Machine learning for ESG data in the financial industry

Jérémi Assael

université PARIS-SACLAY

Machine learning for ESG data in the financial industry

Apprentissage automatique pour les données ESG dans l'industrie financière

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 573: Interfaces: matériaux, systèmes, usages
Spécialité de doctorat: Mathématiques Appliquées
Graduate School: Sciences de l'ingénierie et des systèmes
Référent: CentraleSupélec

Thèse préparée dans l'unité de recherche Mathématiques et Informatique pour la Complexité et les Systèmes (Université Paris-Saclay, CentraleSupélec) sous la direction de Damien CHALLET, Professeur, et la co-supervision de Laurent CARLIER, Head of Data & AI Lab, BNP Paribas CIB Global Markets.

Thèse soutenue à Paris-Saclay, le 19 juin 2023, par

Jérémi ASSAEL

Composition du jury

Membres du jury avec voix délibérative

Antoine MANDEL Professor, Université Paris 1 Panthéon-Sorbonne	Président
Juho KANNIAINEN Professor, Tampere University	Rapporteur & Examinateur
Peter TANKOV Professor, ENSAE, Institut Polytechnique de Paris	Rapporteur & Examinateur
Myriam TAMI Associate Professor, CentraleSupélec, Université Paris-Saclay	Examinatrice

THESE DE DOCTORAT

NNT : 2023UPAST080

université PARIS-SACLAY | ÉCOLE DOCTORALE Interfaces: matériaux, systèmes, usages

Title: Machine learning for ESG data in the financial industry
Keywords: ESG; alternative data; machine learning, interpretability; finance

Abstract: Each of the dimensions of Environment, Social and Governance (ESG) encompasses multiple indicators which provide an ESG profile for a considered company. ESG data are particularly important in the financial industry as they have become a critical tool to evaluate potential investments. Independent data providers assess and aggregate these indicators, but questions remain about the quality and exhaustiveness of the disclosed data as well as the methodologies used for aggregations. Recently, new trends and regulations have resulted in the disclosure of an increasing number of ESG indicators by companies, which resulted in a growing number of available ESG data with quality improvements every year. This thesis proposes solutions to leverage ESG data thanks to carefully adapted machine learning methods and applies them to two case studies relevant to the financial industry. First, using interpretable machine learning, we systematically investigate the relationship between price returns and ESG scores in the European equity market. We show that selected ESG scores can explain a part of price returns not accounted for by classic equity factors. Second, we focus on a specific ESG indicator, greenhouse gas emissions. As greenhouse gas emissions reporting and auditing are not yet compulsory for all companies and methodologies of measurement and estimation are not unified, we propose an interpretable machine learning model to estimate scopes 1 and 2 of these emissions for companies that have not reported them yet. In both cases, a particular emphasis is put on the explainability of the proposed models.

Titre: Apprentissage automatique pour les données ESG dans l'industrie financière
Mots clés: ESG; données alternatives; apprentissage automatique; interpétabilité; finance

Résumé: Chacune des dimensions liées à l'Environnement, le Social et la Gouvernance (ESG) englobe plusieurs indicateurs qui forment le profil ESG d'une entreprise. Les données ESG sont particulièrement importantes dans l'industrie financière: elles sont devenues un outil d'évaluation pour des potentiels investissements. Des fournisseurs de données indépendants évaluent et agrègent ces indicateurs, mais des questions subsistent quant à la qualité et à l'exhaustivité des données publiées et quant aux méthodologies utilisées pour les agréger. Récemment, de nouvelles tendances et réglementations ont conduit à la publication d'un nombre croissant d'indicateurs ESG par les entreprises: ces indicateurs sont de plus en plus disponibles et leur qualité s'améliore chaque année. Cette thèse propose des solutions pour exploiter les données ESG grâce à des méthodes d'apprentissage automatique adaptées et les présente au travers de deux études de cas liées à l'industrie financière. Dans une première partie, à l'aide de méthodes d'apprentissage automatique interprétable, nous étudions la relation entre les rendements de prix et les scores ESG sur le marché européen des actions. Nous montrons que les scores ESG sélectionnés contribuent à expliquer une partie des rendements d'action non expliquée par des facteurs classiques. Dans une seconde partie, nous nous concentrons sur les émissions de gaz à effet de serre, un indicateur ESG. La publication et l'audit des émissions de gaz à effet de serre n'étant pas encore obligatoires pour toutes les entreprises et les méthodologies de mesure et d'estimation n'étant pas unifiées, nous proposons un modèle d'apprentissage automatique interprétable pour estimer ces émissions de scope 1 et de scope 2 pour les entreprises qui ne les ont pas encore publiées. Dans chacune de ces parties, une attention particulière est portée à l'explicabilité des modèles proposés.

Abstract

Each of the dimensions of Environment, Social and Governance (ESG) encompasses multiple indicators which provide an ESG profile for a considered company. ESG data are particularly important in the financial industry as they have become a critical tool to evaluate potential investments. Independent data providers assess and aggregate these indicators, but questions remain about the quality and exhaustiveness of the disclosed data as well as the methodologies used for aggregations. Recent trends and regulatory developments have led to an increase in the disclosure of ESG indicators by companies, which resulted in a growing number of available ESG data with quality improvements every year. This thesis proposes solutions for harnessing ESG data through carefully tailored machine learning methods and applies them to two case studies relevant to the financial industry.

Chapter 1 provides an overview of ESG and highlights its growing significance with the development of regulatory frameworks, particularly within the European Union. This chapter explores various facets of ESG within the financial industry and emphasizes the inherent challenges associated with ESG data, particularly focusing on the divergence of ESG ratings among different data providers. Additionally, the chapter delves into different types of risks associated with ESG, such as transition risks and physical risks. It concludes by underscoring the potential of machine learning to effectively harness ESG data to derive insights and generate value.

Chapter 2 introduces the different machine learning techniques required to effectively exploit ESG data. The chapter begins by providing an overview of machine learning fundamentals pertaining to tabular data and then delves into the specifics of Gradient Boosting Decision Trees algorithms. It concludes by introducing two model-agnostic methods used for interpreting black box models: Shapley values and partial dependence plots.

Chapter 3 investigates the relationship between price returns and ESG scores in the European equity market using interpretable machine learning. We examine whether ESG scores can explain the part of price returns not accounted for by classic equity factors, especially the market one. Thanks to this methodology, we build materiality matrices, showing the most important ESG dimensions broken down by industry and company size. Our findings indicate that the relationship between controversies and price return is the most robust one. The average influence of all the other ESG scores significantly depends on the market capitalization of a company: we find that most of the statistically significantly influential ESG scores weigh negatively on the price returns of small or mid-size companies. Large-

capitalization companies, on the other hand, have significantly advantageous ESG score types. It is important to note that our findings pertain specifically to the Refinitiv ESG dataset employed in this study. Indeed, methodologies to build ESG scores vary across providers and the resulting ESG scores do not necessarily capture the same information. We utilize the MSCI ESG dataset to illustrate the need for caution when generalizing these results to other ESG datasets. The methodology developed to evaluate the relationship between price returns and ESG scores can be applied to different contexts, and we propose various experiments at the conclusion of this chapter to expand upon the framework we have developed.

Chapter 4 is dedicated to the analysis of scope 1 and scope 2 greenhouse gas (GHG) emissions, which are critical and widely used ESG data points. We propose an interpretable machine learning model to estimate scopes 1 and 2 GHG emissions of companies that have not reported them yet. GHG emissions for companies are non-stationary and the quality of reported data can dramatically change from one company to another. By employing suitable machine learning techniques, the resulting models show good out-of-sample performance when assessed globally as well as good and balanced out-of-sample performance when evaluated per sector, country and bucket of revenues. To assess the accuracy of our estimates, we develop a methodology to compare our estimated emissions to those provided by external sources. We found our estimates to be of higher quality. Moreover, our proposed estimations offer better coverage for a broad universe of companies. In the interest of transparency and interpretability, we extensively describe the methodological choices we have made and employ tools based on Shapley values to provide insights into the role played by each feature in the construction of the final estimations. The chapter concludes with multiple additional experiments and robustness checks conducted under alternative settings to further validate our findings and ensure the reliability of our methodology.

In both case studies presented in chapter 3 and chapter 4, we propose a cross-validation methodology allowing the exploitation of ESG data. Indeed, the ESG data used in this thesis is non-stationary, and as its amount, reliability and quality keep increasing, it is imperative to employ robust validation tools to leverage it. We also place a strong emphasis on the interpretability and reproducibility of the machine learning methodologies employed in these studies. This emphasis is crucial in any ESG-related project, as it ensures that the results and insights derived from the analysis can be explained and replicated.

Résumé

Chacune des dimensions liées à l'Environnement, le Social et la Gouvernance (ESG) englobe plusieurs indicateurs qui forment le profil ESG d'une entreprise. Les données ESG sont particulièrement importantes dans l'industrie financière: elles sont devenues un outil d'évaluation pour des potentiels investissements. Des fournisseurs de données indépendants évaluent et agrègent ces indicateurs, mais des questions subsistent quant à la qualité et à l'exhaustivité des données publiées et quant aux méthodologies utilisées pour les agréger. Récemment, de nouvelles tendances et réglementations ont conduit à la publication d'un nombre croissant d'indicateurs ESG par les entreprises: ces indicateurs sont de plus en plus disponibles et leur qualité s'améliore chaque année. Cette thèse propose des solutions pour exploiter les données ESG grâce à des méthodes d'apprentissage automatique adaptées et les présente au travers de deux études de cas liées à l'industrie financière.

Le chapitre 1 offre un aperçu de l'ESG et met en évidence son importance croissante liée au développement de cadres réglementaires, notamment au sein de l'Union Européenne. Ce chapitre explore les différents aspects de l'ESG dans l'industrie financière et met l'accent sur les défis inhérents aux données ESG, en particulier les divergences entre les notations ESG des différents fournisseurs de données. De plus, le chapitre examine différents types de risques associés à l'ESG, tels que les risques de transition et les risques physiques. Il conclut en soulignant le potentiel de l'apprentissage automatique pour exploiter efficacement les données ESG afin d'obtenir plus d'informations et générer de la valeur.

Le chapitre 2 présente différentes techniques d'apprentissage automatique nécessaires pour exploiter efficacement les données ESG. Le chapitre commence par donner un aperçu des fondements de l'apprentissage automatique liés aux données tabulaires, puis approfondit les spécificités des algorithmes de type Gradient Boosting Decision Trees. Il conclut en introduisant deux méthodes indépendantes du modèle utilisées pour interpréter les modèles de type boîte noire: les valeurs de Shapley et les graphiques de dépendance partielle.

Le chapitre 3 examine la relation entre les rendements des prix et les scores ESG sur le marché européen des actions grâce à des méthodes d'apprentissage automatique interprétable. Nous examinons si les scores ESG peuvent expliquer la partie des rendements des prix qui n'est pas expliquée par les facteurs classiques, en particulier le facteur de marché. Grâce à cette méthodologie, nous construisons des matrices de matérialité montrant les dimensions ESG les plus importantes par secteur et taille d'entreprise. Nos résultats indiquent que la relation entre les con-

troverses ESG et le rendement des prix des actions est la plus robuste. L'influence moyenne de tous les autres scores ESG dépend significativement de la capitalisation boursière de l'entreprise: nous constatons que la plupart des scores ESG ayant une influence statistiquement significative ont un impact négatif sur les rendements des prix des petites ou moyennes entreprises. En revanche, les entreprises à grande capitalisation boursière ont des profils de scores ESG nettement plus avantageux. Il est important de noter que ces résultats concernent spécifiquement les données ESG fournies par Refinitiv. En effet, les méthodologies de construction des scores ESG varient d'un fournisseur à l'autre et les scores ESG de différents fournisseurs ne capturent pas nécessairement les mêmes informations. Nous utilisons les données ESG de MSCI pour illustrer la nécessaire prudence lors de la généralisation de ces résultats à d'autres ensembles de données ESG. De plus, la méthodologie développée pour évaluer la relation entre les rendements des prix et les scores ESG peut être étendue à d'autres contextes. Nous proposons diverses expériences à la fin de ce chapitre pour l'illustrer.

Le chapitre 4 est consacré à l'analyse des émissions de gaz à effet de serre de scopes 1 et 2, des points de données ESG particulièrement importants et largement utilisés. Nous proposons un modèle d'apprentissage automatique interprétable pour estimer les émissions de gaz à effet de serre de scopes 1 et 2 pour les entreprises qui ne les ont pas encore publiées. Les émissions de gaz à effet de serre des entreprises sont des points de donnée non stationnaires et la qualité des données publiées peut varier considérablement d'une entreprise à l'autre. En utilisant des méthodes d'apprentissage automatique appropriées, les modèles obtenus présentent de bonnes performances lorsqu'ils sont évalués globalement et lorsqu'ils sont évalués par secteur, pays et catégorie de revenus. Pour évaluer la qualité de nos estimations, nous proposons une méthodologie pour comparer nos estimations d'émissions de gaz à effet de serre avec celles fournies par des sources externes, et nous constatons que les nôtres sont de meilleure qualité. De plus, notre modèle est capable de fournir des estimations pour un large univers d'entreprises. Dans un souci de transparence et d'interprétabilité, nous décrivons en détail les choix méthodologiques que nous avons effectués et utilisons des outils basés sur les valeurs de Shapley pour fournir des explications sur le rôle de chaque variable d'entrainement dans la construction des estimations finales. Le chapitre se conclut par de multiples expériences supplémentaires et des études de robustesse effectuées dans des contextes alternatifs afin de valider nos résultats et garantir la fiabilité de notre méthodologie.

Dans les deux études de cas présentées dans le chapitre 3 et le chapitre 4, nous proposons également une méthodologie de validation croisée permettant d'exploiter les données ESG. En effet, les données ESG utilisées dans cette thèse ne sont pas stationnaires, et à mesure que leur quantité, leur fiabilité et leur qualité augmentent, il est impératif d'utiliser des outils de validation robustes afin d'en tirer

profit. Nous accordons également une grande importance à l'interprétabilité et à la reproductibilité des méthodologies d'apprentissage automatique utilisées dans ces études. Ces dimensions sont cruciales dans tout projet lié à l'ESG car elles garantissent que les résultats et les conclusions tirés de ces études peuvent être expliqués et reproduits.

