A STUDY OF CLASSIFICATION MODELING OF ESG PERFORMANCE ON EUROPEAN FIRMS AND UNVEILING OF INFLUENTIAL FACTORS

MARIA EULÀLIA DOMINGO COTS

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2021549

COMMITTEE

Dr. Gonzalo Nápoles
Prof. Samaneh Khoshrou

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

January 15th, 2024

WORD COUNT

8670

ACKNOWLEDGMENTS

Als meus pares, i al Pau, gràcies per ser-hi sempre.
I would like to thank Professor Nápoles for his dedication, support and advice, mil gracias.

A STUDY OF CLASSIFICATION MODELING OF ESG PERFORMANCE ON EUROPEAN FIRMS AND UNVEILING OF INFLUENTIAL FACTORS

MARIA EULÀLIA DOMINGO COTS

Abstract

This thesis investigates the use of four different machine learning models —Logistic Regression, Decision Tree, Gradient Boosting, and Random Forest— for evaluating Environmental, Social, and Governance (ESG) performance. It identifies Random Forest as the most effective of these models. The study innovatively applies of SHapley Additive exPlanations (SHAP) analysis to ESG data, uncovering the most important factors influencing performance in both Social and Governance datasets. Key variables in the Social data were found to be Product Responsibility Monitoring and Corporate Responsibility Awards. In governance data, factors like Independent Board Members and Executive Members Gender Diversity significantly impact ESG performance. Furthermore, it explores optimal feature retention thresholds in Social data, providing strategic insights for efficient ESG evaluation. This thesis found that approximately half of the variables (16 out of 29) can be discarded while still retaining substantial accuracy in ESG performance prediction. This research not only offers practical guidance for firms to enhance their ESG practices but also makes a significant academic contribution by demonstrating the application of SHAP analysis in this emerging area of study.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

1.1 *Source/Code/Ethics/Technology Statement Example*

Data Source: The ESG Dataset has been acquired from the Refinitiv Database through an online request. The obtained data is anonymised. Work on this thesis did not involve collecting data from human participants or animals. The original owner of the data and code used in this thesis retains ownership of the data and code during and after the completion of this thesis. However, the institution was informed about the use of this data for this thesis and potential research publications. The thesis code will be provided upon request. Part of the code has been adapted by the author from scikit-learn and ChatGPT. The code for this thesis was written in Python (version 3.10.11). The libraries used are NumPy (version 1.24.3), Pandas (version 1.5.3), and Scikit-Learn (version 1.2.1) The reused/adapted code fragments are clearly indicated in the notebook. In terms of writing, the author used assistance with the language of the paper. A generative language model from Grammarly and Notion was used to improve the author's original content, for paraphrasing, spell checking and grammar. No other typesetting tools or services were used.

2 INTRODUCTION

2.1 *Problem Statement*

In a world where the impact of business extends far beyond financial metrics, Environmental, Social, and Corporate Governance (ESG) standards have emerged as the new frontier in assessing a company's contribution to society and the planet. Defined by the European Union as a means to provide "an opinion on a company or financial instrument's sustainability profile or characteristics, exposure to sustainability risks or impact on society and/or the environment" Commission (2023). In the last decade, a growing number of companies have been keen on reporting some of their performance in these fields. However, until this year 2023, the criteria by which this information was reported and the fields on which to report were optional and consequently not standardized. Both the European Union and the United Kingdom are now legislating to standardize and benchmark ESG reporting requirements, establishing clear guidelines for the reporting of sustainable and responsible business practices. KPMG (2023). Many more countries can be expected to follow.

Despite the remarkable surge of interest and investment in ESG-related products, exemplified by a nearly 525% increase in capital allocation from 2015 to 2019 Chung and Michaels (2019), the realm of ESG research remains relatively uncharted. Prior research has mainly explored the relationship between ESG performance and financial outcomes, primarily within the context of investment funds favoring ESG-responsible firms. Existing literature however has largely overlooked the prediction and enhancement of ESG performance as a standalone metric. This quasi-untapped field holds the potential to unravel crucial insights into the future trajectory of ESG practices and their impact.

2.2 *Societal and Scientific Novelty*

This thesis aims to shed some light on the extend to which machine learning algorithms can be successful in predicting future ESG performance, as well as to study which factors play the biggest role in it. This has great social relevance as it can inform investors, policymakers, and industry stakeholders about the sustainability practices in the chosen sector, helping promote responsible investments and sustainable production. To this end, a key aspect of this thesis is determining the optimal feature retention thresholds for ESG performance enhancement without compromising predictive accuracy. This involves identifying the essential features required for accurate ESG score predictions, balancing computational efficiency with cost-effectiveness. This approach not only aims to reduce the financial burden of ESG assessment for firms and investors but also offers strategic direction for companies looking to improve their ESG ratings. By focusing on impactful and efficient ESG evaluation, the thesis aligns with the broader goal of promoting responsible corporate behavior and sustainable business practices.

Furthermore, this thesis will being on scientific novelty by bridging two significant gaps in the current landscape of ESG (Environmental, Social, and Governance) research; Firstly, it narrows its focus exclusively on firm-level ESG metrics, diverging from previous studies that have incorporated broader economic indicators. This targeted perspective aims to unravel the intricate ways in which ESG factors directly impact a company's overall risk profile. Secondly, this thesis pioneers a more thorough and methodical approach to validate feature importance. While some past studies incorporating ESG data have conducted feature analysis, their exploration lacked rigorous examination of these features. By employing advanced machine learning tools such as SHAP analysis and Pixel Flipping Experiment, the thesis aims to elevate the standard of feature analysis in ESG research. This approach provides a clearer, more detailed understanding of how specific ESG factors influence a company's risk profile, offering a significant advancement in the realm of ESG studies.

2.3 *Research Question*

The main goal of this thesis is to answer the following:

> *To what extent can machine learning models accurately classify the ESG performance of European firms while elucidating key factors influencing their performance?*

The sub-questions can be listed separately, as such:

RQ1 *How accurately can we predict ESG performance using machine learning models?* The aim of this subquestion is to explore classification accuracy, using Logistic Regression as a baseline and compare against it the performance of a single Decision Tree, Random Forest and Gradient Boosting. Their performance will be evaluated through the metrics of accuracy and F1. These performance metrics are appropriate for evaluating and comparing classification algorithms, and have been successfully used in a similar setting by Chowdhury et al. (2023). However, contrary to what Chowdhury et al. did in their paper, this thesis will focus on white-box models. Previous literature, which has predominantly employed a predictive regression approach for ESG forecasting, has underscored the value of white-box predictive algorithms. This is echoed in studies like those by Lee et al. (2022) and Laureti et al. (2022), which stress the importance of interpretability and transparency in ESG research as to provide stakeholders with meaningful insights. More detail on the chosen methods will follow in the Methodology section of this thesis.

RQ2 *Which ESG factors play a pivotal role in influencing ESG performance?* While classification was the main focus of the first sub-question, this one focuses on the second half of the main research question; exploring the most influential factors for ESG classification. On the best performing model above, which was Random Forest, this thesis will realize a feature importance analysis to understand which ESG factors have the most significant impact on ESG performance predictions using SHAP. As Gradient Boosting and Random Forest are not easily explainable models, SHAP will provide feature importance scores, bringing to light interesting insights and empower firms to act according to impact importance. Lastly, the validity of the SHAP results will be put to test by performing a pixel flipping experiment.

RQ3 *What are the optimal feature retention thresholds that maximize ESG performance enhancement while maintaining predictive accuracy?*
This thesis will experiment with different feature retention thresholds to assess how selective inclusion of ESG features affects ESG performance predictions. Computing ESG scores can be computationally expensive, and gathering the data is costly too. It would be interesting to investigate what is the minimum number of the features previously found to be important needed by the model to obtain a certain accuracy threshold. Having this information would cheapen the cost for stakeholders to estimate their ESG scores. In the realm of ESG research, it has not been established through existing studies the optimization of feature count required to maintain a specific accuracy threshold in predictions. Hence, this aspect of the thesis is identified as a novel scientific contribution, introducing a unique dimension to the current body of knowledge.

3 RELATED WORK

Current literature surrounding the topic of Environmental, Social, and Governance (ESG) factors mainly focuses on the intersection of ESG with financial performance and sustainable investment, as interest in the subject has grown over the last decade. This research leverages Machine Learning (ML) techniques to investigate the relationships between ESG metrics and financial outcomes. A brief overview of some of the most relevant papers and their findings follows.

The paper "Does Good ESG Lead to Better Financial Performances by Firms?" De Lucia et al. (2020) aimed to predict financial indicators such as Return on Equity (ROE) and Return on Assets (ROA) of public enterprises in Europe based on ESG indicators and economic metrics. Using ML techniques like Random Forest, Support Vector Regression, and others, the paper predicted financial metrics and identified a positive relationship between ESG practices and financial indicators. The results suggested that companies that performed well in ESG metrics also exhibited stronger financial performance. Similarly, the Random Forest method was used by D'Amato et al. (2022) to study how financial sheet data explained ESG scores. Research on how ESG performance is related to Risk Free Securities has been conducted by Khan et al. (2016), MSCI ESG Research LLC (2017) and Melas et al. (2016). These three papers have implemented machine learning methods to deepen their understanding of this complex topic. Their choice of models and discussion of limitations have been very informative and shaped posterior research conducted.

Ariel Lanza, Enrico Bernardini, and Ivan Faiella (2020) presented a distinctive approach in "Mind the Gap! Machine Learning, ESG Metrics, and Sustainable Investment" Lanza et al. (2020). Their paper employed a variety of tree-based approaches to analyze how a variety of ESG metrics to address inconsistencies in ESG scores. An important takeaway for this paper is their discussion on the importance of using white-box models in research concerning ESG's. The interpretability of the models is of outmost importance in order for the findings to be useful to policymakers and stakeholders. It identified ESG indicators that contribute significantly to efficient portfolio construction, particularly those related to environmental factors and climate change risk management. Their research on feature analysis built upon previous research done by Misangyi and Acharya, 2014, which focused on the corporate governance aspect of ESG. This research underscores the potential of ML to enhance ESG-based investment strategies by extracting valuable information from raw ESG data.

Overall, the integration of Machine Learning techniques into the analysis of ESG metrics and sustainable investments offers promising opportunities for investors and researchers. De Lucia et al. (2020) emphasized the relevance of ML in predicting financial outcomes based on ESG practices. Lanza et al. (2020) introduced a unique approach that leverages machine learning to address ESG score inconsistencies and disentangles the contributions of ESG-specific metrics, ultimately contributing to more effective and sustainable investment strategies.

Literature Review on ESG as a standalone metric

Research conducted on ESG as a standalone metric has been limited, with most of the existing literature focusing on ESG performance at the country level rather than at the firm level. Until very recently, companies were not evaluated based on their ESG performance, and therefore, there was less incentive for in-depth research in this area at the firm level. However, with the passage of new European legislation mandating ESG auditing and reporting for firms, there is a growing need to understand and assess ESG practices at the individual company level.

On the country level two interesting studies by the same authors can be found. The first paper, Laureti et al. (2022), focused on corruption while the second one, Laureti et al. (2023) focuses on government effectiveness. The papers use a very similar methodology and approach; firstly it performs a cluster analysis (k-Means), then it tries to predict future value using a variety of machine learning methods, including ANN, tree-based methods and linear regression. After having trained those models it evaluates their performance using R^2, RMSE an MAE, and a discussion on the results naturally follows. A similar approach to compare the performance of different models will be followed in this thesis, except that the evaluation metrics will differ since this thesis is approaching it as a classification problem.

Research Gaps

During the writing of these thesis, a very interesting paper was published by Chowdhury et al. (2023) titled "Environmental, social and governance (ESG) rating prediction using machine learning approaches". This research stands out as it closely aligns with my thesis objective: the classification of ESG as a standalone metric, except for the inclusion of some macroeconomic indicators. Utilizing both firm-specific and macroeconomic predictors, the study discovered that the Random Forest Classifier outperforms other machine learning algorithms by achieving an accuracy of 78.50%. This was measured through various parameters, including Kappa, the area under the curve, receiver operating characteristic, and accuracy. The methodological approach of Chowdhury et al. mirrors what my thesis aspires to accomplish, albeit with differences in ML model selection and metric focus. However, in contrast to the papers by Laureti (Laureti et al., 2022 and (Laureti et al., 2023), Chowdhury et al.'s study does not extend to ranking or a detailed feature analysis study for the chosen models.

While the existing body of literature, such as the studies like Laureti et al. (2022) and Laureti et al. (2023), has made significant strides in integrating ESG data within various predictive and classification models, a discernible gap remains in two critical areas: the focus on analysis in firm-level a sole focus on ESG metrics, and the validation of the feature importance analysis using advanced machine learning techniques.

