sales) were redacted from the final dataset and model, leaving EBIT/NS as the main proxy for profitability due to having proportionally higher prediction power in the models whilst having over 96% correlation with NI/NS.

After omitting variables based on initial model performances and variable correlation, the following is a list of variables used in the models:

- *YEAR:* 2014-2023
- *INDUSTRY:* General industry sector classification variable (range: 1-6). They are transformed into dummy variables in the model.
- *ESGScore:* ESG score from LSEG database.
- *ESGE:* Environmental score from LSEG.
- *ESGS:* Social score from LSEG.
- *ESGG:* Governance score from LSEG.
- *ASSETS:* Total assets of the company.
- *SIZE:* Logarithm of assets.
- *NS/TA:* Net Sales divided by Total Assets.
    - Efficiency ratio (turnover).
    - Measures how efficiently a company uses its assets to generate sales.
    - Indicates operational efficiency and asset utilization; a higher ratio indicates more effective use of assets to drive revenue.
- *EBIT/NS:* The ratio of Earnings Before Interest and Taxes to Net Sales.
    - Profitability Ratio (Operating Margin)
    - The proportion of sales remaining as operating profit before accounting for interest and taxes.
    - Compares operational performance across companies regardless of their financing and tax structures.
- *DIV:* Dividend yield.
    - Income (yield) ratio.
    - The annual dividend per share relative to the stock price.
    - Indicates the cash return on investment and can reflect the company's commitment to returning profits to shareholders.
- *P/E:* Price to Earnings ratio.
    - Valuation Ratio.
    - The market value of a stock relative to its earnings per share.
    - Provides insights into market expectations and relative valuation.
- *CA/CL:* Current Assets to Current Liabilities.

    - Liquidity Ratio (working capital).
    - The ability of a company to cover its short-term liabilities with its short-term assets.
    - Serves as an indicator of short-term financial health and liquidity.
- *TD/TA:* Total Debt to Total Assets.
    - Solvency (Leverage) Ratio.
    - The proportion of a company's assets financed by debt.
    - Evaluates financial risk and solvency; a lower ratio typically implies a more conservative capital structure.
- *ROE:* Return on Equity.
    - Profitability Ratio.
    - The profitability relative to shareholders' equity, reflecting how effectively a company uses equity capital to generate profits.
    - Reflects management effectiveness and overall profitability relative to equity; critical for comparing performance among companies in the same industry.
- *ROA:* Return on Assets.
    - Profitability Ratio (with an efficiency component).
    - How effectively a company generates profit from its total asset base.
    - View of operational efficiency and profitability, valuable for comparing companies irrespective of their financing structures.
- P/E missing flag
    - Categorical variable included in the model where P/E was missing.
- CA/CL missing flag
    - Categorical variable included in the model where CA/CL was missing.

Processed data used in the models consists of ESG score as the main outcome variable, and Environmental, Social and Governmental scores separately, as other dependent variables. Independent variables consist of general industry classification from 1 to 6, Year, Total Assets, Dividend Yield, ROE, ROA and categorical flags for missing P/E and CA/CL

ratios, and financial ratios of Net Sales / Total Assets, EBIT / Total Assets, EBIT / Net Sales, Price / Earnings, Current Assets / Current Liabilities and Total Debt / Total Assets.

Similarly to D'Amato et al. (2021) in this paper "Year" is included as a separate static variable and disregards the year-on-year changes. In the model years 2014 to 2023 are considered, although for year 2023, most of the ESG scores were missing at the time of obtaining the dataset.

Chowdhury et al. (2023) argues that based on variable importance factor, the Lagged ESG score is the most important predictor of ESG, followed by firm size and debt to equity ratio, indicating that previous investments into ESG, firms' total assets and financial leverage are the best predictors of ESG score. In testing and model optimisation process of this paper, Lagged ESG score was also found to be the most important predictor of ESG scores based on variable importances. However, it can be reasoned that this finding is rather obvious, and the variable alone could explain most of the ESG score in the model testing phase, dominating the model and results. As the objective of this paper is to aim to predict ESG scores from financial statement items and reveal information about the underlying influence of financial statement items on ESG scores, including a lagged predictor of the dependant variable itself as independent variable rather works to defeats this purpose, and thus in this paper it is omitted from the models.

3.1 Data Integrity and Initial Screening

The dataset has a panel-like structure, containing annual observations for STOXX Europe 600 companies from 2014 to 2023, but is analysed cross-sectionally. The dataset contains no duplicate values. As first step of the data processing, observations with missing values in ESG Scores were omitted to ensure consistency of the dependent variables. The remaining dataset contains 4937 valid firm-year observations and is described in the following tables in this chapter.

3.2 Currency Standardization

The financial data in the original dataset varied in currency, showcasing different firms' financial information in their local currencies, containing observations with eight different currencies: United Kingdom Pound (GBP), Euro (EUR), Danish Krone (DKK), Swiss Franc (CHF), Hong Kong Dollar (HKD), Norwegian Krone (NOK), Swedish Krona (SEK) and Polish Zloty (PLN). Since most of the variables in the model are ratios, considering compatible values between firms with different currencies only affected total assets. Assets were converted to EUR using the average annual EUR exchange rate corresponding to each observation's year to ensure comparability across firms and time.

3.3 Descriptive Statistics of the Processed Dataset

Table 2. Summary Statistics table**.**

	ESG Score	ESG E	ESG S	ESG G	AS-SETS	EBIT/ NS	TD/ TA	NS/ TA	CA/ CL	DIV Y	P/E	ROE	ROA
count	4937	4937	4937	4937	4935	4901	4935	4934	3807	4916	4562	4887	4870
mean	65.37	64.43	68.84	61.52	85375584	0.23	0.25	0.67	1.60	2.72	38.76	17.53	6.89
std	17.28	23.44	19.87	20.92	342680815	1.02	0.16	0.58	1.32	2.67	206.03	65.76	12.35
min	2.60	0.00	0.25	1.45	36571	-28.28	0.00	-0.05	0.21	0.00	0.20	-262.32	-63.72
25%	54.96	49.44	56.83	46.91	3945650	0.07	0.13	0.23	0.99	1.18	12.50	7.46	2.05
50%	68.55	69.69	73.43	64.95	10395392	0.13	0.24	0.58	1.31	2.35	19.00	13.24	5.38
75%	78.45	83.20	84.41	78.42	40324822	0.23	0.35	0.90	1.80	3.92	28.20	20.91	9.18
max	95.72	99.14	98.20	98.56	6639198547	29.13	1.32	4.41	29.27	111.23	8105.00	2409.86	269.11

The descriptive statistics of all variables are presented in Table 2, summarizing the central tendency and dispersion for ESG scores and the associated financial statement items.

3.3.1 Overview of ESG Scores

The mean ESG Score of the sample is approximately 65.4 with a standard deviation of 17.3, indicating that most companies cluster around mid-to-high ESG performance levels. The environmental (E), social (S), and governance (G) pillars show comparable patterns: the S-score has the highest mean (68.8) and slightly lower dispersion, suggesting more consistent social-performance evaluations across firms, while the G-score shows the greatest variability (SD ≈ 20.9), potentially indicating broader differences in corporate governance practices across Europe. The overall range of scores (minimum ≈ 2.6, maximum ≈ 95.7) shows that the dataset contains both low- and high-performing firms, offering variation for predictive modelling.

The time-series analysis presented later in Table 3 (Section 3.4) reveals an upward trend in ESG means and a gradual reduction in standard deviations over the period 2014–2023. This pattern is consistent with the increasing institutional emphasis on sustainability reporting and improved data coverage in Europe, which have led to more homogeneous ESG assessments in recent years.

3.3.2 Overview of Financial Variables

The financial ratios display considerable heterogeneity, which is expected given the cross-industry composition of the sample. Total assets (ASSETS) vary widely, from roughly hundreds of thousands to over €6 billion, reflecting the coexistence of smaller firms and multinational corporations. Ratios such as EBIT/NS (mean = 0.23, SD = 1.02) and ROE (mean = 17.5, SD = 65.8) show substantial dispersion, partly driven by outliers in profitability and capital structure. This variation underscores the need for robust algorithms and outlier treatment during modelling.

Leverage-related ratios, such as Total Debt to Total Assets (TD/TA), display relatively moderate variation (mean = 0.25, SD = 0.16), suggesting that debt levels among listed European firms are somewhat stable across industries. In contrast, CA/CL exhibits wide variation (mean = 1.6, SD = 1.3) where observed, reflecting differences in liquidity structures, especially between manufacturing firms and financial institutions.

The P/E ratio demonstrates extreme spread (mean ≈ 38.8, SD ≈ 206.0, max > 8,000), indicating the presence of a few extraordinarily high values. Such dispersion arises from low or near-zero earnings denominators, highlighting an example reason for winsorising. Missing or undefined P/E values need to be handled carefully, which will be further discussed in missing-data assessment and the subsequent Imputation chapter.

3.3.3 Interpretation and Relevance for Modelling

The descriptive analysis highlights some considerations for modelling. The dataset is sufficiently diverse with ranging firm sizes, profitability levels and sustainability outcomes for aiming to find relationships between financial statement items and ESG scores.

The magnitudes and dispersion of variables indicate a need for data processing choices, such as winsorisation of extreme values, standardization for linear algorithms, and context-specific handling of missing data.

3.4 Distribution Across Time

Table 3 presents the mean and standard deviation of the ESG score and its subcomponents for years 2014 to 2023. The results show an upward trend in average scores over the period, accompanied by a gradual reduction in dispersion. The mean overall ESG score rises from approximately 58 in 2014 to about 70 in 2021–2022, while standard deviations decline from around 19 to 14. Similar dynamics are observed across the three pillars, although the magnitude and pace of change vary slightly: the Environmental (E)

and Social (S) dimensions increase more steadily than Governance (G), which remains comparatively volatile.

Table 3. Main statistics of the ESG, E, S and G scores distribution by year of the sample of 600 companies listed in the STOXX Europe 600 Index.

	ESG Score		**E Score**		**S score**		**G score**	
Year	Mean	SD	Mean	SD	Mean	SD	Mean	SD
2014	58.52	19.46	61.24	25.34	60.34	23.11	54.76	22.09
2015	60.28	19.57	62.19	25.19	63.52	22.66	55.48	22.32
2016	61.90	18.23	63.74	23.92	65.85	21.57	56.29	21.80
2017	63.37	17.55	63.48	24.28	68.69	19.90	57.08	21.62
2018	64.79	17.53	60.82	25.47	69.26	19.49	61.08	21.05
2019	66.86	16.23	64.25	23.52	70.58	18.73	63.42	19.71
2020	69.30	15.33	66.00	22.28	72.22	17.46	67.43	18.66
2021	70.03	14.44	67.74	20.92	73.00	16.64	67.42	18.26
2022	70.02	13.98	68.86	19.67	73.00	16.67	66.39	18.38
2023	67.17	13.39	65.46	20.27	68.51	16.50	66.21	18.48

The pattern indicates that firms in the STOXX Europe 600 have generally improved their reported sustainability performance during the past decade. The concurrent decline in standard deviations suggests convergence among firms, meaning that extreme low performers have become less frequent while mid-range and high-range scores have become more typical. This aligns with the general observations within sustainability literature, relating to the increasing importance of sustainability, especially in Europe. Whilst, according to LSEG, the ESG scores by LSEG do include country specific characteristics, in general they are still comparable to ESG scores within other continents, and Europe is pioneering as a market in sustainability, in relation to other markets.

Over the years, ESG reporting has benefited from enhanced disclosure requirements and more consistent evaluation frameworks (LSEG), which may also reflect the gradual improvement in the underlying data infrastructure. The increase in mean scores and the narrowing of their spread can both result from a combination of progress in corporate

sustainability and a methodological maturation of ESG measurement. The pattern suggests that the dataset captures broader systemic changes in European sustainability reporting and corporate sustainability in addition to firm-level differences.

3.5 Industry Classification

To capture cross-sectoral patterns in sustainability performance, the dataset is divided into six broad industry groups based on the general business classification used throughout this thesis. The classifications and their definitions are presented in Table 4.

Table 4. General Industry Classification explanation.

1.	Various Industries	This category is a "catch-all" category for companies operating across different sectors, contains a diverging range of industries, including personal goods, pharmaceuticals and biotechnology, technology hardware and equipment, food producers, retail, oil, gas, coal, aerospace, etc.
2.	Electricity and Telecommunications	Companies involved in providing electricity, telecommunications, etc. utilities
3.	Transportation, Travel, Leisure	Companies involved in transportation services, travel, leisure, and related industries such as tourism.
4.	Banks	Banks and financial institutions engaged in commercial banking activities.
5.	Insurance	Companies operating under insurance industry.
6.	Real Estate and Investments	Companies involved in real estate, investment management, investment banking, etc.

The mean and standard deviation of ESG, E, S, and G scores by industry are summarized in Table 5 below.

Table 5. Main statistics of the ESG, E, S, and G score distribution by industry sector of the sample of 600 companies listed in the STOXX Europe 600 Index.

	ESG score		**E score**		**S score**		**G score**	
Sector	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	65.85	17.02	63.25	23.06	70.37	19.69	61.16	21.04
2	69.28	14.72	69.64	19.65	70.98	19.02	64.63	17.14
3	63.53	15.17	64.44	19.49	67.49	16.88	58.58	22.10
4	66.54	17.26	73.92	22.02	69.41	18.12	62.02	21.16
5	64.44	16.05	65.18	23.51	62.08	18.42	69.32	18.50
6	58.04	20.44	58.80	28.06	59.44	21.68	56.30	22.34

Across industries, ESG performance varies notably both in mean values and dispersion. Firms in Electricity and Telecommunications (Industry 2) show the highest average ESG and E-scores, consistent with the sector's strong exposure to environmental regulation and renewable energy transition policies. Conversely, Real Estate and Investment firms (Industry 6) record the lowest overall ESG averages and the widest dispersion, reflecting structural heterogeneity and differing reporting standards within that category.

Banks (Industry 4) exhibit relatively high Environmental (E) scores compared to other sectors, which may stem from their lower direct emissions and increasing engagement in sustainable finance. In contrast, Insurance (Industry 5) firms tend to perform better in Governance (G), likely due to regulatory oversight and mature compliance systems. These cross-industry contrasts confirm that sustainability outcomes are influenced not only by firm-level financial characteristics but also by sector-specific operational and regulatory contexts.

From a modelling perspective, such heterogeneity underscores the importance of including industry identifiers or fixed effects when predicting ESG outcomes. Industry-level variation can capture systematic differences in disclosure norms, business models, and risk exposures. The implications of industry structure for feature importance and model behaviour will be revisited in later chapters, where variable importance and interpretability methods (e.g., SHAP values and partial dependence plots) are discussed.

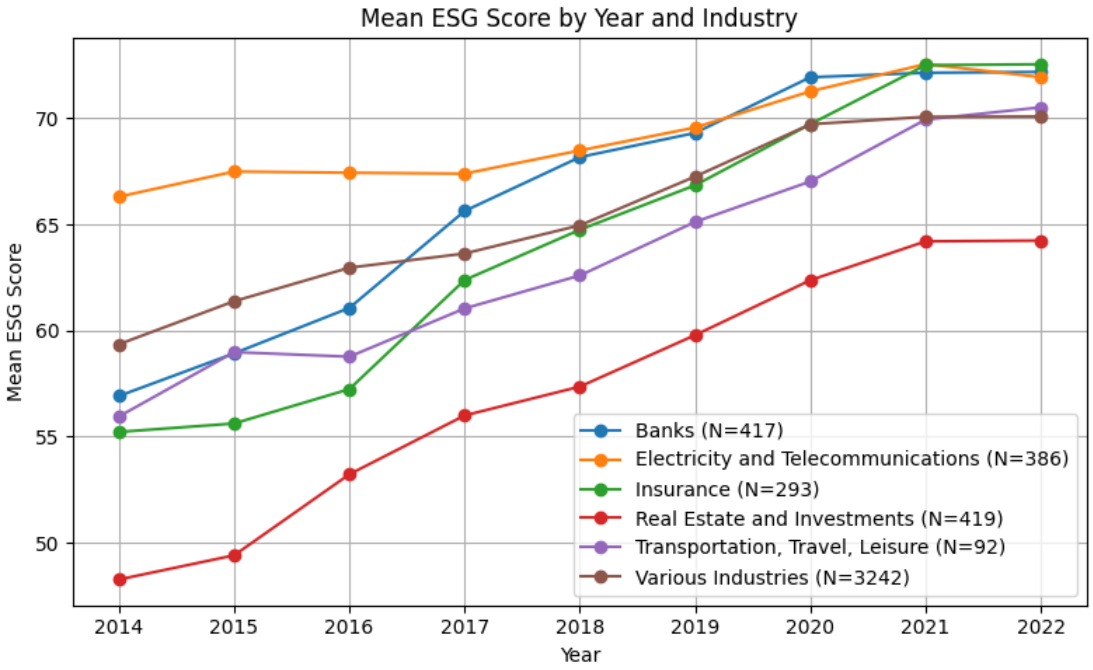

Figure 1. Mean ESG Score by Year and Industry.

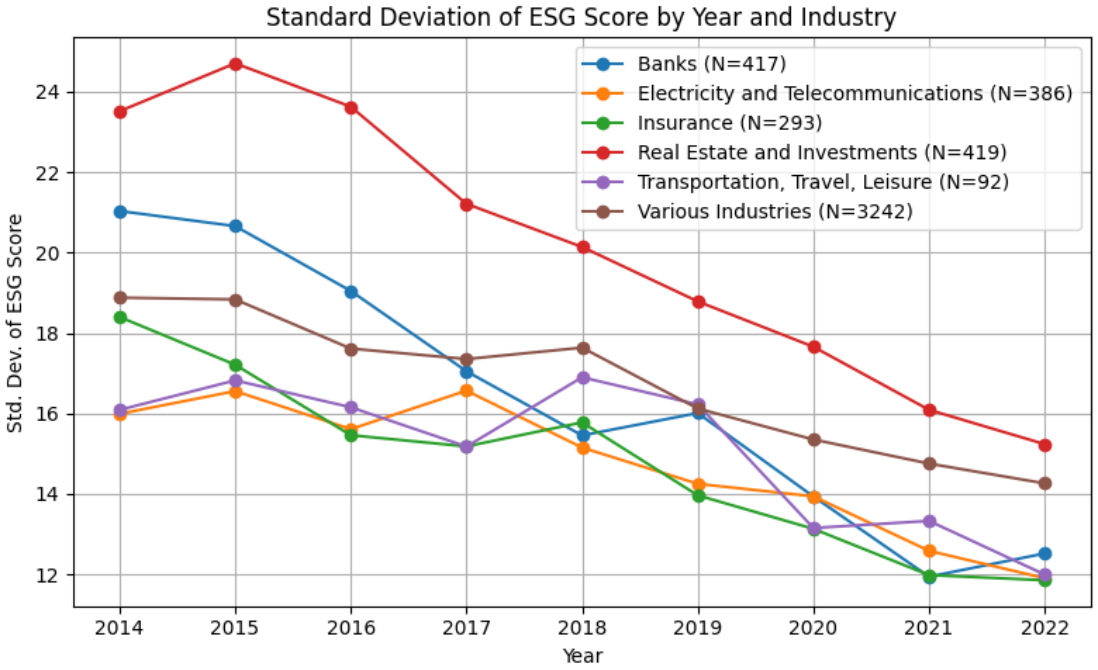

Figure 2. Standard deviation of ESG Score by Year and Industry.

3.6 Correlogram

To examine the relationships among variables and assess potential redundancy among predictors, pairwise correlations were computed using Pearson's correlation coefficient (r). Pearson's r quantifies the linear association between two continuous variables, ranging from –1 to +1, where values near ±1 indicate a strong linear relationship, and values close to zero imply weak or no correlation. Correlation analysis provides a diagnostic step for identifying potential multicollinearity, which can affect model stability and interpretability — particularly for linear estimators.

It should be noted that the variables included in this correlation matrix represent the final selected features after preliminary testing. During the earlier data preparation phase, variables showing near-perfect linear dependence and limited marginal contribution were excluded to reduce redundancy. For example, NI/NS (Net Income / Net Sales) was removed due to its very high correlation ($r \approx 0.96$) with EBIT/NS, while the latter was retained as a more informative profitability measure with higher predictive relevance in preliminary model evaluations. Consequently, the correlogram presented in Figure 3 visualizes correlations among the refined set of predictors that were ultimately used in model training.

To analyze the correlative relationships within variables in the dataset, the variables are plotted in correlograms, which can be found in Figures 3. Positive correlations are illustrated in red while negative correlations are illustrated in blue. Color intensity is proportional to the correlation coefficient.