Acknowledgements

My deepest gratitude goes to Damien Challet, my academic PhD supervisor who has guided me through this thesis for more than three years, who has taken the time to train me with all the academic dimension a PhD thesis involved and who was always available to discuss and to recommend new ideas on the different scientific topic we tackled.

I am extremely grateful to Laurent Carlier, my industrial PhD supervisor at BNP Paribas and head of the Data & AI Lab Global Markets, for providing me almost four years ago with the opportunity to work as a PhD student in his team. Laurent was always keen on discussing further the work done in this thesis and had always propositions to make it move forward. Many thanks for his incredible enthusiasm!

I am also thankful to all of my colleagues in the Data & AI Lab Global Markets, always happy to help, and who provided a motivating and rewarding working atmosphere. Special thanks go to the Paris team with whom I had the privilege of working alongside every day: Baptiste Barreau, William Benhaim, François-Hubert Dupuy and, more recently, Hamza Bodor, Tanguy Colleville, Sarah El Beji, Ashraf Ghiye, Zhen Li, Marouane Maachou and Claire Noot.

I am grateful to the members of the MICS Laboratory at CentraleSupélec, Université Paris-Saclay, for providing me with the opportunity to work with them and particularly to my fellow PhD students Vincent Ragel and Mohammed Salek.

Special thanks to Thibaut Heurtebize, from BNP Paribas Asset Management, for all the thrilling ESG-related discussions we had, his great help as well as his availability.

Many thanks to Aurélie Gonzalez, Jérôme Gava and Julien Turc from BNP Paribas Global Markets, without who this thesis would not have been possible as they have helped me gain access to most of the ESG data used in this thesis. Without them, I would still be struggling to gather it.

I could not have undertaken this thesis without all the people who surrounded me and supported me in my personal life during that time as a PhD student and who were always available for me: my parents, Sabine and Jean-Marc Assael, my brother Alexis Assael, my sister-in-law, Laura Soucheleau Assael and my friends Vincent Auriau, François-Xavier Chamoulaud, Louis de Bellevue and Agathe Llorens.

My thanks also go to all the many colleagues with who I had the pleasure of collaborating in BNP Paribas and who participated in creating such a pleasant and productive working environment.

Contents

1 - ESG in the financial industry

1.1 . ESG definition

Environment, Social and Governance, commonly referred to as ESG encompasses environmental, social and governance issues within a company. They may have an impact on its performance. Each of the three dimensions of ESG can encompass many indicators, giving the ESG profile of the considered company. The environment dimension focuses on areas including but not limited to climate change, waste management, preservation of resources and biodiversity. The social dimension encompasses criteria such as human rights, consumer protection, workforce health and safety, workforce training. The governance dimension includes the assessment of the independence of the board, business ethics or anti-bribery plans. ESG has become in a few years critically important. Financial and non-financial organizations are required to incorporate ESG criteria into their decision-making processes and overall strategies, including investment strategies when applicable. This necessity arises from the need to evaluate companies' activities and investments, the growing interest from investors, the emergence of new regulations and the obligation to communicate transparently about their sustainable approach.

Indeed, the integration of ESG criteria in the strategy and operations of a company is meant to answer the need to fight climate change, for which humans are responsible, by creating a more sustainable business landscape. If nothing is done, the increase in average temperatures could lead to important natural disasters with dramatic consequences. The path towards a more sustainable economy, taking into account ESG performance in addition to financial performance, is meant to lower these physical risks but the transition can have impacts on some activities and assets, with new regulations, obsolete businesses and change of behavior from the population. The urge for actions that leads to taking into account these ESG scores is shifting the ESG landscape from a voluntary framework to a regulatory one.

For a financial institution, adapting itself to this new landscape requires good ESG assessments of companies and financial products, making access to good quality ESG data as well as the capacity to leverage them in meaningful ways a real competitive advantage.

1.2 . ESG reporting and ESG investments: from a voluntary to a regulatory framework

While particularly important, ESG principles and frameworks lack harmonization and convergence. In the past years, frameworks such as the Global Reporting Initiative (GRI) or the Sustainability Accounting Standards Board (SASB) have been developed to create standardization. They were mostly used on a voluntary basis by companies wishing to report on their ESG performance. Given the rising importance of ESG, regulators have seized the topic and we are currently moving from a voluntary to a regulatory landscape.

In Europe, the implementation of the ecological transition has become a priority. Regulators are setting new frameworks to develop ESG standards and harmonize practices within and outside the European Union (EU). Although the EU remains a leader in this transition and the establishment of ESG regulations, it is a global trend: for instance, volumes of sustainable debt issuance were above $1.6 trillion in 2021, more than doubling 2020's end-of-year value according to Bloomberg (BloombergNEF , 2022). While most of the issuers are in Europe, all regions in the world issued sustainable debt.

To successfully tackle the transition towards a low-carbon economy, the European Green Deal was introduced by the European Commission in December 2019 (European Commission, 2019), with three main dimensions:

- Objectives: by 2050 the EU should be the first climate-neutral continent, with no uncompensated greenhouse gas emissions (GHG). To achieve this objective, the EU works at implementing regulations to cut pollution, protect all lives and help companies become world leaders in clean products and technologies while trying to ensure a just and inclusive transition.
- Strategy: the EU sets a milestone in 2030 to reduce net GHG by 55% compared to 1990 levels. All sectors are to take action to meet this target.
- Funding: the EU sets direct flows towards low emissions investments. The European Commission estimates at € 1 trillion the necessary private investments in the next 10 years for a successful transition.

The European Green Deal leads to an ESG regulatory framework, whose objectives are to prevent greenwashing, increase transparency and disclosure and ensure ESG data comparability. Greenwashing is defined by the EU Taxonomy as the practice of gaining an unfair competitive advantage by marketing a financial product as environmentally friendly, when in fact basic environmental standards have not been met (European Parliament and Council of the European Union, 2020). The aim is to encourage sustainable investments by providing investors with appropriate ESG information on their investment choices. This regulatory framework includes several directives and standards, such as the EU taxonomy, the Corporate Sustainability Reporting Directive (CSRD) (European Parliament and Council

of the European Union, 2022) and the Sustainable Finance Disclosure Regulation (SFDR) (European Parliament and Council of the European Union, 2019).

First, the EU Taxonomy, originally published in June 2020, is a common classification system, that enables the identification of economic activities that substantially contribute to one of the following six environmental objectives: climate change adaptation, climate change mitigation, biodiversity and ecosystems, circular economy, pollution and water. It is a transparency tool used by companies and investors to disclose to what extent their activities are sustainable. It also enables the design of credible green financial products. Let us note that while it gives a framework to recognize sustainable investments, it does not legislate on mandatory investments in the ecological transition.

The EU Taxonomy will evolve with time. Currently, only the two environmental objectives related to climate are covered by the EU Taxonomy, the four others being under study. While not all economic activities are currently classified by the EU Taxonomy, the number of covered activities should increase in the near future. The idea of a social taxonomy in addition to this environmental one has also been in development.

In practice, financial and non-financial undertakings have to disclose:

- Taxonomy eligibility: if one wishes to report on a specific activity, it has to be currently covered by the EU Taxonomy, with existing screening criteria. Reporting on the EU Taxonomy eligibility is mandatory from 2022.

- Taxonomy alignment: the activity has to make a substantial contribution to at least one out of the six environmental objectives while not significantly harming any of the others. The EU Taxonomy also imposes to comply with minimum safeguards. Non-financial and financial undertakings respectively have to report on their alignment from January 2023 and January 2024.

- Extent of EU Taxonomy alignment: the proportions of turnover, capital expenditures and operational expenditures linked with the considered activity which are aligned with the EU Taxonomy have to be disclosed, from respectively January 2023 and January 2024 for non-financial and financial undertakings. In particular, banks will have to disclose their Green Asset Ratio (GAR), showing the proportion of EU Taxonomy-aligned assets and investments against total assets and investments. To report on the GAR, banks will need accurate data from their clients.

From January 2024 in the EU, the CSRD is expected to require a large universe of companies to disclose a large set of ESG criteria in a standardized way. This will include information on the company's business model and strategy, on its transition, on social matters, on respect of human rights, on board diversity, on ESG risks and opportunities... The CSRD replaces the Non-Financial Reporting Directive (NFRD) of 2014 and will be applied to a larger scope of companies, including

non-European ones having a presence in the EU: it is expected to cover around 50,000 companies in comparison to the around 11,000 covered by the NFRD. Any company meeting two out of the three following criteria will have to abide by the CSRD: turnover larger than 40 million euros, value of assets larger than 20 million euros or number of employees larger than 250.

The SFDR in the EU requires, from January 2023, European financial institutions or non-European ones having a presence in the EU, to disclose ESG information at both entity and product levels. In particular, the SFDR seeks to improve transparency regarding the ESG profile of investment portfolios through a classification under Article 6, Article 8 or Article 9, depending on how ambitious the portfolio is in this regard. Financial institutions will also have to disclose Principle Adverse Impact (PAI) indicators of their investments, a set of mandatory indicators which seeks to show financial market participants the potential sustainability risks involved with these investments. Investors should then have access to more transparent portfolios to make their investment decisions.

The European Union has implemented several regulations to advance toward its ecological transition. We presented here the EU taxonomy, CSRD and SFDR. Many other regulations, labels and frameworks have been voted by the EU such as the Renewed Sustainable Finance Strategy in July 2021 (European Parliament, 2021), aiming at improving the financing of sustainable economic activities, or the European Green Bond Standards (EuGBS) (European Commission, 2023b), a voluntary standard that seeks to raise the environmental ambitions of the green bond market by ensuring that new green bonds will finance activities that are aligned with the EU Taxonomy.

In Asia, some taxonomies were also developed but are not harmonized between countries. The EU and China specifically show some efforts towards obtaining a common taxonomy through an in-depth comparison highlighting commonalities and differences between the EU and China's green taxonomies. In the United States, in March 2022, the SEC proposed a new rule requiring listed companies to disclose climate-related information alongside their financial information (Securities and Exchange Commission, 2022). Overall, there are international efforts towards a minimum disclosure baseline of ESG information.

1.3 . ESG data and indicators

1.3.1 . The ESG data needs in the financial industry

In the financial industry, access to reliable ESG data is particularly important to achieve tangible impacts on sustainability. Without data, a practitioner risks being seen as a greenwasher with only theoretical commitments. Data is needed for regulatory requirements such as reporting green assets and for assessing ESG commitments and controversies. It can also help identify the positive or negative impacts of an activity, build sustainable investment products, and align portfolios

with net zero goals.

ESG data are available at numerous levels, from corporate to transactional data, and can be historic, projected, or real-time. ESG data can be built and stored in-house, obtained through providers or directly from companies' communications following disclosure regulations.

It is very important regarding ESG data to make the distinction between raw indicators and complex scores, often computed with specific methodologies using these raw indicators.

1.3.2 . The divergence of ESG scores according to the chosen data provider

There are, in 2023, many ESG data providers offering ESG scores for companies. These various providers, including Vigeo Eiris (Moody's), Refinitiv, Sustainalytics, or MSCI, often provide their customers with historical data for raw ESG metrics (greenhouse gas emissions, employees training hours, forestation commitments, turnover rates, the inclusion of ESG-linked criteria in the compensation of top management...) as well as global and granular ESG scores built using these raw ESG indicators. In the absence of regulatory modeling standards, each of them has its own methodology and vision of ESG, which can lead to significantly different scores, even though they are based on the same ESG indicators. Some providers base their scores primarily on analyst studies, while others use artificial intelligence or rule-based data-driven methods. Some also use a combination of these approaches. We illustrate the difference in the construction methodology of the different providers by focusing on two examples, the MSCI and Refinitiv methodologies.

The MSCI ESG Ratings model, developed in MSCI ESG Research (2020), evaluates the ESG risks and opportunities that companies in a particular industry face, considering both large-scale trends and the nature of their operations. To do so, a quantitative model is used to identify the material risks and opportunities associated with the industry. Key issues are then assigned to each industry and company. For the environmental and social pillars, key issues are assessed based on the company's exposure to these key issues and how it manages them. Management assessment is impaired by any controversy that has occurred within the last three years, the resulting score decrease being a factor of the nature of the controversy and its potential materiality. Weights associated with key issues are determined by the industry's impact on the environment or society and by the timeline for the materialization of risks or opportunities. The governance pillar score represents an absolute assessment of a company's governance. Each company starts with a perfect score, and deductions are made based on the assessment of key metrics related to selected key issues. To determine the final ESG rating, the weighted average of individual key issue scores is normalized relative to the ESG ratings of industry peers.

Refinitiv uses a different method for constructing ESG scores (Refinitiv, 2020). While analysts are also involved in collecting and standardizing ESG indicators, scores are built using a disclosed formula that ranks companies against each other, leaving less room for the variability of studies between analysts. Refinitiv collects over 400 indicators for all companies and activates between 70 and 170 depending on the company's industry. They are then divided into ten categories used to calculate the pillar scores defined in Fig. 1.1. The formula used to rank companies is first applied to the selected indicators to obtain a score per indicator, and then to the sum of indicators scores to obtain the pillar scores. By assigning to each pillar a weight proportional to the number of indicators available in the market for that pillar, Refinitiv aggregates the pillar scores into Environmental, Social, and Governance ones, and then into an overall ESG score. Refinitiv also constructs a score reflecting the number and severity of controversies experienced by a company in a given year. It is worth noting that companies are compared within the same sector for indicators related to the environment and social issues, and within the same country for governance indicators.

MSCI and Refinitiv methodologies are thus very different and their scores capture different elements of the ESG profile of a company. Questions arise on the comparability of the resulting ESG scores, at various levels of granularity and on the impact of the choice of a particular provider for the desired use case. The difference in providers' methodology and the lack of correlation between resulting ESG scores have already been addressed in the literature by Berg et al. (2022). They propose several experiments on six ESG data providers, exhibiting the correlations between ESG scores according to the selected levels of granularity. They show the different indicators chosen by each provider in each ESG pillar and their estimated weight. As a result, they identified three sources of divergence between the ESG scores of the different providers:

- Scope divergence, where scores are based on different sets of attributes.
- Measurement divergence, where providers measure the same attribute using different indicators.
- Weight divergence, when providers have different views on the relative importance of attributes.