This thesis aims to pioneer in these two novel fronts. Firstly, it will concentrate exclusively on ESG metrics at the firm level, seeking to provide a nuanced understanding of how ESG factors influence overall ESG risk. Secondly, this thesis seeks to establish a robust framework for assessing the veracity and effectiveness of feature analysis in ESG models, a critical aspect currently underexplored in the literature. Using advanced machine learning techniques like SHAP and pixel flipping will be instrumental in not only identifying but also validating the significance and reliability of the features derived from ESG metrics. This dual-novelty approach is poised to contribute a substantial advancement in the realm of ESG research, providing a comprehensive and robust framework for future studies to build upon.

4 METHOD

The methodology section, detailed in the diagram below, outlines the investigation approach employed in this thesis. A discussion on each section of the methodology will follow.

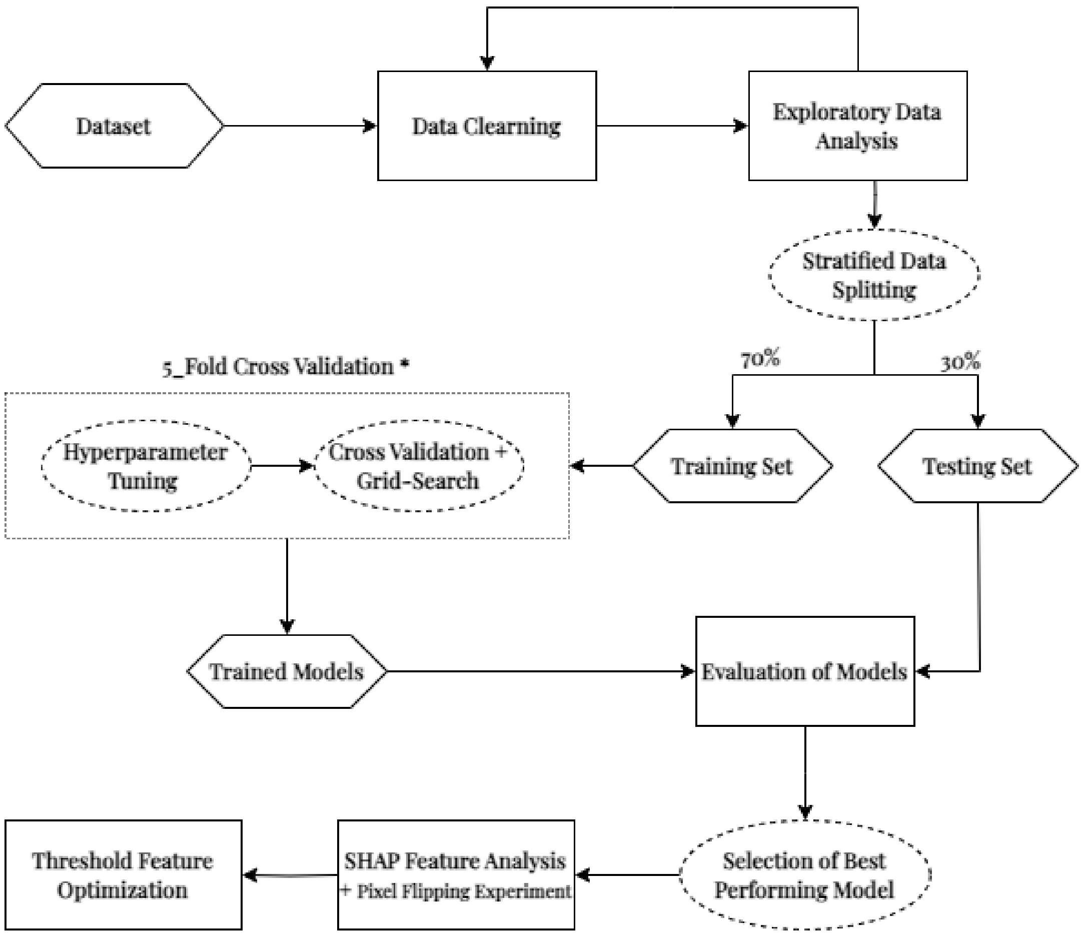
Dataset
Data Clearning
Exploratory Data Analysis
Stratified Data Splitting
70%
30%
5_Fold Cross Validation *
Hyperparameter Tuning
Cross Validation + Grid-Search
Training Set
Testing Set
Trained Models
Evaluation of Models
Threshold Feature Optimization
SHAP Feature Analysis + Pixel Flipping Experiment
Selection of Best Performing Model

Figure 1: Experimental Setup Flowchart: Starts with the initial Dataset that is first Cleaned and then subjected to Exploratory Data Analysis (EDA). Post-EDA, the data undergoes Stratified Data Splitting, 70% of the data going towards the Training Set and remaining 30% towards the Testing Set. The Training Set is then used for Hyperparameter Tuning, informed by problem complexity and utilizing 5-Fold Cross-Validation and Grid Search, five models will be obtained per algorithm. Best model (combination of hyperparameters) of the five will be selected per algorithm and subsequently trained on the entirety of Training Set. These Trained Models are evaluated to determine their performance. Selection of the Best Performing Model is based on the evaluation results. The best model's interpretability is enhanced by SHAP Feature Analysis coupled with a Pixel Flipping Experiment. Finally, as Post-Modeling Analysis, there will be an exploration on the minimum amount of features needed to achieve certain accuracy or F1 thresholds.

4.1 *Data Exploration and Pre-Processing*

4.1.1 *Initial Exploratory Data Analysis*

The initial phase of the research involved extracting Environmental, Social, and Governance (ESG) data relevant to the classification problem at hand. The data was taken from the Refinitv Database. Though lack of standardization of ESG data across-industry and national regulation remains a challenge for conducting this type of studies, Refinitiv offers reputable, standardized and rigorous data.

Data extracted contained information on the Environmental, Social and Governance aspects of each firm, however, this thesis will focus on the Societal and Governance parts. Since the focus if narrowed down geographically, to European firms, rather than by industry or sector, the Social and Governance aspects make for a more fair and equitable comparison. The environmental footprint of each firm is highly variable across industry, Morningstar Sustainalytics, 2023, and therefore comparisons, conclusions and advice on that field should be in the context of a certain industry rather than globally; banking and energy production cannot be held to the same standards.

The data extracted spans five years per firm, ranging from 2018 to 2022, for the 76 largest European firms. The Governance data initially contained 145 features while the Social data contained a total of 237 features.

The initial dataset was characterized by a granular classification of firms into multiple categories, ranging from A+ to lower tiers like D+, comprising a total a dozen distinct classes. This high-level of detail in their labeling led to significant class imbalances, particularly for lower-tier categories with scarce representation. The large number of classes also made the classification task complex and impractical. Reducing the number of classes served two purposes; to create a manageable classification problem and to alleviate the class imbalances. Drawing inspiration from the approach of Chowdhury et al. (2023), which effectively clustered ESG ratings into four roughly balanced groups, this thesis also implemented a similar strategy of class aggregation. Consequently, the data was reorganized into three classes for Social data and four classes for Governance data, guided by the original distribution present in the data, shown in Figure 2. While some class imbalance remains, it has been substantially mitigated as can be seen in Figure 3.

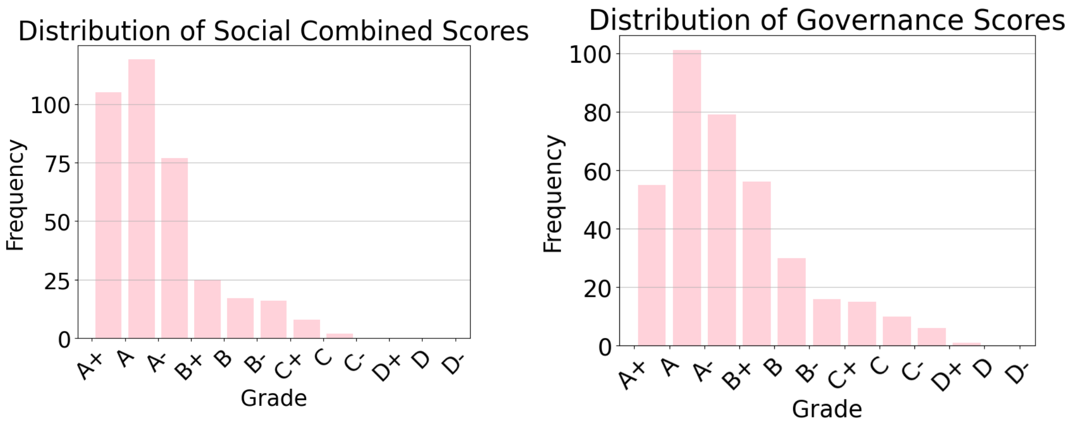

Figure 2: Distribution of Social data on the right and Governance data on the left. Notice the positive skewness in the data.

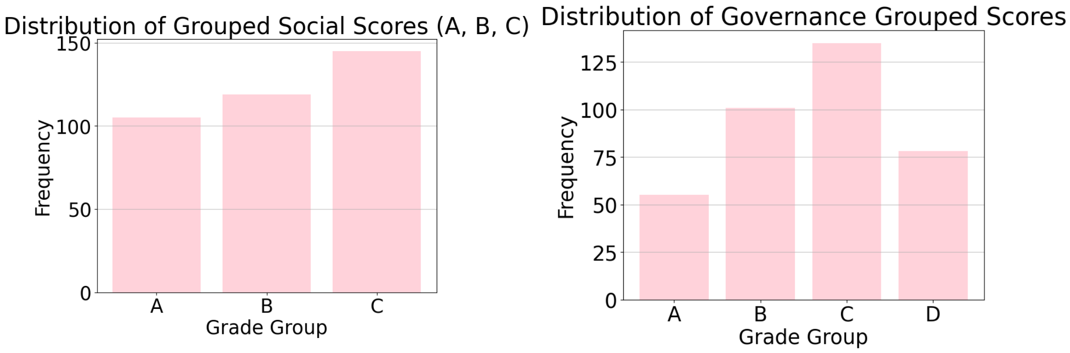

Figure 3: Distribution of Social data on the right and Governance data on the left after re-grouping

This consolidation is conceptually aligned industry practice, ESG performance is often evaluated relative to industry peers or standards, making the notion of absolute scores less relevant. This resource has been commonly used in academic literature concerning ESG like Gao et al. (2023) and Chowdhury et al. (2023), as it helps both in the analysis and also improves the interpretability, which is more insightful than scores void of context. In the context of this thesis the new categories should be interpreted as follows: 'A' indicates "High Relative Performance," indicating firms that are leaders in ESG practices. 'B' signifies "Above Average Performance," for firms doing better than most but not the top performers. Similarly, 'C' denotes "Below Average Performance" where firms meet the industry average, and 'D' could represent "Relatively Poor Performance" indicating a need for improvement. In Social Data, where only A, B and C categories are found, C represents "Relatively Poor Performance".

4.1.2 *Pre-Processing and Final Data Analysis*

After the data had been downloaded from Refinitiv, aware of the potential for high levels of missing data due to non-mandatory ESG reporting, this thesis embarked on pre-processing the data before any further exploration, in order to ensure that the data fed into the models would be clean and lead to fruitful and reliable results. To make data usable, boolean True/False variables were hot encoded into binary dummy-variables, taking values of zero and one, and numerical variables were stripped of text like percentage signs (%). During pre-processing, columns with null values and low variability were dropped, as they contribute little to no predictive power and could skew the model's results. After all these pre-processing, only 18 (down from 145) features remained in the Governance dataset, but 85 out of the 237 initial features remained in the Social Dataset.

Therefore, a factor importance analysis was conducted to identify the most relevant columns for the Social dataset. In doing so, the dimensionality of the dataset was reduced to only the 30 most important features, improving the computational efficiency, and potentially enhancing the model's performance by eliminating noise-inducing variables. Lastly, as the Logistic Regression is a model sensitive to the scale of data, the numerical variables were normalized and used for all models going forward.

The following page Table 1 presents the specific features extracted from the Social and Governance data categories used in this study. Find in the Appendix the Feature Importance Analysis done for the final Social Data and Governance data for further information on how the features ranked at this stage in the process. A brief overlook on the variables follows.

For Social data features, the focus is primarily on workforce-related metrics such as the Number of Employees, which can indicate a company's size and capacity for impact. The inclusion of Women Employees and Net Employment Creation reflects a commitment to diversity and job creation, which are pivotal in assessing social responsibility. Health and safety measures are spotlighted through features like Policy Customer Health and Safety and Employees Health, signaling a company's dedication to well-being. Other features like Corporate Responsibility Awards and various quality management systems underscore a company's engagement with sustainable practices and ethical standards.

On the Governance side, the presence of Independent Board Members and Executive Members Gender Diversity suggests a progressive approach to leadership composition and decision-making. Financial accountability is also considered, with variables such as Total Senior Executives Compensation To Revenues. Board structure and processes are further examined through features like Board Gender Diversity and Audit Committee Independence, which are indicative of balanced governance practices. Finally, the ESG Reporting Scope and Board Member Term Duration point to the transparency and longevity of governance commitment.