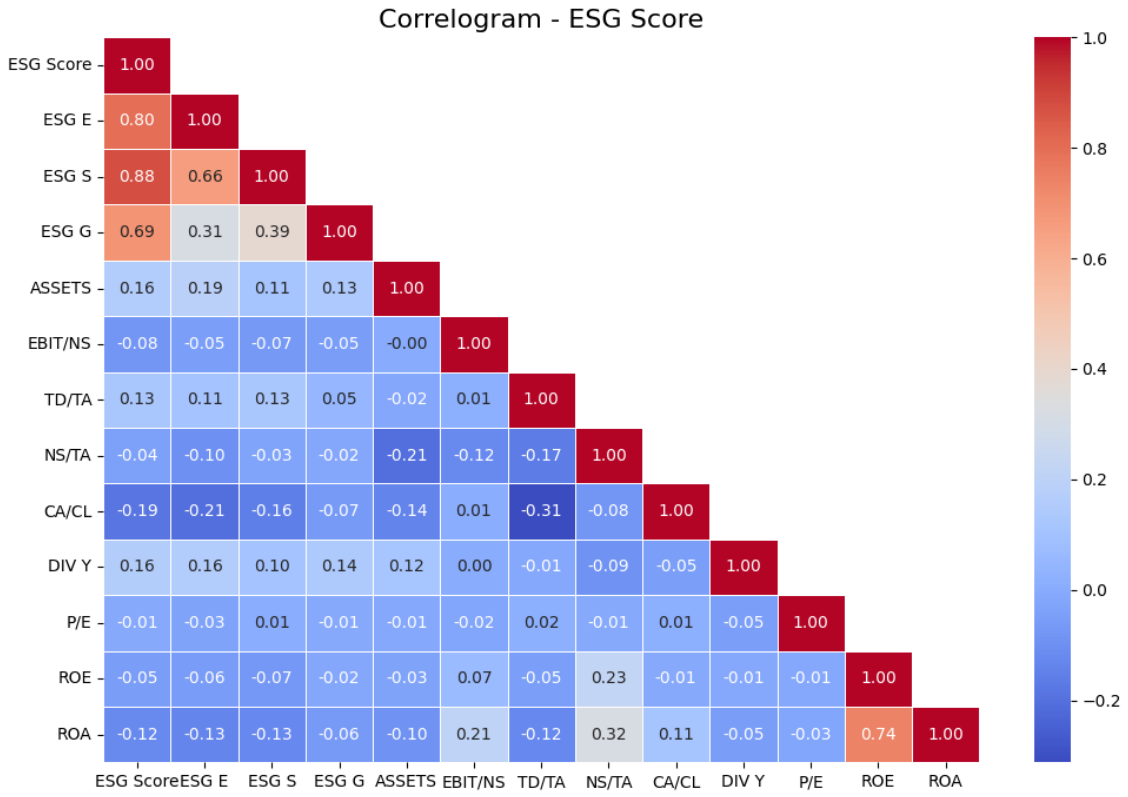

Figure 3. Correlogram.

The strongest positive associations appear between Return on Equity (ROE) and Return on Assets (ROA), both profitability-based measures driven by net income performance. Similarly, EBIT/NS exhibits a strong positive correlation with ROE, indicating that firms with higher operating profitability typically achieve higher returns on equity. Moderate negative correlations emerge between leverage and profitability measures, such as between Total Debt to Total Assets (TD/TA) and ROA, suggesting that higher leverage is, on average, associated with lower returns. Liquidity ratios such as CA/CL show weaker or more heterogeneous relationships with profitability, indicating that short-term solvency conditions vary independently from performance and sustainability factors across sectors.

Overall, the correlation results indicate moderate interdependencies but no critical multicollinearity among the retained predictors. Tree-based models such as Random Forest and XGBoost, which form the primary modelling techniques in this study, are robust to

the remaining correlations due to their hierarchical structure. For linear models such as Ridge and Lasso regression, regularization further mitigates residual multicollinearity, which will be further discussed.

From a broader perspective, the observed positive associations among size and profitability ratios and ESG outcomes suggest that larger and more profitable firms can achieve higher sustainability ratings. While this does not imply causality, it hints that financially healthy firms may possess greater resources and incentives for sustainability practices.

3.7 Density Functions of ESG Scores

Figure 4 displays kernel density estimates (KDE) for the overall ESG score and its E, S and G score subcomponents. The dataset contains 4,937 firm-year observations. KDE is a non-parametric method for estimating the probability density function of a random variable, providing a representation of how values are distributed across the sample (Silverman, 1986).

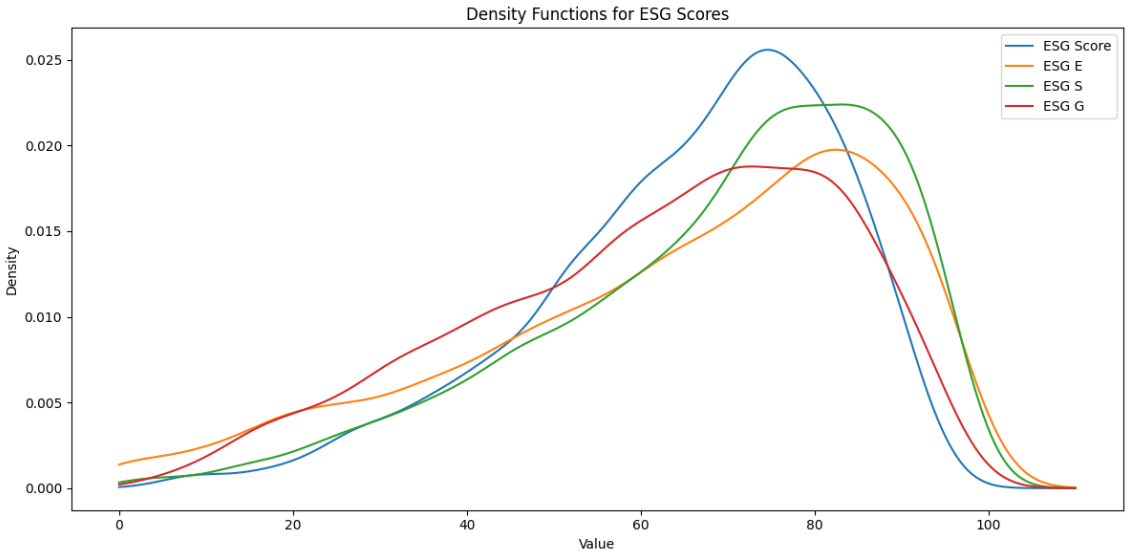

Figure 4. Density functions of ESG score, E, S and G.

Table 6. Density functions of ESG score, E, S and G.

	ESG score	ESG E	ESG S	ESG G
Sample Size	4937	4937	4937	4937
Bandwidth	3.15344	4.27896	3.62623	3.8181

The bandwidth determines how much the curve is smoothed: A smaller bandwidth produces a curve that follows local fluctuations more closely, and a larger bandwidth smooths over wider ranges, emphasizing the general shape of the distribution but potentially hiding smaller peaks or irregularities.

Bandwidths are determined using Silverman's rule of thumb, resulting in values between approximately 3.1 and 4.3 for the ESG variables. The values in the figure mean that, for example, the ESG Score distribution is smoothed over windows of about three units on the 0–100 scale, while ESG E is smoothed slightly more broadly.

The density curves show how ESG scores are not uniformly distributed. Most observations cluster in the 60-85 range, while relatively few observations locate at extremes. E and G distributions appear broader and more dispersed, and S has slightly sharper peak at high values. The density functions matter as they highlight both the central tendency and dispersion of ESG scores in the dataset.

3.8 Missing Data Assessment and Handling

It is important to examine the extent and nature of missing data before modelling, as missingness can affect both the reliability and interpretability of results. In the dataset, missing values occur across several variables with differing magnitudes and underlying causes. Table 7 summarizes the proportion of missing observations for each variable by industry classification. The missingness pattern is unevenly distributed, both in terms of variables and industries, suggesting that values are not missing completely at random.

This section focuses on describing the patterns and implications of missing values within the dataset, while the specific imputation procedures and justifications are presented separately in Section 3.9.; Pre-Modelling Procedures and Pipeline Design.

Table 7. Missing values per industry table.

	Industry 1 Missing	Industry 2 Missing	Industry 3 Missing	Industry 4 Missing	Industry 5 Missing	Industry 6 Missing
ESG score	0	0	0	0	0	0
ESG E	0	0	0	0	0	0
ESG S	0	0	0	0	0	0
ESG G	0	0	0	0	0	0
ASSETS	2	0	0	0	0	0
EBIT/TA	14	0	0	14	6	0
EBIT/NS	15	0	0	14	6	1
TD/TA	2	0	0	0	0	0
NS/TA	3	0	0	0	0	0
CA/CL	2	0	0	424	293	411
DIV Y	18	1	0	1	0	1
P/E	251	32	9	46	15	22
ROE	46	2	0	1	0	1
ROA	10	1	0	55	0	1

3.8.1 Extent and Distribution of Missingness

Table 7. reports the proportion of missing observations for each variable across the six industry classifications. The pattern is uneven and concentrated in specific variables and sectors. The Current Assets to Current Liabilities (CA/CL) ratio shows the highest level of missingness, with approximately one-quarter of observations absent overall. This absence particularly occurs in the financial and real-estate sectors—industries 4 (Banks), 5 (Insurance), and 6 (Real Estate and Investments)—where in several cases all CA/CL values are missing.

Another variable affected by substantial missingness is the Price-to-Earnings (P/E) ratio. Inspecting the data reveals that most missing P/E entries appear with negative Earnings Before Interest and Taxes (EBIT), making the ratio undefined rather than simply unreported. Missingness is therefore embedded in the accounting structure of the variable. For the remaining financial ratios, missing values are comparatively rare and irregular, suggesting minor gaps in firm-level reporting rather than systematic omissions.

3.8.2 Mechanisms of Missing Data

These patterns imply that missingness is not Missing Completely at Random (MCAR), where the probability of missingness is independent of any variable in the dataset. Rather, the missingness mechanisms align with Missing at Random (MAR) or Missing Not at Random (MNAR).

The CA/CL variable follows an industry-dependent pattern consistent with MAR, since the likelihood of missing values depends on a categorical factor (industry classification) observable in the data. In contrast, the P/E variable is closer to MNAR, because missingness is systematically related to the unobserved (negative) earnings values that make P/E undefined. These mechanisms imply that missingness has economic meaning which needs to be considered.

3.8.3 Implications for Data Integrity and Modelling

Recognizing the origins of missingness has implications for how these gaps should be treated. Excluding all observations with missing values would remove entire industries and firms with negative earnings from the analysis, introducing selection bias and reducing the sample size, and simple mean imputation would ignore economically meaningful differences between industries or profitability levels. Therefore, a more context-sensitive approach is needed, which allows the dataset to remain as complete as possible while preserving the interpretability of model relationships.

3.8.4 Treatment Principles

In this thesis, for models that need complete data, the missing data are handled using variable-specific strategies that depend on the structure and origin of the missingness. For variables such as CA/CL and P/E, missingness itself carries interpretive value, indicating the presence of unique financial structures or earnings conditions. The chosen methods aim to retain their informational content while ensuring that models requiring complete inputs can still be trained. The details of these procedures—including the use of sentinel values paired with missingness indicators, the rationale for applying them only to selected industries, and how they are implemented separately within training and test partitions—are presented in the “Pre-Modelling Procedures and Pipeline Design” chapter 3.9. For variables with minor or unsystematic missingness, such as ROE, ROA, or Dividend Yield, a more conventional industry-mean imputation approach is applied, as described there as well.

In contrast, models such as XGBoost inherently manage missing values during tree construction and therefore do not require these imputations.

3.9 Pre-Modelling Procedures and Pipeline Design

Before applying imputation methods, the data is randomly split into training (80%) and testing (20%) sets, to prevent data leakage.

Next step of the data processing is to deal with the rest of the missing values. When choosing an imputation strategy, it should be considered why values are missing, and how extensive the missingness is.

Examining the data, it was discovered that missingness is most prevalent in CA/CL and P/E - CA/CL having approximately quarter of the observations missing, which are occurring in banks, insurance, and real estate and investment companies (Industries 4 to 6).

Since all the CA/CL observations in industries 4 and 6 are missing, imputation methods such as MissForest or mean imputation would not produce desirable outcomes for them, as the industries do not have values to interpret from. Due to this, in some models, for industries 4 and 6, an absolute value of -999 was imputed where CA/CL was missing, with missing flag to highlight the missingness. This compromise was the result of testing different methods to deal with this issue – including omitting CA/CL as variable. As a result, the models where this data processing was used were able to recognize this solution as a variable without real weight for predicting ESG Score, as the value of -999 is absurdly large in comparison to average CA/CL values, paired together with the missingness flag and only appearing in two industries. This allowed for keeping the variable in the models that cannot internally handle missing values, as for real CA/CL values the variable was a contributing predictor for ESG Score.

Another variable with high missingness was P/E. In the case of P/E, lots of the observations flagged for missing value also had negative Earnings Before Interest and Taxes. For those companies, similarly, due to most of the existing P/E values being positive, and missing values being negative by their real nature, typical imputation methods do not provide desirable outcomes. Because of this, for models XGB3 and RF2 missing P/E values were replaced with figure *-999*, when their EBIT was negative, and flagged with categorical variable of 1 to identify them, as most missing P/E values resulted from the company having negative earnings, and the models require a value for optimal performance, without having to sacrifice the predictor itself from the analysis, or the corresponding rows of the missing variable. This method did not appear to bias the results, and it provided higher prediction accuracy for Random Forest models than RF models that handled missing values internally, as with the missing flag, the model able to disregard imputed values that were way off from real ones, as well as being able to interpret the positive link between sustainability score and simply having a positive P/E value. This is due to the problem where when there are no negative P/E values, and thus the model is only comparing positive ones, the predictive power over sustainability score is low. The main predictive power of this variable only showed up when observations with missing value,

as in negative earnings, can be included. For observations with positive EBITs, missing P/E values were imputed with industry means.

For other variables, the general missingness was low, and they were imputed with mean values of their respective industries, computed on the training set. The imputation results were then applied to the test set.

It should be noted that the motivation behind these imputation choices for specific models was an aim to get the best results from models that have theoretical potential for competitive prediction accuracy but have limited functionality to accommodate for the weaknesses of the dataset, in terms of missing values. While these imputations improved the performance of random forest models, they were ultimately outperformed by XGBoost model that handled missingness internally and did not require previously discussed imputations.

3.10 Winsorisation

Winsorisation is the process of moving outlier values to match values of a specific quantile (Ranta, 2023, Ch. 9). Analysing the dataset, a winsorisation threshold of 1%-99% is selected, similarly to, for example, Chowdhury et al. (2023), as it appears as a literature standard for firm-level financial datasets of similar scale.

For Random Forest, winsorisation was performed after the imputation stage to the training set to ensure that outlier caps were computed using the complete imputed distributions of each variable. P/E was capped only on the upper tail at the 99^{th} percentile among positive values leaving sentinel unchanged, while other continuous variables were symmetrically capped at the 1^{st} and 99^{th} percentiles. The same percentile thresholds, estimated from the training set, were applied to the test set to ensure consistent feature ranges without leaking target information. Finally, industry dummies were created and

aligned across splits. The adopted order maintains consistent data ranges and prevents the reappearance of outliers introduced by later steps.

For XGBoost, winsorisation was applied to all continuous financial ratios to mitigate the influence of extreme outliers while preserving the overall rank order and structure of the data. ASSETS were winsorised, and only after turned into LOG(ASSETS). The 1st and 99th percentile thresholds were computed from the training set and then applied to both the training and test set, ensuring consistent feature ranges and preventing data leakage. As XGBoost natively handles missing values, winsorisation was applied only to the non-missing values, leaving NaN entries untouched for internal treatment by the model.

4 Methodology

This chapter provides a short overview of the different machine learning techniques used in the thesis, and follows by reviewing the model specifications, such as data processing withing the model and model parameters.

4.1 Machine Learning Models – general information on models used

This subchapter provides short overview of the different machine learning techniques used in the thesis.

4.1.1 Random Forest

Introduced by Leo Breiman in 2001, Random Forest is an ensemble learning method, that constructs a multitude of decision trees and combines their results for predictions (Ramchandran et al., 2021). The random forest algorithm selects a random subset of features for each weak estimator known as random subspace method (Ranta, 2023, Ch. 8). The idea of random forest is to reduce overfitting and improve accuracy by averaging many uncorrelated decision trees; While a single decision tree can be prone to bias, a random forest reduces this by training multiple trees on random subsets of data, then aggregating their outputs (IBM, n.d.-a). This so called bagging, or bootstrap aggregating approach with feature randomness makes that each tree provides a unique result, and their average is more generalizable than any tree individually (IBM, n.d.-a). In practice, each tree is grown on a bootstrap sample of the training dataset, and at each split only a random subset of features is considered as candidates for splitting (Breiman, 2001). This process introduces diversity among the trees, preventing any single feature or data point from dominating the model. After training, the forest makes predictions by aggregating the trees' outputs. Random forest is capable of both classification and regression tasks, and for a regression problem, the model averages the predicted values from all trees, and for classification, it takes a majority vote among the trees (IBM, n.d.-a).

4.1.2 XGBoost

Like random forests, gradient boosting is an ensemble method that combines multiple decision trees, but unlike random forest, it's boosting builds trees sequentially rather than in parallel (NVIDIA, n.d.). In gradient boosting, each new tree is trained to correct the errors (residuals) of the combined previous trees, gradually improving the model's performance (NVIDIA, n.d.). The model can be seen as a weighted sum of multiple weak learner trees, which together form a model with strong predictive power. The final model comes together by combining many small decision trees, each one correcting the mistakes of the previous ones. In contrast to random forests, which reduce variance by averaging many independent trees, boosting reduces bias by gradually improving the fit through a sequence of corrections (NVIDIA, n.d.).

XGBoost (Extreme Gradient Boosting) is one of the most successful implementations of gradient boosting tenchniques (Ranta, 2023, Ch. 8). It extends the basic gradient boosting framework with engineering improvements and regularization, enhancing model speed and accuracy (NVIDIA, n.d.). XGBoost builds decision trees using a novel level-wise parallelization strategy: instead of the strictly sequential tree-by-tree boosting process, XGBoost can grow trees in parallel by processing multiple splits at once (NVIDIA, n.d.). It also includes regularized learning objective, which adds a penalty for model complexity to the loss function, helping prevent overfitting and improving generalization (NVIDIA, n.d.). In practice, XGBoost's objective at each iteration minimizes a combination of the gradient-based loss (e.g. mean squared error) and a regularization term that penalizes overly complex trees (NVIDIA, n.d.).

4.1.3 RIDGE and LASSO Regression

Ridge and Lasso regression extend the scope of linear regression models by including regularization (Ranta, 2023, Ch. 9). The methods constrain the coefficient estimates and force them towards zero. The models address common regression problems of multicollinearity and overfitting, improving robustness of the models compared to traditional

regression. Ridge and Lasso are considered as best regularization techniques (Ranta, 2023, Ch. 9)

In a standard linear regression, the model finds coefficients that minimize the residual sum of squares, fitting the data as closely as possible. Both ridge and lasso methods add a penalty to the size of the regression coefficients to the loss function, which prevents the model from relying heavily on any specific variable, making the results more stable (IBM, n.d.-b).

Ridge regression (L2 regularization) shrinks coefficients towards - but never fully to zero, retaining all predictors in the model, but reducing their influence (IBM, n.d.-b). Ridge is useful when predictors are correlated, as it spreads their effect more evenly (IBM, n.d.-b).

Lasso regression (L1 regularization) can also shrink coefficients to zero. In practice, Lasso can automatically select a smaller subset of predictors, that can improve model interpretability and efficiency (IBM, n.d.-b). Lasso is valuable when identifying and selecting the most important predictors is a priority. Both models balance model complexity and prediction accuracy in different ways (IBM, n.d.-b).

4.2 Trained Models

This subchapter further reviews the model specifications, such as data processing with-ing the model and model parameters for trained models. The models were implemented and executed in Python using the Google Colab environment.

4.2.1 XGB1

XGBoost model trained on original data without imputation of missing variables, as the model can internally deal with missing values. This was best performing model with both

ASSETS and SIZE (LOG(ASSETS)), and performed best after retraining best parameters for model using SIZE around prior best searches, with:

Parameter distributions of = {

```
  "max_depth": randint            [6, 18],
  "min_child_weight": randint     [1, 10],
  "learning_rate": loguniform     [0.01, 0.3],
  "n_estimators": randint         [150, 600],
  "subsample": uniform            [0.6, 0.4],
  "colsample_bytree": uniform     [0.6, 0.4],
  "gamma": uniform                [0.0, 2.0],
  "reg_alpha": loguniform         [1e-4, 10.0],
  "reg_lambda": loguniform        [1e-3, 30.0],
},
```

yielding in best parameters of:

```
{learning_rate≈0.024, max_depth=16, n_estimators=356, sub-
sample≈0.643,     colsample_bytree≈0.889,     gamma≈0.472,
min_child_weight=9, reg_alpha≈0.014, and reg_lambda≈0.0013}.
```

The parameter distribution is applicable for following XGboost models as well.