The challenges inherent to ESG data are related to the different methodologies used by data providers and, as well, to companies' communication of their ESG data. Kotsantonis and Serafeim (2019) highlight four main difficulties when studying ESG data:

- The variety and inconsistency of data communication among different companies.

Environment	**Resource Use**	Reduce the use of natural resources and find more eco-efficient solutions by improving supply chain management.
	Emissions	Commitment and effectiveness towards reducing environmental emissions in the production and operational processes.
	Innovation	Reduce the environmental costs for customers, thereby creating new market opportunities through new environmental technologies and processes or eco-designed products.
Social	**Workforce**	Job satisfaction, healthy and safe workplace, maintaining diversity and equal opportunities, development opportunities for workforce.
	Human Rights	Respecting the fundamental human rights conventions.
	Community	Commitment towards being a good citizen, protecting public health and respecting business ethics.
	Product Responsibility	Producing quality goods and services integrating the customer's health and safety, integrity and data privacy.
Governance	**Management**	Commitment and effectiveness towards following best practice corporate governance principles. Composition, remuneration, transparency of the board.
	Shareholders	Equal treatment of shareholders, use of anti-takeover devices.
	CSR Strategy	Integration of social and environmental dimensions into the day-to-day decision-making processes, in addition to economic and financial ones.

Figure 1.1: Refinitiv ESG methodology - Ten pillar scores definition.

- How providers group companies to compare them: by country, by industry, or both. This can significantly affect a company's score, for example, if a region has much stricter environmental regulations.
- Differences in the treatment of missing data by providers. Refinitiv, for example, considers missing data for an indicator to correspond to a score of 0 for that indicator.
- The more an enterprise reveals information about its ESG practices, the more providers will tend to offer significantly different scores, indicating real differences in the interpretation of data and its relative importance.

These data issues raise some very interesting research questions from an ESG perspective. Is it possible to understand the implications of differences in ESG data from different providers? Can we identify the best data and the best level of granularity for a specific task? Should we work with ESG scores or directly with the underlying indicators? Moreover, to use ESG ratings, suitable ESG data processing methods have to be developed, capable of taking into account all the inherent characteristics of this data: many levels of granularity, non-stationary because of evolving laws and communication practices, few historical data but with strong growth, many missing data points, low frequency with typically only one data point per year per company.

For these reasons, some researchers focus on ways to become independent of the different ESG providers by directly using the very sources of providers, such as Corporate Social Responsibility (CSR) reports or any other company's communications and the news. Using directly these documents necessitates leveraging Natural Language Processing (NLP) methods, directly linking an embedding of the relevant part of the documents to the task at hand. For instance, to identify the relevant paragraphs in these documents from an ESG perspective, NLP techniques such as Word2Vec word embedding models, used in Jeunesse et al. (2020) to find out whether a paragraph in a document addresses a UN Sustainable Development Goal, or more advanced transformer-based models, as in Luccioni et al. (2020) or Nugent et al. (2020) to classify ESG controversies, are particularly useful.

While this solution may seem attractive at first, it suffers limitations such as the number of available company reports dealing with ESG topics (some companies may have few documents), and the difficult choice of ESG categories or parameters to select to obtain an embedding. However, it can be interesting to use these NLP approaches in addition to ESG data from providers. These latter data are usually annual and lack reactivity to new situations. Adding controversies data, for example, is interesting as it allows for direct action based on current events. Once the embedding of a controversy is constructed, adding it to the provider's ESG dataset could bring more reactivity to the built models.

1.3.3 . GHG emissions data

Past human activities or "footprints" are commonly held responsible for the current pollution of the environment. The human footprint is measured by how fast humans consume resources and generate waste versus how fast Earth can absorb their waste and generate resources (Wackernagel and Rees, 1998). When it comes to air emissions footprint, greenhouse gas (GHG) emissions are the most widely analyzed.

Climate change and GHG emissions

Evidence shows that our planet has been getting hotter. Pörtner et al. (2022) insist on the current and future increase in temperature and the resulting consequences. This 2022 Intergovernmental Panel on Climate Change (IPCC) report shows the impact of GHG emissions on climate change as it is driven by the higher level of GHG emissions in the atmosphere.

Practically, GHG emissions allow quantifying global warming of the Earth, enabling the calculation of radiative forcing: when this metric is positive, the Earth system captures more energy than it radiates to space (Hansen et al., 2005). The calculation of GHG emissions tends to account for all GHG emissions caused by an individual, event, organization, service, place, or product. It is expressed in units of carbon dioxide equivalent (CO_2-eq), to provide a common scale for measuring the climate effects and global warming impacts of different gases, such as water vapor, carbon dioxide or methane. In the Global Warming Potential (GWP) framework, for any gas, CO_2-eq is calculated as the mass of carbon dioxide (CO_2) which would warm the Earth as much as the mass of that gas.

The annual meetings of the United Nations Climate Change Conference at the World Conferences of the Parties (COP) review the objectives of the global effort to fight climate change. They assess GHG footprints at the global level and gather engagement of countries to limit CO_2 emissions for fighting global warming and its impact on biodiversity. In line with these engagements, new definitions, laws, and methodologies for calculating and limiting these GHG emissions are voted at the country level, creating a new framework applicable to companies, the underlying hypothesis being that a country's emissions are the sum of emissions coming from its inhabitants and its companies.

Measuring GHG emissions for companies

Following engagements to limit climate change, listed and unlisted companies started reporting their emissions in their extra-financial communication. According to Wiedmann et al. (2009), the carbon footprint of a company depends on the total amount of CO_2-eq that is, directly and indirectly, caused or accumulated over the life stages of its products. From the company's point of view, the assessment of its GHG footprint can be useful not only for regulatory or accounting disclosure but also for implementing strategies designed to mitigate and reduce its emissions. All frameworks such as carbon pricing policies, measuring alignment to climate scenario with the Paris Agreement Capital Transition Assessment (PACTA), or moving toward net zero GHG emissions via Net Zero Banking Alliances (NZBA), need a correct GHG emissions baseline. This momentum will be emphasized by the new CSRD coming into force in 2024 for the largest companies and in 2026 for Small and Medium-sized Enterprises (SMEs) in the EU. This directive will also apply to non-European companies. Companies will need to report audited GHG emissions as well as a quantitative pathway and remediation plan to cancel their net emissions.

At this stage, for a company, reporting GHG emissions is either voluntary or mandatory depending on its location and is linked to defined nomenclatures mostly taking into account activity types and company size. The calculation methodology is often defined along with the regulation and specified at the sector level. The heterogeneity of these methodologies can sometimes make comparisons among companies in different countries or sectors difficult and thus creates biases. Moreover, not only may calculation methodologies vary, they are often not even documented in the companies' reports.

In practice, measuring the GHG emissions of a stakeholder requires much more information. To standardize these methodologies of calculation, the GHG Protocol, first published in 2001 (Ranganathan et al., 2015), is used by large companies, the World Business Council for Sustainable Development (WBCSD) and the World Resources Institute (WRI). Even if, in some cases, companies report according to the ISO 14064 standards or the carbon-balance tool used in France, the GHG protocol has become the most widely used methodology in the world to assess GHG emissions. The GHG inventory is divided into three scopes corresponding to direct and indirect emissions:

- Scope 1: sum of direct GHG emissions from sources that are owned or controlled by the company: stationary combustion, e.g., burning oil, gas, coal, and others in boilers or furnaces; mobile combustion, e.g., from fuel-burning cars, vans, or trucks owned or controlled by the firm; process emissions, e.g., from chemical production in owned or controlled process equipment, such as the emissions of CO_2 during cement manufacturing; fugitive emissions from leaks of GHG gases, e.g., from refrigeration or air conditioning units.

- Scope 2: sum of indirect GHG emissions associated with the generation of purchased electricity, steam, heat, or cooling consumed by the company.
- Scope 3: sum of all other indirect emissions that occur in the value chain of the company, including financed emissions via investments.

Most current regulatory standards make reporting on scope 1 and scope 2 mandatory for large companies. Reporting on scope 3 is mostly optional or to be reported later in 2023 or 2024, even if scope 3, also referred to as value chain emissions, is often the largest component of a company's total GHG emissions, especially for some industries such as automakers or financial institutions.

Achieving convergence in emissions calculation methods within a particular industry simplifies not only the computation but also the comparison of emissions between companies. To guarantee data quality of reported GHG emissions of companies, independent bodies, such as the Carbon Disclosure Project (CDP), a not-for-profit charity that runs the global disclosure system, or external auditors in extra-financial CSR reports, are increasingly involved, leading to an increase in convergence of methodologies and controls.

Focus on the financial industry

Abilities to measure GHG emissions properly have several applications in the financial industry. For investment purposes, financial institutions need to be able to aggregate GHG emissions at the portfolio level for several companies and thus need homogeneous methodologies. GHG emissions assessments measure exposure to transition risk (see section 1.4.2) and negative cash flows coming from fines or outflows to competitors with greener footprints. They are useful for fundamental financial analysis and slowly implemented in corporate valuation methodologies, at least for the most vulnerable sectors. Financial institutions can also propose financial products directly linked with the GHG emissions of a project or a company (see section 1.5).

Financial institutions including banks will be required by recent regulations, especially in the EU, to disclose their scope 3 GHG emissions including their financed emissions: GHG emissions associated with investment activities are considered part of a financial institution's carbon footprint. Some frameworks such as the Partnership for Carbon Accounting Financials (PCAF) (PCAF, 2022), officially recognized by the GHG protocol, make it possible to measure scope 1 and scope 2 emissions of investee companies to allocate them to a financial institution. This allocation is done using the ratio of the investment made by the financial institution in the company divided by the Enterprise Value Including Cash (EVIC) of the investee. Moreover, PCAF allows the use of estimates to measure these financed emissions. They have developed a scoring system, assessing the quality of each GHG emission data according to the method of calculation.

1.4 . ESG and risks

The use and disclosure of ESG-related data and products come with three types of risk: legal, transition and physical risks. These latter are interconnected and should be monitored closely. Moreover, geographic heterogeneity creates an important variability regarding these risks.

For financial institutions, ESG risks provide opportunities to work with clients, for instance, to help them define solid ESG strategies and their need for sustainable financing, highlight the physical risks of their investments or evaluate GHG emissions reduction paths (D'Orazio and Valente, 2019; Gourdel et al., 2022).

1.4.1 . Legal risks

Legal risks include all the risks of financial loss, reputational loss and legal actions that can result from the lack of awareness, misunderstanding, disregard, or denial of integration of ESG issues by a company's policy. Legal risks should be monitored continuously as regulations keep growing and evolving, requiring constant adaptations.

Legal risks can arise because of several factors.

- An ESG integration strategy that does not reflect the real functioning and activities of a company because of a lack of information may result in greenwashing, which could in turn result in reputational loss.

- The growing number of regulations, sometimes technical and challenging to implement, may also be a factor of legal risks, that could be mitigated by anticipating increased scrutiny from regulators.

- The actions of activists could also be a source of legal risks that can influence how ESG is taken into consideration within a company strategy.

- Legal actions can arise against companies in order to obtain a change of behavior and monetary compensation for activities deemed damageable to the environment or human rights for instance. These legal actions sometimes find their ground in the non-binding guidelines from the Organisation for Economic Co-operation and Development (OECD), establishing responsible business conduct obligations at a global level (OCDE, 2011).

- Financial institutions can face legal actions from investors judging that they have been provided with misleading or incomplete information.

- Companies taking extra-financial commitments, even voluntary and non-binding, create accountability and expectations. Failure to comply may create reputational risks and legal consequences.

Companies can mitigate these risks with robust ESG integration and due diligence procedures that require good ESG data and meaningful processing methodologies.

1.4.2 . Transition risks

Efforts to move towards a greener economy result in some business sectors facing important shifts in asset values and higher costs of doing business. Cambridge Centre for Sustainable Finance (2016) defines transition risks as "risks which arise from efforts to address environmental change, including but not limited to abrupt or disorderly introduction of public policies, technological changes, shifts in investor sentiment and disruptive business model innovation". Transition risks thus represent institutional, financial or reputational damages that can result directly or indirectly from the process of adjustment towards a lower carbon and more environmentally friendly economy. Some instances of transition risks that can impact companies are the introduction of carbon taxes that can modify behavior and local demands. Some assets can also no longer be viable and become stranded, such as coal-fired power plants. Stranded assets for a company can have a strong impact on its valuation, leading to a reduction of external investments in this company (Baldwin et al., 2020; Rozenberg et al., 2020; Cahen-Fourot et al., 2021).

Mitigating transition risks in the financial industry requires taking action, mainly through risk assessments and data-driven modeling (stress tests). Such studies could help the institution and its clients to transition by rebalancing some portfolios and seizing new opportunities.

ESG data providers often propose different metrics and indicators to assess transition risks for companies, such as GHG emissions, presence of specific countries' regulations or compliance with voluntary reporting.

1.4.3 . Physical risks

Cambridge Centre for Sustainable Finance (2016) defines physical risks as "risks which arise from the impact of climatic (i.e. extremes of weather) or geologic (i.e. seismic) events or widespread changes in ecosystem equilibria, such as soil quality or marine ecology. [...] They can be event-driven ('acute') or longer-term in nature ('chronic')". Extreme weather events, such as heatwaves, wildfires, storms or floods, become more frequent and more acute because of climate change, resulting in important financial and human losses. Chronic changes include rising sea levels, rising average temperature and ocean acidification whose consequences are long terms.

Physical risks may weaken the performance of a company through high impact on asset values, revenue disruptions, interruption of operations under certain types of events or disastrous human and material damages (Pinchot et al., 2021). Studying and modeling physical risks, their likelihood and their magnitude can help identify them and take the right course of action. Companies should measure their

asset exposure to physical risks as well as their asset vulnerability. Such indicators are often proposed by ESG data providers.