Table 1: Features used in Social and Governance Data

Social Data Features	**Governance Data Features**
Number of Employees	Independent Board Members
Number of Employees from CSR reporting	Executive Members Gender Diversity, Percent
Women Employees	Board Member Affiliations
Net Employment Creation	Board Specific Skills, Percent
Policy Customer Health & Safety	Total Senior Executives Compensation To Revenues in million
Six Sigma and Quality Mgt Systems	Board Gender Diversity, Percent
Product Responsibility Monitoring	Auditor Tenure
Human Rights Breaches Contractor	Audit Committee Independence
Employee Resource Groups	Non-Executive Board Members
Corporate Responsibility Awards	Board Size More Ten Less Eight
Product Access Low Price	Anti Takeover Devices Above Two
Quality Mgt Systems	ESG Reporting Scope
Product Sales at Discount to Emerging Markets	Board Member Term Duration
Day Care Services	Voting Cap Percentage
HIV-AIDS Program	Nomination Committee Independence
Internal Promotion	Nomination Committee NonExecutive Members
ISO 9000	Audit Committee NonExecutive Members
Flexible Working Hours	ESG Period Last Update Date
Policy Freedom of Association	
Supply Chain Health & Safety Training	
Targets Diversity and Opportunity	
Employees Health & Safety OHSAS 18001	
Policy Supply Chain Health & Safety	
OECD Guidelines for Multinational Enterprises	
Policy Fair Competition	
Nuclear 5% Revenues	
Improvement Tools Business Ethics	
Diseases of the Developing World	
Supplier ESG training	
Announced Layoffs To Total Employees	

4.2 *Data Splitting*

Given the slightly unbalanced nature of the dataset, a stratified splitting approach was implemented to divide the data into a training set (70%) and a testing set (30%). Maintaining the proportion of the classes in the dataset across both training and testing sets leads to more reliable and generalizable model performance assessments. The models were trained with the training set, which would be further split into a Validation and Training set during the tuning of the hyperparameters. Later in the process, once the models were trained, they were be tested with the previously unseen testing set, ensuring the validity of the evaluation results.

4.3 *Model Training and Selection*

In the model training phase, four machine learning algorithms were trained and compared. The models selected were Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting. Table 2 below shows the hyperparameters chosen to be tuned per model, a discussion on which proved to be optimal will be had in the results section. Each model was chosen for specific characteristics that it brings to the classification task:

Logistic Regression is one of the basic machine learning models, known for its simplicity and interpretability. Though initially conceived as a binary classifier, it is not limited to that and it has been proven to be able to handle multi-class classifications through a cross-entropy cost function as explained in Bisong, 2019. In this thesis, it will serve as a benchmark against which to compare the performance of the other models. However, its simplicity can be a drawback when dealing with complex, non-linear relationships, which are often present in ESG data. Furthermore, it is sensitive to class imbalance, which is present in this thesis' dataset. Therefore, it would be reasonable to expect other models which allow for more complex interactions in the data to outperform the Logistic Regression.

Decision Tree models mimic human decision-making processes by splitting data into branches based on feature values. It can handle non-linear relationships better than Logistic Regression, but due to its simplicity it remains an intuitive and easy to interpret model. As explained in Brown and Mues, 2012, a decision tree is a good fit for the problem at hand as is reasonable to expect that our data will contain non-linear relationships and slight class imbalance. Furthermore, feature importance can be easily derived from it. However, it is prone to overfitting, especially if the tree is allowed to grow complex without pruning, and it is often biased towards the dominant classes. To prevent overfitting from happening, a penalty parameter will be included in the hyperparameters to be tuned.

Random Forest is an ensemble learning method that builds multiple decision trees and uses them to get a more accurate and stable prediction. With the right tuning of its hyperparameters, it is robust against overfitting and is capable of handling a large number of features, complex data structures and is an appropriate model to use in classification problems, Liaw and Wiener, 2002 and Chowdhury et al., 2023. The randomness injected in the model helps in dealing with variance, making it a strong candidate for the dataset at hand. However, its ability to handle increased complexity comes at a cost of a more complex, less interpretable and computationally expensive model. To aid with the interpretability of its results and be able to extract feature importance analysis from it, SHAP will be performed, more details about this will be given below.

XGBoost stands for Extreme Gradient Boosting, an advanced implementation of gradient boosted decision trees designed for speed and performance. It is recognized as a good candidate for problems with unbalanced data, T. Chen and Guestrin, 2016, as it is able to prioritize the misclassified points during training. However, similar to the Random Forest discussed before, Gradient Boosting can be computationally intensive, requires careful tuning to prevent overfitting and ranks low on interpretability. It will be interesting to see which of these two latter mentioned models performs best with the characteristics of the ESG dataset.

4.4 *Hyperparameter Optimization*

These four models were trained on the dataset discussed above, and hyperparameter tuning was conducted to find the optimal settings for each model. The tuned hyperparameters can be found in Table 2 below:

Table 2: Hyperparameters for Different Models

Model	Hyperparameters	Values
Logistic Regression	logisticregression__C	0.01, 0.1, 1, 10, 100
	logisticregression__solver	liblinear, lbfgs
Decision Tree	criterion	gini, entropy
	max_depth	3, 5, 10
	min_samples_split	2, 4, 6, 8
	min_samples_leaf	1, 2, 3, 4
Random Forest	criterion	gini, entropy
	max_depth	3, 5, 10
	min_samples_split	2, 4, 6, 8
	min_samples_leaf	1, 2, 3, 4
Gradient Boosting	gb__n_estimators	100, 200, 500
	gb__learning_rate	0.01, 0.1, 0.2
	gb__max_depth	3, 4, 5
	gb__min_samples_split	2, 5, 10
	gb__min_samples_leaf	1, 2, 4

The best combination of hyperparameters for each model is then selected based on the results of a 5-Fold Cross-Validation (CV) process. 5-Fold CV works by shuffling and splitting the data into five groups, and sequentially using one as validation set while the other four are conform the training set (Brownlee (2023)). This process is repeated five times and then the results of each round are aggregated. CV reduces bias and ensures that the models' evaluation performances are not due to chance but are reliable and can be generalized (Bates et al. (2021)). The chosen hyperparameters will be discussed in the Results section of this thesis.

4.5 *Model Evaluation*

Once trained, the models are evaluated based on their performance on the testing set. This stage assesses the models' ability to generalize to new, unseen data, which is crucial for practical applications. The performance metrics that will be used are two; Accuracy, which reflects the proportion of total predictions that were correct, and the F1-score, a harmonic mean of precision and recall that is particularly informative for imbalanced datasets. The choice of these metrics is substantiated by their utilization by Chowdhury et al. (2023), which also dealt with a classification problem for ESG data classification. While a wider selection of evaluative metrics exists and could potentially yield a more nuanced understanding of model performance, adhering to these two metrics was done in the spirit of keeping within the thesis' scope.

It is important to note that in this context not all mistakes are equal. Classifying a very well-performing firm as a very poorly performing is a more costly mistake than for instance classifying it as a average performing firm. The further away the diagnosis is from the truth, the more it misleads the company into either taking the unnecessary actions or or refraining from the necessary ones. Therefore, in this context it is important to both train and evaluate the models using a cost-matrix. This technique is often used in medical and economic research (Argyrides et al. (2009)). The idea is to assign a larger penalty to the model for a more severe misclassification error. The implementation of this idea is done through a cost-matrix which translates into weighted classification errors. The models train with this weighted errors and later, both the accuracy and the F1 evaluation metrics, also take these weights into account. Find below the cost matrix used in this thesis. This presents an innovation with respect to current ESG related literature, as evaluation methods used do not yet incorporate cost matrices despite it being very suitable to the problem at hand.

$$\text{cost\_matrix} = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 3 & 2 & 1 & 0 \end{pmatrix}$$

4.6 *Post-Modeling Analysis*

After selecting the best performing model, a deeper analysis was conducted to understand the model's decision-making process:

SHAP Feature Analysis: SHAP is based on game theory and provide aimed to provide a measure of the impact of an individual player in the collaborative team S. Chen, 2021. Generalizing this intuition, SHAP provides the impact of each feature on the model's prediction. Chen's paper explain how SHAP sees the predictive value of the model (its accuracy, for instance) as the sum of the attribution value of all its features. SHAP offers both global interpretability—the overall importance of features across the dataset—and local interpretability—the contribution of each feature to individual predictions. As discussed above, it is challenging to understand and interpret complex models like Random Forest and Gradient Boosting, and SHAP is a useful tool to overcome this opaqueness. S. Chen, 2021. Furthermore, S. Chen, 2021 offers insight on how SHAP can be used in a multiclassification problem like the one this thesis faces. SHAP values will be presented in a class level and in a wholistic level separately, as feature importance might differ between classes.

Pixel Flipping Experiment: Although typically used in image processing, the Pixel Flipping experiment is adapted here as a novel approach to test the robustness of the SHAP results. By systematically removing feature values and monitoring the impact on the models accuracy, this experiment assesses feature importance empirically. The name "Pixel Flipping" is metaphorical in this context; instead of pixels, feature values are altered. The results will be represented graphically, showing how the accuracy decreases at each feature removed, starting from the feature deemed most contributing by SHAP to least one. Therefore, one would expect a decreasing curve. To establish a baseline against which to compare SHAP's feature importance results, this thesis will also conduct the Pixel Flipping Procedure using a randomly generated feature sequence and use it as a baseline to compare SHAP's ranking against.

These post-modeling analysis techniques are pivotal in validating the reliability of the model, ensuring the robustness of its predictions, and providing transparency into the model's behavior, which is critical for stakeholders making decisions based on the model's insights.

5 RESULTS

In this section there will be a discussion of the results obtained from the models implemented in this thesis. The models' performance on Governance and Social data will be compared, along with an analysis of hyperparameter optimization for each model. The effectiveness of the models will be evaluated based on two metrics, F1 scores and accuracy during validation, training, and testing phases. Furthermore, the interpretation of models using SHAP values and Pixel Flipping will be presented to understand feature importance.

5.1 *Selected Hyperparameters*

A discussion on the chosen hyperparameters follows. Figure 4 reveals that the social data responds best to a moderate to high level of regularization (Log Reg C) in logistic regression, while the governance data requires a significantly lower regularization level. Interestingly, the optimization algorithm that proved most effective for the governance data is 'lbfgs', known for its efficiency in smaller datasets. In contrast, the 'liblinear' algorithm was optimal for social data, suggesting its suitability more complex datasets.

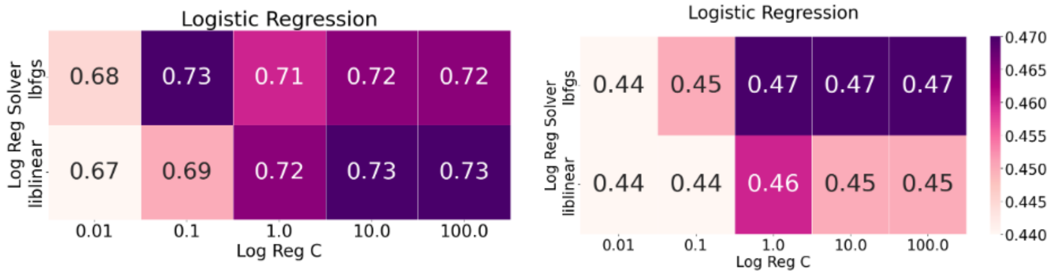

Figure 4: Logistic Regression Hyperparameter Graphs, Social data on the left, Governance data on the right

According to Figure 5, both social and governance data sets achieved their best performance with a Decision Tree using the entropy criterion. This criterion, focusing on maximizing information gain, was complemented by a tree depth capped at a moderate level, ensuring a balance between model complexity and overfitting risk. Notably, both data sets favored minimal restrictions on the sample size for leaf nodes and splits, indicating a preference for capturing greater nuance in the tree's decision-making process.