Hyperparameter optimization for the XGBoost model was conducted using scikit-learn's `RandomizedSearchCV`, which samples random combinations of parameter values from pre-specified ranges and evaluates each using five-fold cross-validation `(cv=5)`. Sixty parameter combinations were evaluated `(n_iter=60)` and the configuration yielding the lowest cross-validated RMSE was selected. The model was then refitted on the full training set using these optimal parameters `(refit=True)` to obtain the final tuned estimator. This process of hyperparameter optimization is applicable for the following XGBoost and Random Forest models as well.

Table 8 below describes what the XGBoost hyperparameters control in the model, and their effects on model learning, based on the official XGBoost documentation (XGBoost Developers, 2024).

Table 8. Descriptions of XGBoost hyperparameters (XGBoost Developers, 2024).

Parameter	Description
learning_rate	Shrinkage factor applied to each tree's contribution. A smaller value slows learning and requires more trees, reducing overfitting.
max_depth	Max depth of individual trees. Deeper trees capture more complex nonlinear relationships but increase overfitting risk.
n_estimators	Number of boosting rounds. Determines total model complexity with learning rate: more trees compensate for lower learning rate.
subsample	Fraction of the training data randomly sampled for each tree. Introduces randomness improving generalization and reducing overfitting.
colsample_bytree	Fraction of features randomly selected for each tree. Reduces feature correlation effects and increases model diversity.
gamma	Minimum required loss reduction to make a further split. Acts as a regularization term: higher values make the algorithm more conservative, pruning weak splits.
min_child_weight	Minimum sum of instance weights in a child node. Prevents the creation of nodes representing too few samples or low variance, stabilizing deep trees and controlling overfitting.
reg_alpha	L1 regularization term on leaf weights. Encourages sparsity in tree leaf weights, reducing the number of active leaves and simplifying the model.
reg_lamba	L2 regularization term on leaf weights. Penalizes large weight magnitudes to reduce model variance and improve generalization stability.

4.2.2 XGB2

Same data and model as XGB1, but uses ASSETS instead of SIZE, and with that has different best parameters.

Best Parameters:

```
{'colsample_bytree': 0.9, 'learning_rate': 0.1,
'max_depth': 15, 'min_child_weight': 5, 'n_estima
tors': 290, 'subsample': 0.8}
```

4.2.3 XGB3

Trained on data with same preprocessing steps as RF2.

Best Parameters:

```
{'colsample_bytree': 0.9, 'learning_rate': 0.1,
'max_depth': 15, 'min_child_weight': 5, 'n_estima
tors': 290, 'subsample': 0.8}
```

4.2.4 RF1

This RandomForest model trained on data where CA/CL is dropped, as approximately 25% of observations are missing. For low missingness, uses mean industry imputation. For P/E, mean imputation where EBIT is positive, -999 where EBIT is negative. P/E missing flag added for missing observations. Hyperparameters tuned with RandomizedSearchCV, yielding:

Best Parameters:

```
{'max_depth': 14, 'min_samples_leaf': 1, 'min_samp
les_split': 4, 'n_estimators': 140}
```

4.2.5 RF2

CA/CL imputed with -999 for industries 4-6 when missing. For missing in industries 1-3, mean industry imputation is applied. CA/CL missing flag added for missing observations. For P/E, mean imputation is applied where EBIT is positive, -999 where EBIT is negative. P/E missing flag added for missing observations. Hyperparameters were searched using RandomizedSearchCV, with:

Parameter distributions of ={

```
"n_estimators":              [300, 500, 800],
"max_depth":                 [None, 8, 12, 16, 20],
"min_samples_split":         [2, 4, 6, 10],
"min_samples_leaf":          [1, 2, 4, 8],
"max_features":              ["sqrt", "log2", 0.3, 0.5, 0.7,
1.0],
```

```
"bootstrap":                    [True],
"max_samples":                  [None, 0.6, 0.8],
"criterion":                    ["squared_error","absolute_er-
                                ror"],
"min_impurity_decrease":        [0.0, 1e-6, 1e-5, 1e-4],
"max_leaf_nodes":               [None, 128, 256],
"ccp_alpha":                    [0.0, 1e-4, 1e-3]
}
```

Resulting in best parameters of:

```
{'n_estimators': 800, 'min_samples_split': 6, 'min_sam-
ples_leaf': 1, 'min_impurity_decrease': 1e-05, 'max_samples':
None, 'max_leaf_nodes': None, 'max_features': 1.0,
'max_depth': 16, 'criterion': 'squared_error', 'ccp_alpha':
0.0001, 'bootstrap': True}.
```

Table 9 below describes what the RandomForest hyperparameters control in the model, and their effects on model learning, based on The RandomForestRegressor implementation in *scikit-learn* (scikit-learn developers, 2024).

Table 9. Descriptions of RandomForest parameters (scikit-learn developers, 2024).

Parameter	Description
n_estimators	The number of trees in the ensemble. Larger numbers reduce variance through averaging to a certain point.
max_depth	Maximum depth of individual trees. Controls the detail of each tree's partitions. Low depth simplify relationships and reduce overfitting; deeper trees allow more complex patterns but increase variance.
min_samples_split	Minimum number of samples required to split an internal node. Higher values: fewer splits and smoother predictions. Lower values: deeper, more detailed trees.
min_samples_leaf	Minimum number of samples required to form a leaf node. Prevents leaves that represent very small sample subsets. Increase smooths prediction.
max_features	Proportion of predictors randomly considered at each split. Smaller values increase model diversity and reduce overfitting; larger values allow each tree to fit more accurately to the training data.
bootstrap	Whether to sample training observations with replacement when building each tree. The default "True" enables classical bagging, producing more diverse trees and allowing estimation of out-of-bag error for validation.

max_samples	If set (<1.0), limits the proportion of the training data used for each bootstrap sample. Using less than the full sample increases tree diversity and training speed but can increase bias.
criterion	Measure split quality. "squared_error" minimizes mean-squared deviations (default for regression); "absolute_error" is more robust to outliers and yields median-based predictions.
min_impurity_decrease	A split is performed only if it decreases impurity (MSE or MAE) by at least this value. Acts as a small regularization term, preventing, low-gain splits and reducing overfitting.
max_leaf_nodes	Caps the number of terminal nodes in each tree. Provides a direct upper bound on tree complexity, similar to limiting max_depth. Smaller values simplify the model and control overfitting.
ccp_alpha	Cost-complexity pruning parameter. After trees are built, branches that contribute less than ccp_alpha to overall model performance are pruned. Larger values yield simpler, more regularized models.

4.2.6 RF3

Instead of Sklearn library's RandomForestRegressor, this model used Xgboost library's XGBRFRegressor. It uses random forest-style bagging instead of boosting residuals and can natively handle missing values, with

Param grid: {

```
        "n_estimators":             [100, 200, 300, 500],
        "max_depth":                [3, 5, 7, 9, 12],
        "subsample":                [0.6, 0.8, 1.0],
        "colsample_bynode":         [0.6, 0.8, 1.0],
        "reg_alpha":                [0.1, 0.5, 1],
        "reg_lambda":               [1, 2, 3],
}
```

Resulting in best parameters of:

```
{'n_estimators': 200, 'max_depth': 12, 'subsample': 1.0,
'colsample_bynod': 0.6, 'reg_alpha': 1, and 'reg_lambda': 1}
```

4.2.7 RIDGE

L2 Regularization. Ridge minimizes the squared error with a penalty on the squared magnitude of coefficients. It helps prevent overfitting by shrinking coefficients, especially useful when predictors are correlated. Ridge regression natively handles multi-target outputs.

Variables with excessive missingness dropped: CA/CL, P/E. For others with low missingness, industry mean imputation was applied. Because asset values typically span several orders of magnitude and can be right-skewed, a logarithmic transformation was applied. model was trained with `sklearn.linear_model.Ridge`. Features were standardized using `StandardScaler`, and the model used a fixed penalty `term alpha=1.0`.

4.2.8 LASSO

L1 Regularization for Multi-Output. LASSO encourages sparsity by penalizing the absolute values of coefficients, which can set some coefficients exactly to zero. The MultiTaskLasso variant enforces a common sparsity structure across all output variables, which is useful when similar features drive multiple ESG metrics.

Variables with excessive missingness dropped: CA/CL, P/E. For others with low missingness, industry mean imputation was applied. Because asset values typically span several orders of magnitude and can be right-skewed, a logarithmic transformation was applied. model was estimated using `sklearn.linear_model.MultiTaskLasso`. Preprocessing mirrored Ridge regression with standardized features. The penalty parameter was set at `alpha≈0.1`.

5 Results

In this chapter model outputs and obtained results are outlined and analyzed.

Table 10. Model performance comparison.

MODEL	ESG Score			ESG E			ESG S			ESG G		
	MSE	MAE	R^2	MSE	MAE	R^2	MSE	MAE	R^2	MSE	MAE	R^2
XGB1	***105.1595***	***7.9780***	***0.6433***	***198.619***	***10.8325***	***0.6278***	***160.0157***	***9.6094***	***0.5918***	***269.8518***	***13.1475***	***0.3729***
XGB2	111.9203	7.9959	0.6203	204.2286	10.8775	0.6172	166.9638	9.7034	0.5741	282.0085	13.3166	0.3446
XGB3	112.5980	8.1956	0.6180	218.3546	11.2473	0.5908	166.8371	9.8598	0.5744	272.7754	13.2920	0.3661
RF1	124.7424	8.7185	0.5768	225.2110	11.7319	0.5779	184.4019	10.4137	0.5296	297.6851	13.9232	0.3082
RF2	***120.8700***	***8.5900***	***0.5900***	***225.9264***	***11.6951***	***0.5766***	***179.8529***	***10.2343***	***0.5412***	***286.6319***	***13.6902***	***0.3339***
RF3	127.3670	9.2377	0.5679	228.0585	11.8150	0.5726	189.5273	10.5340	0.5165	299.5708	13.9441	0.3038
RIDGE	186.8742	8.7077	0.3661	358.0796	15.1353	0.3289	268.1764	12.8060	0.3159	368.4399	15.6844	0.1437
LASSO	186.9609	10.6435	0.3658	358.1047	15.1250	0.3288	268.3426	12.8152	0.3155	368.4728	15.6947	0.1437

The performance metrics used to compare the models are Mean Squared Error (MSE), Mean Absolute Error (MAE) and Coefficient of Determination (R^2).

MSE

MSE measures the average of the squared differences between predicted values and actual values, formally:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \quad (1)$$

In the MSE equation, y_i is the actual value and $\hat{y}_i$ is the prediction. Because errors are squared, MSE penalizes large errors more heavily than small ones, making it sensitive to outliers (BMC, n.d.). A lower MSE indicates a better fit — 0 meaning perfect predictions, while higher MSE indicates that the model's predictors deviate substantially from actual scores on average.

MAE

MAE is the mean of the absolute differences between predicted and actual values, formally:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i | . \quad (2)$$

In the MAE equation, y_i is the actual value and $\hat{y}_i$ is the prediction. MAE gives the average magnitude of errors in the same units as the target variable. Compared to MSE, MAE is less sensitive to outliers because it doesn't heavily penalize large errors. MAE reveals the typical prediction error of the model, setting an expectation on how far a predicted ESG score is from true value on average.

R^2

R^2 represents the proportion of variance in the target variable that is explained by the model's predictions. An R^2 of 1.0 (or 100%) indicates that the model perfectly explains the variation in ESG scores, whereas a value of 0 indicates no improvement over predicting the mean, and negative values indicate the model performs worse than that baseline on the test set. R^2 is calculated as:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}. \quad (3)$$

In the R^2 equation, y_i is the actual value, $\hat{y}_i$ is the prediction and $\bar{y}_i$ is the mean of the observed values y_i, and thus $\sum(y_i - \hat{y}_i)^2$ is the residual sum of squares, i.e., the total

squared error between predictions and actual values, and $\sum(y_i - \bar{y}_i)^2$ is the total sum of squares, i.e., the total variance in the observed data relative to the mean.

By these metrics, the best performing XGBoost, - and overall, model was XGB1, with MSE of 105.16, MAE of 7.98 and R^2 of 0.64. The model was best by every measured metric. The best performing random forest model was RF^2 (R^2 of 0.59), which included the previously discussed P/E and CA/CL imputations. The best performing model (XGB1) is the subject for more in-depth explainable AI analysis. The analysis focuses on ESG Score instead of its subcategories. Further results considering other models can be found in Appendix 2 for comparison.

5.1 Observed and Residuals vs. Predicted ESG Score

Chapter 5.1. starts by explaining how to interpret the observed vs. predicted ESG score, and residuals vs. predicted ESG score graphs, and then outlines their results.

5.1.1 Observed vs. predicted plot

An observed vs. predicted plot is a scatter plot of true target values against the model's predicted values for the same observations. It visualizes how well the predictions align with real values. In a perfectly optimized model, all the points lie on the 45° line. The closer the points lie of the 45° line, the more accurate the prediction. When the model is unbiased and accurate, the cloud of points will be symmetric around the diagonal line. Systematic deviations from the diagonal reveal biases: for example, if points tend to fall above the line at one end and below at the other, the model might be under-predicting high scores and over-predicting low scores. A tight clustering of points along the line implies low error variance, whereas a wide scatter indicates high prediction variability.

5.1.2 Residuals vs. predicted plot

Residual plot displays the residuals (prediction errors) on the vertical axis against the predicted values. In a well-behaved model, residuals should scatter randomly around 0 (the horizontal axis) with no clear pattern, indicating the model's errors are unbiased and have constant variance across predictions. If the spread of residuals increases or decreases with the predicted value, it indicates heteroscedasticity. For Random Forests or XGBoost, which are flexible models, a residual plot can reveal if there are regions where even these models have systematic errors, possibly pointing to segments of the data with different patterns or the need for additional features.

5.1.3 Observed vs. Predicted ESG Score

Below are figures for Observed vs. Predicted ESG Score, and Residuals vs. Predicted ESG Score for XGB1.

Figure 5. XGB1 Observed vs predicted ESG score and residuals.

The Observed vs. Predicted ESG Score scatterplot compares observed ESG scores on the horizontal axis with the model's predicted scores on the vertical axis. Each point corresponds to a firm-year observation from the test sample.

The overall shape of the scatter is elongated along the diagonal, indicating a clear positive relationship between observed and predicted ESG scores. This alignment means that the model's predictions generally increase as the actual ESG score increases

The density of points is highest in the middle range, approximately between observed ESG scores of 50 and 80, which corresponds to where most firms are concentrated in the dataset. In this region, the points are relatively close to the reference line, showing that the model predicts mid-range ESG values with relatively low deviation.

At the lower end of the scale (roughly below 40), the scatter becomes more dispersed. The predicted values vary more widely around the reference line, with several points located well above or below it. This shows that prediction errors are larger and more variable among low-scoring firms. This might indicate that the financial characteristics of low-ESG firms are more heterogeneous, making their ESG outcomes harder to estimate from the used financial statement items alone. Another factor which might affect this is the lower observations pool of lower scoring firms among the STOXX Europe 600 Index dataset.

At the upper end of the distribution (above roughly 85), most points fall slightly below the reference line. This may reflect a mild underestimation of the best-performing firms, possibly because the model compresses extreme values toward the mean or because financial ratios alone capture only part of the drivers behind very high ESG ratings.

Across the full range, no clear horizontal or vertical clustering appears, implying that the relationship between observed and predicted values remains rather monotonic and consistent rather than segmented or scattered.

In summary, the figure shows that XGB1 reproduces the general structure of ESG variation in the data, with close alignment through the middle range and greater dispersion at the extremes. The widening variance at both tails likely reflects the bounded nature

of ESG scores and the limited explanatory scope of purely financial indicators in predicting very low or very high sustainability performance.

5.1.4 Residuals vs. Predicted ESG Score

The Residuals vs. Predicted ESG Score displays the model residuals ($e_i = y_i - \hat{y}_i$) on the vertical axis and the predicted ESG scores on the horizontal axis. The red dashed line represents the zero-residual line, where predicted and observed scores coincide.

The residuals are distributed roughly symmetrically around zero across the full range of predicted ESG scores. No clear curvature or systematic slope is visible in the scatter, which indicates that deviations between observed and predicted values do not vary in a directional way across the prediction range. Most of residuals fall within approximately ±20 ESG points, with only a small number of observations extending beyond ±40.

The horizontal spread of residuals appears widest between predicted values of about 40 and 70, after which the scatter gradually narrows toward the higher end of the predicted range. This pattern might indicate mild heteroscedasticity, where prediction errors are somewhat larger in the mid-range of the model's fitted values and smaller for firms that are predicted to have very high ESG scores.

There is no visible grouping or banding of points that would indicate separate firm types or missing categorical variables. The residuals do not show any clear shape or curve, which means the model's structure fits the data reasonably well, and most of the remaining variation might likely reflects random noise rather than a systematic modelling error.

A small number of observations have residuals larger than ±40 ESG points. These cases are scattered and do not occur within a specific range of predicted values. They might represent firms whose ESG scores are influenced by factors not captured in financial statement items, such as controversies, or unique governance situations.

5.2 Feature Importances

Tree-based models like Random Forests and XGBoost can provide intrinsic feature importance measures based on their internal structure. These importance scores reflect how useful each feature was in reducing prediction error across all the trees. They highlight which features the model relied on most during training.

A high importance means the feature was frequently and effectively used to split data in the trees (thus reducing error), but it does not directly imply a causal influence on the target. Moreover, intrinsic importances can be biased: features with many categories or a wide range can appear artificially more important, and if two features are highly correlated, importance may split between them or be concentrated on one arbitrarily.

5.2.1 Permutation Importances

Permutation importance is a technique to assess the importance of a feature by evaluating how much the model's performance deteriorates when the feature's values are randomly shuffled (Molnar et al., 2023); if a feature is truly important, scrambling its values should cause an increase in prediction error (e.g. MSE). It is computed by permuting the values of a feature, in trained model, across all samples and measuring the performance again, against the original performance values. The importance is quantified by the increase in error. This approach quantifies how much each variable contributes to the model's overall predictive ability.

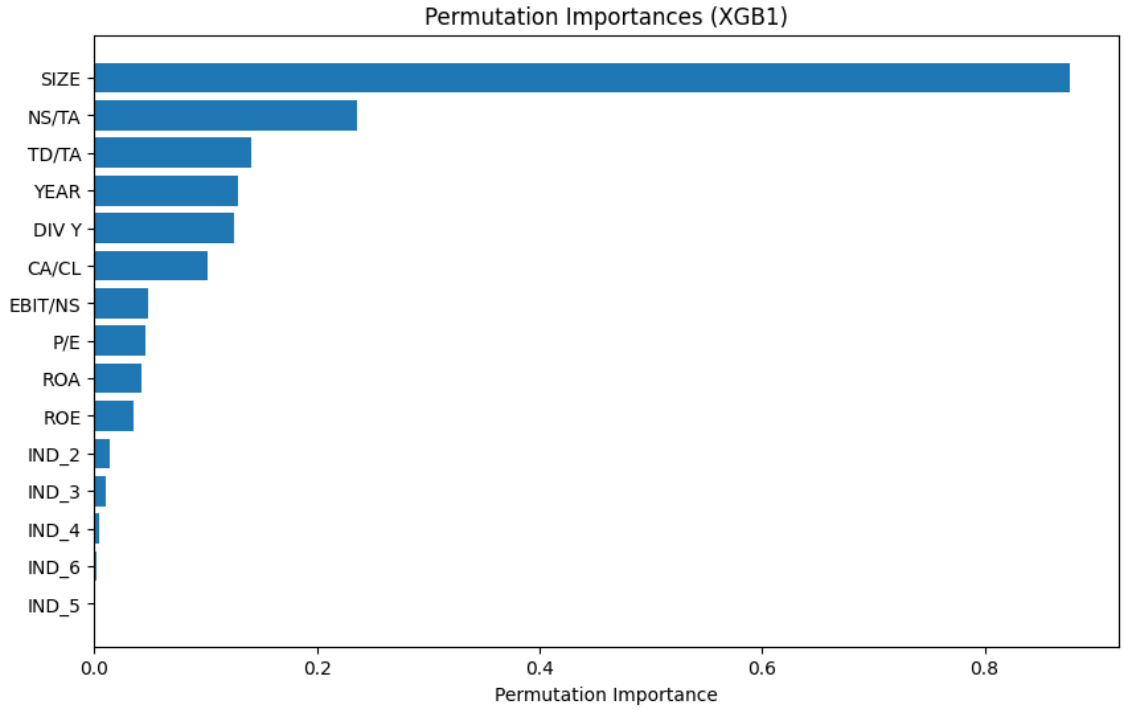

Figure 6. Permutation Importances (XGB1).

The bar chart presents the permutation importances of all predictor variables in the XGB1 model. The x-axis measures the average reduction in predictive performance when each feature's values are randomly shuffled, while the y-axis lists the features in descending order of importance.