1.5 . ESG financial products

1.5.1 . Equity related ESG financial products

The simplest equity-related products that can be qualified as ESG products are funds in which stocks have been carefully selected to match ESG criteria. Different approaches in ESG integration are possible and are more and more regulated, for instance in the EU through SFDR. These approaches are often based on ESG data including ESG scores (Verheyden et al., 2016; Branch et al., 2019; Autorité des Marchés Financiers, 2020).

The 'best-in-universe' approach consists of selecting the best companies with the best ESG profiles in all available business sectors. Portfolios built this way often have strong biases towards some specific "virtuous" industries. To mitigate these biases, the 'best-in-class' approach identifies the best companies in the selected ESG fields, for every business industry. The ESG profile of companies in 'best-in-class' portfolios is compared to the ESG profile of its competitors within the same industry so as to exclude the ones with the lowest standards. These approaches can be complemented by imposing minimum ESG standards required to select an asset, and by the 'best-effort' approach, in which portfolios are built by favoring companies whose ESG performance has improved over time.

More mathematically-driven methodologies enable building portfolios by directly optimizing some selected ESG criteria in addition to the portfolio's expected return and risk (Gasser et al., 2017; Chan et al., 2020; Chen et al., 2021; Prol and Kim, 2022).

EU regulations and specifically the EU taxonomy also define what sustainable investments are and insist on data-driven methodologies to identify them. In particular, an investment in economic activity is sustainable if it contributes to an environmental or social objective, provided that it does not significantly harm any other environmental or social objectives and that the investee company follows good governance practices.

These methodologies of portfolio construction require a good knowledge of ESG data and the ability to select the most meaningful and comparable indicators.

1.5.2 . Sustainable bonds and other credit financial products

Sustainable bonds help issuers define their sustainability strategy and their commitment to sustainable projects. The goal of sustainable bonds is to finance the economy as well as the ecological transition. They cover both environmental and social dimensions, in comparison to green and social bonds. There are two types of sustainable bonds.

- A use-of-proceeds sustainable bond can be used to finance projects showing serious effort towards transitioning to a lower carbon economy. The issuer of the bond must report on the advances of the project and where the funds have been allocated. The company usually has to report annually on the allocation of the use of proceeds until full allocation.

- In Sustainability-Linked Bonds (SLB), there is no consideration of how the funds were allocated. Instead, a Key Indicator of Performance (KPI) is selected and the company has to achieve during the life of the bond a public target on this KPI, reporting annually. If the target is not achieved in the set time frame, there is usually a coupon step-up. The selected KPI can be environmental (GHG emissions, waste, pollution, energy use, ...), social (employee's health and safety, training hours, the share of women on board, ...) or linked to governance practices (compliance, measure against conflicts of interest, anti-money laundering policy, ...). It should be adapted to the company, benchmarked as much as possible, ambitious enough and sourced from external providers or externally audited.

Similar products exist for loans and repurchase agreements, i.e. use-of-proceeds loans and repurchase agreements and sustainability-linked loans and repurchase agreements with similar processes.

1.5.3 . Carbon offsets

To tackle climate change, GHG emissions in the atmosphere have to be reduced. Companies not only need to decrease their GHG emissions but also compensate for their unavoidable emissions to reach net zero. Reaching net zero refers to achieving a balance between the amount of GHG produced and the amount removed from the atmosphere, adding no more emissions than the ones removed. This process of compensated GHG emissions can be done through carbon markets.

In 1997, following the Kyoto Conference, carbon markets were created. With the Kyoto Protocol, national and international transfers of emissions among market participants were allowed. Carbon markets are divided between compliance markets and voluntary markets.

Compliance markets are driven by regional, national or international policy: companies are required to achieve binding emission reduction targets. They use

emission allowances to comply with these requirements. Compliance markets include in some regions of the world Emissions Trading Systems (ETS): it is for instance the case in the EU. In the EU ETS, the EU sets a limit on the total amount of GHG that can be emitted by the operators covered by the ETS, and this limit is reduced over time so that total GHG decreases. Within this limit, operators can trade emissions allowances as desired but, after each year, they must be able to surrender enough allowances to fully compensate for their emissions (European Commission, 2023a).

Voluntary markets enable companies to voluntarily offset their GHG emissions through the purchase of carbon credits. A carbon offset occurs when a company buys a carbon credit to compensate for its emitted GHG. In this system, the transaction money must be used to fund projects removing the same amount of GHG in the air, through natural (planting trees) or technological methods. Standards have been put in place such as the Verified Carbon Standards (VCS), a certification program ensuring the quality, credibility, and transparency of carbon offset projects. The certified projects issue Verified Carbon Units (VCU), carbon credits that can be used in most voluntary carbon markets to offset GHG emissions.

1.6 . ESG and machine learning

So far, this introduction highlighted the rising significance of data in the different dimensions of ESG in the financial industry. Thanks to the growing number of ESG data providers and the expansion of available data points, coupled with enhanced data quality resulting from regulations on ESG disclosure, it seems natural that data-intensive methodologies, and specifically machine learning, could help leverage ESG data to produce meaningful results. The aforementioned introduction aimed to convey the increasing number of potential applications for these methodologies:

- comparison of ESG datasets, evaluation of the redundancy of information and selection of the most relevant data points.
- data imputation, estimating ESG data not yet reported by companies.
- identification or prediction of controversies a company may face.
- assessment and modeling of the potential physical risks a company may face according to its region.
- portfolio optimization based on the investigation of the explanatory and predictive power of ESG on returns.
- identification of material ESG issues for specific companies or specific sectors.

- estimating GHG emissions for companies that have not yet reported them to get insights on the number of necessary carbon credits to offset their emissions.

- predicting ESG indicators in the future to monitor the performance of the issuer of sustainability-linked bonds on the selected KPI.

This thesis proposes solutions to leverage ESG data, using carefully adapted machine learning methods, through two case studies relevant to the financial industry.

Chapter 2 presents the different machine learning tools needed to exploit ESG data.

Chapter 3 investigates the relationship between price returns and ESG scores in the European equity market using interpretable machine learning. We examine whether ESG scores can explain the part of price returns not accounted for by classic equity factors, especially the market one. Thanks to this methodology, we build materiality matrices, showing the most important ESG dimensions broken down by industry and company size.

Chapter 4 focuses on particularly important and used ESG data points, the scope 1 and scope 2 GHG emissions. We propose an interpretable machine learning model to estimate scopes 1 and 2 GHG emissions of companies that have not reported them yet. We propose experiments and discussions on the interpretability of the model as well as its accuracy when input data are missing.

In particular, we propose a cross-validation methodology allowing the exploitation of ESG data and which is used in both case studies. Indeed, the ESG data used in this thesis is non-stationary, and its amount, reliability and quality keep increasing: we need powerful validation tools to exploit it. We also put a particular emphasis to the interpretability and reproducibility of the considered machine learning methodologies, a prerequisite in any ESG-related project that will be discussed in depth in this dissertation.

2 - Elements of interpretable machine learning for ESG data

2.1 . Principles of supervised learning for tabular data

2.1.1 . Definition

The goal of supervised learning is to learn a mapping h between a dataset of input $\mathbf{X}$ and an output $\mathbf{y}$ so that this mapping is generalizable to previously unseen data. h is a parameterized mapping, also called a model and which can take different forms from linear models to more complex algorithms such as neural networks or gradient boosting (see section 2.2). The parameters of h are learned by optimizing a loss function linking the output $\mathbf{y}$ to the predicted output $\hat{\mathbf{y}} = h(\mathbf{X})$. The choice of the loss function is specific to the problem that we want to solve. The main ones used in this thesis are detailed in sections 2.1.3 and 2.1.4.

The data is a collection of labeled samples $(\mathbf{x_i}, y_i)_{i\in\{1,N\}}$. The vectors $\mathbf{x_i}$ of size P are called feature vectors. Each coordinate of these vectors can refer for instance to ESG scores, ESG indicators or fundamental data describing a specific sample.

The label y_i, associated with the feature vector $\mathbf{x_i}$, is what we want to infer from the input data. It can be any variable of interest such as the return of the associated stock at the date of reporting or the associated GHG emissions.

This setting is called a tabular setting. When the labels are discrete values, we are in a classification setting. In particular, when the label can only take two values, it is a binary classification problem. When the labels take continuous values, we are in a regression setting.

More theoretically, let us assume our labeled data points $(\mathbf{x_i}, y_i)$ are sampled from a joint probability distribution $\mathcal{X} \times \mathcal{Y}$. Let us assume we have a loss function $\mathcal{L}$ measuring how different the prediction $\hat{y}$ is from the true value y. The goal of supervised learning is to learn a mapping h minimizing the risk associated with h, which is the expectation of the loss function, $\mathbb{E}\left[\mathcal{L}(\hat{y}, y)\right]$. The mapping h is a parameterized function that can take different forms. Usually, we choose one class of mapping for the considered problem, called $\mathcal{H}$.

The supervised learning problem consists of solving the following optimization problem

$$\underset{h\in\mathcal{H}}{\operatorname{argmin}}\, \mathbb{E}\left[\mathcal{L}(\hat{y}, y)\right]. \tag{2.1}$$

This problem cannot be solved theoretically as the joint distribution $\mathcal{X} \times \mathcal{Y}$ is unknown. The supervised learning framework approximates the solution by minimizing the empirical risk which is the average of the loss function on a dataset

called the training set, composed of the labeled data points $(\mathbf{x_i}, y_i)$. Then, in practice, supervised learning consists of solving this second optimization problem

$$\underset{h \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(h(\mathbf{x_i}), y_i). \tag{2.2}$$

However, it is often easy to create a mapping h driving the empirical risk of equation (2.2) to a minimum by matching exactly the training dataset. It is called overfitting and it is not what we seek as it leads to very poor generalization capacities for the trained model: the mapping h has only captured a part of the distribution $\mathcal{X} \times \mathcal{Y}$ and cannot generalize beyond unseen data points. Indeed, the minimization of the empirical risk is only an approximation, made because it is the only metric we can have access to. The ultimate goal is to minimize the risk of equation (2.1), so that the mapping h capture the entirety of the distribution $\mathcal{X} \times \mathcal{Y}$ and generalize to unseen data: the learned mapping h should lead to a low loss-function for all data points sampled from $\mathcal{X} \times \mathcal{Y}$, whether they are seen during training or not. This can be achieved by splitting the collected labeled data into different splits.

2.1.2 . Training, validation and test sets

Test set

Section 2.1.1 shows that a trained model h working well (meaning that the average of the loss function is low) only in the training set is not enough: this good performance could be due to overfitting and the model may not be able to generalize well. The most common way to prove the generalization capacities of a model is to set aside a fraction of the labeled data from the training set: this new dataset is called the test set. The test set should only be looked at the end of the training process and used to assess the final performance of the trained model: it is important not to create any leakage between the training process and the test set. Most of the data should remain in the training set so that the model can see enough labeled examples to be well-fitted. If we have enough data, both training and test sets are good representations of the data distribution $\mathcal{X} \times \mathcal{Y}$. In the context of ESG, it is not always the case as we are limited by the quantity of available data: in this thesis, we present some methods developed to mitigate this issue.

Validation set

The current framework consists of learning a model $h \in \mathcal{H}$ on a training set of labeled data by minimizing the empirical risk and checking the generalization capacities of the learned model on a separate test set of labeled data.

$\mathcal{H}$ is a class of parameterized models. These parameters are determined by fitting the model, meaning by solving the optimization problem in (2.2). However, very often, a model in $\mathcal{H}$ is not only determined by learned parameters but also by carefully chosen hyperparameters: it is for example the case for decision trees and gradient boosting methods. Hyperparameters are parameters whose values control the learning process. For instance, it is the maximum depth of a decision tree, the fraction of training samples used to build a model, the number of iterations in a gradient boosting algorithm or a regularization coefficient.

Choosing these hyperparameters by minimizing the empirical risk on the training set is not a good idea as it may lead to overfitting. We can not choose them either on the test set as it will create some leakage between the training process and the test set and skew the evaluation of the final performance. A third set of labeled data is needed. It is called the validation set and is used to tune the hyperparameters associated with the model h. Nonetheless, removing a part of the training set to become the validation set is not always desirable, especially when the number of available labeled samples is small. It could lead to a loss of important information contained in these samples, on which the model will not be trained. Moreover, the distribution of samples on the training and validation sets might be slightly different. We require a methodology that leaves enough data for the training set while providing enough data for the validation set. This is what K-fold and K-times repeated random sub-sampling cross-validation achieve.

K-fold and K-times repeated random sub-sampling cross-validations

In K-fold cross-validation, the basic idea is to divide the training set into K non-overlapping subsets of approximately equal sizes. Each of the K subsets is used once as the validation set of a model trained on the $K-1$ other subsets. This requires training K models whose performance on the K subsets used as validation are averaged to select the hyperparameters. Every data sample is in a validation set exactly once and is used $K-1$ times in training: this method allows for more robust performance than just having a separate validation set. However, especially for small datasets, the validation performance estimated via K-fold cross-validation can be noisy from one run to another. Indeed, if different splits of the dataset into K-folds are implemented, it may lead to a different distribution of performance for the different folds, depending on their distribution of samples.

K-times repeated random sub-sampling cross-validation creates K random splits of the data into training and validation sets. The idea is to randomly select a proportion of samples of the training set to be the validation set, with the remaining samples still being used to train the model. This is done K times. The proportion

is set by the user. A model is trained on each split and performance on the different validation sets is averaged to select the hyperparameters. In comparison to the K-fold cross-validation procedure, the number of samples in the sets is not dependent on the number of splits, making it possible to set a proportion so that enough data remains in both training and validation sets. However, if the number of splits is too low, the model may not be both trained and validated on all samples. This is not an issue if the proportion of samples in both sets is high enough so that their distribution of data could be assumed to be the same.

Training, validation and test sets in the context of temporal tabular data

In an ESG setting, as companies are only reporting ESG-related data once a year, each sample in the input dataset is referring to a company and a specific year. In some cases, the frequency of reporting of the data can be higher: the sample is then identified by the company identifier and its date. Machine learning in the ESG field is then strongly related to machine learning for time series.