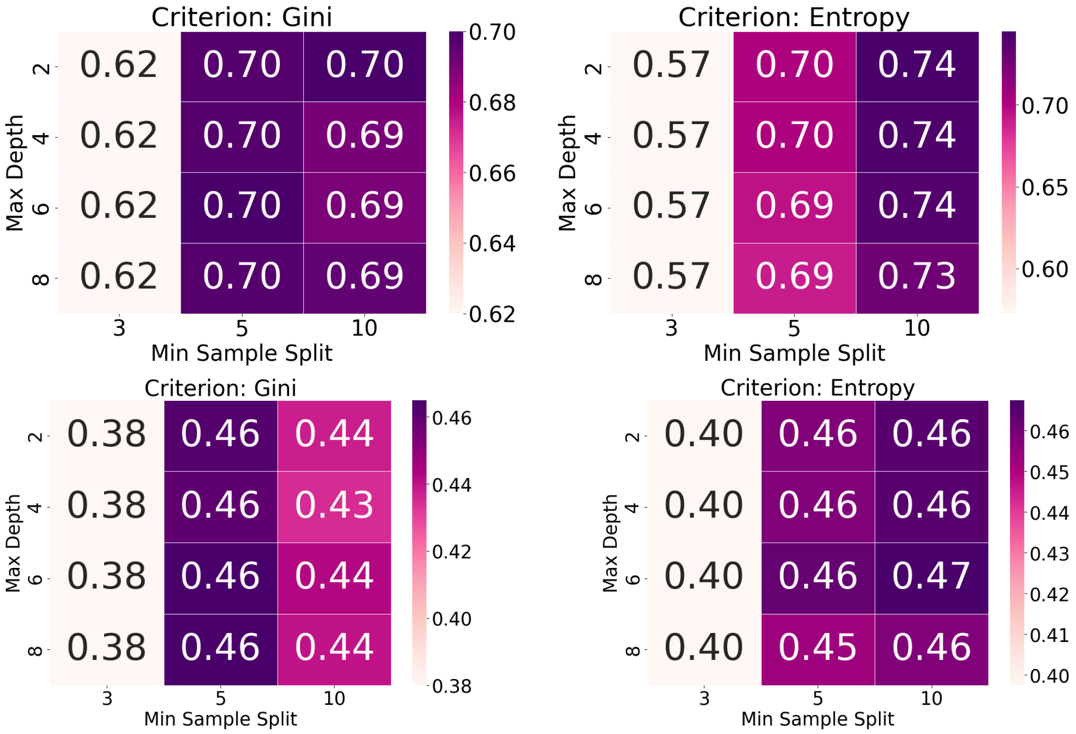

Figure 5: Decision Tree Hyperparameter Graphs, Social data above, Governance data below

As depicted in Figure 6, the Random Forest model parameters for social and governance data showed intriguing differences. Both data types benefited from the entropy criterion and a similar tree depth, which suggests a consistent approach to managing model complexity across datasets. However, the number of estimators, representing the total trees in the forest, varied significantly. The social data achieved optimal results with a lower number of estimators, while the governance data required a substantially higher count, indicating the need for more diverse decision-making perspectives to capture the complexity in the governance data.

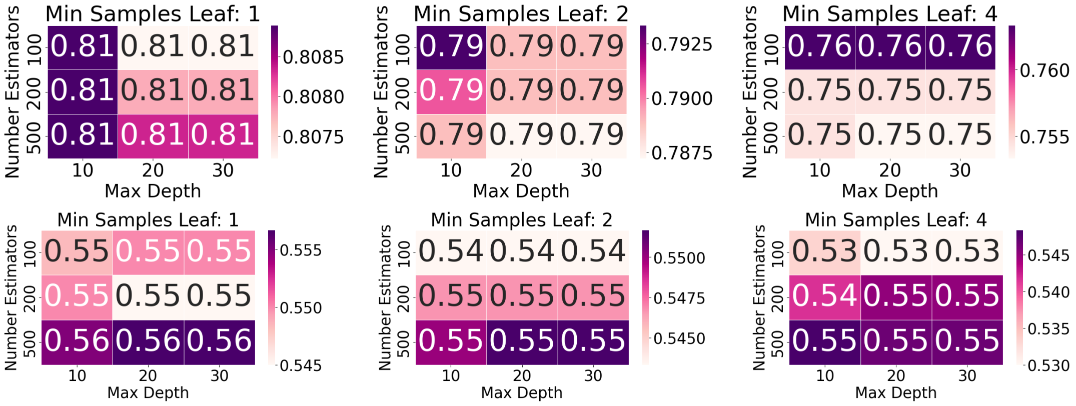

Figure 6: Random Forest Hyperparameter Graphs, Social data above, Governance data below

Figure 7 highlights a distinction in the Gradient Boosting approach between the two datasets. The social data responded best to a learning rate and tree depth combination that allowed for moderate learning progression and complexity. In contrast, while the governance data shared the same learning rate and tree depth, it required adjustments in the minimum samples for leaf nodes and splits. This distinction underscores the different data characteristics, where the governance data perhaps demanded a more cautious approach to splitting, avoiding overfitting while still capturing essential patterns.

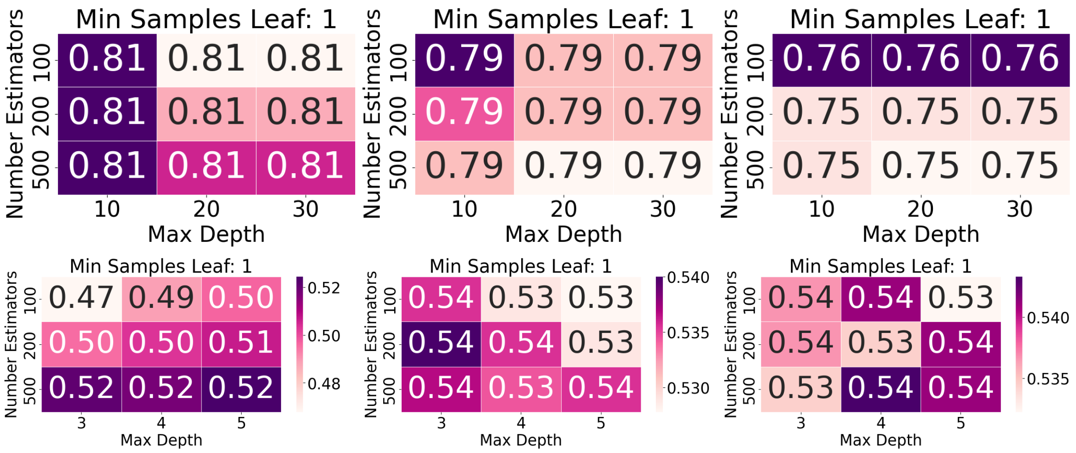

Figure 7: Gradient Boost Hyperparameter Graphs, Social data above, Governance data below

For completion, find below in Table 3 all the optimal hyperparameter combinations chosen during the 5-Fold Cross Validation and subsequently used in training the models in this thesis.

Table 3: Optimal Hyperparameters for Social and Governance Data

Model	Social Data	Governance Data
Logistic Regression	Regularization (C): 10 Solver: lbfgs	Regularization (C): 100 Solver: liblinear
Decision Tree	Criterion: Entropy Max Depth: 10 Min Samples Leaf: 1 Min Samples Split: 2	Criterion: Entropy Max Depth: 10 Min Samples Leaf: 1 Min Samples Split: 2
Random Forest	Criterion: Entropy Max Depth: 10 Max Features: auto Min Samples Leaf: 1 Min Samples Split: 2 Estimators: 100	Criterion: Entropy Max Depth: 10 Max Features: auto Min Samples Leaf: 1 Min Samples Split: 2 Estimators: 500 (Higher)
Gradient Boosting	Learning Rate: 0.2 Max Depth: 4 Min Samples Leaf: 1 Min Samples Split: 5 Estimators: 100	Learning Rate: 0.2 Max Depth: 4 Min Samples Leaf: 2 Min Samples Split: 10 Estimators: 100

5.2 *Model Performance*

The two tables, Table 4 and Table 5, presented below showcase the accuracy and F1 scores for various machine learning models applied to Governance and Social datasets. In Table 4, while the Logistic Regression shows relatively closer training and validation scores, the Decision Tree, Random Forest, and Gradient Boost models exhibit a significant discrepancy; the training scores for these models are near perfect, with Random Forest and Gradient Boost almost reaching 1, but their validation scores are markedly lower. This is a sign of overfitting, so these models are expected to not generalize effectively to new, unseen data. The Table 5 follows a similar pattern. Again, Logistic Regression does not show signs of overfitting, however the Decision Tree, Random Forest, and Gradient Boost models show extremely high training scores yet low validation scores.

Attempts to rectify this overfitting, such as forceful parameter tuning punishing complexity have been made. However, these attempts led to a worsening of the overall performance, indicating the challenge of finding a balance between model complexity and generalization ability.

The test results for Social and Governance models in Table 6 are consistent with the previously discussed pattern of overfitting, with lower scores in testing compared to training. In both datasets, Random Forest outperforms other models, showcasing its superior capability in handling diverse ESG data. The Social dataset exhibits overall higher accuracy and F1 scores across all models, indicating a relatively clearer pattern or less complexity in the data compared to the Governance dataset, where scores are notably lower.

Table 4: Accuracy and F1 scores for Governance Models

Governance Models				
	Accuracy		F1	
Model	Training	Validation	Training	Validation
Logistic Regression	0.6076	0.4613	0.6050	0.4601
Decision Tree	0.9854	0.4753	0.9864	0.4723
Random Forest	1	0.5705	1	0.7610
Gradient Boost	.9986	0.5436	0.9978	0.7235

Table 5: Accuracy and F1 scores for Social Models

Social Models				
	Accuracy		F1	
Model	Training	Validation	Training	Validation
Logistic Regression	0.8497	0.7056	0.8475	0.7045
Decision Tree	0.9970	0.7638	0.9970	0.7504
Random Forest	1	0.8472	0.1	0.8155
Gradient Boost	1	0.8256	0.9998	0.8145

Table 6: Test Results for Social and Governance Models

Models	Social Test		Governance Test	
	Accuracy	F1 Score	Accuracy	F1 Score
Logistic Regression	0.6936	0.6951	0.5135	0.5063
Decision Tree	0.66	0.65	0.50	0.5021
Random Forest	0.8234	0.7923	0.5765	0.5762
Gradient Boost	0.8018	0.7845	0.5585	0.5486

6 CONFUSION MATRICES OF MODELS

The following figures illustrate the confusion matrices for the Logistic Regression, Decision Tree, and Random Forest models, respectively. These matrices provide insight into the performance of each model with respect to the true classifications versus the predicted classifications.

Beginning with the Logistic Regression Confusion Matrix, as shown below in Figure 8, for the Social data (left) indicates that the logistic regression model performs well in classifying class C for social data but confuses classes A and B to a higher degree.

For Governance data (right) it shows a more even performance across classes A, B, and C, yet it struggles with class D, suggesting that perhaps additional features may be necessary to improve classification accuracy for this category.

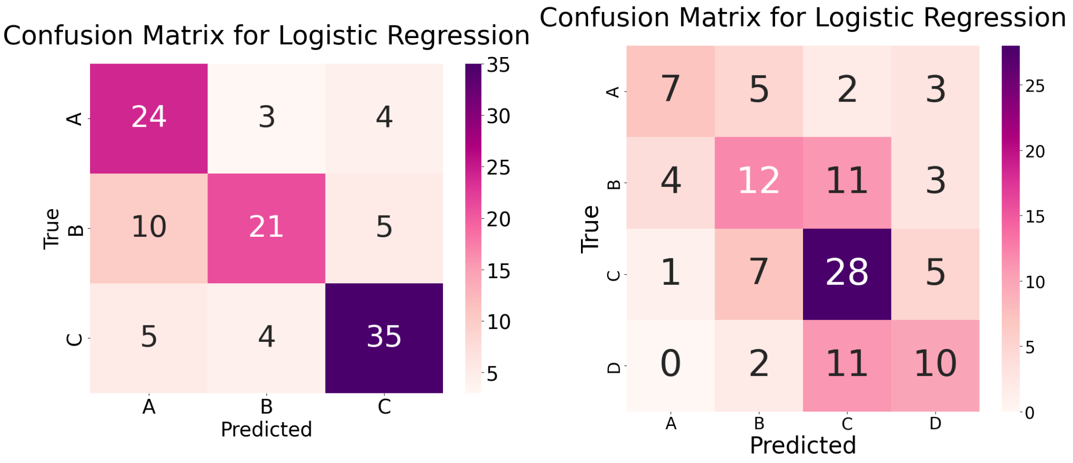

Figure 8: Confusion matrix for the Logistic Regression model, Social data on the left and Governance data on the right.

The Decision Tree Confusion Matrix, Figure 9, reveals a quite balanced but not very accurate classification for social data, as all classes have many mistakes. For governance data the model performs the worst in classes B and C, and performs best for class D. Surprisingly, class A gets confused often for class C in this model, which does not happen in other confusion matrices or in social data.