SIZE has the highest importance value, followed by NS/TA and TD/TA. YEAR and DIV Y form the next group, showing moderate contribution levels. CA/CL, EBIT/NS, and P/E follow with smaller effects, while ROA, ROE, and the industry dummies (IND_2 to IND_6) have limited impact on model performance.

This pattern suggests that firm size (SIZE) is the dominant factor used by the model when estimating ESG scores, while operational efficiency (NS/TA) and leverage (TD/TA) provide secondary but meaningful information. The relatively low importance of profitability ratios (EBIT/NS, ROA, ROE) implies that these variables, while economically relevant, add

little unique predictive signal once firm size and balance sheet structure are accounted for.

The ranking indicates that firm size (SIZE) provides the most influential source of predictive information for ESG scores, while sales efficiency (NS/TA) and leverage (TD/TA) also contribute meaningfully to model accuracy.

Profitability ratios (EBIT/NS, ROA, ROE) appear less influential, suggesting that once firm scale and balance-sheet structure are accounted for, these variables add little independent predictive value.

YEAR appears in the middle range, but its higher gain value (coming up in next subchapter) suggests that temporal variation adds explanatory value beyond what permutation results alone imply.

Similarly, while the industry dummies rank low in the permutation chart, their higher gain and cover scores indicate that sector membership occasionally plays an important role in explaining broader differences in ESG levels across industries, even if such effects are not strong drivers of overall predictive performance.

Overall, the permutation importance indicates that the model's prediction is driven primarily by scale-related variables, with size effects dominating the learned relationships between financial indicators and ESG outcomes.

5.2.2 Feature Importances by Weight, Cover and Gain

XGBoost provides three types of feature importance metrics derived from the trained model (Ahmed et al., 2024):

- **Weight** refers to the number of times a variable is used to split the data across all trees. It reflects how frequently a feature appears in the model's decision rules,

but not the size of its effect. Variables with high weight values are often used early or repeatedly in the tree-building process.

- **Cover** is the proportion of observations affected by a feature's splits. It shows how many observations are, on average, affected by a feature's splits. A higher cover value means that the feature influences a larger share of the dataset, indicating that it plays a broad role across many predictions.

- **Gain** measures the average improvement in predictive accuracy achieved when a feature is used to split the data. It measures the effectiveness rather than frequency of the feature's use and is the most direct indicator of predictive contribution.

5.2.2.1 Feature Importances by Weight

The bar chart (Figure 7) displays the frequency with which each variable is used to split nodes across all decision trees in the XGB1 model. Higher values indicate that the variable was selected more often during tree construction, reflecting its relative contribution to the model's structure. For example, for SIZE, importance score of 7378 means that summed over every tree and depth, the model chose a split on SIZE 7378 times.

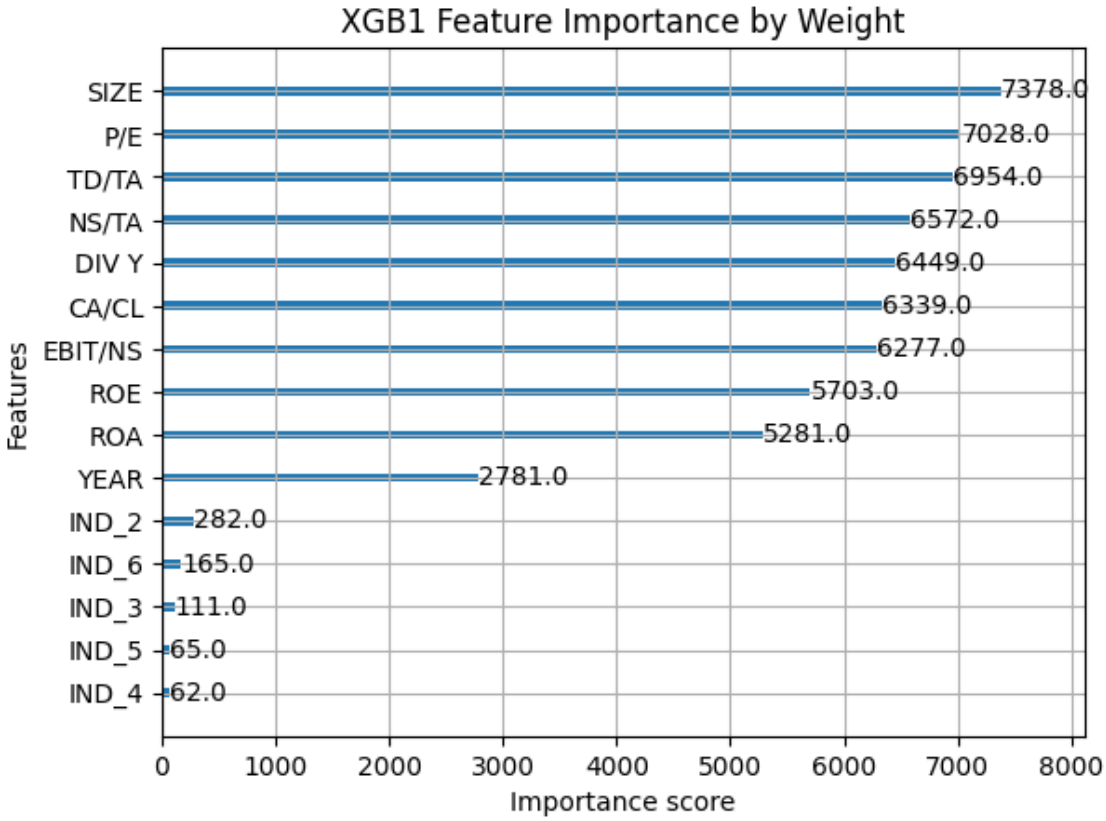

Figure 7. Feature Importance by Weight (XGB1).

SIZE has the highest importance by weight, followed by P/E, TD/TA, and NS/TA. DIV Y, CA/CL, and EBIT/NS appear next with comparable frequencies. ROE, ROA, and YEAR have lower but still visible scores, while all industry dummy variables (IND_2–IND_6) show minimal participation in tree splits.

The frequent use of SIZE indicates that firm scale is one of the most consistently applied variables during tree construction. The model often relies on SIZE for early partitioning of the data, suggesting that firm size provides a general basis for distinguishing ESG score levels among firms.

P/E ranks second by weight but substantially lower in gain and cover, meaning that it is used frequently yet contributes limited incremental improvement to predictive accuracy.

This pattern is typical for features that interact with others or appear in deeper parts of the tree but have weaker independent effects.

TD/TA and NS/TA follow closely, indicating that leverage and operational efficiency also appear regularly in the model's decision structure. These variables likely help refine ESG predictions once firm size-related differences have been accounted for.

The low weights of the industry dummies (IND_2–IND_6) show that sector is seldom used for splitting, but when these variables are used, they tend to affect a large share of the observations (cover) and provide noticeable improvements in predictive accuracy (gain). This suggests that industry information helps the model capture broad differences in ESG scores between sectors, while most of the detailed variation within industries is explained by continuous firm-level financial ratios.

5.2.2.2 Feature Importances by Cover

The figure shows how many observations, on average, are affected by each feature's splits across all trees in the model. For example, for SIZE, importance score of approximately 221 means that when the model does split on SIZE, that split is on average reached by 221 weighted training instances.

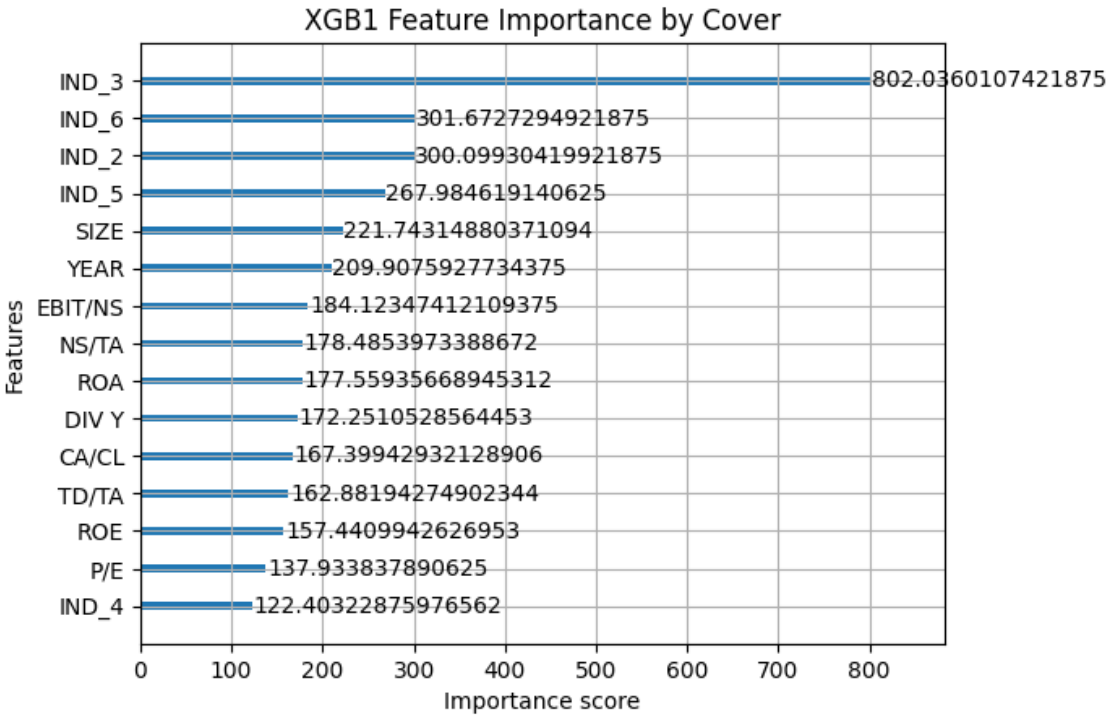

Figure 8. Feature Importance by Cover (XGB1).

IND_3 has the highest cover value, followed by IND_6 and IND_2. IND_5 and SIZE occupy the next positions with moderately lower scores. YEAR, EBIT/NS, NS/TA, and ROA appear in the middle range, while variables such as DIV Y, CA/CL, TD/TA, and ROE are slightly below them. P/E and IND_4 have the lowest coverage values among all features.

The relatively high cover values of several industry indicators (IND_3, IND_6, IND_2, and IND_5) suggest that, when these variables are used for splits, they apply to broad subsets of the sample rather than narrow partitions. This pattern indicates that industry classification captures general structural differences across firms, affecting large groups of observations in a single split.

SIZE and YEAR also display moderate cover, implying that their splits influence a substantial share of firms while still contributing to more refined partitioning than industry-level variables. In combination with their high gain values, this suggests that these variables

not only affect many observations but also meaningfully improve model accuracy when used.

The remaining financial ratios show lower coverage, which is typical for continuous predictors that the model applies selectively across specific regions of the data. These features are more likely to refine predictions within industries and time periods, rather than driving broad segmentation across the sample.

5.2.2.3 Feature Importances by Gain

The figure ranks variables by their average contribution to reducing model error when used for data splits. Higher values indicate features that yield greater improvement in predictive accuracy on average. For example, for SIZE, importance score of approximately 696 means that on average, every split on SIZE reduced the training objective (sum of squared-error loss) by about 696 units. The unit corresponds to the squared-error loss used in the objective function (`reg:squarederror`), meaning Gain quantifies the average reduction in total training MSE.

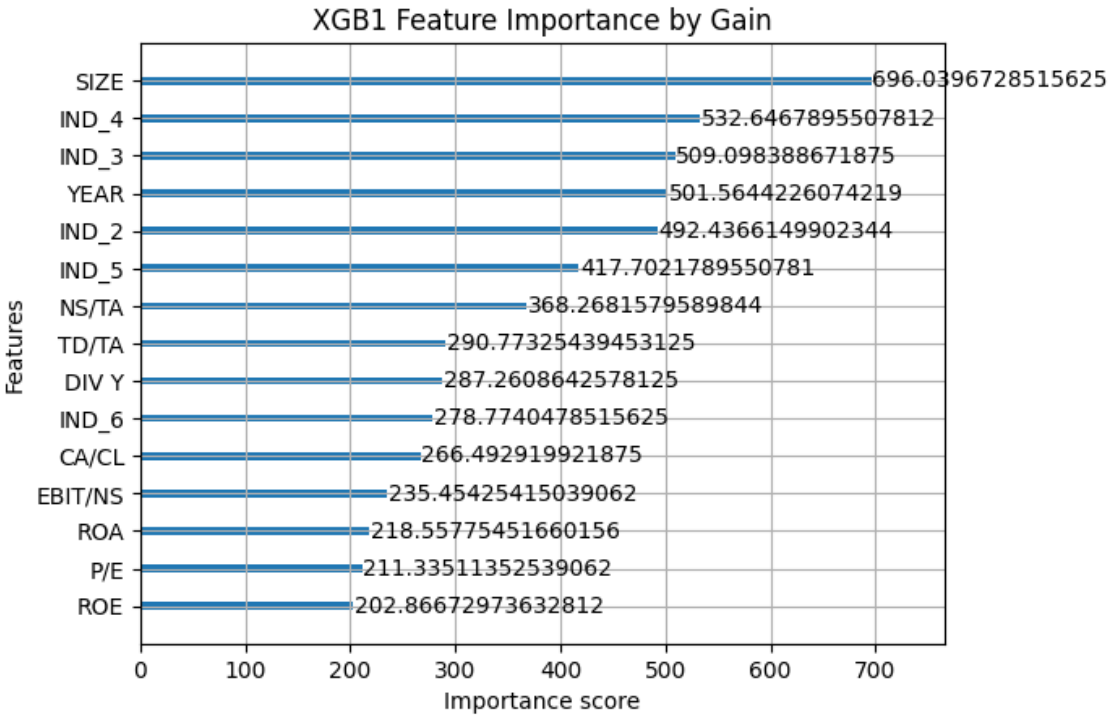

Figure 9. Feature Importance by Gain (XGB1).

SIZE has the highest gain score, followed by IND_4, IND_3, YEAR, and IND_2. IND_5 and NS/TA are next with moderately lower scores. TD/TA, DIV Y, IND_6, CA/CL, and EBIT/NS follow with slightly smaller contributions, with ROA, P/E and ROE having the lowest gain values among the features.

The high gain of SIZE indicates that splits based on firm size tend to produce the largest reductions in prediction error. This suggests that SIZE not only appears frequently in tree construction but also provides the most effective partitioning of ESG outcomes when used.

The relatively high gain values of several industry dummies (IND_4, IND_3, IND_2, IND_5) contrast with their low frequency in the weight metric. This implies that while these industry variables are used less often, the splits they generate contribute meaningful local improvements in prediction accuracy. When industry effects are included, they are likely

to capture broader sector-level differences that the financial statement items do not explain alone.
YEAR also has a relatively high gain despite not being used frequently. It might suggest that temporal variation contributes meaningful information to the model. This may reflect the general upward trend in ESG scores over time, indicating that the inclusion of YEAR helps the model account for systematic changes in sustainability reporting or performance across the sample period.

Profitability ratios (ROA, ROE) and valuation measures (P/E) yield low gain values, suggesting that their inclusion offers limited incremental improvement once the larger effects of SIZE, leverage, and industry segmentation are accounted for.

5.2.2.4 Cross-metric feature comparison

Together weight, cover and gain describe which variables the model relies on, and how and to what extent they form predictions.

SIZE consistently ranks among the top variables, appearing most frequently in tree construction (weight) generating the largest average reduction in prediction error (gain) and affects a relatively large number of observations when it is used for splitting (cover). This pattern indicates that firm size is the model's primary organizing variable, defining broad partitions in ESG scores and improving predictive accuracy when included. It is likely to reflect relationships between resources to invest in sustainable practices and more structured sustainability reporting practices.

P/E shows high weight but low gain and cover. This means that the variable is frequently used in the tree structure but does not contribute much to improving prediction accuracy or affecting large parts of the dataset. It may function as an interaction variable that is useful in certain contexts or in combination with other financial statement items but alone contributes limited predictive power.

Both TD/TA and NS/TA locate around middle positions across the metrics. They are used rather regularly and provide moderate improvements when included, suggesting that financial structure and operational efficiency add consistent but secondary information, and help in refining predictions rather than defining major data partitions, after considering the effect for firm size.

YEAR has relatively low frequency but high gain and moderate cover, implying that temporal information meaningfully improves model accuracy when used, reflection systematic upwards trend in ESG scores over the sample, functioning as a contextual adjustment feature within the model.

The industry indicators show low values in weight, but rather high values in gain and cover, meaning that the model rarely uses industry variables to form splits, but when they are used, those splits affect a large portion of the sample and provide accuracy improvements, therefore capturing sector-level differences in ESG performance, complementing firm-level predictors that explain variation within industry.

EBIT/NS, ROA and ROE have moderate to low weight, gain and cover. Such weaker presence across the metrics suggests that while profitability is economically relevant, it does not contribute substantial new information for ESG score prediction after other aspects like size, leverage and efficiency are included.

Liquidity (CA/CL) and dividend yield (DIV Y) are located lower to middle ranks across metrics, indicating supporting roles for improving predictions but not acting as main features. Their effects may be more context specific, such as influencing predictions within certain financial profiles or industries.

5.3 SHAP Metrics

To complement the intrinsic importance measures of XGBoost, this study applies SHAP (SHapley Additive exPlanations) values to quantify and interpret the contribution of each financial variable to the predicted ESG scores. SHAP is based on Shapley values from cooperative game theory, which allocates the model's total prediction among the input variables in a way that satisfies fairness and additivity principles (Lundberg et al., 2019). Each feature thus receives a numerical value representing its individual contribution to increasing or decreasing the prediction for a given observation (Lundberg et al., 2019).

SHAP provides a local explanation that can be aggregated to measure global importances by incorporating SAGE (Shapley Additive Global Importance) framework (Ranta, 2023, Ch. 12), to explore, for example, feature importances or dependence plots. Rather than quantifying how features contribute to a specific prediction, SAGE estimates how each feature contributes to reducing the model's loss function (Ranta, 2023, Ch. 12). Because the approach is grounded in Shapley values, it inherently accounts for complex nonlinearities and interaction effects between features (Ranta, 2023, Ch. 12). This approach allows for each variable's influence in the model both individually, and for example, in presence of correlated financial indicators.

Similarly to prior feature importances, SHAP values presented here were computed on the training set to interpret the relationships that the model learned during the fitting process, since the focus is on understanding how financial statement variables contribute to the model's decision structure. Using the training data ensures that the interpretation reflects the complete learning signal available to the model.

5.3.1 Directional SHAP Feature Importance Metrics

Directional SHAP feature importance metrics are derived from aggregating SHAP values across all observations. The directional feature importance bar indicates both magnitude and direction of features effect on the predicted ESG Score. This differs from

permutation or intrinsic importances, which only measure how much a variable affects prediction accuracy without describing whether its effect is positive or negative. Below is the directional SHAP feature importance metric bar for the model XGB1.

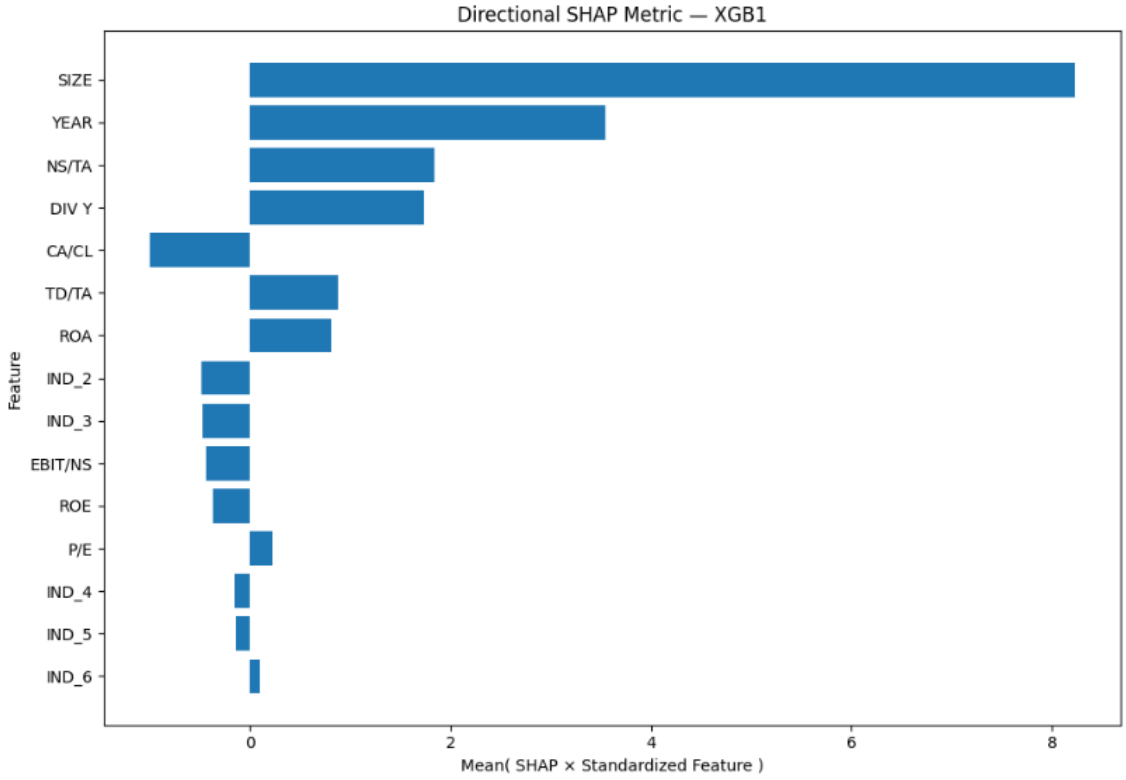

Figure 10. Directional SHAP feature importance Metric Bar (XGB1).