When working with temporal tabular data, composed of one or more time series of data, the training, validation and test set cannot be composed at random. The standard procedure consists of splitting the data into causal consecutive train, validation and test data sets so that there is no leak of the future into the validation set and more importantly into the test set. Sometimes, margins (subsets of unused data at specific dates) can be added between the different temporal sets to prevent with more confidence any potential leakage.

Regarding training and validation sets, Bergmeir and Benítez (2012) discuss different cross-validation methodologies in the context of stationary time series. When used time series are not stationary, as in the case of ESG data, these methods should be improved on and propositions of methodologies are made in this thesis.

Checking the evolution of model performance with time is also interesting. It can be done using rolling calibrations. It consists of training different models, associated with different test sets, always set at a date after the training/validation sets. A representation of these rolling calibrations is shown in Fig. 2.1: data are available yearly, from 2002 to 2020, and five rolling calibrations are used. Hyperparameters can be selected per model, using a different set of hyperparameters for each of the rolling calibrations or globally by measuring the average validation performance across all calibrated models. This latter method may introduce leakage from the future into the model trained on only the oldest data. Measuring the evolution of performance as well as the interpretability for each of these models can yield meaningful results with the potential emergence of trends.

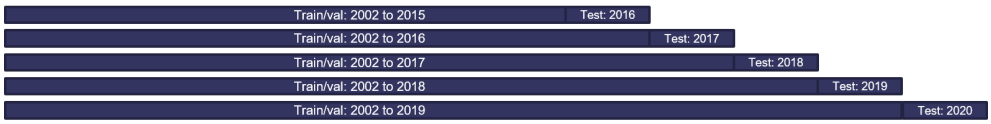


Figure 2.1: Illustration of rolling calibrations.

2.1.3 . Metrics for classification problems

Cross-entropy

Suppose we have trained a binary classifier on two classes, the positive one, 1, and the negative one, 0. Usually, a binary classifier does not output directly the predicted class for the sample but an estimation of the probability that the sample belongs to class 1. The cross-entropy, also known as the logloss, is defined as

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(p_i) + (1 - y_i) \log(1 - p_i), \tag{2.3}$$

where p_i is the model probability that sample i belongs to class 1, $y_i \in \{0, 1\}$ is the true class and N the number of samples it is computed on.

The cross-entropy is often chosen as the loss to optimize in classification settings. It is always positive and the closer to 0 it is, the better the performance of the model is.

Confusion matrix for binary classification and associated metrics

A confusion matrix adapted to a binary classifier shows how well the classifier is performing. Using the same positive and negative classes, it displays:

- The number of true positives: samples predicted positive by the model and actually positive.
- The number of false positives: samples predicted positive by the model and actually negative.
- The number of true negatives: samples predicted negative by the model and actually negative.
- The number of false negatives: samples predicted negative by the model and actually positive.

An illustration of a confusion matrix is available in Tab. 2.1.

Using the number displayed by the confusion matrix, several metrics can be built to evaluate binary classifiers.

- **Accuracy** It shows the number of rightly predicted samples out of the total number of samples. It is an indicator between 0 and 1, 1 meaning perfect accuracy.

		Actual values	
		Positive	**Negative**
Predicted values	**Positive**	True Positives (TP)	False Positives (FP)
	Negative	False Negatives (FN)	True Negatives (TN)

Table 2.1: Confusion matrix for a binary classification problem.

$$\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.4}$$

A drawback of accuracy is that, if classes are unbalanced, accuracy could be quite high if the model predicts everything as belonging to the densest class. A way to remedy that is to used the balanced accuracy metric, the average of the sensitivity and the specificity.

- **Sensitivity** Also called true positive rate, it is the ratio of true positives divided by the total number of actual positives. It is an indicator between 0 and 1.

$$\text{sensitivity} = \frac{TP}{TP + FN} \tag{2.5}$$

- **Specificity** Also called true negative rate, it is the ratio of true negatives divided by the total number of actual negatives. It is an indicator between 0 and 1.

$$\text{specificity} = \frac{TN}{TN + FP} \tag{2.6}$$

- **Balanced Accuracy**: Balanced accuracy is defined as the average of sensitivity and specificity. By definition, it is an indicator between 0 and 1 which has a value of 0.5 if the model did not learn anything significant. In some figures of this document, the balanced accuracy is sometimes abbreviated Bal_Acc.

$$\text{balanced accuracy} = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \tag{2.7}$$

2.1.4 . Metrics for regression problems

In this section, we suppose a regression model h, trained on N samples $\mathbf{x_i}$ each associated with a continuous label y_i.

Mean-square error and root-mean-square error

The mean-square error also referred to as the MSE, is defined as

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}(y_i - h(\mathbf{x_i}))^2. \tag{2.8}$$

The root-mean-square error (RMSE) is simply the square root of the MSE.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - h(\mathbf{x_i}))^2} \tag{2.9}$$

MSE and RMSE vary between 0 and infinity, a value of 0 meaning that the model is perfectly accurate. The MSE is often chosen as the loss to optimize regression models.

Mean-absolute error

Another measure of performance is the mean-absolute error, referred to as MAE and defined as

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - h(\mathbf{x_i})|. \tag{2.10}$$

MAE varies between 0 and infinity, a value of 0 meaning that the model is perfectly accurate.

Let us note here that MSE and RMSE penalize more large errors than MAE because the errors between prediction and actual values are squared. MAE is more robust to outliers than these two metrics.

R-squared

The R-squared metric noted R^2 is commonly used in regression problems. It is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - h(\mathbf{x_i}))^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}, \tag{2.11}$$

with $\bar{y}_i$ is the average of the ground truth labels y_i.

The R^2 metric varies between $-\infty$ and 1, a value of 1 meaning that the model is perfectly accurate.

Using metrics from classification problems for regression ones

When the predicted targets can be discretized, it is possible to use metrics usually used for classification with the discretized predicted values from the regression model. For instance, in section 3.7.3, we use the balanced accuracy metric on the sign of the estimated return.

2.2 . Gradient Boosted Decision Trees

Gradient boosting is a non-linear machine learning algorithm for both regression and classification tasks. It is based on the concept of ensemble learning, which combines the predictions of multiple weak learners to yield a strong learner with improved accuracy. The advantage of such methods with respect to linear regression is that they are able to learn more generic functional forms. When the weak learners used are decision trees, the algorithm is called Gradient Boosted Decisions Trees (GBDT). This algorithm iteratively adds decision trees to the model, each tree learning from the errors of the previous ones to improve predictions. In this section, we provide a mathematical description of the GBDT algorithm.

The state of the art for regression and classification problems on tabular data is provided by gradient boosting models from Friedman (2001), as shown for instance in Shwartz-Ziv and Armon (2022). Several studies, such as Schmitt (2022) have shown that gradient-boosted models are at least as effective as deep neural networks for classification purposes in the context of tabular data. Gradient-boosted models are typically much faster to train than deep neural networks.

2.2.1 . A specific class of functions

As in any supervised machine learning model, the goal is to find the best approximation $h \in \mathcal{H}$ of the true model, solution to the following optimization problem

$$\underset{h \in \mathcal{H}}{\operatorname{argmin}} \, \mathbb{E}\left[\mathcal{L}(h(\mathbf{x}), y)\right], \tag{2.12}$$

with $\mathcal{L}$ the chosen differentiable loss function.

In the context of GBDT, $\mathcal{H}$ is a specific class of functions, a weighted sum of M binary decision trees, d_m. Using similar notations as Friedman (2001), d_m being a decision tree, it partitions the input space J_m into several disjoint regions R_{j_m}, $j \in [1, J_m]$, and predicts the constant value b_{j_m} for the associated region. We have

$$d_m(\mathbf{x_i}) = \sum_{j=1}^{J_m} b_{j_m} \mathbf{1}_{R_{j_m}}(\mathbf{x_i}). \tag{2.13}$$

Then, $h \in \mathcal{H}$ can be written in the following form:

$$h(\mathbf{x_i}) = \sum_{m=1}^{M} \rho_m \sum_{j=1}^{J_m} b_{j_m} \mathbf{1}_{R_{j_m}}(\mathbf{x_i}), \tag{2.14}$$

where ρ_m is a parameter selected to optimize the loss function which is dependent on the considered weak learner. We can simplify equation (2.14), incorporating b_{j_m} into ρ_m.

$$h(\mathbf{x_i}) = \sum_{m=1}^{M} \sum_{j=1}^{J_m} \gamma_{j_m} \mathbf{1}_{R_{j_m}}(\mathbf{x_i}) \tag{2.15}$$

This can be interpreted as using an optimal coefficient γ_{j_m} for each region of each fitted tree instead of using a unique one ρ_m per fitted tree.

GBDT being an iterative model, this equation can be rewritten in a recursive form, giving the parametrization of functions $h \in \mathcal{H}$.

$$h_m(\mathbf{x_i}) = h_{m-1}(\mathbf{x_i}) + \sum_{j=1}^{J_m} \gamma_{j_m} \mathbf{1}_{R_{j_m}}(\mathbf{x_i}) \tag{2.16}$$

2.2.2 . Finding h_m

The GBDT algorithm is first proposed by Friedman (2001). Using the empirical risk form of equation (2.2), considering a training set composed of N labeled data points $(\mathbf{x_i}, y_i)$ and $\mathcal{L}$ the chosen differentiable loss function, the algorithm is initialized with a constant value γ such that

$$h_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{N} \mathcal{L}(y_i, \gamma). \tag{2.17}$$

For the next steps $m = 1$ to $m = M$, the goal is to find a weak learner which will minimize the residual between $h_{m-1}(\mathbf{x_i})$ and y_i for each point in the training set. The GBDT algorithm proposes to find this weak learner with the following steps:

1. Pseudo-residuals r_{i_m} between y_i and $h_{m-1}(\mathbf{x_i})$ are computed for each point in the dataset.

$$r_{i_m} = -\left[\frac{\partial \mathcal{L}(y_i, h(\mathbf{x_i}))}{\partial h(\mathbf{x_i})}\right]_{h(x)=h_{m-1}(x)} \quad \text{for } i = 1, \ldots, N. \tag{2.18}$$

2. A decision tree is trained to fit these residuals, using the whole training set $(\mathbf{x_i}, y_i)_{i=1,\ldots,N}$.

3. Compute γ_{j_m} solving the following optimization problem:

$$\gamma_{j_m} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{jm}} \mathcal{L}(y_i, h_{m-1}(\mathbf{x_i})) + \gamma). \tag{2.19}$$

Let us note here that we are only using the regions found when fitting the tree. The b_{j_m} coefficients are discarded in favor of the γ_{j_m}.

4. The model is then updated as

$$h_m(x) = h_{m-1}(x) + \mu \sum_{j=1}^{J_m} \gamma_{j_m} \mathbf{1}_{R_{j_m}}(x). \tag{2.20}$$

μ represents the learning rate, a hyperparameter that can be determined by cross-validation. It controls the contribution of each weak learner in the final model and helps prevent overfitting.

After M steps, the final model $h_M(x)$ is output. The number of iterations, a hyperparameter of the model, is often tuned using a validation set and monitoring the evolution of the loss on it: when the loss does not evolve following a fixed number of iterations, the training is considered complete.

2.2.3 . The LightGBM GBDT implementation

Different implementations of the Gradient Boosted Decision Trees algorithm exist, each having some specificities, advantages and drawbacks. XGBoost (Chen and Guestrin, 2016), LightGBM (Ke et al., 2017) and CatBoost (Prokhorenkova et al., 2018) are some of them. In this thesis, we use the LightGBM implementation of GBDT, which offers some convenient features, increasing the accuracy and speed performance of the algorithm as well as the type of data it is capable to handle. Specifically:

- LightGBM overcomes the bottleneck of building decision trees (finding the right split between left and right nodes) by always using the histogram-based algorithm to find the split between left and right nodes which leads to speed gains. This histogram-based algorithm is referenced in Ke et al. (2017).

- LightGBM uses a technique called "leaf-wise" growth to construct decision trees. In traditional GBDT algorithms, decision trees are grown in a "level-wise" manner, where each level of the tree is expanded before moving to the next level. It can lead to many small leaf nodes, which can be inefficient and reduce the model's accuracy. Leaf-wise growth first expands the leaf node which reduces the loss the most, resulting in a deeper and more accurate tree. With this method, hyperparameters should be carefully tuned to prevent overfitting, especially the maximum depth of the tree.

- LightGBM is also able to handle missing values, allocating them to whichever node reduces the loss the most.
- Categorical features are also handled, without the need for one-hot encoding which can dramatically increase the dimension of the features space.

2.3 . Elements of interpretable machine learning

In the context of ESG data in the financial industry, with the emergence of new regulations and to make machine learning more largely adopted, models require to be explainable: a user has to understand why a model output a prediction and how the input data influences this prediction. An important challenge is to find the right balance between model accuracy and model interpretability. Indeed, simple models like linear regressions, logistic regressions or decision trees are easy to interpret, but their performances are limited as they are not able to capture very complex relationships in the training data. On the other hand, models, including Gradient Boosting Decision Trees or deep neural networks, are usually more accurate but are not directly interpretable. We review in this section two model-agnostic methods for interpreting such black box models: Shapley values and partial dependence plots.

2.3.1 . Shapley values

Definition of Shapley values

Shapley values were first introduced in the context of game theory by Shapley (1953). In a coalition game of cooperative game theory, a group of players cooperates to achieve a specific goal whose payoff is distributed among them. Shapley values provide a way to fairly allocate this payoff among the players based on their contributions to the achievement of the goal.

The Shapley value of a player is calculated by considering all possible coalitions of players and determining the player's marginal contribution to each coalition. The Shapley value is then the weighted sum of the marginal contributions over all possible coalitions. Shapley (1953) summarizes this definition as

$$\phi_i(v) = \sum_{S \subseteq P \setminus \{i\}} \frac{|S|!\,(p - |S| - 1)!}{p!} (v(S \cup \{i\}) - v(S)), \tag{2.21}$$

with ϕ_i being the Shapley value for player i, p the total number of players, P the total number of subsets of players, $|S|$ the number of players in subset S and v a function mapping a subset of player to a real value representing the obtained payoff for this subset.

The term $v(S \cup \{i\}) - v(S)$ simply consists of computing the marginal contribution of player i for a coalition S by differentiating the payoffs between S with player i and S alone.