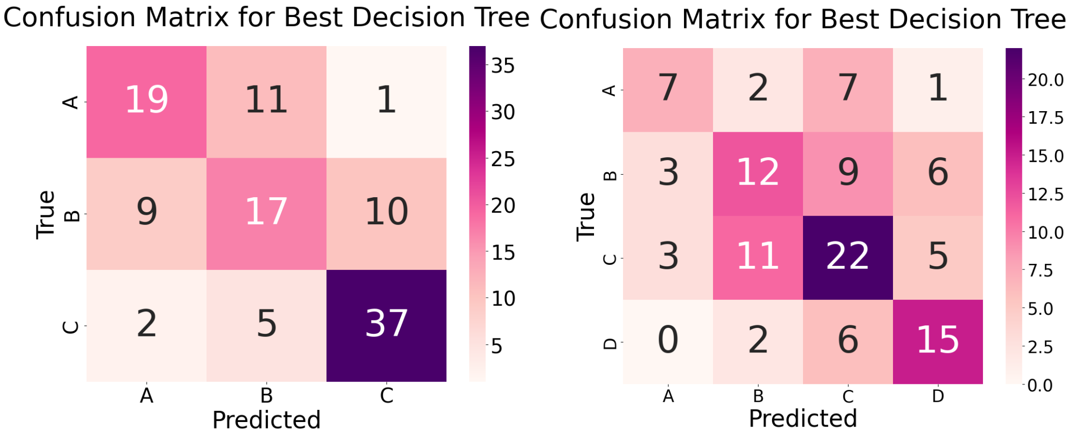

Figure 9: Confusion matrix for the Decision Tree model, Social data on the left and Governance data on the right.

Figure 10 depicts the Random Forest Confusion Matrix. Social data performed quite well across classes, while for Governance data it now struggled with class B but improved significantly in class C. Overall, it is a notable improvement from previous confusion matrices, specially regarding the upper right and lower left squares, where the most severe mistakes are located, there is only 2 missclassifications for the social data and none for the governance.

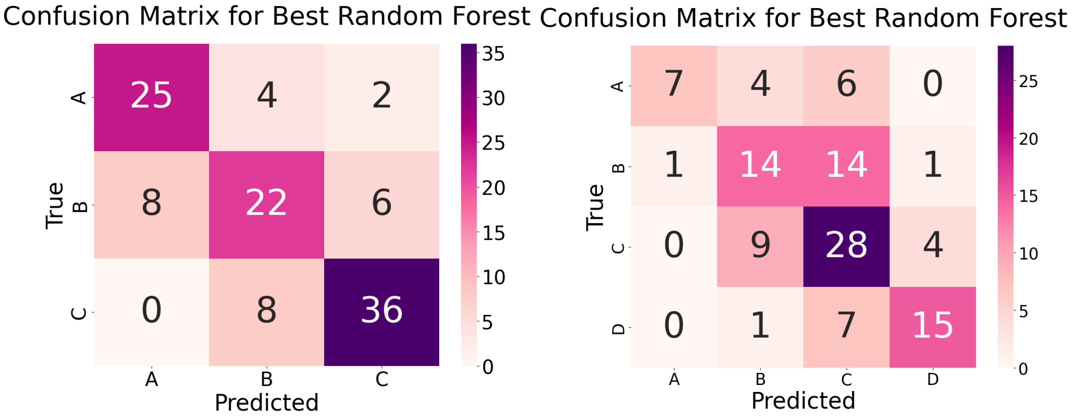

Figure 10: Confusion matrix for the Random Forest model, Social data on the left and Governance data on the right.

Figure 11 depicts the Gradient Boosting Confusion Matrix. While it performed very similarly to the Random Forest model, albeit overall slightly worse. Most importantly, it does make more severe mistakes, as we see four class A's being classified as C for social data as well as two extreme mistakes in the Governance data.

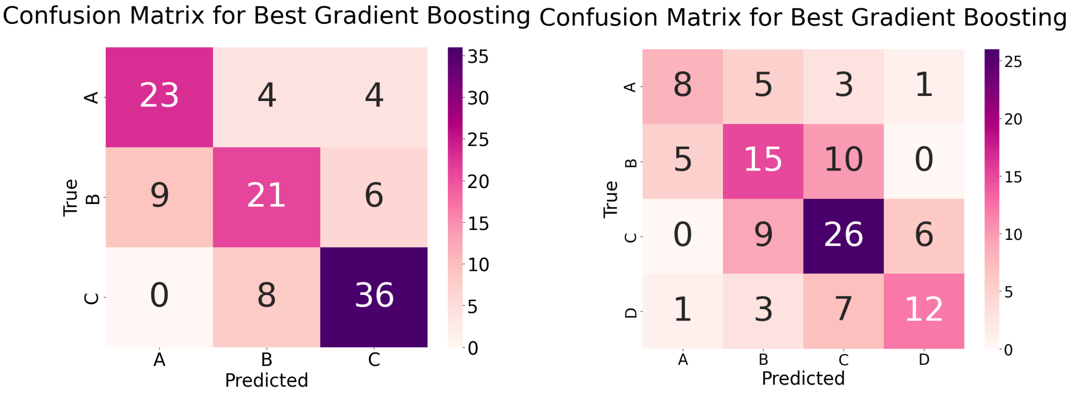

Figure 11: Confusion matrix for the Gradient Boosting model, Social data on the left and Governance data on the right.

6.1 *SHAP Value Analysis*

SHAP values were computed to interpret the models' predictions, since Random Forest is not an easily interpretable model. Figure 12 and Figure 13 present the SHAP value for the aggregated classes, indicating which features have the most significant impact on model predictions. Individual graphs per each class can be found in the Appendix, belonging to Figure 17 to 19 for Social data and 20 to 23 for Governance data. As it can be seen below, figures that contribute the most to the overall accuracy of the model vary per class, not all features hold the same importance in each class. In the Discussion section of this thesis, a more comprehensive analysis of the key features identified through SHAP values will be provided. It's important to note that the data presented here represents the absolute mean SHAP values, which do not indicate whether a feature correlates positively or negatively with good ESG performance. Consequently, this ranking should be interpreted as a guide to identifying crucial features rather than a definitive checklist indicating the direction of their contribution to a firm's ESG classification.

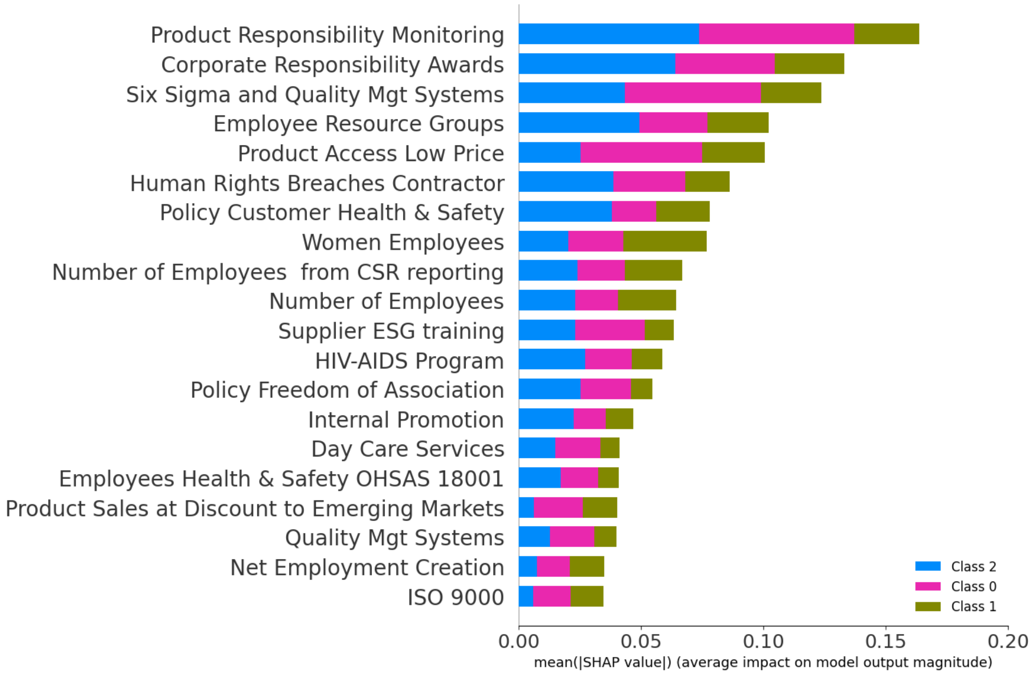

Figure 12: Absolute Mean SHAP Values for Social data aggregated for all classes. Note class A is 0, B is 1, C is 2.

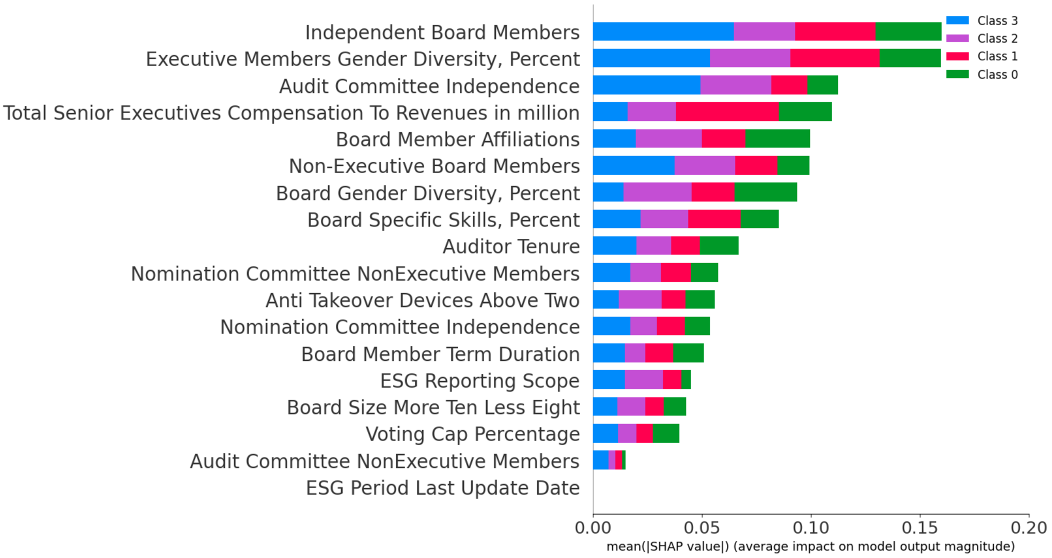

Figure 13: Absolute Mean SHAP Values for Governance data aggregated for all classes. Note class A is 0, B is 1, C is 2 and C is 3.

6.2 *Pixel Flipping Experiment*

The pixel flipping experiment, as depicted in Figure 14 below, presents insightful data on the area under the curve (AUC) for both SHAP values and randomly shuffled feature curves as a baseline. For Social data focusing on accuracy, the SHAP AUC is approximately 0.537, while the corresponding value for the randomly shuffled curve is about 0.623. In terms of F1 scores for Social data, the SHAP AUC stands at roughly 0.474, with the shuffled feature curve showing a higher value of approximately 0.602. For Governance data, a similar pattern emerges. The accuracy metric yields a SHAP AUC of about 0.218, with the shuffled feature curve at approximately 0.265. The F1 score for Governance data shows a SHAP AUC of roughly 0.312 and a shuffled curve AUC of about 0.442. This trend, where the randomly shuffled curves have higher AUC values than the SHAP values, suggests SHAP is doing a good job, as the smaller the area under the curve the better.

In pixel flipping experiments, one typically expects to see a downward sloping or plateauing curve. However, the presence of peaks in these curves, as observed in this analysis, might indicate a potential miscalculation in the SHAP values. This irregularity suggests that the SHAP analysis may not have perfectly captured the feature importance or the interactive effects between features, calling for a cautious interpretation of these results.

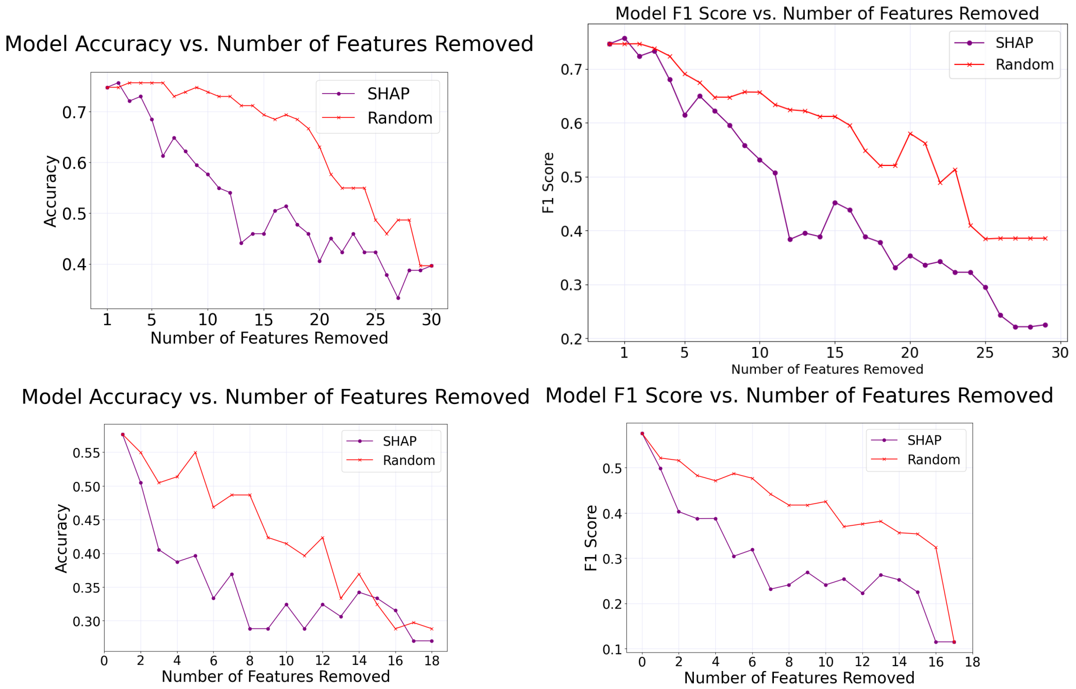

Figure 14: Pixel Flipping Experiment curves. Above, for Social Data, and below, for Governance Data.