The figure presents the mean of SHAP values multiplied by standardized feature values for the XGB1 model. Positive bars indicate that higher values of a variable tend to increase predicted ESG scores, while negative bars indicate the opposite.

SIZE has the largest positive contribution, followed by YEAR, NS/TA, and DIV Y. Smaller positive effects appear for TD/TA, ROA, P/E, and IND_6. The most clearly negative value is observed for CA/CL, while smaller negative contributions are visible for IND_2, IND_3, EBIT/NS, ROE, IND_4, and IND_5.

The dominant positive direction of SIZE indicates that larger firms are systematically predicted to have higher ESG scores. This aligns with the notion that firm scale captures both greater visibility and greater capacity for sustainability reporting or investment.

The positive effect of YEAR restates the earlier observation on how ESG scores contain an upwards trend during the sample period.

The positive effects of NS/TA and DIV Y imply that firms demonstrating higher operational efficiency and consistent dividend payouts tend to receive slightly higher ESG predictions. These characteristics may act as proxies for financial stability or organizational maturity.

CA/CL shows a clear negative direction, indicating that higher liquidity ratios correspond to lower predicted ESG scores. Similarly, the modest negative effects observed for EBIT/NS and ROE suggest that marginal profitability does not drive ESG performance within this dataset.

Industry effects remain limited but display mixed directions. IND_6 (Real Estate and Investments) contributes weakly positively, whereas IND_2 to IND_5 show slightly negative contributions. This pattern suggests that once financial characteristics are accounted for, sectoral differences exert only marginal influence on predicted ESG scores.

Overall, the directional SHAP metric indicates that the XGB1 model primarily links higher ESG predictions to firm scale, and indicators of financial stability. Conversely, high liquidity and certain profitability or industry characteristics are associated with lower predicted ESG scores. The results suggest that the model emphasizes long-term structural and size-related determinants of ESG performance rather than short-term financial outcomes.

5.3.2 SHAP Summary Dot Plot

SHAP summary dot plot visualizes every observation's SHAP value for each feature, sorted by overall importance as mean absolute SHAP value. The x-axis shows each feature's SHAP value (how much the variable pushes a single prediction up or down). The color gradient indicates whether the features value is high or low, red representing high, and blue representing low. The plot shows both direction and dispersion, highlighting nonlinearities, asymmetry and overlap between low and high feature values. Basically, each dot in the SHAP summary plot represents one individual firm-year observation datapoint in the training set, and how much the feature pushes the model's predicted ESG score up or down, relative to the model's average prediction.

Where the directional SHAP feature importance bar answers the question "On average, to which direction and by how much does a variable push the ESG score prediction?", the SHAP summary dot plot answers the question "How does each variable affect individual predictions, and how consistent is that effect across the dataset?". Basically, the directional SHAP metric establishes the overall direction and average strength of a features effect, and the SHAP summary plot provides more detailed, observation-level results behind those averages.

Below is the figure for SHAP feature importance summary dot plot.

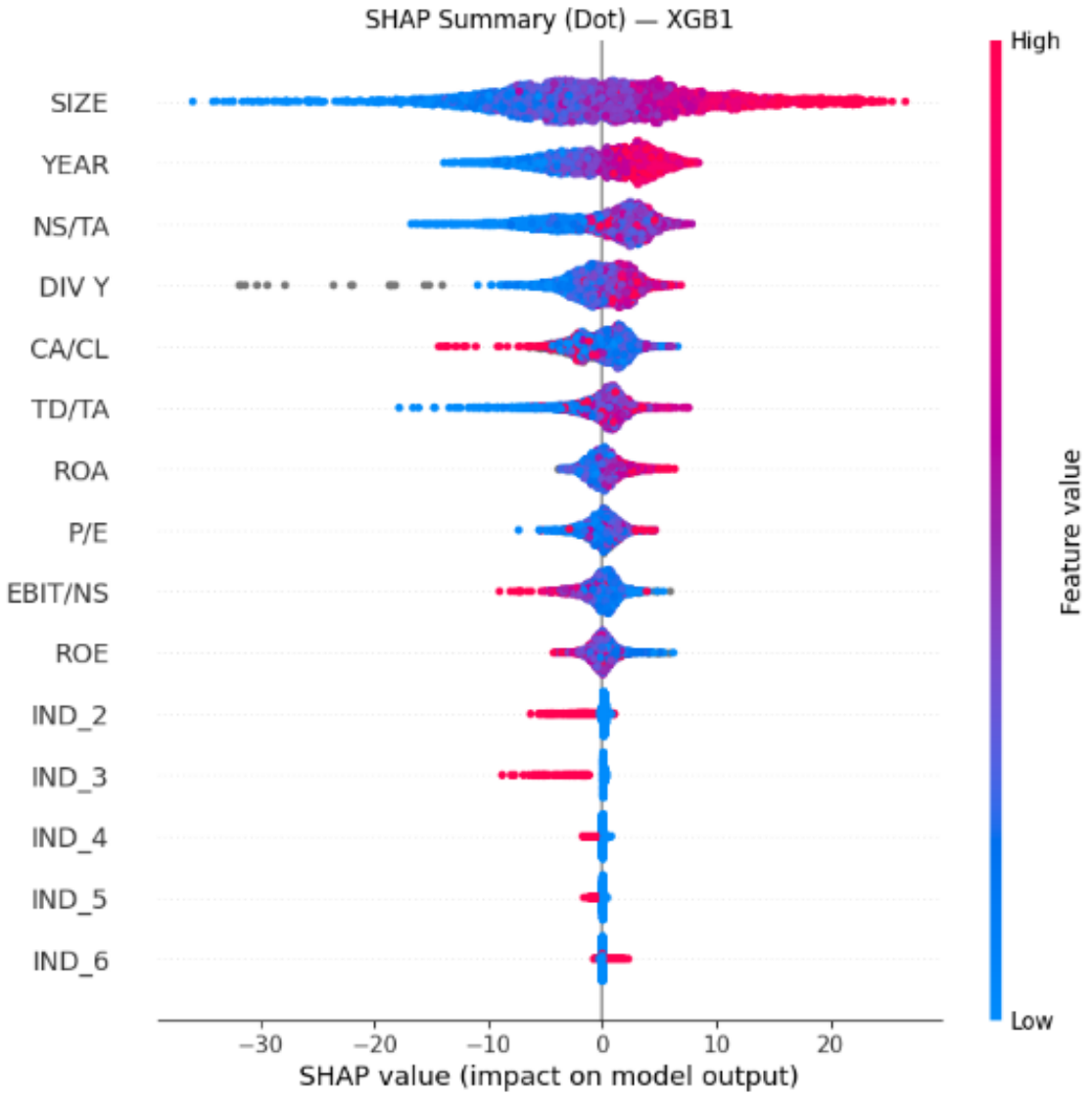

Figure 11. SHAP Feature Importance Summary Dot (XGB1).

For SIZE, high value (red) observations locate on the right (positive SHAP value), while low-value observations (blue) locate on the left. The results on the figure indicate that similarly to earlier findings, SIZE is the most dominant predictor, and while larger firms consistently push ESG predictions upward, firms with smaller size push prediction downward, showing positive link between firm size and ESG ratings.

For NS/TA, the pattern shows that higher values generally have positive SHAP value, while lower ratios shift predictions downward, suggesting that asset efficiency contributes positively within the model.
DIV Y shows values clustering near zero, but higher values mostly on the positive side.

CA/CL shows skew where higher values make up the most negative SHAP value impacts, while lower feature values appear slightly on the right side.

TD/TA shows clear color gradient, where lower feature values fall on the negative side, and higher feature values lean towards positive, indicating that higher leverage was linked with slightly higher ESG scores and lower leverage with lower ESG scores.

Rest of the features showed rather modest effects. ROA also has monotonic color gradient indicating that higher ROA increased ESG score prediction slightly. P/E has a small and rather neutral spread besides both tails. EBIT/NS has a rather compact distribution where higher values are slightly on the negative side, while lower values appear slightly positive. ROE is centred around zero with no clear colour gradient besides tails. Industry dummies have small spreads around zero with some visible baseline shifts by industry, but small in magnitude.

Overall, the figure shows rather monotonic or strong gradients for financial statement items of SIZE, NS/TA, DIV Y and ROA, and a bit more mixed or weak gradients for TD/TA, CA/CL, P/E, EBIT/NS, ROE and industry variables. The tails of the features consistently show monotonic colour gradients. The colour gradient within the tails clarifies the direction of relationships the model learned during training. The length of each tail indicates the maximum effect that feature reached during model fitting. Longer tails mean the model assigned larger positive or negative contributions for some training observations. Even for variables that did not have large impacts on the model, their tails indicate the direction where extreme values of those variables are pushing a prediction.

5.4 Cross-validated Partial Dependence Plots

Partial dependence plots (PDPs) are technique for global model interpretation introduced by Friedman (2001). A PDP visualizes the average relationship between explanatory and target variables, showing the marginal effect of features on the model prediction. “Specifically, PDP is calculated by changing the level of one or two explanatory variables of interest for the entire dataset, while keeping other variables constant, and then taking the arithmetic mean of the model predictions derived for each level of the variables of interest” (Hashimoto et al., 2023).

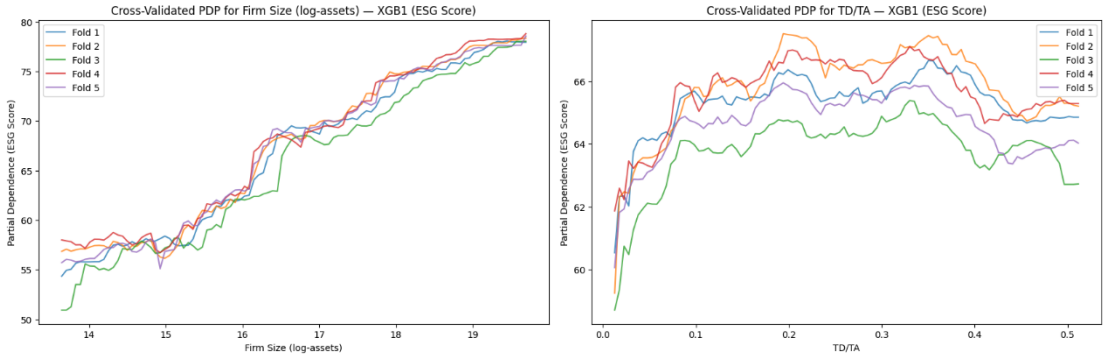

Figure 12. XGB1 Cross-Validated PDPs for SIZE and TD/TA against ESG Score.

The cross-validated PDP illustrates how the model’s predicted ESG score changes with explanatory feature across the five cross-validation folds. Each line represents one fitted model trained on a different training subset and evaluated on its corresponding validation fold.

The relationship between firm size and ESG score is strongly positive and rather monotonic. As firm size increases, the model consistently predicts higher ESG scores. The close alignment of across folds indicates high stability and reproducibility of this relationship across data splits. There is minor variation in the lower range (below log-assets of 15) suggesting slightly more uncertainty among smaller firms, but overall, the trend is monotonic.

Overall, the pattern shows that during cross-validated estimation, the model repeatedly learned a systematic positive association between firm size and ESG performance, implying that firm scale is a consistently influential factor in ESG score prediction.

PDP for TD/TA shows that the marginal effect of total debt to total assets on the ESG score. The relationship shows a slightly nonlinear and concave pattern as the ESG score initially increases as TD/TA rises from zero to approximately 0.2 – 0.3, after which the curve slightly declines. This shape indicates that the model learned an optimal mid-range leverage level associated with the highest predicted ESG scores.

At very low leverage (TD/TA < 0.05), ESG scores are notably lower, suggesting that firms with minimal debt may correspond to less mature or smaller entities that the model associates with weaker ESG performance. As leverage increases moderately, ESG predictions rise, possibly reflecting that moderately leveraged firms tend to be larger, better established, or more capital-efficient, which are characteristics that were linked to higher ESG scores in the training data. Beyond the turning point (≈0.3-0.35), the predicted ESG score levels off and even slightly declines, implying that excessive leverage may start to link with slightly lower ESG performance.

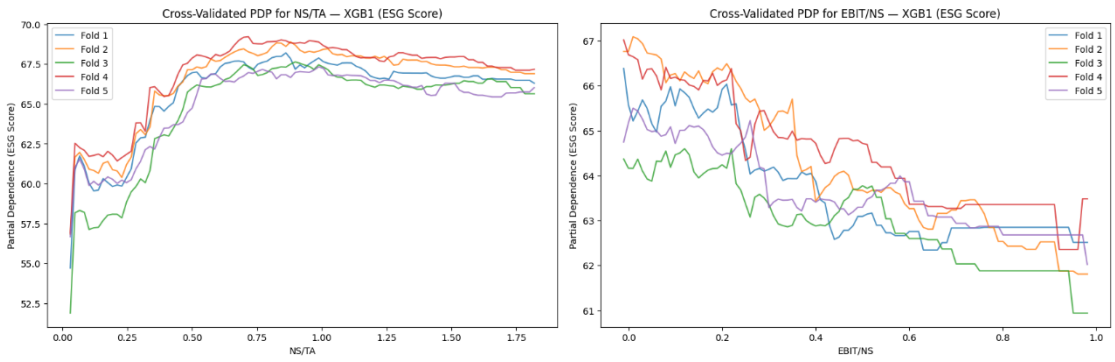

Figure 13. XGB1 Cross-Validated PDPs for NS/TA and EBIT/NS against ESG Score.

For NS/TA, the plot shows nonlinear and concave relationship with ESG score. The plot indicates that firms which generate very low net sales against their asset base are predicted to have noticeably lower ESG scores in the model. Firms with higher NS/TA are

associated with higher ESG scores, but after approximately 0.75 point, the predicted ESG scores decrease slightly, suggesting that after a certain level of asset utilization, additional increases in NS/TA provide diminishing or slightly negative marginal effects on the predicted ESG score.

For EBIT/NS, the plot shows negative relationship with ESG score. The plot indicates that firms with higher profitability relative to net sales tends to be associated with lower predicted ESG score in the model.

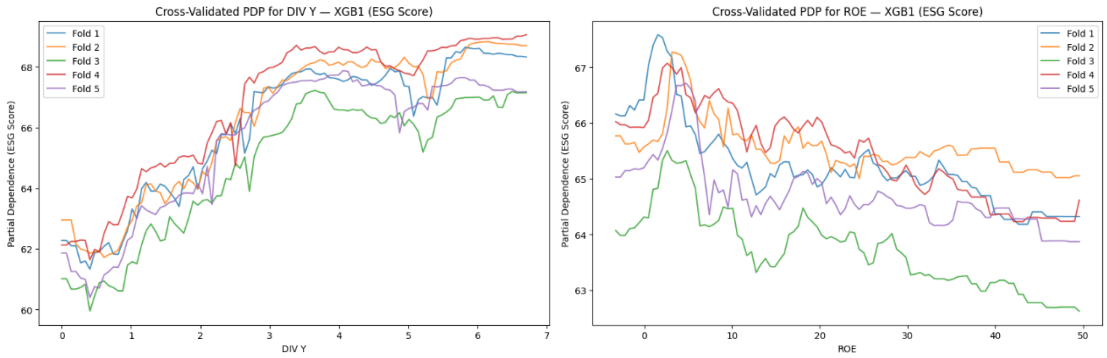

Figure 14. XGB1 Cross-Validated PDPs for DIV Y and ROE against ESG Score.

For DIV Y, the plot shows positive relationship between dividend yield and ESG score. For very low or zero dividend yields, the predicted ESG scores are lowest. As dividend yield increases toward approximately 2-3%, ESG predictions rise sharply, and then gradually level off between 4% and 7%, stabilizing around 67-69 ESG points. This indicates that within the training data, the model systematically associates higher dividend yields with higher predicted ESG scores, though the marginal effect diminishes at the upper end of the range.

For ROE, the plot shows nonlinear and slightly inverse relationship with ESG score. In the lowest ROE range (0-10), ESG predictions are highest. Beyond this range, as ROE rises, the predicted ESG score starts to decline by 2-4 points across the folds. The fold-specific curves follow the same general pattern but with slight variation in magnitude, reflecting moderate cross-validation stability.

All folds capture a peak at low-to-moderate ROE, followed by a downward slope, suggesting that the model repeatedly learns a concave relationship between profitability and ESG prediction across the training subsets. Because the PDP is computed on the training folds, it reflects the average fitted dependence within the observed ROE range. The overall shape implies that the model associates very high equity returns with somewhat lower fitted ESG scores, which can reflect the model's learned tendency to link extreme profitability with less sustainable firm profiles within the training data.

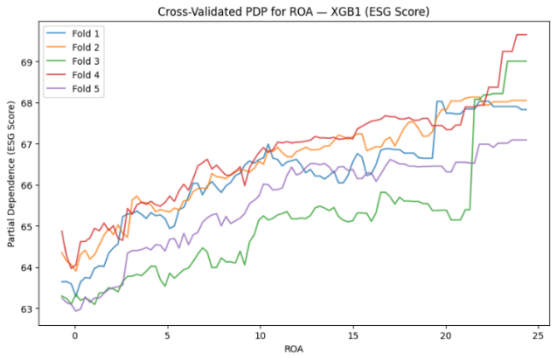

Figure 15. XGB1 Cross-Validated PDPs for ROA against ESG Score.

For ROA, the plot shows positive relationship with ESG score, indicating that firms with higher returns on assets tend to be associated with higher predicted ESG scores, with ESG score rising by approximately 4 points across the folds when increasing ROA from 0 to 20.

Overall, the cross-validated partial dependence plots support and clarify the relationships identified in the SHAP analyses and feature importance results. The PDPs indicate that several financial variables have systematic effects on the model's fitted ESG predictions. Across interpretability methods, firm size shows stable, monotonic positive associating with predicted ESG score, indicating that the model systematically links larger firm scale with higher ESG performance. Profitability measures like ROA similarly display positive relationship, while EBIT/NS and ROE reveal inverse or concave patterns, suggesting that higher profit margins or returns are linked with lower ESG scores. The NS/TA

follows a concave form with the highest predicted ESG levels at moderate values, and TD/TA exhibits a slightly similar nonlinear effect with a turning point at medium leverage levels. The overall cross-validation stability of these relationships implies that the model's learned dependencies are robust across data splits. Taken together, the PDP results complement the SHAP and feature importance analyses by providing a clear visualization of how the model interprets firm-level financial statement items in relation to ESG scores, highlighting that the model's fitted relationships are both statistically stable and interpretable.

5.5 SHAP Dependence plots and 2D partial dependence

The preceding tools have mainly described marginal effects of the predictors. Tree-based models can also learn non-additive relationships, where the effect of one variable on the prediction can depend on the level of another, which is explored in this chapter using SHAP dependence plots and two-dimensional partial dependence surfaces.

For investigating interactions between two features, PDPs can be extended to two dimensions. A two-dimensional partial dependence plot visualizes the joint effect of two features on the predicted outcome (Hashimoto, 2023). The 2D PDPs are displayed on coloured surface, where the axes represent the values of the two features, and the colour represents the model's prediction.

Here the SHAP dependence plots and 2D PDPs are highlighted for the five highest interacting pairs. The interaction pairs shown were chosen by computing a pairwise SHAP interaction values (TreeSHAP) for the financial statement items and the pairs were ranked by their mean absolute interaction strength across observations. This section continues with analysis of the figures starting from the highest ranked interaction pair.