The term $\frac{|S|!\ (p-|S|-1)!}{p!}$ is the weight assigned to the marginal contribution computed previously and is the probability of occurrence of the coalition S.

Application of Shapley values to machine learning - From Shapley values to SHapley Additive exPlanations

Shapley values can be applied to a machine learning problem to explain its individual predictions with the following analogies, as shown in Lipovetsky and Conklin (2001) or Štrumbelj and Kononenko (2014):

- The prediction task of the machine learning model for every single sample of the dataset is a game.
- The players are the different features taken as input by the model, a coalition referring to a specific set of features.
- The payoff is the difference between the actual prediction made by the model h and the average prediction made by the model for all samples in the dataset.

Different papers and books such as Shapley (1953), Sundararajan and Najmi (2020) or Molnar (2020) show that Shapley values are the only attribution method satisfying specific desirable properties. Shapley values are a strong explanation method, with a solid theory.

Let us denote by h the machine learning model, $\phi_{j,i}$ the Shapley value of feature j for a sample $\mathbf{x_i}$ and $\mathbb{E}_X[h(X)]$ the average prediction made by the model for all samples in the dataset. The model used P features.

- **Efficiency:** the features contributions, i.e. the Shapley values, of each feature, for a sample $\mathbf{x_i}$ must add up to the difference between prediction for $\mathbf{x_i}$ and the average prediction made by the model for all samples in the dataset.

$$\sum_{j=1}^{P} \phi_{j,i} = h(\mathbf{x_i}) - \mathbb{E}_X[h(X)] \tag{2.22}$$

- **Symmetry:** if two features j and k contribute equally to all possible coalitions, their Shapley values are the same.

  Using notations from 2.21, for a sample i, if $v(S \cup \{j\}) = v(S \cup \{k\})$ for all $S \subseteq P \setminus \{j, k\}$, then $\phi_{j,i} = \phi_{k,i}$.

- **Dummy:** if a feature j does not change the predicted value for any coalition it is added to, this feature has a Shapley value of 0.

  Using notations from 2.21, for a sample i, if $v(S \cup \{j\}) = v(S)$ for all $S \subseteq P$, then $\phi_{j,i} = 0$.

- **Linearity:** This is a very interesting property in a machine learning context: if a final model consists of averaging intermediates models, the Shapley values of the final model are the average of the Shapley values of these intermediate models.

  Suppose two models are trained h and h', that are linearly combined into one final model $g = \alpha h + \beta h'$, with α and β two reals. The associated Shapley values for feature j and sample i for models h, h', g are respectively $\phi_{j,i}$, $\phi'_{j,i}$ and $\psi_{j,i}$ where $\psi_{j,i} = \alpha\phi_{j,i} + \beta\phi'_{j,i}$.

The main difficulty when computing Shapley values for a machine learning model is in the computation of the feature contribution in a coalition excluding certain features: how is it possible to exclude some features in a model calibrated on the full set of P features?

Similarly to what is done in Štrumbelj and Kononenko (2014), simulating that a feature value is missing from a coalition can be done by marginalizing the feature by sampling values from the feature's empirical marginal distribution. This simply consists of approximating the effect of removing a variable from the model by sampling a value for the missing feature from its samples in the training dataset. Better results can be achieved by repeating this sampling step and averaging the output.

Thus, instead of computing the payoff in equation (2.21) using $h(\mathbf{X})$, we use the conditional expectation $\mathbb{E}_X[h(X)|X_S = \mathbf{x_S}]$ with h the explained model, X the random variable representing the distribution of features across samples and X_s the random variable representing the distribution of the subset of present features in coalition S whose values are $\mathbf{x_S}$. These are SHapley Additive exPlanations, also referred to as SHAP and introduced by Lundberg and Lee (2017). SHAP values inherit the same desirable properties as Shapley values.

This process suffers important drawbacks: if features in the dataset are correlated, the sampling procedure can lead to training examples that do not make any sense (associating, for instance, a company with a very low environment, social and governance score with a strong ESG one).

Moreover, considering all possible coalitions of features to calculate the exact Shapley value is computationally expensive, as the number of possible coalitions exponentially increases as more features are added.

TreeSHAP implementation

TreeSHAP is a specific variant implementation of SHAP values, designed for tree-based models such as GBDT and which is trying to overcome some of the drawbacks of SHAP values. The TreeSHAP implementation was proposed in Lundberg et al. (2018).

In the classic SHAP implementation, to simulate a missing feature from a coalition, we sample it from the feature's empirical marginal distribution. The payoff function is thus defined as $\mathbb{E}_X[h(X)|X_S = \mathbf{x_S}]$ with h the explained model, X the random variable representing the distribution of features across samples and X_s the random variable representing the distribution of the subset of present features in coalition S whose values are $\mathbf{x_S}$.

In the TreeSHAP implementation, to simulate a missing feature from a coalition, we sample it from the feature's empirical conditional distribution. Using the same notations, the payoff function is thus defined as $\mathbb{E}_{X|X_S}[h(X)|X_S = \mathbf{x_S}]$. Intuitively, this means that, in the decision tree, for a specific sample, we average the predictions weighted by the leaves sizes (the number of training samples in the leaf) only for leaves that are reachable given S. If we come across a node whose split depends on a feature that is in contradiction with those in S, the node is ignored and the sample is propagated through the two child nodes.

This procedure has to be applied to each possible subset S of the features and each feature value inside those subsets. Lundberg et al. (2018) proposes a fast exact implementation of this methodology.

Choosing SHAP values as the interpretation method

The use of SHAP values enables the computation of feature-specific explanations for individual samples, providing insight into the contribution of each feature to a given prediction. This approach is founded on sound theoretical principles and constitutes an exact method. TreeSHAP implementation is a fast method for computing SHAP values. Although other techniques for explanatory purposes exist, they lack the beneficial properties of SHAP values. Let us discuss two such methods, namely the feature importance derived from the LightGBM model and Local Interpretable Model-Agnostic Explanations (LIME).

As a gradient-boosted model, LightGBM allows the derivation of feature importance metrics directly from the trained model. Typically, this is computed as the number of times a particular feature is utilized in building each of the trees comprising the LightGBM model. However, such an approach to computing feature importance yields only a global measure and does not allow the derivation of per-sample feature importance. This shortcoming renders it impossible to ascertain if a given feature can yield divergent effects depending on the sample being evaluated. Consequently, the feature importance derived from LightGBM is usually deemed unsatisfactory for explaining models because of its lack of granularity.

Other machine learning interpretation methods that provide a per-sample expla-

nation are available. One such method is LIME, developed by Ribeiro et al. (2016). To explain a unique sample, LIME consists of creating a whole new dataset: samples are obtained by perturbing the original sample in its neighborhood, targets are infered using the trained model on these pertubed samples. A new explainable linear model is trained on this built dataset, using sample weights dependent on the proximity of each perturbed sample to the original one. Molnar (2020) discusses this methodology, highlighting its advantages and drawbacks. Although LIME provides per-sample explanations, it lacks robustness and does not constitute an exact method. Defining what the neighborhood of a sample is is not straightforward. LIME defines it using the kernel width hyperparameter: changing the value of this parameter can change the explanations. Additionally, the explanations derived via LIME are unstable, as they can vary upon running the methodology on different occasions: models trained on different perturbations of the same sample can lead to different explanations. Moreover, the LIME methodology is irrelevant when the model being explained is not locally linear. These drawbacks are illustrated in the work of Alvarez-Melis and Jaakkola (2018), which discusses the limitations of LIME concerning robustness.

2.3.2 . Partial dependence plots

Partial dependence plots, first introduced by Friedman (2001), show the marginal effect of a group of features on the predictions made by the model. It is a way of understanding the relationship inferred by the model between a group of selected features and the target: it can show if this relationship is linear, monotonic or more complex.

The main idea behind the partial dependence plot for a given feature of interest is to marginalize the predicted output over the values of all other input features.

Let us suppose that we want to build a partial dependence plot for a selected set of S features out of the P features used by the model. Let us denote by h the machine learning model and H the partial dependence plot function. H is a function of the selected subset of features $\mathbf{x_S}$. Usually, the number of features in S is restricted to one or two numerical or categorical features to be able to visualize the partial dependence plot function. Mathematically and using the idea of building H by marginalizing the predicted output over the values of all other input features, we have for any $\mathbf{x_S} \in S$

$$H(\mathbf{x_S}) = \mathbb{E}_X[h(X)|X_S = \mathbf{x_S}]. \tag{2.23}$$

Let us denote by $\mathbf{x_C} \in C$ the remaining features used by the model so that $C \cap S = \emptyset$. Equation (2.23) can be translated as

$$H(\mathbf{x_S}) = \mathbb{E}_{X_C}[h(\mathbf{x_S}, X_C)], \tag{2.24}$$

with X_C a random variable representing the distribution of the features in C.

The function H can be estimated by a Monte-Carlo method: for any vector of features of interest $\mathbf{x_S}$, we sample and average over all possible training examples $\mathbf{x_C}^i \in C$. Let us denote by N the number of samples in the training set. We have

$$H(\mathbf{x_S}) = \frac{1}{N} \sum_{i=1}^{N} h(\mathbf{x_S}, \mathbf{x_C}^i). \tag{2.25}$$

Using this Monte-Carlo method leads to a similar drawback for partial dependence plots that the one we noted in section 2.3.1 for SHAP values: the combinations of $\mathbf{x_S}$ and selected $\mathbf{x_C}$ might be unlikely or impossible if features used by the model are correlated.

Computing and plotting H for all possible values of $\mathbf{x_S}$ allow visualizing the marginal effect of the group of features S on the prediction made by the model.

Further mathematical details on the partial dependence plot are available in the original paper Friedman (2001) and in Molnar (2020).

Published papers

Chapters 3 and 4 of this dissertation develop two case studies on which solutions to leverage ESG data using machine learning methods are proposed and applied. The research in these chapters has been presented at conferences and published in academic journals.

- Sections 3.1 to 3.7.1 of chapter 3 were the object of a publication in the *Journal of Risk and Financial Management* (Assael et al., 2023a). This work was orally presented at the 15th Financial Risks International Forum organized by the Institut Louis Bachelier.
- In chapter 4, sections 4.1 to 4.7 and the first paragraphs presenting the methodology to remove outliers in section 4.10.6 were published in the journal *Sustainability* (Assael et al., 2023b).

3 - Dissecting the explanatory power of ESG features on equity returns by sector, capitalization, and year with interpretable machine learning

3.1 . Context

Investing according to how well companies do with respect to their ESG scores has become very appealing to a growing number of investors. Beyond moral criteria, such kinds of investments may increase the value of high-ESG-scoring companies, which will attract even the non-ESG-minded investor, thereby starting a virtuous circle both for the investors and for the beneficiaries of high ESG scores. It may also lead to successful impact investing whereby an investor generates positive environmental or societal impact while targeting a specific level of return (Townsend, 2020; Grim and Berkowitz, 2020).

From a quantitative point of view, ESG scores raise the question of their information content: do these scores contain some signal to estimate the company's fundamental or market information? Restricting themselves to the study of the explanatory and predictive power of ESG scores regarding financial performance, Friede et al. (2015) aggregate the results of more than 2200 studies: 90% of them showed a non-negative relationship between ESG and corporate financial performance measures, a majority displaying a positive relationship. However, more recently, Cornell and Damodaran (2020), Breedt et al. (2019), and Margot et al. (2021) reached less clear-cut conclusions.

The confusion surrounding this question is mostly caused by the nature of ESG data: (i) they are quite sparse before 2015, as the interest in even computing such scores is quite recent; (ii) they are usually updated yearly; (iii) the way they are computed often changes as a function of time and may depend on the way companies disclose data; (iv) human subjectivity may be involved to a large extent in the computation of the scores, according to the methodology chosen by a given data provider. Findings are therefore inevitably data-vendor dependent. While data consistency and quality can only be solved at the data provider level, points (i) and (ii) require a tailored approach.

In this chapter, we argue that settling this issue requires a globally robust and consistent methodology. We discuss how to solve each of the two remaining problems listed above and propose a methodology that combines a novel cross-validation procedure for time series with increasingly reliable data, explainable machine learning, and multiple hypotheses testing. Although we focus on explaining companies' price returns with ESG scores, this methodology can be easily adapted

and extended to different settings, as shown at the end of this chapter.

Another crucial ingredient of our approach is to focus on the simplest possible question. Instead of performing sophisticated regressions, we seek to explain the sign of excess price returns. From an information-theoretic point of view, this means that we focus on a single bit of information (the sign) instead of many bits (full value), which yields significant and robust results that then can be interpreted as a function of market capitalization, business sector and country.

Specifically, we propose in this chapter the current research contributions:

1. We focus on the sign of returns discounted by the market factor and use state-of-the-art classification machine learning models.

2. We propose a company-wise cross-validation scheme that makes it possible to train and validate models with the most recent (and thus most reliable) data. From this validation scheme, we keep the models with the five best validation scores.

3. We show that, according to the selected ESG dataset, the fitted models explain the sign of excess returns in test periods well. We also show that models trained with ESG scores increasingly outperform models trained with fundamental data only.

4. Finally, we show how each individual ESG score contributes to the overall performance of our algorithm and the evolution of their explanatory power as a function of time. We propose a new way to build a so-called materiality matrix based on the interpretability of the chosen machine learning models, showing that the importance of ESG scores depends on both the business sectors and market capitalization.

In the remainder of this chapter, the terms ESG scores and ESG features are used interchangeably. This chapter ends with the description of additional experiments realized in this study, in which we explore the enlargement of the proposed framework or its application to new datasets.

3.2 . Literature review

3.2.1 . Asset selection, investment strategies, and portfolios

According to Chen and Mussalli (2020), ESG integration into investment strategies mainly consists in integrating the investors' values into their own strategies. The scientific literature describes three main ways to achieve it: filtering companies based on their ESG scores, directly looking for alpha in ESG data, or measuring ESG impact on other risk factors.

ESG scores can offer a systematic approach to screen out controversial industries, commonly referred to as "sin industries", including but not limited to tobacco,

alcohol, pornography, weapons, etc. For example, some studies advocate for selecting companies with ESG scores surpassing specific thresholds (Schofield et al., 2019). While this method yields good portfolios ESG-wise, Alessandrini and Jondeau (2020) argue that this may lead to underperforming portfolios because of the reduction in the investment universe and the potentially higher returns generated by "sin industries" because of their very exclusion.