7 DISCUSSION

In this section will discuss the results presented in the previous section, and attempt to answer the proposed research questions for this thesis.

7.1 *First Sub-Question: how accurately can we predict ESG performance using machine learning models?*

The initial baseline established with Logistic Regression was notably outperformed by the other models under consideration: Decision Tree, Random Forest, and Gradient Boosting. When ranked based on performance, the models align in the order of Random Forest, Gradient Boost, Decision Tree, and finally, Logistic Regression. It is observed that the performance of Random Forest and Gradient Boost is closely aligned, indicating only a subtle improvement in the former over the latter. Similarly, Decision Tree and Logistic Regression exhibited comparable levels of performance. This ranking suggests that ensemble methods like Random Forest and Gradient Boosting tend to outperform simpler models like Decision Trees and Logistic Regression in this context. This was to be expected as ESG data presented complex, imbalanced, non-linear data, it unlikely that rather complex models could handle it well. This finding aligns with current academic literature, Bentéjac et al. (2021) found Gradient Boosting and Random Forest were among best methods to use across a wide range of differently imbalanced datasets. Furthermore, Murorunkwere et al. (2023) found Random Forest to outperform all other tree-based models in his study on detection of tax fraud. More importantly, this thesis findings corroborate Chowdhury et al. (2023) which also found Random Forest to outperform other models for ESG performance classification problems. This could be attributed to their ability to capture complex patterns in the data, a feature particularly beneficial in ESG (Environmental, Social, and Governance) data analysis.

The two metrics chosen to evaluate the models, accuracy and F1 score, yielded consistent results across the models, reaffirming the reliability of the performance assessment. In their study in tax fraud, Murorunkwere et al. (2023) also found accuracy and F1 to agree on which were the best performing models, and findings in Chowdhury et al. (2023) further corroborate this. Notably, these results were consistent across different types of data - social and governance. In her paper, Sokolova and Lapalme (2009) addresses which scenarios could cause F1 and accuracy scores to be very similar; when the dataset is fairly balanced, or if the model is effectively addressing class imbalance, the accuracy and weighted F1 scores. The cost-matrix used in this thesis alongside the subtle class imbalance could conduce to those two measures being closely similar.

However, a stark contrast was observed in the scores of accuracy and F1 between the Social and Governance data, with the former exhibiting significantly higher scores. This disparity could be attributed to the comparative simplicity of the social data, characterized by fewer features and a simpler 3-class classification problem, as opposed to the more complex, feature-rich 4-class classification challenge posed by the governance data.

The fourth class, labeled 'D', deserves special mention. It was formed by grouping various smaller, poorly performing subclasses. This amalgamation likely introduced a higher degree of diversity and complexity within the class, making it more challenging to classify accurately. This observation underscores the nuanced nature of ESG data, where the amalgamation of disparate elements can significantly alter the difficulty of the classification task. On the other hand, because class D encapsulated the far end of the skew, the most dissimilar elements were present in it, making it easier to tell apart. Class B for instance was quite close to class A in the original data, and many overlap characteristics can be found between firms belonging to either of those categories. Ideally, the grouping would be made with less skewed data, where more defined classes with higher separation between them can be formed.

The dataset used in this study comprised just over 300 instances. While this limited size offers an initial insight, training the models on a larger dataset is expected to yield improved results. The small size of the dataset, coupled with the significant number of metrics dropped due to incomplete data (as discussed in the Methodology section), suggests that key insights might have been missed. Additionally, Governance data had more features to begin with, therefore a higher degree of information loss happened, perhaps explaining its poorer performance comparably. This limitation highlights the critical importance of comprehensive and accurate reporting in ESG.

As the field of ESG reporting evolves, with more regulations and standardization, it is anticipated that the quality and completeness of the data will improve. This progress is expected to enhance the accuracy of similar models in future research. Nevertheless, the findings of this study lay down a valuable framework for future research, indicating the potential of machine learning models in analyzing complex and multifaceted ESG data.

7.2 *Second Sub-Question: Which ESG factors play a pivotal role in influencing ESG performance?*

This thesis aimed to understand which ESG factors are pivotal in influencing overall performance by drawing on insights from SHAP results. As discussed in the Methodology section, SHAP analysis offers a nuanced perspective, highlighting how different factors bear varying levels of significance across social and governance data. The Social and Governance results will be discussed separately below:

7.2.1 *Interpreting Social Data*

The Social aspect of business puts the 'S' on ESG, and encompasses a company's management of relationships with employees, suppliers, customers, and the communities where it operates, focusing on issues like human rights, labor standards, and diversity and inclusion. The SHAP results for social data reveal a complex interplay of factors influencing ESG performance. Looking separately at the SHAP rankings for each class one notices that across the board certain factors repeatedly emerge as significant, indicating their overarching influence in shaping ESG outcomes. Notably, Product Responsibility Monitoring stands out as a critical element. This factor's prominence underscores the importance of how companies manage their product-related responsibilities, resonating with the idea that robust product is integral to high ESG performance.

Another consistent factor across classes is the emphasis on Corporate Responsibility Awards. This finding suggests that external recognition of a company's corporate responsibility efforts plays a significant role in shaping its ESG profile. It could be inferred that such awards not only reflect a company's commitment to ESG principles but also enhance its reputation and stakeholder trust. As governments seek to make businesses more ESG-friendly, it is good to highlight that incentive systems seem to work in this regard, offering national or international (for instance at the European Union level), awards and recognition could be a way to stir companies in the right direction in a less coercive way than through obligatory regulations and litigation.

Operational excellence, as indicated by the importance of Six Sigma and Quality Management Systems, also emerges as a key determinant. This aligns with the notion that operational efficiency and adherence to quality standards are crucial for sustainable ESG practices. Furthermore, factors like Employee Resource Groups and Human Rights Breaches Contractor highlight the importance of human-centric approaches and ethical practices in businesses, illustrating a comprehensive view of what constitutes social responsibility in the corporate world.

7.2.2 *Governance Data Insights*

Turning to governance data, the SHAP analysis brings to light a different set of factors. The prominence of Independent Board Members points to the critical role of governance structure in ESG performance. This aligns with governance literature that views independent directors as essential for mitigating conflicts of interest and enhancing decision-making (Author, 2023).

The significance of Executive Members Gender Diversity Percentage aligns with contemporary views on diversity and inclusion, suggesting that gender diversity in leadership is not just a matter of equity but also a potential driver of enhanced corporate performance. This is in line with the extensive research that has been conducted on men and women's different leadership styles, like highlighted in the paper by Bajcar and Babiak, 2019. More points of view, a balance between the need for innovation and consolidation, and a diverse interest focus can lead to better business outcomes.

Audit Committee Independence and Total Senior Executive Compensation to Revenues are also identified as influential factors. These findings echo the growing focus on financial ethics and transparency in governance, indicating that how companies manage their financial oversight and executive remuneration is closely scrutinized in ESG assessments. Lastly, Board Member Affiliation emerges as an influential factor, hinting at the impact of the broader network and affiliations of board members on governance practices.

The SHAP analysis, while revealing, also brought forth certain doubts on the validity of its results. The pixel flipping experiment curves did not display a consistently downward trend, hinting at the complex relationships between features. This could be due to the varying importance of features across different classes or potential limitations in the SHAP methodology.

Nevertheless, when comparing SHAP results against a random shuffle baseline, the analysis outperforms the latter. This is evidenced by the slower descent and larger area under the curve for random shuffle values, affirming that despite some limitations, SHAP provides valuable insights into the dynamics of ESG performance. This trend is true for both the SHAP when showcasing accuracy metric models and also when using F1 as the evaluation metric.

In conclusion, the SHAP analysis has unraveled a rich tapestry of factors influencing ESG performance. It offers a foundational understanding that can be further expanded in future studies, potentially with larger datasets and refined methodologies, to continue exploring the complex landscape of ESG factors in corporate performance

7.3 *Third Sub-Question: What are the optimal feature retention thresholds that maximize ESG performance enhancement while maintaining predictive accuracy?*

The third subquestion of this thesis addresses a crucial aspect of ESG analysis: determining the optimal number of features needed to maintain a desired impact performance metrics (accuracy and F1 scores). For firms and other interested parties, this means they can focus on a reduced set of key features, minimizing the time and resources spent on data collection and analysis. This streamlined approach not only makes ESG reporting more accessible and actionable, but also ensures that attention is concentrated on the most impactful areas By identifying the threshold at which feature removal begins to), the study offers a more cost-effective approach to ESG analysis.

While this framework could potentially be applied to both Social and Governance data, the decision to focus solely on Social data is driven by the relatively low accuracy and F1 scores observed in Governance data. The lower performance metrics in Governance data suggest a higher complexity and variability in factors influencing ESG performance, thus making it less suitable for this analysis. Dropping the accuracy and F1 scores further does not serve any purpose in the Governance case, although it would be interesting to do in the future with better data which presented higher scores in the evaluation metrics.

In the Social data analysis, current values of accuracy and F1 scores are close to 0.8. The objective was to identify how many and which features could be dropped before reaching two chosen thresholds: 75% and 70%. The plot in Figure 15 shows how many features (starting with the ones deemed to contribute less by the SHAP analysis) need to be dropped for the evaluation metric to reach the aforementioned thresholds.

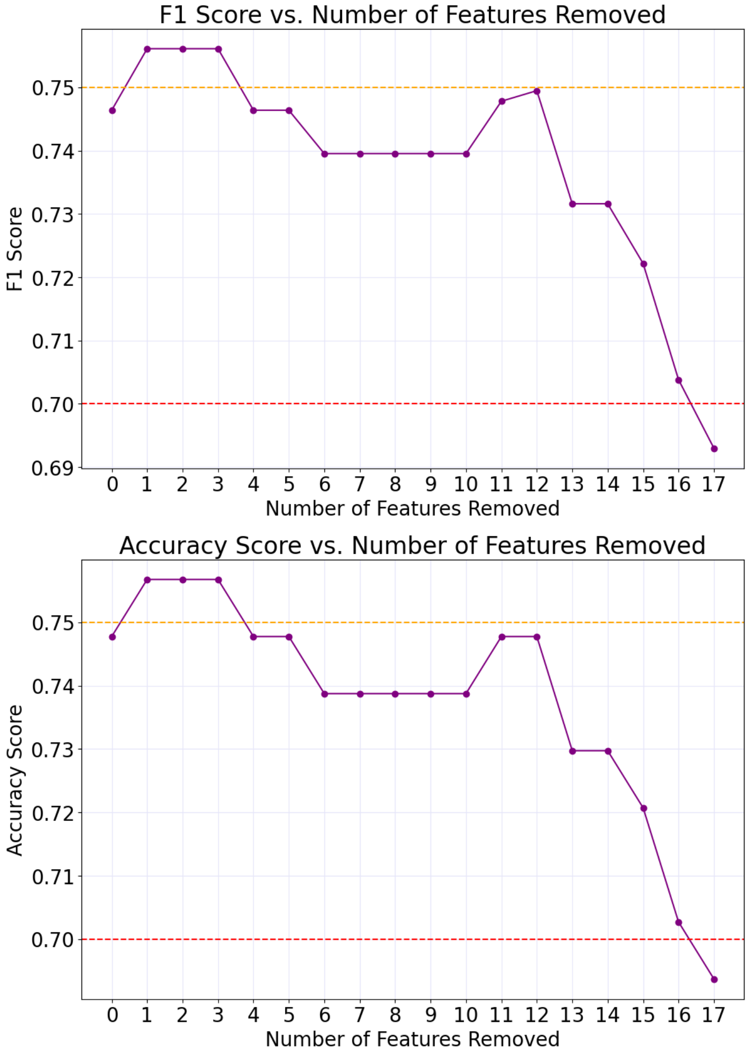

Figure 15: Graphs showing feature reduction and performance thresholds for F1 (above) and accuracy (below) metrics

Analysis revealed that upon removing four features, the 75% threshold for both accuracy and F1 was breached. When extending to sixteen features removed, the performance dropped below the 70% threshold. The features removed included metrics such as 'Announced Layoffs To Total Employees', 'Nuclear 5% Revenues', and 'Improvement Tools Business Ethics', among others. The impact of removing these features indicates their relative importance in maintaining predictive accuracy, underscoring their significance in evaluating social aspects of ESG performance.