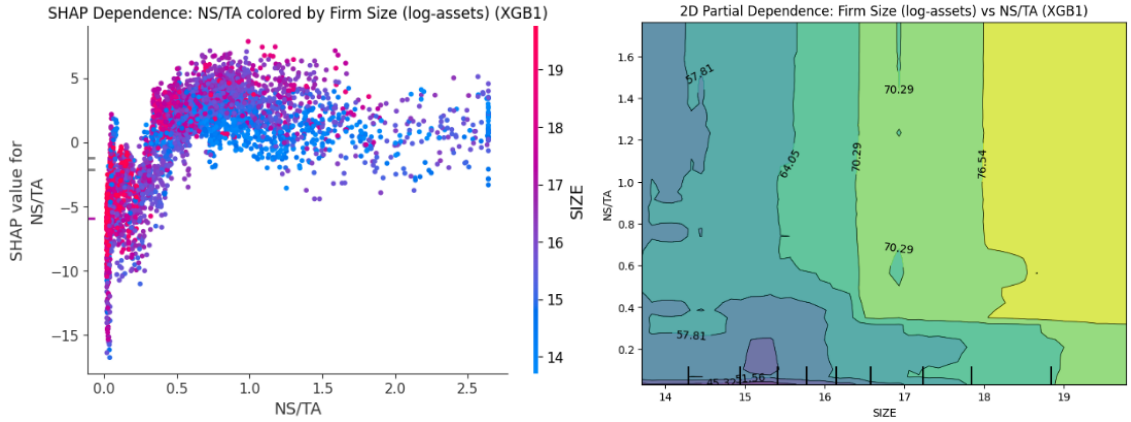

Figure 16. SHAP Plot and 2D Dependence for NS/TA and SIZE.

The interaction between firm size and asset turnover provides further insight into the joint structure of the model's predictions. The SHAP dependence plot indicates a nonlinear relationship between NS/TA and its SHAP values, that increases from strongly negative values at near-zero to positive values at the mid-range, after which the effect becomes more dispersed. The color gradient, which represents firm size, shows that larger firms (depicted in red and pink) generally correspond to higher SHAP values at a given level of NS/TA, implying that greater positive influence is attributed to sales efficiency when it occurs withing larger firms.

The corresponding two-dimensional partial dependence plot supports this interpretation. The contour levels are predominantly vertical, confirming that firm size is the dominant driver of predicted ESG scores within this interaction pair. However, the contours also show a low-value region at very small NS/TA ratios, where ESG predictions decline noticeably, followed by a broad zone of relatively stable or slightly increasing partial dependence as NS/TA rises to moderate levels.

Taken together, these two visualizations suggest that the model captures a complementary relationship between firm scale and asset utilization: higher asset efficiency enhances ESG scores most strongly for larger firms, while very low turnover ratios are associated with reduced predictions regardless of size.

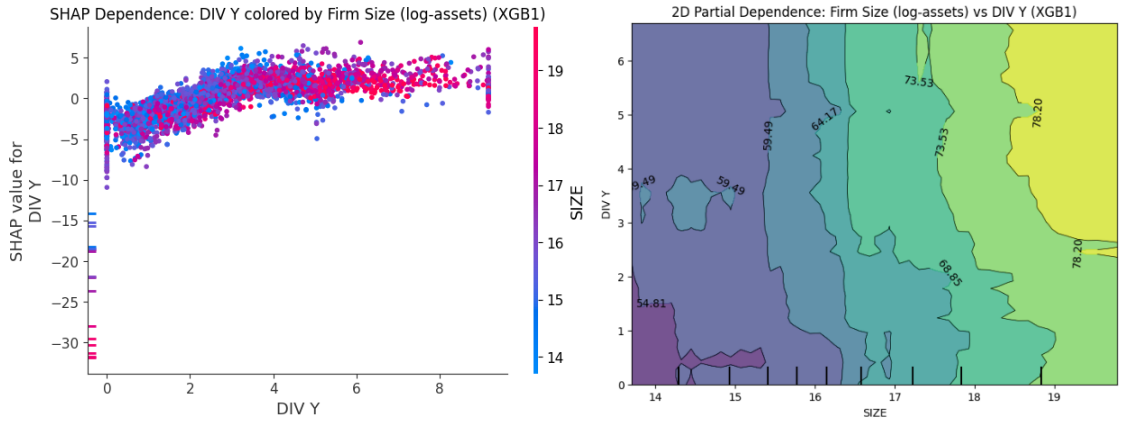

Figure 17. SHAP Plot and 2D Dependence for DIV Y and SIZE.

The interaction between dividend yield and firm size shows how payout policies and firm scale jointly influence ESG predictions in the model. As dividend yield increases from zero, SHAP values show a rising trend, but imply diminishing marginal influence at higher yields. The colour gradient reveals that larger firms tend to occupy the upper section of the SHAP distribution, especially at moderate to high dividend yields. The pattern suggests that the negative effect of low or missing dividends is more prominent for smaller firms, while larger firms experience a milder reduction in ESG predictions even at low dividend levels.

The 2D PDP shows rather vertical contours, highlighting size as dominant influence, and indicating that ESG scores remain consistently higher for large firms across the dividend yield range, reflecting weaker marginal sensitivity to payout level. However, the plot also shows a distinct gradient along the dividend yield axis for smaller firms. Dividend yield positively contributes to ESG predictions primarily for smaller firms, where the absence of dividends is associated with lower scores. In contrast, larger firms achieve high ESG predictions regardless of their dividend yield.

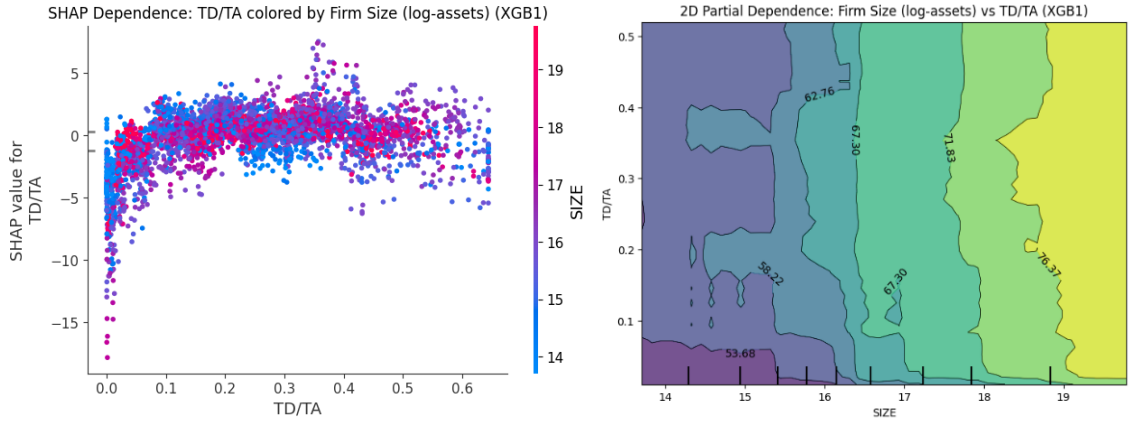

Figure 18. SHAP Plot and 2D Dependence for TD/TA and SIZE.

The TD/TA – SIZE interaction suggests that very low leverage is associated with reduced ESG scores, and that moderate leverage most favored, but these effects are amplified for smaller firms. For larger firms ESG score remains primarily determined by size with TD/TA having very subtle, concave role within the prediction.

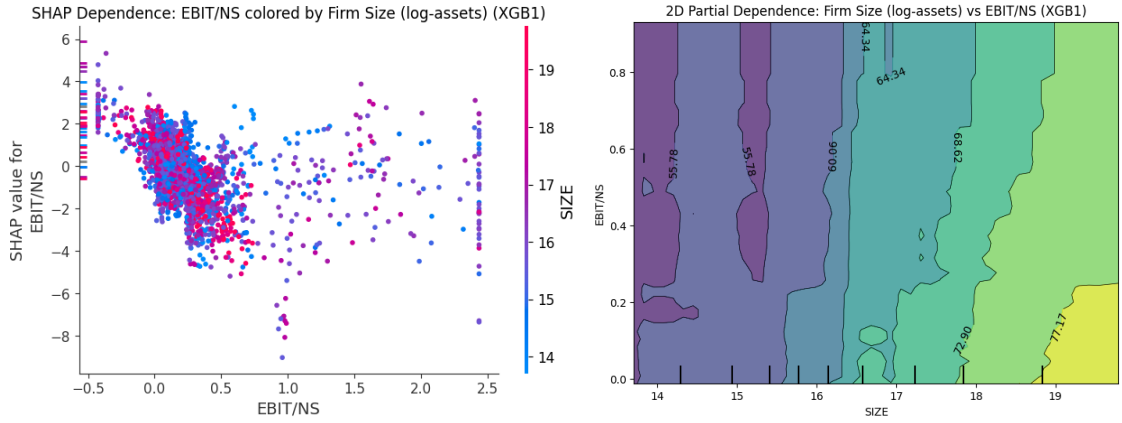

Figure 19. SHAP Plot and 2D Dependence for EBIT/NS and SIZE.

The EBIT/NS – SIZE interactions suggests that while firm size remains as the primary predictor for ESG score, profitability margin exerts diminishing or weakly negative effect on ESG score within each size category. Although the influence is very small compared to size itself, the slight downturn of the contours at higher EBIT/NS ratios implies that the model associates very high profit margins with lower ESG scores, which could imply an

association between aggressive cost efficiency or profit orientation and lower investment in sustainability practices.

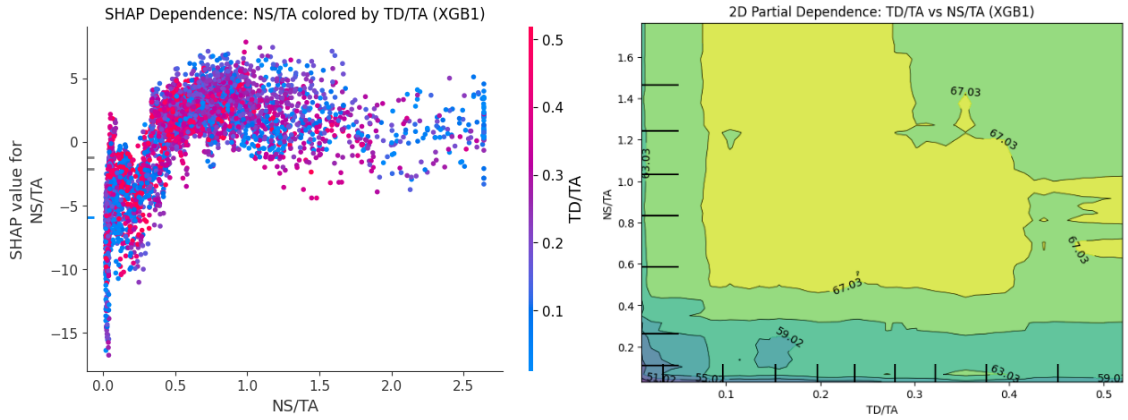

Figure 20. SHAP Plot and 2D Dependence for NS/TA and TD/TA.

The interaction between NS/TA and TD/TA suggests complementary relationship between operating efficiency and leverage. Firms exhibiting both moderate to high asset turnover and moderate leverage levels receive most favorable ESG scores in the model, while those with low values of both ratios tend to be assigned with lower predictions. The pattern implies that combination of financial activity and capital intensity is associated with higher ESG performance.

Overall, the predominantly vertical surfaces of the 2D PDPs indicate that firm size sets the level of ESG score prediction, while the partner variables contribute modest, value-dependent adjustments. NS/TA shows concave profile with penalty for very low turnover and diminishing gains beyond mid-range; low dividend yield also shows a small penalty for varying firm sizes; and TD/TA improves prediction when moving from very low to moderate leverage. For EBIT/NS conditional effects are small but rather consistent and tend to be flat to slightly negative as margins increase, most visibly for smaller sized firms. Outside of firm sizes influence, highest interaction was among NS/TA and TD/TA, where predicted ESG scores were the lowest when both ratios were low, rising sharply once NS/TA increases to moderate range.

5.6 Further Analysis of Negative Directions

The negative SHAP directions observed for CA/CL, EBIT/NS, and ROE suggest that higher liquidity and profitability ratios are associated with lower predicted ESG scores in the model. While this relationship might appear counterintuitive, some explanations can be theorised for why these variables exhibit a negative rather than positive predictive direction.

One plausible interpretation relates to sectoral and structural composition. Industries that tend to maintain high liquidity or profitability, such as energy or financial sectors, can face greater environmental externalities or weaker ESG disclosure practices, at least in the past. Conversely, sectors characterized by lower liquidity ratios, such as technology or utilities, might have demonstrated more sustainability governance and reporting practices. In this sense, the negative contribution of CA/CL and profitability indicators may partly reflect sector-linked financial structures rather than firm-specific ESG behaviours.

A second explanation concerns the financial conservatism hypothesis. A high current ratio can indicate excess liquidity or underutilized capital, implying managerial caution or a reluctance to invest in long-term projects. Firms with conservative balance-sheet management may be less likely to allocate resources toward sustainability initiatives or innovation, resulting in a weaker ESG orientation. Similarly, the negative directions of EBIT/NS and ROE may capture the tension between short-term profitability and long-term ESG orientation. High profitability can sometimes result from cost-cutting strategies, resource extraction efficiency, or reduced expenditure on environmental and social programs. Firms emphasizing immediate financial performance may therefore priorize less ESG activities, leading to an inverse association between profitability and ESG predictions. This interpretation aligns with some literature suggesting that sustainability initiatives often require upfront investment that can temporarily suppress profitability indicators.

Another perspective links these results to the firm lifecycle hypothesis. Firms in earlier growth stages often display high profitability ratios due to expansion momentum yet may lack established ESG policies or reporting frameworks. In contrast, mature firms with moderate profitability may have more formalized governance and sustainability structures. Under this interpretation, the negative SHAP direction of profitability variables reflects differences in organizational maturity, not necessarily a causal link between profitability and ESG performance.

It is also possible that these patterns reflect ESG disclosure dynamics. Highly profitable or liquid firms may perceive less reputational or regulatory pressure to strengthen ESG transparency, relying instead on their financial reputation. Meanwhile, less profitable firms may use ESG engagement as a strategic signaling mechanism to attract investors or improve legitimacy. This compensatory behavior can produce a negative empirical relationship between profitability and ESG outcomes.

Finally, model-specific considerations should be acknowledged. Tree-based models capture complex feature interactions and conditional dependencies. The negative SHAP values for profitability and liquidity may therefore emerge in cases where these variables overlap in predictive information with others, such as firm size or leverage. The resulting direction reflects interaction effects rather than a purely economic mechanism.

A related distinction can also be made between EBIT and EBIT/NS, as these variables represent different economic dimensions. While EBIT in absolute terms captures firm scale and total operating income, EBIT/NS measures relative efficiency or operating margin. Large firms typically exhibit high total EBIT values even with moderate margins, reflecting scale advantages and greater capacity to engage in sustainability initiatives. By contrast, small or niche firms may display high EBIT/NS ratios but lack comparable ESG structures or disclosure frameworks. Consequently, EBIT as a scale variable would likely contribute positively to ESG predictions, while EBIT/NS displayed a negative direction due to its potential association with smaller, high-margin firms. This distinction

underscores that the model interprets scale-related profitability as an indicator of sustainability capacity, whereas efficiency-based profitability may signal limited ESG integration.

In summary, the negative contributions of CA/CL, EBIT/NS, and ROE can stem from a combination of sectoral structure, managerial conservatism, firm lifecycle dynamics, and interaction effects within the model. The results suggest that high liquidity and profitability, while conventionally seen as indicators of financial strength, may not correspond to proactive ESG engagement. Instead, the model associates stronger ESG predictions with firms that potentially balance financial stability with long-term strategic investment capacity.

6 Discussion

This thesis started with motivation towards investigating to what degree can basic financial information explain ESG ratings - influential synopses of firm sustainability, yet partly "black boxes" and non-standardized, - and what financial factors are the most influential using machine learning models and explainable AI tools for model interpretation. This discussion interprets what the best performing model learned about ESG score predictability from financial information and compares the main findings thematically with prior research, acknowledges this research's limits and provides suggestions for future research.

6.1 Findings and Previous Literature

The results obtained showed greater performance of nonlinear models over regularized linear models, which is consistent with the findings of D'Amato et al. (2022) and Del Vitto et al. (2023), and compatible with the assumption that linear models are useful baselines to compare the performances of nonlinear models to.

The results in this thesis showed that the model explained rather significant portion of the variation in ESG scores but did not manage to achieve as high R^2 scores as other studies referred to, such as Lin & Hsu (2023), but was rather close in predictive power with, for example, D'Amato et al. (2021).

In the model, firm size was identified as the most influential predictor of ESG scores, with a clear positive association – larger firms consistently have higher predicted ESG score. This aligns with findings of D'Amato et al. (2021), who found company size to be a key determinant of ESG ratings, and with Chowdhury et al. (2023), where firm size was the second most important variable in predicting a firm's ESG rating. The consistent importance of size suggests that bigger companies, likely due to greater resources and stakeholder scrutiny, tend to achieve better ESG performance in both the literature and the findings of this thesis. The papers by D'Amato et al. will be the most compared

sources in this discussion, as their chosen features, dataset and findings are the most similar to ones in this thesis, out of the papers introduced.

In the findings of Chowdhury et al. (2023), the most important variable was prior ESG rating, which was also the most important variable in this thesis' preliminary variable testing, but in this thesis, the variable was dropped from the modelling, because the aim of the thesis was to investigate the impacts of financial items on the prediction, and the influence of prior ESG score was rather obvious dominant variable for the prediction, and was thought to distract value from potential less obvious variables with impact. Including this variable would have, ceteris paribus, resulted in an R^2 of over 0.9 in the model.

This thesis found a nuanced relationship between profitability, and marginal profitability and ESG. Return on Assets (ROA) displayed a positive relationship with ESG scores, indicating that moderately higher profitability supports better ESG outcomes. However, EBIT/Net Sales (operating margin) and Return on Equity (ROE) showed an inverse or concave effect – beyond a certain point, very high profit margins corresponded to slightly lower ESG scores. D'Amato et al. (2021) likewise identified profitability as a significant predictor of ESG score. In their paper, Net Income to Sales ratio was highlighted as the most important financial variable, and they observed increase in NI/S to cause a reduction in the ESG score. The findings of this thesis and their paper could align with the "trade-off" hypothesis, where implementation of sustainable practices costs resources or can be more expensive than less sustainable practices, and firms with extremely high profitability might engage less with sustainable practices cutting their operational costs. Prior research by Lin et al. (2019) observed a negative link between extreme financial performance (high ROE, ROA, ROI) and corporate social responsibility, supporting the idea that aggressively maximizing profits can come at the expense of sustainability efforts. Both the prior literature and this thesis find a link between profitability and ESG score, with a detailed difference of this thesis finding that moderate profitability appears beneficial for ESG, whereas overly high profitability margins can be negatively linked.

The model indicated a non-linear pattern considering capital structure: firms with very low debt-to-assets ratio were linked with lower ESG score, while those with moderate leverage achieved highest ESG scores. Chowdhury et al. (2023) found a firm's debt-to-equity ratio to be the third most influential contributor to ESG rating predictions. D'Amato et al. (2021) - rather similarly to this thesis' findings, - found an initial increase in leverage (from 0 to 0.35) to improve ESG score, but when TD/TA gets past around 50% mark, the ESG score worsens.

Operational efficiency showed concave relationship with ESG score in this thesis' findings. Very low asset turnover was linked to lower ESG prediction, increasing to a moderate turnover led to highest ESG scores, and beyond middle range, the marginal ESG benefit levelled off. The core literature discussed in this thesis did not explicitly discuss asset turnover as an ESG driver. For example, studies like D'Amato et al. (2021) and Del Vitto et al. (2023) focused on broader financial ratios and did not single out efficiency metrics. The findings of this thesis imply that how effectively firms utilize their assets can influence ESG score. One possible explanation is that moderate asset efficiency reflects sound management and balanced operations that support sustainability efforts, whereas extremely high turnover might be characteristic resource-intensive business models that do not prioritize ESG.

The model assigned small secondary role to liquidity, measured in CA/CL, and found an inverse relationship with ESG score. The negative tail of SHAP values and importance metrics hinted at a penalty in ESG score prediction for very high liquidity ratios. In the literature D'Amato et al. (2021) noted negative correlation between ESG score and liquidity ratio but didn't find it statistically significant in their linear regression and didn't further discuss its effect. In this thesis' model, its effect was slightly more impactful for the prediction than, for example, TD/TA, which was second highest feature in variable importance in their paper.

In this thesis, industry sector membership had real but limited effect on ESG score prediction. The industry indicators had low frequency, but higher gain and cover values, meaning that distinguishing the firm sector could explain residual ESG variance. Del Vitto et al. (2023) explicitly accounted for sector groupings in their ESG prediction study, finding that companies in certain sectors have inherently varying levels of ESG profiles. This thesis findings implied a similar pattern where the model rarely uses industry variables to form splits, but they affected large portions of the sample and provide accuracy improvements, therefore capturing sector-level differences in ESG performance, complementing firm-level predictors that explain variation within industry. This suggests that investors or researchers should control for sector in decision making if they are using ESG scores in the decision-making process.