Chen and Mussalli (2020) propose a Markowitz-like optimization method by defining an ESG-compatible efficient frontier. Similarly, Hilario-Caballero et al. (2020) add a third term to the mean-variance cost function, the portfolio exposure to carbon risk, and use a genetic algorithm to solve this three-criterion optimization problem. This method is equivalent to optimizing ESG criteria under the constraint of specific risk and return levels. Schofield et al. (2019) also note that the resulting portfolio can have a good global ESG score while containing assets with bad ones.

Finally, Alessandrini and Jondeau (2020) elaborate on "smart beta" strategies, in which investors build portfolios whose assets are not weighted according to their market capitalization but rather to their exposure to some specific risk factors. Bacon and Ossen (2015) explain that integrating ESG into investment strategies can be simply achieved by tilting the asset weights according to their ESG scores while controlling the portfolio exposure to other risk factors. This procedure raises the question of whether ESG is a new risk factor or if optimizing ESG scores amounts to exposing the portfolio to well-known ones. It is indeed a crucial point to explore when attempting to improve portfolio performance with ESG scores (Anson et al., 2020): instead of trying to obtain a premium by finding a suitable ESG factor, it is more judicious to understand the impact of ESG data on the exposure to well-known risk factors.

3.2.2 . ESG scores: risk and returns

Reaching a consensus on the nature of the links between ESG and returns is hard. Friede et al. (2015) aggregate more than 2200 studies on the topic: 41% did not find any ESG impact on returns, 48% found these impacts to be positive and 9% negative. Alessandrini and Jondeau (2020) and Anson et al. (2020) stress the fact that filtering a portfolio on ESG scores leads to improved durability of the investment but does not yield a positive alpha. However, they did not find any proof of negative alpha either. Thus, there may be no added value in integrating ESG data into portfolio construction from an alpha point of view.

Plagge and Grim (2020) find no statistically significant underperformance or overperformance of different equity funds specialized in ESG investing. They argue that since the ESG scores are not of economic nature, they should not have any impact on the portfolios: any information contained in ESG data should already be contained in other risk factors. However, Lee et al. (2022) find, using machine learning methods, that ESG granular data of considered equity funds provide information on the annual financial performance of these funds. Pástor et al.

(2022) use MSCI ESG ratings to evaluate the greenness of US stocks. They find that, between 2012 and 2020, stocks with high environmental ratings outperform stocks with low environmental ratings. They show that these high returns were unexpected in theory and due to raising concerns about environmental issues.

This lack of consensus on the links between ESG and returns may be due to the use of different assessment methodologies (including ESG scores from different data providers) or to a wrong use of the ESG scores (Anson et al., 2020). Indeed, Margot et al. (2021) show that because ESG data have a very low signal-to-noise ratio, using aggregated ESG scores leads to a high loss of information. It is then necessary to use more granular scores to obtain more relevant results. Moreover, they emphasize that the links between ESG and returns are highly dependent on the considered business industry and region. Cappucci (2018) finds that ESG scores lack information on asset price returns and that a better indicator of returns is the progress made by companies in the different ESG sub-fields.

Only a few papers are devoted to the relationship between ESG scores and risk. Guo et al. (2020) train a deep learning model to predict a company's volatility using ESG news. Chen and Mussalli (2020) show that focusing on ESG investments can reduce the risk of underperformance as companies with good ESG scores can be less exposed to both systemic and idiosyncratic risks.

Risk factors

Many studies, such as Renshaw (2018), find that ESG scores and well-known risk factors, such as size, are partially redundant. Anson et al. (2020) and Konqui et al. (2019) study the variation of portfolio exposure to well-known risk factors when one integrates ESG data in portfolio construction: the impact varies according to geographical regions, which reduces the significance of global studies. Similarly, Alessandrini and Jondeau (2020) explain that the discrepancies in ESG portfolio performance in different regions and industries can be attributed to different exposures to risk factors. Furthermore, Breedt et al. (2019) argue that most of the financial performance of a portfolio can be explained by well-known factors and that the residuals cannot be explained by any other factors. For Breedt et al. (2019), the environmental and social aspects of ESG are noise and the governance part is strongly correlated to the quality factor. However, enriching ESG data with other types of information, or preprocessing it, can bring added value. In the same vein, Bacon and Ossen (2015) decorrelate the ESG scores from the other risk factors before integrating them into strategies and are able to obtain added value from ESG scores.

Materiality of ESG data

For a better ESG integration, it is important to understand which ESG features are the most material, i.e. have the largest impact on the financial performance of a company. According to Anson et al. (2020) and Margot et al. (2021), materiality is highly dependent on the chosen asset class, region and industry. Bacon and Ossen (2015) build a materiality matrix using the LASSO method (Tibshirani, 1996). Their matrix is specific to an industry and shows the magnitude of the impact of a specific ESG feature on a company's financial performance versus the probability of this feature having an impact.

Temporality of ESG data

Alessandrini and Jondeau (2020) warn that their results were obtained in a period when a large amount of money was poured into ESG funds, which could have increased their respective performance. Margot et al. (2021) also stress that their study was realized between 2009 and 2018 during a period when the market was particularly bullish, which may have affected the overall strength of ESG-based funds. For Drei et al. (2019), the impact of ESG scores differs not only by region and by industry but also according to the testing strategy and the selected time period. That is why Renshaw (2018) argues that any methodology that treats historical ESG data in the same way for every period is, likely, not relevant. A solution is to use back-testing on several time periods, with several universes, to validate the results (Anson et al., 2020). Finally, Margot et al. (2021) and Plagge and Grim (2020) apply the efficient markets theory in the context of ESG investing: it is possible that investor awareness rises as a function of time and the information included in ESG data is already included in the prices of assets, leading to a loss of predictive power of ESG features and thus of the embedded alpha.

3.3 . Datasets

3.3.1 . Financial data

In this chapter, the following datasets are used:

- Stock prices. We use daily close prices, adjusted for dividends and foreign exchange rates. BNP Paribas internal data sources.
- Market capitalization. BNP Paribas internal data sources.
- Fama–French market, size and value factors: these factors are taken from the online French data library (Fama and French, 2021). They are all computed according to the Fama and French methodology exposed in Fama and French (1993).
- Risk-free rate: these data are also taken from Fama and French (2021) and computed according to the Fama and French method.

In addition, metadata such as The Refinitiv Business Classification (TRBC) sectors at levels 1, 2 and 3 of granularity and the country of incorporation are used. They come from Refinitiv data sources.

3.3.2 . ESG data

ESG data are provided by Refinitiv. Their database alleviates some of the challenges listed above:

1. The coverage of the dataset is sufficient to extract meaningful results. Figure 3.1 shows the number of samples in the geographical regions as defined by Fama and French (Fama and French, 2021): Europe (EUR), North America (NAM), Japan (JPN), Asia-Pacific excluding Japan (APexJ) and emerging countries (EMERGING). Refinitiv ESG data starts in 2002 and the number of samples per year increases several-fold until 2019, as shown in Fig. 3.2. The drop in 2020 is due to the fact that not all the ESG scores had been computed by Refinitv when we had access to the dataset (many companies had not yet published enough data).

2. Scores are built with a well-documented methodology explained in Refinitiv (2020). Every ESG score ranges between 0 and 1, with 1 being the best score. In addition, the same methodology is used throughout the years, yielding consistent data. A score is always provided even when there is missing information: the dataset does not contain any missing data.

3. Human intervention is limited to the gathering of the initial indicators and some quality checks.

4. Scores can be updated up to 5 years after the first publication, which is beneficial in an explanatory setting, as the data become more accurate. In a purely predictive setting, however, this adds noise and look-ahead bias as we do not have point-in-time data, i.e. we do not know the initial and intermediate ESG estimates at their first publication date.

Refinitiv ESG data includes samples from different regions of the world. Each region has specific regulatory frameworks and ESG transparency rules. This is why this paper focuses on the European region and includes all the companies in the Refinitiv ESG dataset whose country of incorporation is in Europe or a European-dependent territory.

The European ESG dataset contains 20,509 samples for 2429 companies uniquely identified by their ISIN. The time evolution of the number of samples per year is reported in Fig. 3.3. All the sectors have enough data, with the notable exception of the Academic and Educational Services sector as shown in Fig. 3.4.

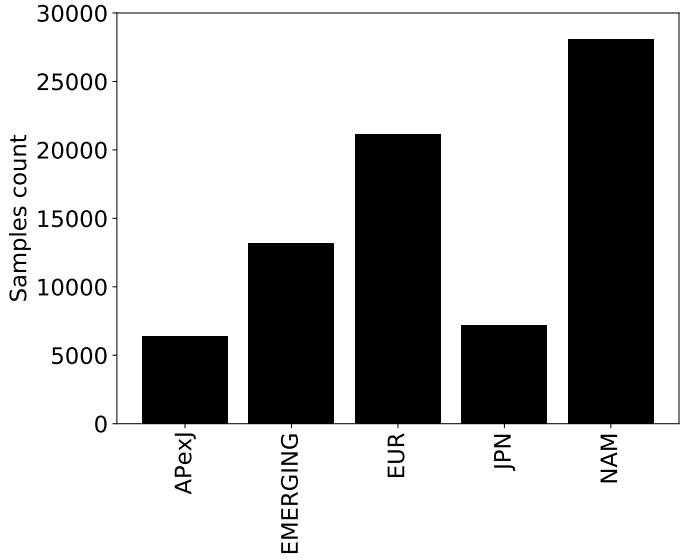


Figure 3.1: Number of samples in each Fama-French region in the Refinitiv ESG dataset.

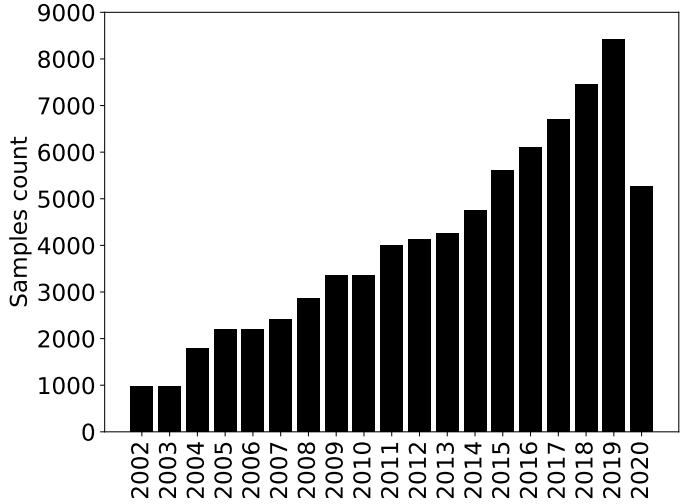


Figure 3.2: Time evolution of the number of samples per year in the Refinitiv ESG dataset.

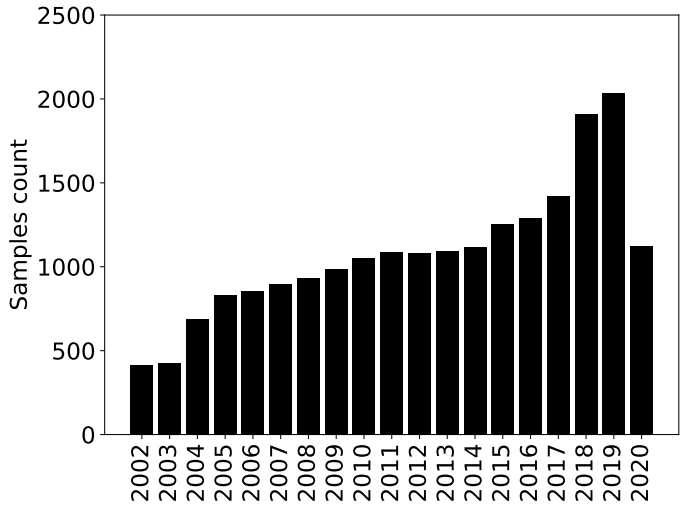


Figure 3.3: Time evolution of the number of samples per year in the Refinitiv ESG dataset - Europe.

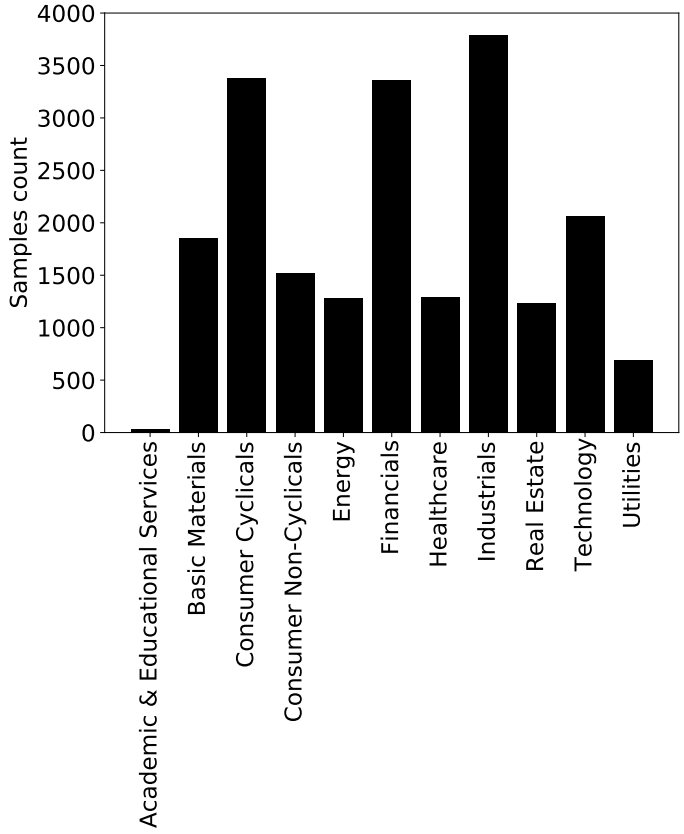


Figure 3.4: Number of samples per TRBC sector of level 1 in the Refinitiv ESG dataset - Europe.