7.3.1 *Limitations and Further Research*

The study's limitations primarily revolve around its dataset and scope. The research was based on data from 76 of the largest European firms, which may limit the generalizability of the findings. ESG performance is deeply influenced by regional, cultural, and developmental factors, so the results specific to these large European firms might not universally apply. Future research would benefit from a broader and more diverse dataset, encompassing firms from various regions and developmental stages, to enhance the universality and applicability of the findings. Additionally, the focus on the Social and Governance dimensions of ESG, to the exclusion of the Environmental aspect due to its sector dependency, presents a limitation in the comprehensiveness of the study. Further research could look into replicating the methodology framework but focused on sector-specific Environmental data.

The necessity to extensively clean the data, particularly in dealing with null values, also suggests that future research could benefit from larger, more comprehensive datasets. A silver lining on this front is the evolving legislative landscape mandating more robust ESG reporting. Such developments could provide richer and more unbiased data, enabling deeper insights in future studies. The richer and more complete data will perhaps unveil the importance of features that have now been dismissed, or relationsihps in the data which currently cannot be found. Perhaps difficulties in the thesis like the inconsistencies found through the pixel flipping experiment will resolve themselves in the future with better quality data.

In essence, while this thesis offers significant contributions to the understanding of ESG performance evaluation, its limitations highlight the importance of data completeness and availability.

8 CONCLUSION

In conclusion, this thesis has presented a comprehensive analysis of ESG performance using four machine learning models and subsequently performing a SHAP analysis on the best performing one, the Random Forest. The study was structured around three critical sub-questions, each contributing uniquely to the understanding of ESG performance evaluation and presenting a novel framework on this field.

The first sub-question focused on identifying the most effective machine learning model for ESG performance analysis. Among the four models tested – Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting – Random Forest was found to be the best model, closely followed by Gradient Boosting. This finding highlights the model's robustness and effectiveness in handling complex ESG data, aligning with existing literature on the superiority of ensemble methods in complex classification tasks.

The second sub-question delved into the key ESG factors influencing performance. Through SHAP analysis, the study revealed that factors like Product Responsibility Monitoring and Corporate Responsibility Awards were significant in the social context, while governance data highlighted the importance of Independent Board Members and Executive Members Gender Diversity Percentage. These insights offer a nuanced understanding of the different factors driving ESG performance in various domains.

The third and final sub-question explored the optimal feature retention thresholds in ESG performance enhancement. Focusing on Social data due to its higher predictive accuracy, the study identified key features that, when retained, maintained performance above critical thresholds. This finding make ESG assessment processes more accessible to firms, offering a method to focus on the most impactful aspects of their ESG performance.

Methodologically, the study's application of a rigorous approach in the selection and evaluation of machine learning models to ESG performance as a stand-alone metric, as well as the performance of SHAP analysis to ESG data represents an academic novelty. This thesis establishes a useful framework to delve into into the influence of individual features on model predictions.

Overall, despite its datas' limitations, this thesis contributes to the growing body of knowledge in ESG performance evaluation, offering practical insights for companies and providing a foundation for future research in this rapidly evolving field.

REFERENCES

Argyrides, C., Pradhan, D. K., & Kocak, T. (2009). Matrix codes for reliable and cost efficient memory chips. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, *19*(3), 420–428.

Author, U. (2023). Composition and activity of the board of directors: Impact on esg. *MDPI*. https://www.mdpi.com/2071-1050/12/4/1436

Bajcar, B., & Babiak, J. (2019). Gender differences in leadership styles: Who leads more destructively?

Bates, S., Hastie, T., & Tibshirani, R. (2021). Cross-validation: What does it estimate and how well does it do it? https://ar5iv.org/abs/2104.00673

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*, 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Bisong, E. (2019). *Logistic regression*. Apress. https://doi.org/10.1007/978-1-4842-4470-8_20

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, *39*(3), 3446–3453. https://doi.org/https://doi.org/10.1016/j.eswa.2011.09.033

Brownlee, J. (2023). A gentle introduction to k-fold cross-validation. https://machinelearningmastery.com/k-fold-cross-validation/

Chen, S. (2021). Interpretation of multi-label classification models using shapley values.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Chowdhury, M., Abdullah, M., & Azad, M. e. a. (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

Chung, J., & Michaels, D. (2019). Esg funds draw sec scrutiny. *The Wall Street Journal*.

Commission, E. (2023). *Sustainable finance – environmental, social and governance ratings and sustainability risks in credit ratings*.

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

De Lucia, C., Pazienza, P., & Bartlett, M. (2020). Does good esg lead to better financial performances by firms? machine learning and logistic regression models of public enterprises in europe. *Sustainability*, *12*(13), 5317. https://doi.org/10.3390/su12135317

Gao, S., Meng, F., Wang, W., & Chen, W. (2023). Does esg always improve corporate performance? evidence from firm life cycle perspective. *Frontiers in Environmental Science*, *11*. https://doi.org/10.3389/fenvs.2023.1105077

Khan, M., Serafeim, G., & Yoon, A. (2016). Corporate sustainability: First evidence on materiality. *The Accounting Review*, *91*(6), 1697–1724.

KPMG. (2023). *Esg ratings — the eu's journey to regulation begins*.

Lanza, A., Bernardini, E., & Faiella, I. (2020). Mind the gap! machine learning, esg metrics, and sustainable investment. (561).

Laureti, L., Costantiello, A., & Leogrande, A. (2022). The fight against corruption at global level. a metric approach. https://doi.org/10.2139/ssrn.4315359

Laureti, L., Costantiello, A., & Leogrande, A. (2023). The role of government effectiveness in the light of esg data at global level. https://doi.org/10.2139/ssrn.4324938

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

Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. *R News*, *2*(3), 18–22.

Melas, D., Nagy, Z., & Kulkarni, P. (2016). Factor investing and esg integration. *Journal*.

Misangyi, V., & Acharya, A. (2014). Substitutes or complements? a configurational examination of corporate governance mechanisms. *The Academy of Management Journal*, *57*, 1681–1705. https://doi.org/10.5465/amj.2012.0728

Morningstar Sustainalytics. (2023). The state of esg risk across industries: Three key takeaways from our annual industry reports [Accessed: [insert date of access]].

MSCI ESG Research LLC. (2017). Foundations of esg investing / part 1: How esg affects equity valuation, risk, and performance. *Journal*.

Murorunkwere, B. F., Ihirwe, J. F., Kayijuka, I., Nzabanita, J., & Haughton, D. (2023). Comparison of tree-based machine learning algorithms to predict reporting behavior of electronic billing machines. *Information*, *14*(3), 140. https://doi.org/10.3390/info14030140

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing Management*, *45*(4), 427–437. https://doi.org/https://doi.org/10.1016/j.ipm.2009.03.002

9 APPENDIX

9.1 *Feature importance Analysis*

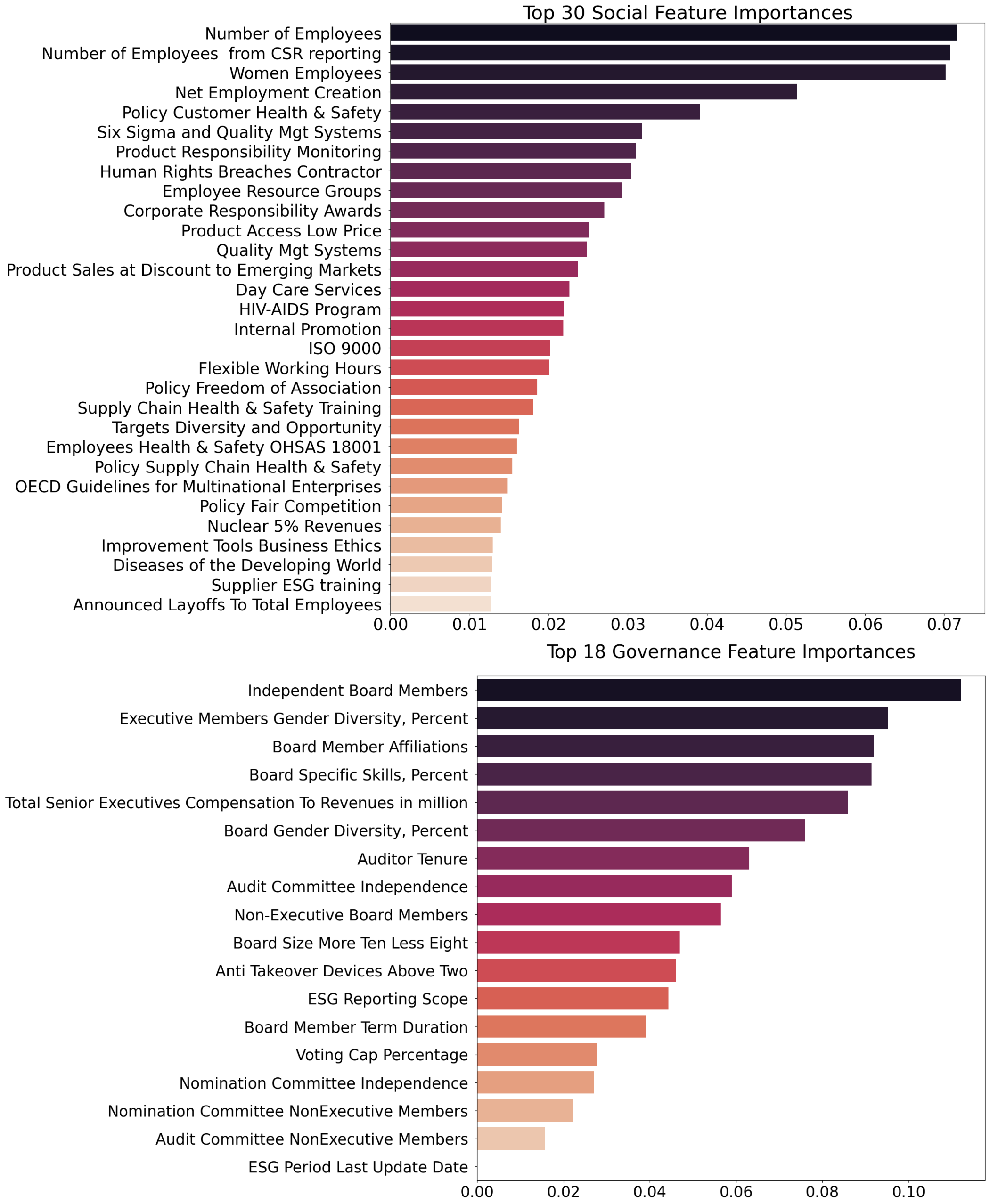

Figure 16: Preliminary Feature Importance Analysis of Social data on the top and Governance data on the bottom made during pre-processing

9.2 *SHAP Results*

Find below the SHAP results for the Social data per individual class.

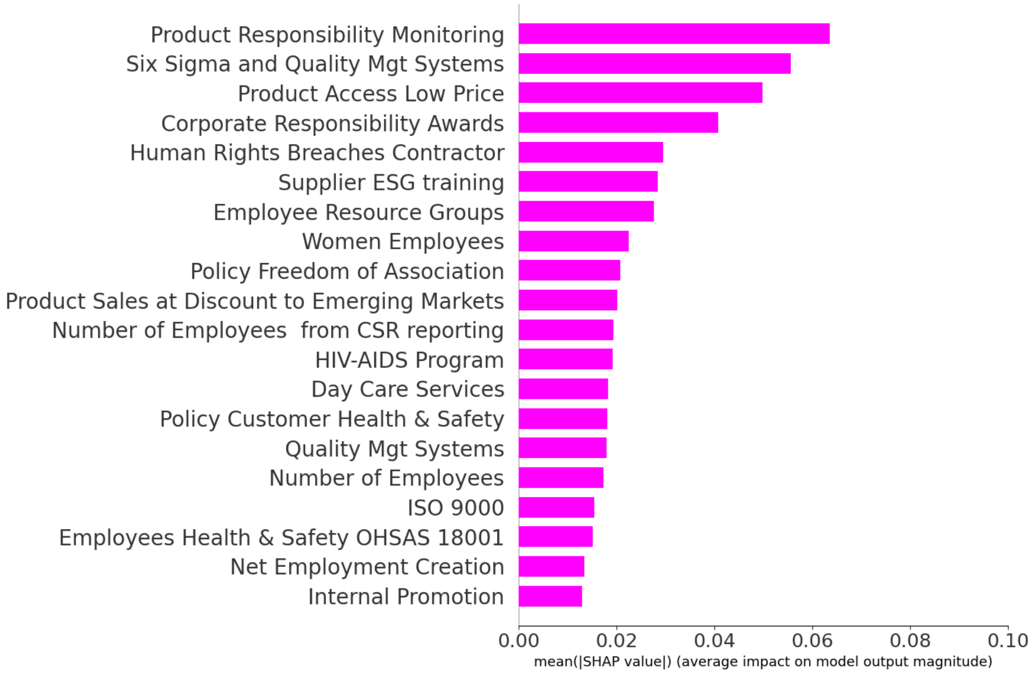

Figure 17: Absolute Mean SHAP Values for Social data for class A.