Additional robustness testing by creating a model without total assets or firm size as a variable, with ceteris paribus methodology as XGB1, resulted in R^2 of 0.4135). As that model is more focused on financial ratio predictors and ignores the most dominant predictor (firm size), the results highlight more clearly the impacts of financial ratios and industry sector memberships, especially the findings similar results with Del Vitto et al. (2023). However, the main model chosen for analysis in this thesis (XGB1) which included firm size as a predictor was still chosen, because even though this robustness testing model shows more clearly the impacts of financial ratios by assigning higher relative predictive power to them (by removing a variable of higher importance), the independent results and directions of the ratios are still similar. Furthermore, the impacts of the ratios vary considering their interaction with firm size, yielding more nuanced results.

In general, the results from this study and previous literature suggest that financial characteristics with sector membership explain a notable proportion of ESG score variance. The findings imply that for ESG score relying on sustainability comparisons across firms, different characteristics and industry membership should be taken into consideration for more accurate comparison. For example, based solely on the findings in this study and ignoring other potential factors, if comparing firms A and B withing the same industry

membership exhibit the same ESG score, yet A is characterized by, for example, smaller size, ROA and leverage, with higher marginal profitability, it could be hypothesized that firm A could be proportionally more responsible in its operations in practice.

6.2 Limitations

This subchapter discusses the limitations of comparing the obtained results with previous literature, the technical and methodological constraints, data representativeness, construct validity, feature scope, interpretability, and external generalizability.

6.2.1 Comparisons to Previous Literature

Comparing the results and methodologies against previous studies highlights how feature sets and domain scope can lift headline accuracy beyond what is attainable from financial ratios alone.

Even though a large portion of the predictable signal in overall ESG was captured by non-linear models using only standard financial data, meaningful residual remains unexplained. This is expected and consistent with the literature, considering the data and chosen methodology.

While results can be measured by the same metrics (MSE, MAE, R^2, RSME), headline performance across papers can be difficult to directly compare. Different limitations of the thesis methodology, findings and their comparability with prior research arise:

Rating provider idiosyncrasies should be noted. Depending on the ESG rating provider and their methodology for computing ESG score, the scored used the different studies can vary.

Sample composition matters for obtained results. Heterogeneity or homogeneity of sample (universe and size mix, industries, geography, time window), inclusion or

exclusion rules for variables, treatment of outliers, imputation choices and dataset structure are examples of characteristics that impact model outcome.

Research purpose behind variable selection should be considered. Including highly correlating variables, such as lagged ESG scores or other highly collinear proxies increase prediction accuracy but do not necessarily provide useful information, depending on the research aim.

6.2.2 Data and construct validity

ESG data reflect proprietary scoring methodologies that are not fully transparent. Because the LSEG rating framework partly relies on company self-disclosures and weighting schemes unknown to the public, the models may have learned patterns specific to LSEG's internal scoring logic rather than objective sustainability performance. Thus, as a result, to be specific, findings should be interpreted as revealing the drivers *of* LSEG-assessed ESG scores, not necessarily of true environmental or social outcomes. It would technically be possible for all the findings in this study to simply reflect a bias within LSEG assessment weights. This limitation is one of the reasons why it is important to have similar studies that use differing sources and time periods of ESG scores in the literature, to compare obtained results based on data from different providers.

ESG ratings aggregate diverse environmental, social, and governance aspects into a single composite indicator. Using only the aggregated ESG score as the main target variable may mask offsetting dynamics between pillars—strong governance can compensate for weaker environmental performance and vice versa. While pillar-specific models were included, the focus on aggregate predictability limits granularity in understanding domain-specific drivers of sustainability performance.

6.2.3 Study Design

A limitation of this study is that it is based on annual, static observations and does not model temporal dependence or firm-specific effects. While the dataset spans nearly a decade, the analysis treats each firm-year as an independent observation to focus on cross-sectional prediction and model interpretability. This design limits the ability to infer causality or dynamic adjustment processes between financial health and ESG performance. Because firm characteristics and ESG practices evolve gradually, ignoring within-firm persistence may obscure feedback mechanisms—for instance, whether financial improvements lead to better ESG outcomes or vice versa. Although appropriate for the study's predictive and explanatory objectives, this cross-sectional approach captures association rather than direction of influence.

6.2.4 Feature Score and Omitted Variables

The study focuses exclusively on quantitative financial ratios, omitting non-financial factors such as ownership structure, board composition, or country-level governance indicators that could influence ESG assessments. The restricted feature set ensures comparability and interpretability but may leave out relevant determinants that could further improve model accuracy and explanatory depth.

6.2.5 Technical Constraints

Other imitations include computational issues relating to larger datasets and machine learning as a method. One methodological limitation is computational scalability as feature count grows. More variables inflate hypothesis space and interaction depth, driving up training time, memory use, and the variance of estimates, which increases overfitting risk and making hyperparameter search computationally expensive. High dimensionality also worsens multicollinearity, which can destabilize permutation importances and SHAP attributions. Adding further features, dummies or imputation flags helps modelling but further increases depth requiring computing resources. Practically, this forces different modelling compromises like fewer CV folds, coarser hyperparameter grids or

regularization and mitigations in feature screening, for example, it is not feasible to start feature screening from including every financial metric, instead, initial feature screening relies on combinations of findings in previous literature.

6.2.6 Interpretability Caveats

Although SHAP and PDP methods enhance transparency, they rely on model-specific approximations and assume feature independence during aggregation. Interactions or correlations between variables may therefore distort the absolute magnitude or direction of individual attributions. Consequently, interpretability outputs should be viewed as indicative of model behaviour, not as direct causal mechanisms.

6.2.7 External Validity and Regulatory Shifts

The analysis is limited to European listed companies within the STOXX Europe 600. Differences in accounting standards, disclosure quality, and ESG regulation across regions may restrict the generalizability of the findings. The relationships between financial structure and ESG scores observed in Europe might not fully apply to firms operating under different regulatory or market contexts, such as North America or emerging markets.

The period 2014–2023 covers major developments in sustainability reporting, such as the introduction of the EU Taxonomy and the CSRD. These evolving disclosure frameworks may have influenced ESG scores over time, introducing latent structural breaks that static models cannot explicitly capture.

6.3 Suggestions for Future Research

Future research could extend this work by incorporating the temporal dimension of ESG development through longitudinal or panel-data approaches. Using the same firm-year dataset in a panel framework would enable estimation of firm fixed effects and lag

structures to capture within-firm changes over time. Such models could test whether shifts in profitability, leverage, or efficiency precede subsequent improvements in ESG scores, or alternatively, whether enhanced ESG performance leads to measurable changes in financial outcomes. Combining interpretable machine learning with panel-econometric techniques would allow both higher predictive accuracy and stronger causal insight, providing a more comprehensive view of how financial and sustainability dynamics interact over time.

Future studies could enrich the predictive framework by integrating non-financial data such as textual disclosures, sustainability reports, or ESG-related news sentiment. These unstructured sources often capture qualitative dimensions of sustainability—governance controversies, environmental commitments, or stakeholder engagement—that are not reflected in financial ratios. Combining structured financial variables with text-derived indicators through natural-language-processing techniques could reveal whether ESG ratings primarily respond to firms' reported narratives or to their underlying financial fundamentals. Such multimodal approaches would bridge the gap between disclosure quality and real sustainability performance.

While this thesis employed tree-based and regularized linear models for interpretability, explainable deep learning could extend these findings. Neural architectures such as attention-based networks or hybrid models combining gradient boosting with deep representations can capture higher-order nonlinearities while remaining partially transparent through integrated-gradient or SHAP-Deep explanations. The space of deep learning applications is continuously evolving, and applying these techniques or novel methodologies in the future to ESG prediction would test whether additional predictive accuracy can be achieved without sacrificing interpretability.

As sustainability reporting frameworks such as the European Sustainability Reporting Standards (ESRS) and the Corporate Sustainability Reporting Directive (CSRD) become increasingly operational, cross-regional comparisons will gain importance. Future

research could examine how these regulatory developments influence the relationship between financial structure and ESG scores across jurisdictions. Comparing European data with markets operating under different disclosure mandates would clarify whether observed financial–ESG linkages are global or shaped by institutional environments. Such analysis would also shed light on how regulatory convergence affects ESG rating consistency and the economic meaning of sustainability scores over time.

Conclusion

The objective of this study was to investigate the extent to which company-level financial information can explain variation in sustainability performance, as reflected by ESG scores, and to identify which financial characteristics most strongly influence these scores. The research contributes to the literature on sustainable finance by applying interpretable machine learning methods to examine whether financial statement items contain predictive information about corporate sustainability ratings.

The study used a dataset of firms included in the STOXX Europe 600 index between 2014 and 2023, obtained from the LSEG database. The dependent variable – ESG score, was a subject of predictive modelling using a set of independent profitability, leverage, liquidity and valuation ratio variables, together with firm size, industry, and year identifiers.

The empirical analysis employed a range of supervised machine learning models to predict ESG outcomes. Tree-based ensemble models were used to capture nonlinear relationships and feature interactions, while regularized linear models provided benchmarks. Model performance was assessed using mean squared error, mean absolute error, and the coefficient of determination on the test set. Explainable-AI techniques were incorporated to interpret behaviour and feature effects on the best performing model, which allowed for evaluating predictive accuracy, direction and stability of the learned financial – sustainability relationships.

This study shows that financial statement information explains a meaningful share of the cross-sectional variance in ESG scores. Among the evaluated models, XGBoost delivered best out-of-sample performance for the overall ESG score ($R^2 \approx 0.64$; MSE ≈ 105; MAE ≈ 7.98). Regularized linear baselines underperformed nonlinear methods, showing the presence of nonlinear and interaction effects in the data.

Multiple interpretability approaches converge on the set of key drivers. Firm size is the dominant positive predictor, shaping predictions broadly (top by permutation

importance and gain) and interacting most with other variables. Operational efficiency (NS/TA) and capital structure (TD/TA) add secondary but stable signal. NS/TA exhibits a concave relation: very low turnover depresses predicted ESG, mid-range is most favourable, and gains flatten at higher levels. TD/TA is also non-linear: moving from near-zero leverage to moderate leverage (~0.2–0.35) raises predicted ESG, with a mild fade beyond that. Dividend yield (DIV Y) is modestly positive, especially for smaller firms.

Profitability measures split in direction. ROA is positively associated with ESG predictions, suggesting that broad, asset-based profitability aligns with higher sustainability assessments. In contrast, EBIT/NS (margin) and, to a lesser extent, ROE show inverse or concave patterns: very high margins/returns are linked to lower predicted ESG. This is consistent with a trade-off narrative where maximizing short-term profitability does not move in alignment with the ESG signal captured by ratings. Liquidity (CA/CL) contributes little in level but shows a negative direction at higher values, indicating that excess liquidity does not translate into higher ESG score in this sample.

Industry and time effects matter but do not overturn firm-level patterns. Industry memberships (high gain/cover) capture sectoral differences that financial ratios alone do not explain. SHAP interactions and 2D PDPs show mostly vertical contours, reinforcing that size sets the baseline while other ratios apply context-specific adjustments.

In general, financial statement items with identifiers can predict a substantial portion of ESG scores with size, efficiency, and moderate leverage doing the heavy lifting, while extreme profitability and high liquidity do not correspond to higher predicted ESG in the dataset.

For analysts or investors, the partial replicability of ESG scores from financials implies that the ratings are not statistically independent from fundamentals. ESG-focused investment screens and tilts may unintentionally embed exposures to firm size, balance-sheet strength, and operating efficiency, reflecting underlying financial structures rather than

purely sustainability attributes. For issuers, improvements in operational efficiency and balanced leverage appear aligned with higher assessed sustainability, whereas extreme profitability and excess liquidity do not. The results suggest that rating agencies and regulators should enhance transparency in scoring methodologies and distinguish between disclosure practices and real sustainability outcomes to limit financial-bias effects in ESG ratings.

The study has different limitations considering data and construct validity, feature scope, cross-sectional design, external validity and interpretability caveats, detailed in the section 6.2. For future research, moving from association to dynamics and causality, expanding beyond financial inputs and testing cross-regional settings are logical next steps.

Overall, the evidence indicates that financial structure and scale anchor much of what current ESG scores capture. Nonlinear machine learning can recover this signal, but genuinely multidimensional sustainability assessment will require models and data that go beyond conventional financial indicators.

References

Ahmad, H., Yaqub, M., & Lee, S. H. (2024). *Environmental-, social-, and governance-related factors for business investment and sustainability: A scientometric review of global trends.* Environment, development and sustainability, 26(2), 2965-2987. https://doi.org/10.1007/s10668-023-02921-x

Ahmd, U. et al. (2024). *Investigating boosting techniques' efficacy in feature selection: A comparative analysis.* Energy Reports, Vol 11, p.3521-3531. DOI:10.1016/j.egyr.2024.03.020

Alsayyad, M., & Fadel, S. M. (2025). *Predicting ESG Scores Using Firms' Financial Indicators: A Machine Learning Regression Approach.* Preprints. https://doi.org/10.20944/preprints202503.0096.v1

Berg, F., Kölbel, J., & Rigobon, R. (2022). *Aggregate Confusion: The Divergence of ESG Ratings*. Review of Finance, Volume 26, Issue 6, 2, pp. 1315–1344, https://doi.org/10.1093/rof/rfac033

Billio, M., Costola, M., Hristova, I., Latino, C., & Pelizzon, L. (2021). *Inside the ESG ratings: (Dis)agreement and performance.* Corporate Social Responsibility and Environmental Management, 28(5), 1426–1445.

BMC. (n.d.). *Scikit learn guide.* BMC Machine Learning & Big Data Blog. https://www.bmc.com/blogs/categories/machine-learning-big-data/

Breiman, Leo. (2001). *Random Forests.* Machine Learning 45, 5–32. https://doi.org/10.1023/A:1010933404324

Choi, J.-C., Chen, Q., & Lee, S.-J. (2024). *Predicting ESG Ratings by Machine Learning and Analyzing Influencing Factors by XAI.* In Proceedings of the 2024 International Conference on Software Engineering and Information Management (ICSIM)

Chowdhury, M.A.F., Abdullah, M., Azad, M.A.K. et al. (2023). *Environmental, social and governance (ESG) rating prediction using machine learning approaches*. Annals of Operations Research. https://doi.org/10.1007/s10479-023-05633-7

Cini, F., & Ferrari, A. (2025). *Towards the estimation of ESG ratings: A machine learning approach using balance sheet ratios.* Research in International Business and Finance, 73, 102653

D'Amato, V., D'Ecclesia, R. & Levantesi, S. (2022*). ESG score prediction through random forest algorithm. Comput Manag Sci* 19, 347–373. https://doi.org/10.1007/s10287-021-00419-3

D'Amato, V., D'Ecclesia, R., & Levantesi, S. (2021). *Fundamental ratios as predictors of ESG scores: A machine learning approach*. Decisions in Economics and Finance, 44(2), 1087–1110https://doi.org/10.1007/s10287-021-00419-3

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

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of statistics, 29*(no. 5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Hashimoto, R., Miura, K., & Yoshizaki, Y. (2023). *Application of Machine Learning to a Credit Rating Classification Model: Techniques for Improving the Explainability of Machine Learning.* Bank of Japan Working Paper Series No. 23-E-8. https://www.boj.or.jp/en/research/wps_rev/wps_2023/data/wp23e08.pdf https://doi.org/10.48550/arXiv.1802.03888

Huber, B., Comstock, M., Polk, D., & LLP, W. (2017). *ESG reports and ratings: What they are, why they matter. Harvard Law School*. https://corpgov.law.harvard.edu/2017/07/27/ESG-reports-and-ratings-what-they-are-why-they-matter/.

IBM. (n.d.-a). *What is random forest?* IBM Think. https://www.ibm.com/think/topics/random-forest

IBM. (n.d.-b). *What is ridge regression?* IBM Think. https://www.ibm.com/think/topics/ridge-regres-si-on

Lee O, Joo H, Choi H, Cheon M. (2022). *Proposing an Integrated Approach to Analyzing ESG Data via Machine Learning and Deep Learning Algorithms.* Sustainability, 14(14):8745. https://doi.org/10.3390/su14148745

Licari, J., Loiseau-Aslanidi, O., Piscaglia, S., & Solis Gonzalez, B. (2021). *ESG Score Predictor: Applying a Quantitative Approach for Expanding Company Coverage*. Moody's Anal.

Lin, H.-Y., & Hsu, B.-W. (2023). *Empirical Study of ESG Score Prediction through Machine Learning—A Case of Non-Financial Companies in Taiwan.* Sustainability, 15(19), 14106. https://doi.org/10.3390/su151914106

Lin, W. L., Law, S. H., Ho, J. A., & Sambasivan, M. (2019). *The causality direction of the corporate social responsibility–corporate financial performance nexus: Application of panel vector autoregression approach.* The North American Journal of Economics and Finance, 48, 401–418. https://doi.org/10.1016/j.najef.2019.03.004

Lins, K. V., Servaes, H., & Tamayo, A. (2017). *Social capital, trust, and firm performance: The value of corporate social responsibility during the financial crisis*. Journal of Finance, 72(4), 1785–1824

LSEG. (n.d.). ESG scores. LSEG Data & Analytics. https://www.lseg.com/en/data-analytics/sustainable-finance/esg-scores

Lundberg, S., Erion, G., & Lee, S.-I. (2019). *Consistent individualized feature attribution for tree ensembles. arXiv preprint* arXiv:1802.03888. https://arxiv.org/abs/1802.03888

Molnar, C. et al. (2023). *Relating the Partial Dependence Plot and Permutation Feature Importance to the Data Generating Process.* In: Longo, L. (eds) Explainable Artificial Intelligence. xAI 2023. Communications in Computer and Information Science, vol 1901. Springer, Cham. https://doi.org/10.1007/978-3-031-44064-9_24

NVIDIA. (n.d.). Glossary: XGBoost. NVIDIA. https://www.nvidia.com/en-us/glossary/xgboost/

OECD. (n.d.). Corporate sustainability. OECD. https://www.oecd.org/en/topics/sub-issues/corporate-sustainability.html#:~:text=Corporate%20sustainability%20,company%27s%20business%20strategy%20and%20operations

Pilz, G. (2024, April 23). *Accelerating the pace of sustainability transformations in U.S. publicly held companies*. Harvard ALI Social Impact Review. Retrieved from https://www.sir.advancedleadership.harvard.edu/articles/accelerating-pace-sustainability-transformations-us-publicly-held-companies

Ramchandran, M., Mukherjee, R., & Parmigiani, G. (2021). *Cross-Cluster Weighted Forests. arXiv preprint* arXiv:2105.07610. https://doi.org/10.48550/arxiv.2105.07610

Ranta, M., (2023). *Introduction to Data Analytics in Accounting and Finance.* (Ch. 8-12). https://mranta-ai.github.io/Data_analytics_in_accounting/book_intro.html

Scikit-learn developers. (2024). *RandomForestRegressor — scikit-learn 1.5 documentation*. scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall/CRC.

Svanberg, J., Ardeshiri, T., Samsten, I., Öhman, P., Neidermeyer, P. E., Rana, T., Semenova, N., & Danielson, M. (2022). *Corporate governance performance ratings with machine learning.* Intelligent Systems in Accounting, Finance and Management, 29(1), 50–68. https://doi.org/10.1002/isaf.1505

Taskın, D., Sariyer, G., Acar, E., & Cagli, E. C. (2024). *Do past ESG scores efficiently predict future ESG performance? Research in International Business and Finance, 74,* 102706. https://doi.org/10.1016/j.ribaf.2023.102706

XGBoost Developers. (2024). *XGBoost parameter documentation (v2.1)*. XGBoost Project. https://xgboost.readthedocs.io/en/stable/parameter.html

Note: In addition to the references above, artificial intelligence was utilized during the writing processes. ChatGPT was used to improve writing quality and to assists with the data analysis process.

Appendices

Appendix 1. ESG Report/Rating Summary Table by Huber et al. (2017)

Table 11. ESG Report/Rating Summary Table by Huber et al. (2017).