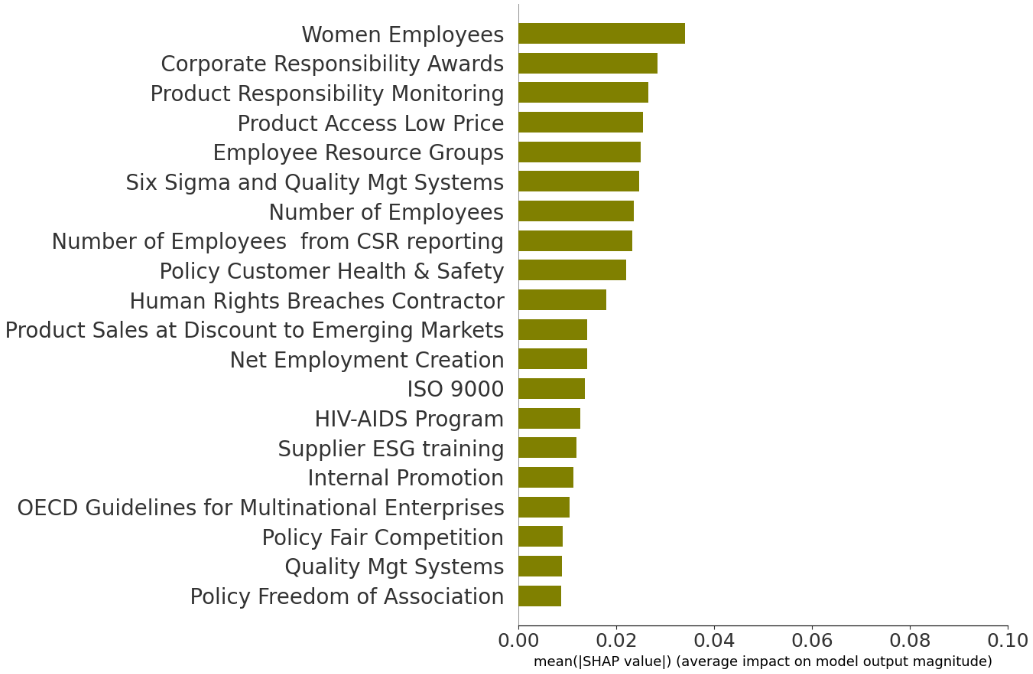

Figure 18: Absolute Mean SHAP Values for Social data for class B.

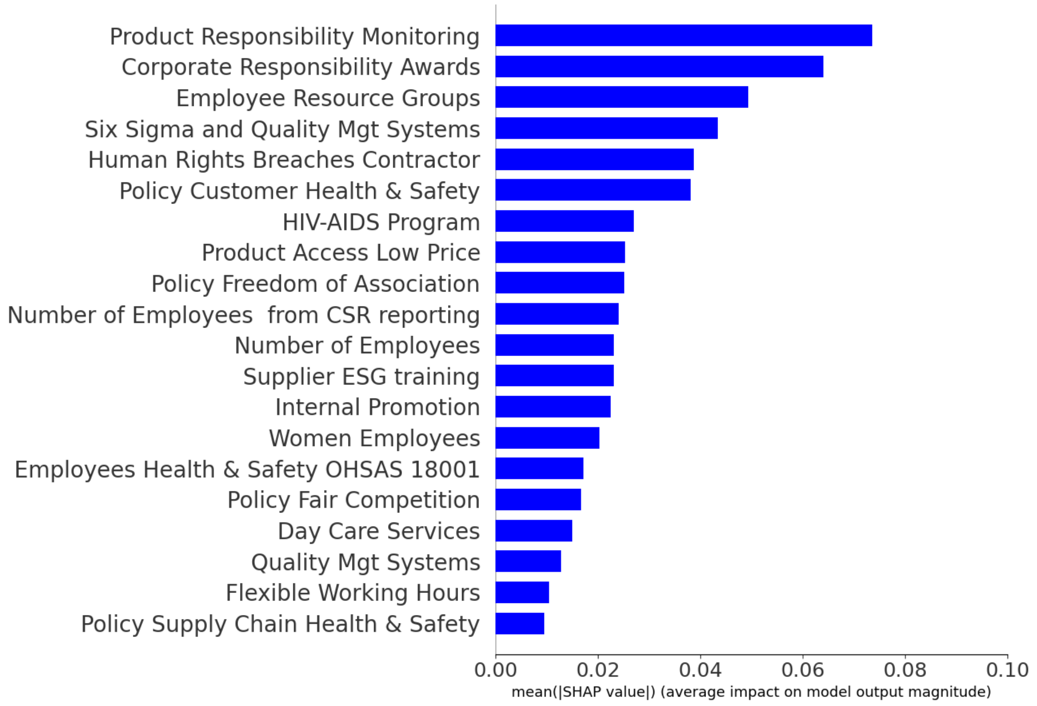

Figure 19: Absolute Mean SHAP Values for Social data for class C.

Find below the SHAP results for Governance data per individual class.

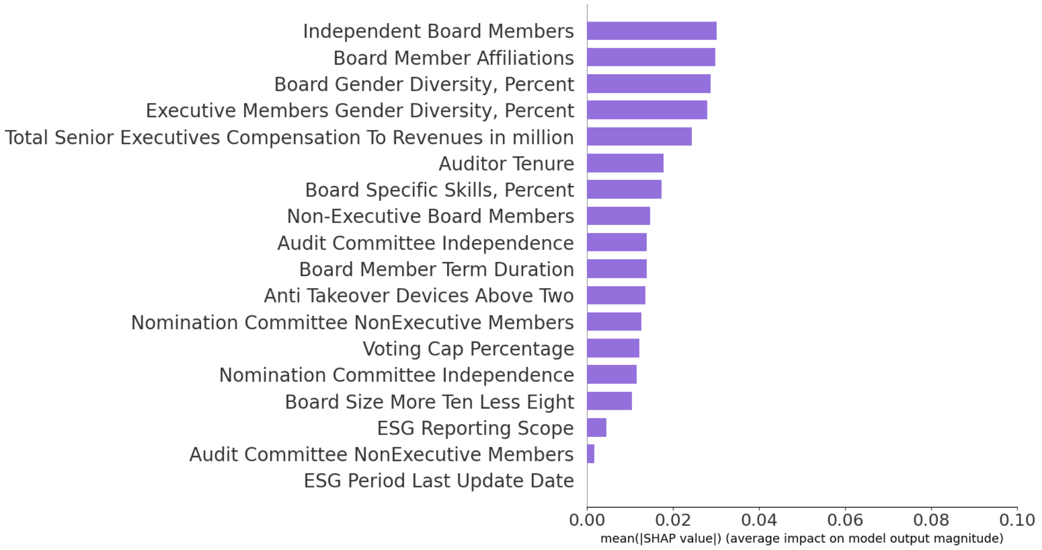

Figure 20: Absolute Mean SHAP Values for Governance data for class A.

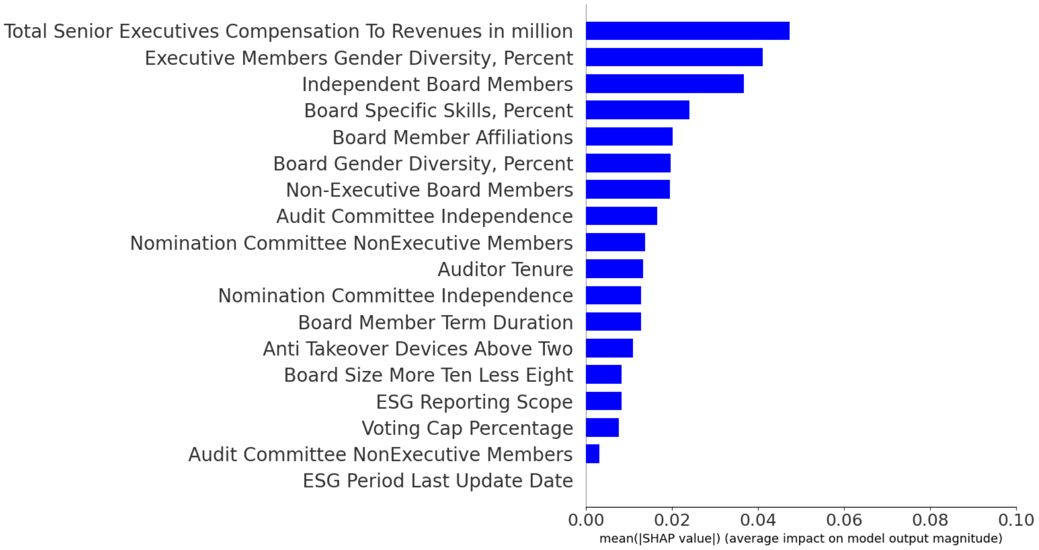

Figure 21: Absolute Mean SHAP Values for Governance data for class B.

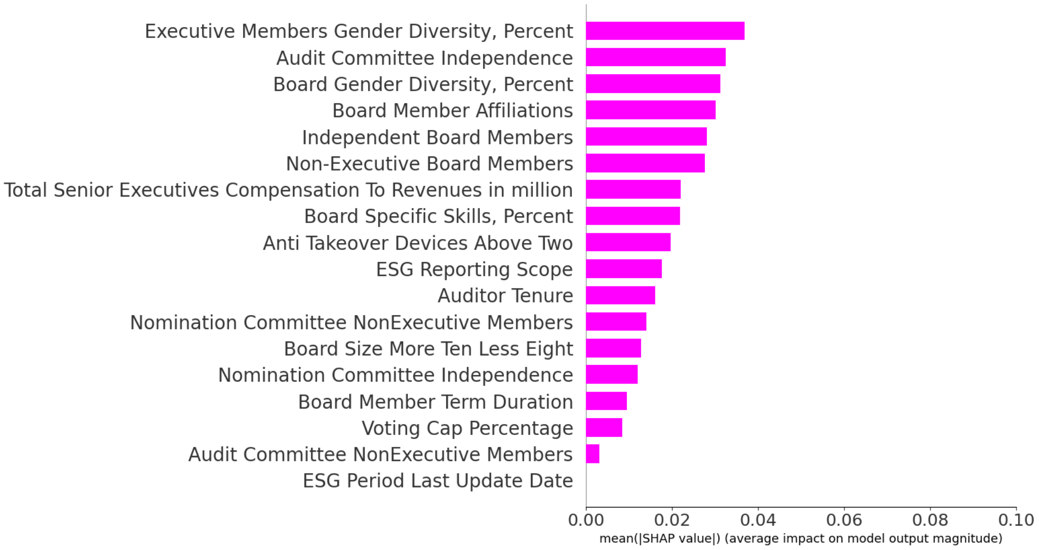

Figure 22: Absolute Mean SHAP Values for Governance data for class C.

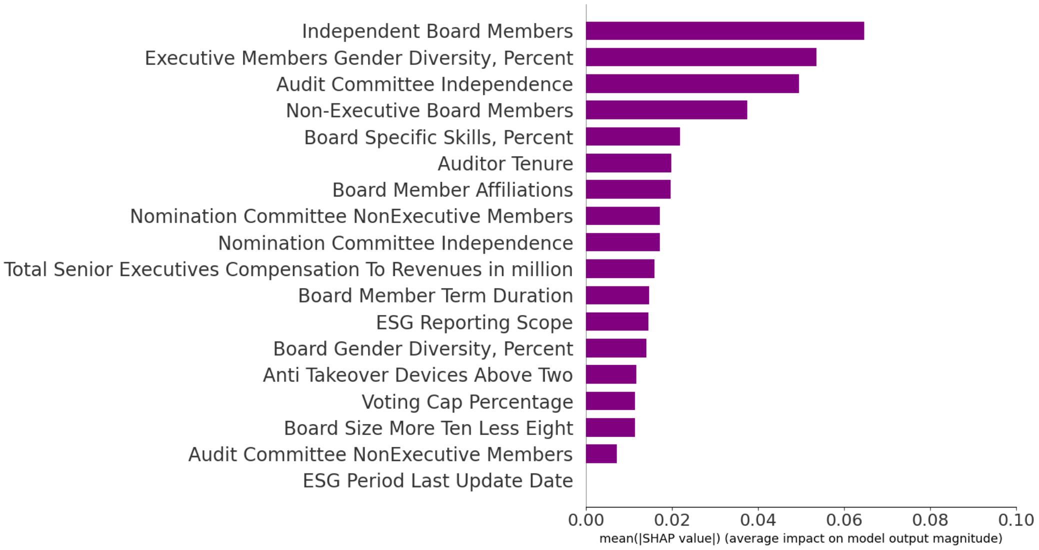

Figure 23: Absolute Mean SHAP Values for Governance data for class D.