ESG Report Provider	Background	Rating Scale	Methodology	Usage and Reputation
Bloomberg ESG Data	Collects ESG Data for over 9,000 companies; integrated into Bloomberg Equities and Intelligence Services. International scope.	Out of 100	Provides scores from third-party rating agencies; looks at 120 ESG indicators.	In 2016, Bloomberg had over 12,200 ESG customers.
Corporate Knights Global 100	Publishes an annual index of the Global 100 most sustainable corporations in the world.	Out of 100	Ranked against other companies in their industry group; 14 key performance indicators. Companies only scored on relevant performance indicators for their specific industry.	Of the top 10 corporations listed on the 2017 'Global 100,' 4 of them published a press release about their listing.
DJSI	First global index to track sustainability-driven companies based on RobecoSAM's ESG analysis. Broken down into DJSI World, DJSI Regions, and DJSI Country. International scope.	Out of 100	Ranked against other companies in their industry. Industry-specific questionnaire covering relevant economic, environmental, and social factors (80–120 questions). Updated annually. Partnered with RobecoSAM.	Of the 10 Industry Group Leaders listed on the 2016 DJSI, all 10 published a press release about their listing.
ISS	Acquired Ethix SRI; partnered with RepRisk to provide ESG and SRI research. Solutions also include climate change data and analytics (via acquisition of Climate Neutral Investments). ISS QualityScore provides corporate governance reports	ISS QualityScore: 1–10; Climetrics Score: 1–5 'green leaves'	ISS QualityScore: covers board structure, compensation/remuneration, shareholder rights, and audit/risk oversight; updated on an ongoing basis. ISS-Ethix: provides research, screening, and analysis on SRI topics.	A leading provider.

	on over 5,600 public companies. In July 2017, ISS-Ethix and CDP launched 'Climetrics.' International scope.			
MSCI ESG	Provides ratings for over 6,000 companies and 350,000 equity and fixed-income securities. International scope.	AAA to CCC	Looks at 37 key ESG issues, with data collected from publicly available sources. Companies are monitored on an ongoing basis and receive an annual in-depth review. Includes iShares MSCI EAFE ESG Select ETF and MSCI EM ESG Select ETF.	Used by institutional investors such as Legal & General Investment Management, Morgan Stanley, Northern Trust Asset Management, and PIMCO.
RepRisk	Founded in 1998. Provides ESG reports on over 84,000 private and public companies across 34 sectors. International scope.	AAA to D	Looks at 28 ESG issues (mapped to the Ten Principles of the UN Global Compact) plus additional 'Hot Topics.' Updated daily. Partnered with the UN-supported Principles of Sustainable Investment.	Used by institutional investors including Amundi and APG. Partnered with ISS (Institutional Shareholder Services).
Sustainalytics	Formed in 2008 through a consolidation of DSR, Scoris, and AIS. Covers over 6,500 companies across 42 sectors. International scope.	Out of 100	Sector/industry-based comparison. At least 70 industry-specific ESG indicators per sector. Also looks at systems to manage ESG risks and disclosure of ESG issues and performance.	Strategic relationships with BNY Mellon, City of London Investment Management, Columbia Threadneedle, Norwegian Government Pension Fund, and Prudential Fixed Income.
Thomson Reuters ESG Research Data	Thomson Reuters acquired Asset4 in 2009. Provides ESG data on over 6,000 companies. International scope.	Percentile rank (D- to A+)	Covers 400 ESG metrics, selecting 178 of the most relevant data points. Categories are weighted. Updated every 2 weeks. Comprehensive database.	ESG Scores are available on Thomson Reuters Eikon platform.

Appendix 2. Results from other trained models

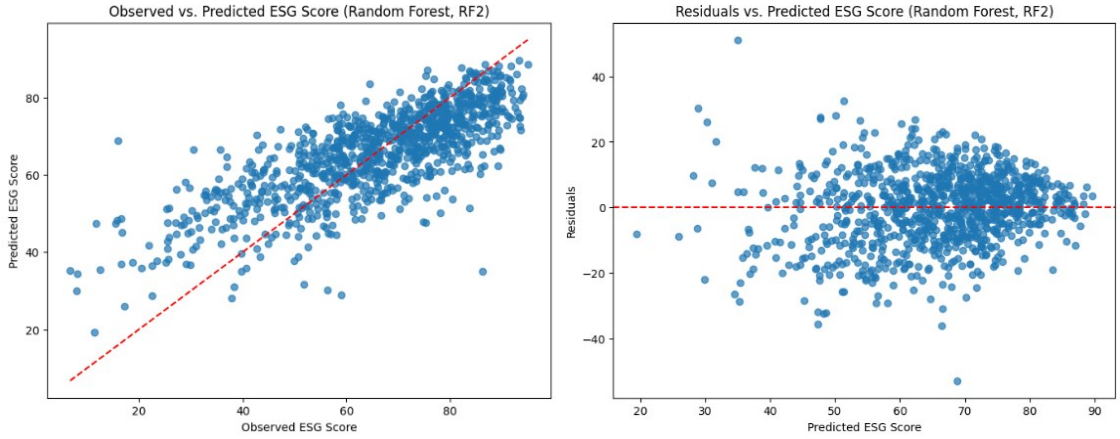

Figure 21. RF2 Observed vs predicted and residuals.

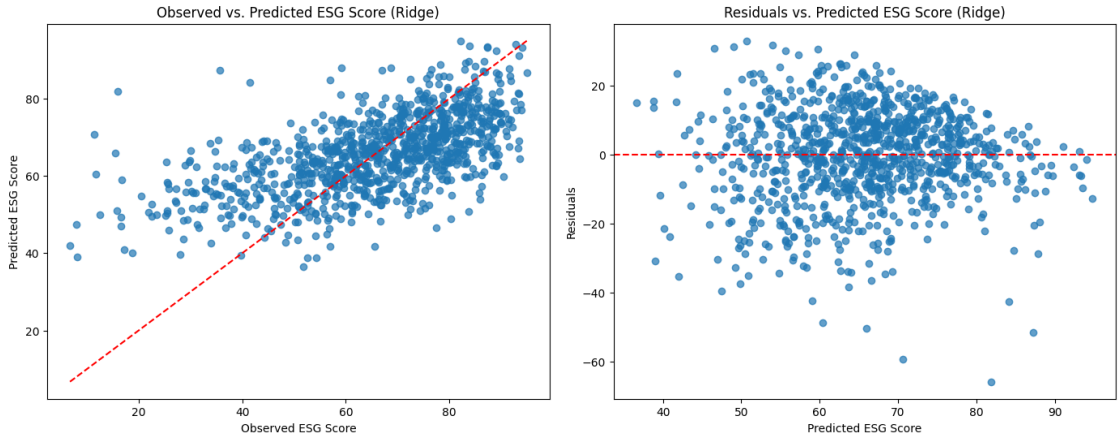

Figure 22. RIDGE Obtained vs predicted and residuals.

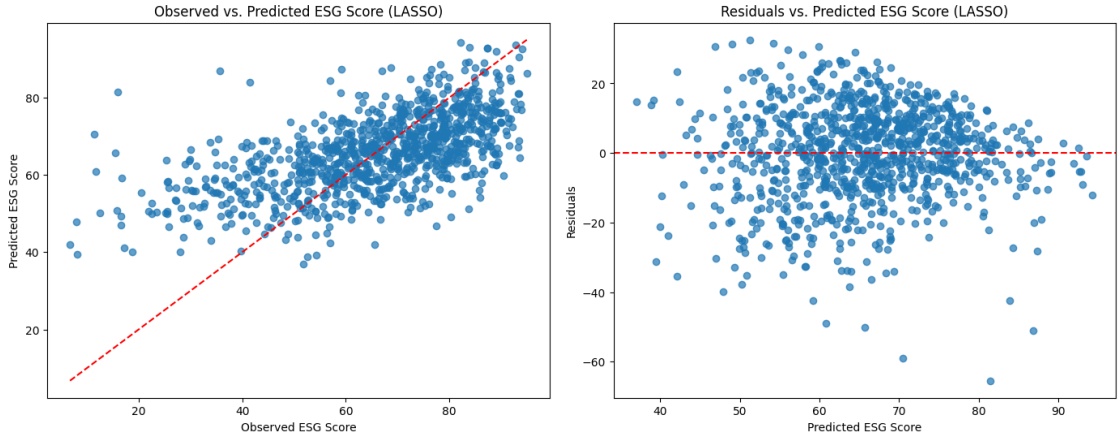

Figure 23. LASSO Obtained vs predicted and residuals.

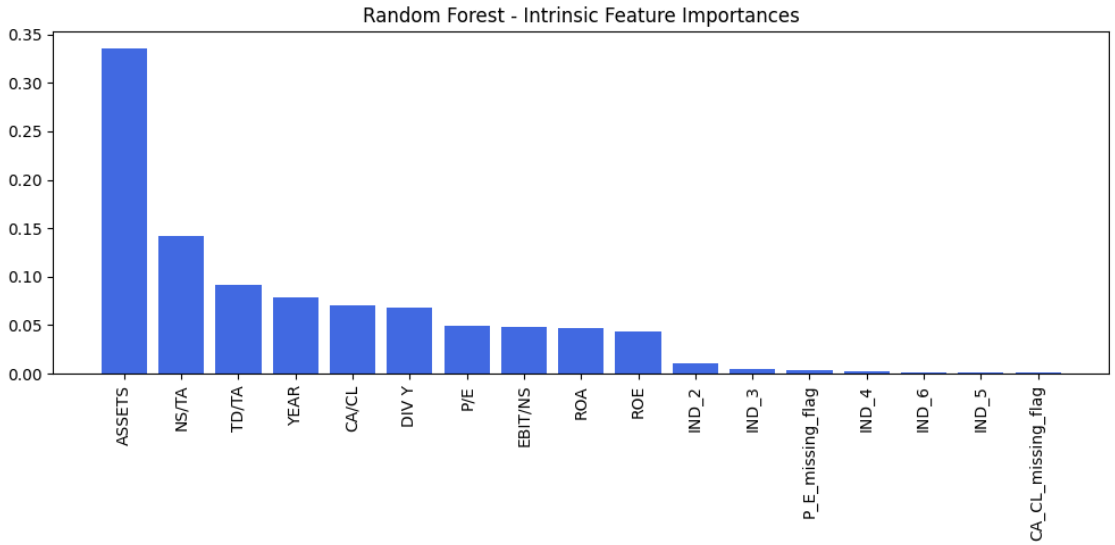

Figure 24. RF2 Intrinsic Feature Importances.

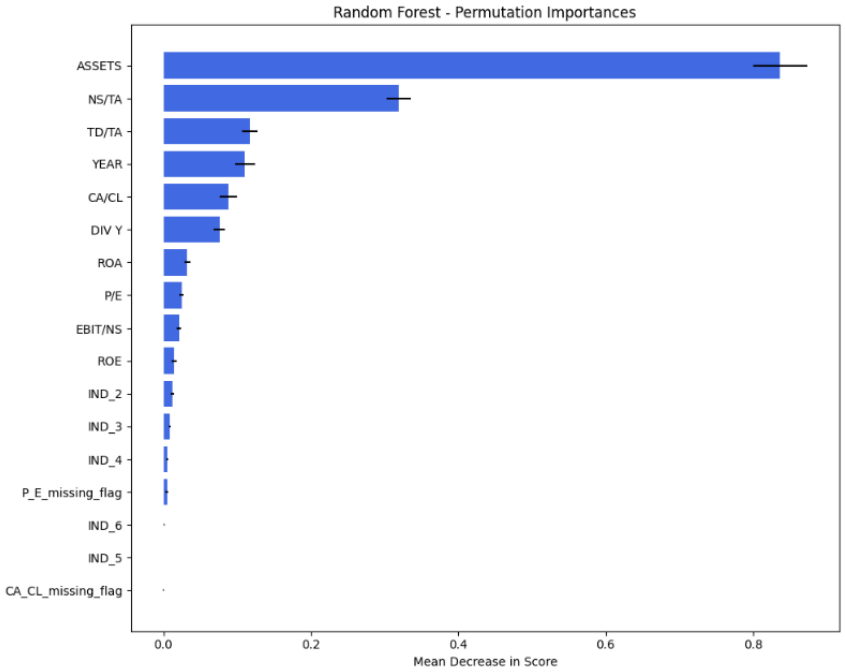

Figure 25. RF2 Permutation Importances.

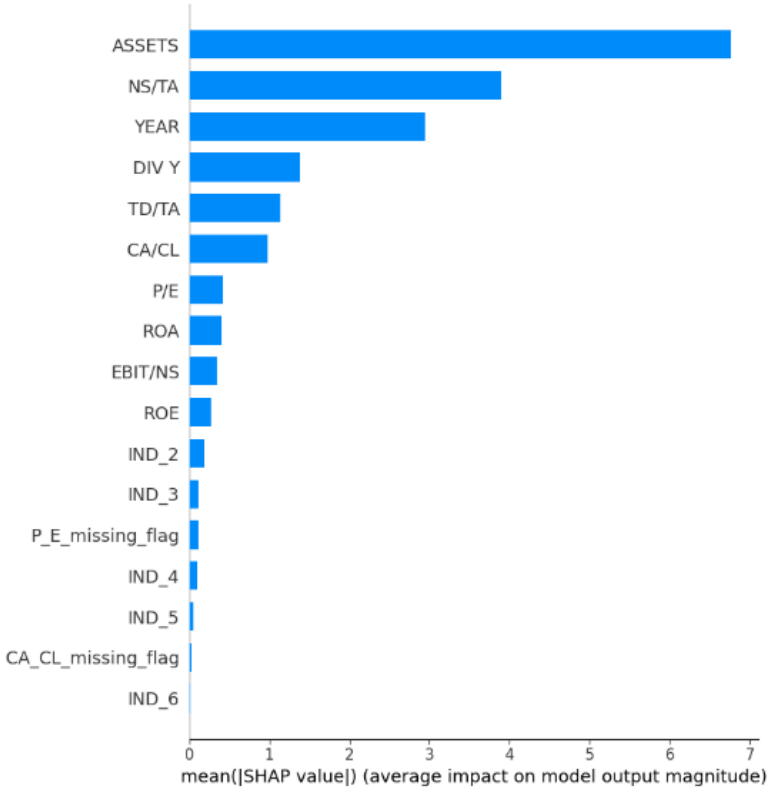

Figure 26. RF2 SHAP Feature Importances bar.

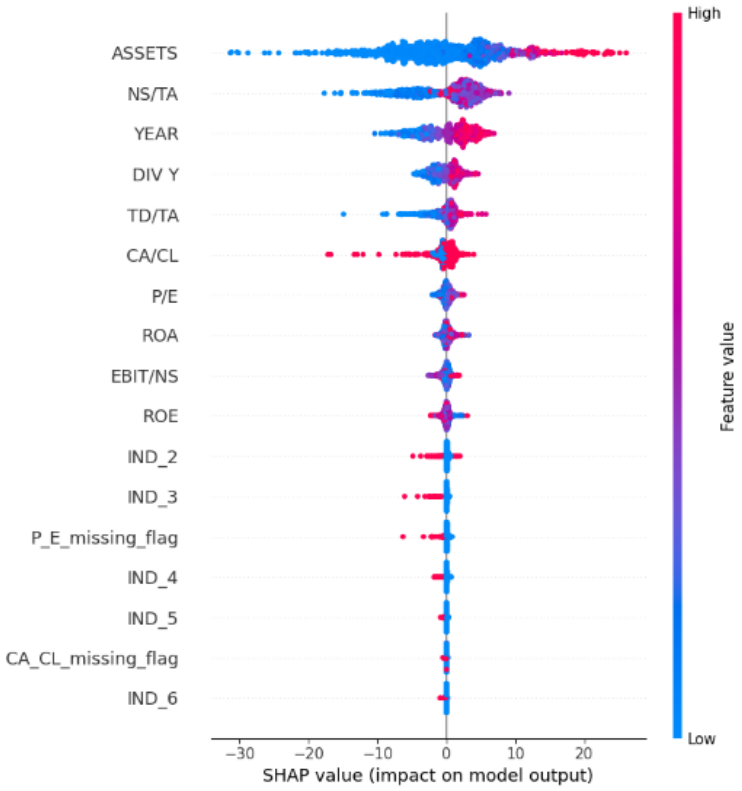

Figure 27. RF2 SHAP Feature Importances plot.

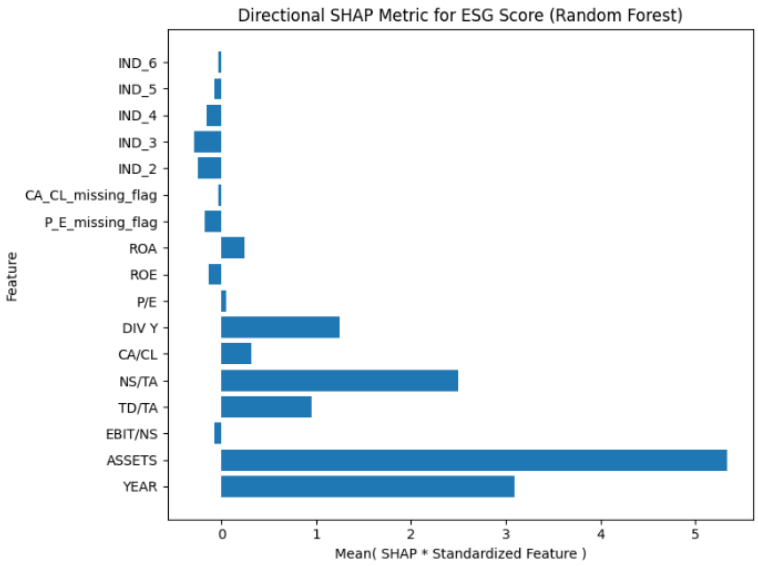

Figure 28. RF2 SHAP Directional metrics.

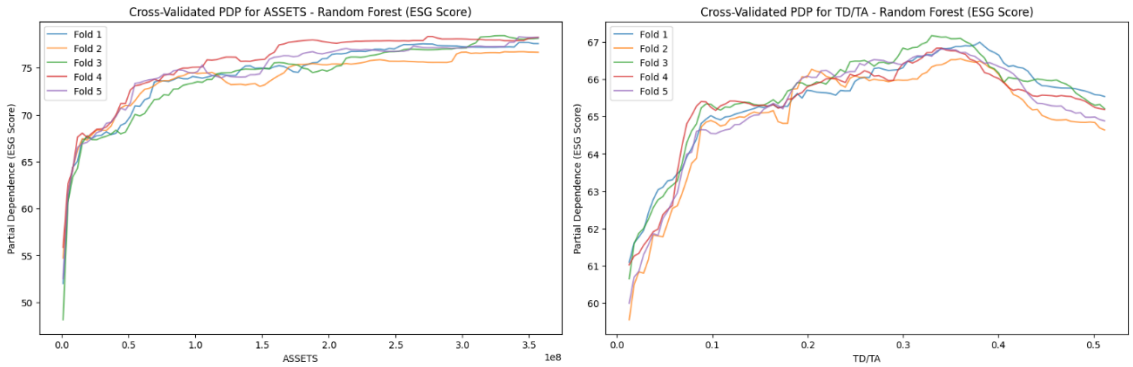

Figure 29. RF2 Cross validated PDPs for ASSETS and TD/TA.

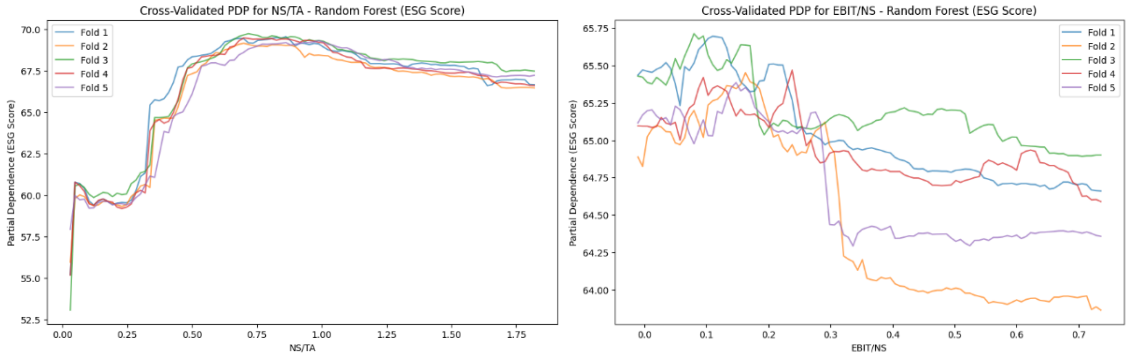

Figure 30. RF2 Cross validated PDPs for NS/TA and EBIT/NS.

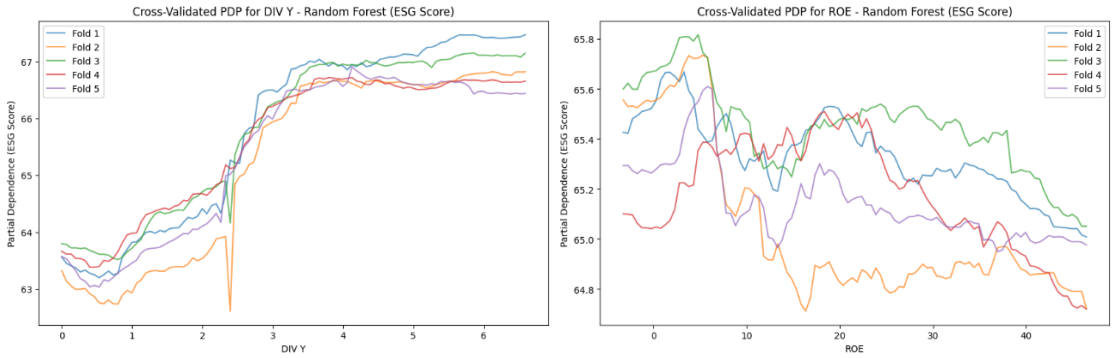

Figure 31. RF2 Cross validated PDPs for DIV Y and ROE.

Figure 32. RF2 Cross validated PDPs for P/E and ROA.