3.4 . Methods

3.4.1 . Problem settings

Our goal is to understand how and what ESG features participate in the formation of price returns. Specifically, we seek to investigate whether ESG features help capture information to explain the parts of stock returns realized at the time of the publication of the ESG data that are not accounted for by well-known equity factors, especially the market, size and value factors. In a multi-factor model, one writes at time t

$$r_{i,t} = r_{f,t} + \sum_k w_{i,k} F_{k,t} + \alpha_i + \epsilon_{i,t}, \tag{3.1}$$

where $r_{i,t}$ is the return of asset i, $r_{f,t}$ the risk-free rate, $F_{k,t}$ the value of factor k at time t and $w_{i,k}$ is the factor loading; the idiosyncratic parts are α_i, the unexplained average return, and the zero-average residuals $\epsilon_{i,t}$. In this work, we use the Capital Asset Pricing Model (CAPM), relying on the market factor (r_m), and its extension, the Fama-French 3-factor model that includes, in addition to the market factor, the size Small Minus Big (SMB) and value High Minus Low (HML) factors (Fama and French, 1993).

ESG data are neither abundant nor of constantly high quality. Directly estimating the explanatory power of ESG features on price returns by estimating the idiosyncratic part of equation (3.1), $\alpha_i + \epsilon_{i,t}$, is a challenging task. Therefore, we settle in this chapter for a less ambitious goal. Specifically, we investigate whether ESG features help explain the sign of the idiosyncratic part of price returns. Mathematically, one needs to explain

$$Y_{i,t} = \begin{cases} 1 & \text{if } \operatorname{sign}(\alpha_i + \epsilon_{i,t}) = 0, \\ \frac{1+\operatorname{sign}(\alpha_i+\epsilon_{i,t})}{2} & \text{otherwise,} \end{cases} \tag{3.2}$$

with the candidate ESG features. Equation (3.2) means that the chosen target is 0 if the sign of the idiosyncratic parts $\alpha_i + \epsilon_{i,t}$ is negative and 1 if this sign is positive or null.

This work takes a machine learning approach to this problem and treats it as a classification problem, using definitions from section 2.1. $Y_{i,t}$ defines two classes as it can take two values: it is a binary classification setting with tabular data.

The state-of-the-art for machine learning with tabular data is gradient boosting models. Gradient boosting models and specifically GBDT are developed in section 2.2 of this dissertation. We use here the GBDT algorithm and its LightGBM implementation.

The models are trained to minimize the cross-entropy loss, defined in section 2.1.3. This type of loss implicitly assumes that both classes appear with roughly similar frequency in the training set, which is the case with 51.7% of samples belonging to class 1 and 48.3% to class 0.

3.4.2 . Training features

The Refinitiv ESG dataset contains several levels of granularity. We choose to train our models with the 10 pillar scores, described in section 1.3.2, Fig. 1.1, to which we add the aggregated Controversy score. This level of granularity is a good compromise between the limitation of the number of features and the necessary granularity to extract meaningful information.

We add five non-ESG features: market capitalization, country of incorporation and TRBC sectors at levels 1, 2 and 3. These features, capturing the size, country and main activity of the considered companies, provide the benchmark features needed to settle the question of the additional information provided by ESG features.

3.4.3 . Target computation

We compute the coefficients of the regression defined in equation (3.2) with monthly factors, available online at Fama and French (2021), and monthly price returns over periods of 5 civil years. For instance, the regression coefficients used to compute the 2017 target, possibly explained by 2017 ESG features, are computed with historical data ranging from 2013 to 2017. We then compute targets over the year corresponding to the year of the publication of the ESG features: as we are in an explanatory setting, we aim to explain the return of a company for a specific year using the ESG profile of this company during the same year.

3.4.4 . Cross-validation and hyperparameter tuning in an increasingly good data universe

The usual strategy of a single data split into causal consecutive train, validation and test data sets may not be fully appropriate for the currently available ESG features. This is because the amount of data grows from a very low baseline, both quantity- and quality-wise, which was not exploitable, to an amount that more likely is. Thus, not only are the data non-stationary but their reliability and quality keep increasing. As a consequence, the cross-validation time-splitting schemes known to work well in the context of non-stationary time series (Bergmeir and Benítez, 2012) may be improved upon, as exposed in section 2.1.2.

For this reason, we experiment with K-times repeated random sub-sampling company-wise cross-validation, where 75% of companies are randomly assigned to the training set and the remaining 25% to the validation set (see Fig. 3.5). In other words, there are K different train+validation sets. For each of the K train sets, we train 180 models, varying 12 hyperparameters of the LightGBM (maximum tree depth, learning rate, etc.) with a random search, and pool the five best ones according to model performance in the respective validation sets. We use in addition early stopping with a patience of 50 iterations. In this way, models are trained with most of the most recent (hence, more relevant) data while also being validated with the most recent and best data. If the dependencies completely

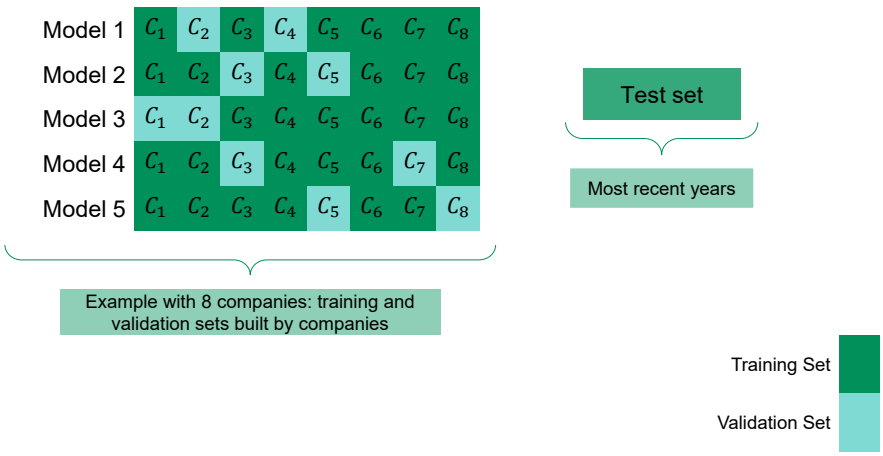


Figure 3.5: 5-times repeated random sub-sampling company-wise cross-validation: the validation sets consist of randomly selected companies, which allows training to account for most of the most recent data.

change every year, this validation scheme is bound to fail. As we shall see, this is not the case. We take $K = 5$.

In addition, we use the rolling calibrations described in section 2.1.2. They are expanding (train+validation)-test windows, using the last year as the test window: it allows us to perform a time-wise analysis of the performance of the models and thus of the explanatory power of ESG features. Because data are insufficient before 2015, we have five different periods: the first test year is 2016 and the last one is 2020.

Using the 5-times repeated random sub-sampling company-wise cross-validation, pooling the five best models in validation and accounting for the rolling calibrations, we then train 125 models.

For each testing period, we will compare the performance of the company-wise random splits with that of the standard temporal split (75% train/25% validation).

3.5 . Results

We investigate the results of the standard temporal split and the 5-times repeated random sub-sampling company-wise cross-validation method for a target computed using the CAPM model, as described in section 3.4.1.

We first assess the quality of the models according to the cross-entropy loss, using their direct probability outputs. We also assess the end result, i.e. the predicted class. As it is usual, we map the output, a probability p_i, to classes 0 and 1 with respect to a 0.5 threshold. This allows us to compute the balanced accuracy.

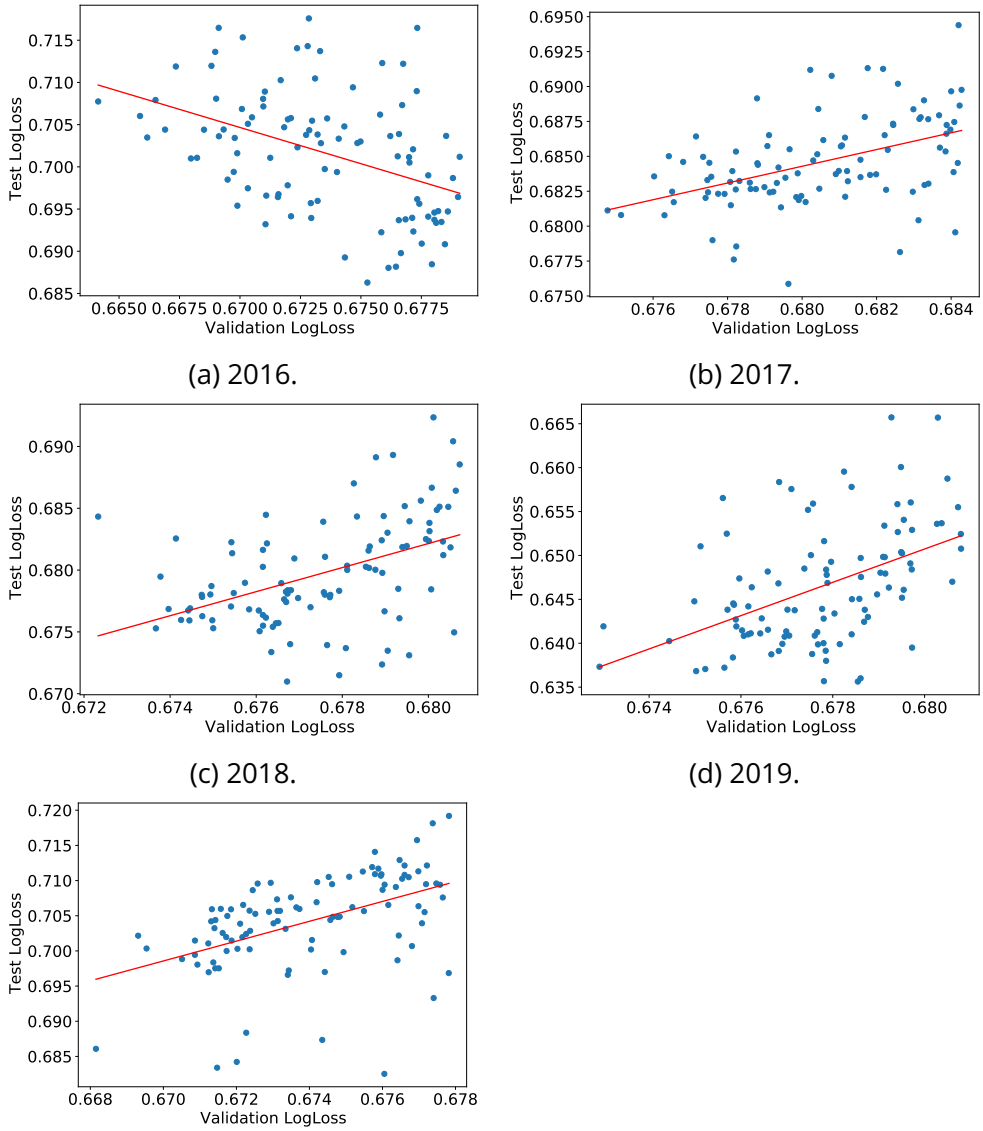
An advantage of balanced accuracy over classical accuracy is that balanced accuracy accounts for class imbalance in the test set. By definition, it assigns a score of 0.5 if the model did not learn anything significant. These metrics are further detailed in section 2.1.3.

We check that the performance of the models in the test sets bears some relationship with their performance in the validation sets. More precisely, for each (train+validation)-test period, we investigate the dependence between the cross-entropy losses in the validation and test sets, respectively noted $\mathcal{L}_m^{\text{validation}}$ and $\mathcal{L}_m^{\text{test}}$, for the best models trained during the hyperparameters random search. It makes it possible to characterize the training quality year by year. A significantly positive relationship shows that these models did learn persistent relationships, i.e., something useful. Mathematically, we assess the relationship $\mathcal{L}_m^{\text{test}}$ versus $\mathcal{L}_m^{\text{validation}}$ for each model m, selecting the 100 models with the best validation cross-entropy losses for each of the five sets of (train+validation)-test sets. Figure 3.6 shows these relationships for the standard time-splitting scheme and adds a linear fit. Figure 3.7 displays these relationships for the company-wise cross-validation scheme and adds a linear fit. Generally, both test and validation cross-entropy losses are positively correlated, except for 2016. We believe that this comes from the fact that ESG data were of insufficient quality before that date. The year 2020 is also special: in addition to the coronavirus crisis, the data for 2020 were obtained at the beginning of 2021 when not all companies had ESG ratings, leading to a smaller dataset and a (mostly likely) biased test set.

We compute the Pearson correlation, the R^2 of the linear fit, the Kendall tau and its p-value for the standard temporal split and the 5-times repeated random sub-sampling company-wise cross-validation, which are reported in Tab. 3.1. This table allows us to compare the respective advantages and disadvantages of each validation strategy. All the dependence measures increase significantly from 2017 to 2019 for company-wise splits. The case of the temporal split shows the limitations of this approach: the performance measures are roughly constant, which is consistent with the fact that adding one year of data to the train+validation dataset does not lead to much change in what the model learns.

Our second and most important aim is to establish that ESG data contains additional valuable and exploitable information on price returns in comparison to a set of benchmark features defined in section 3.4.2. To this end, for each training period defined above, we train a model with both ESG and benchmark features and another model with benchmark features alone. We assess both the absolute performance metrics of the models and the extent of additional information provided by ESG features by calculating the difference in performance metrics on the test sets.

The company-wise splits make it easy to compute error bars on various metrics: instead of training $K = 5$ models, we train 100 of them and then compute the median performance on 100 random subsets of size $K = 5$ among these 100



(a) 2016.

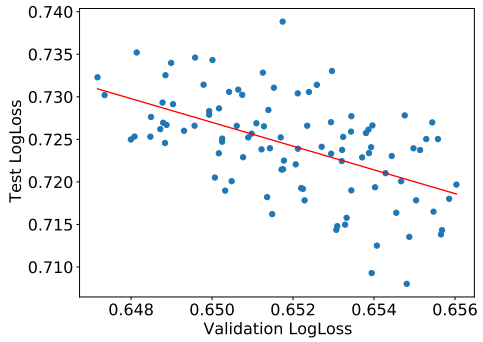
(b) 2017.

(c) 2018.

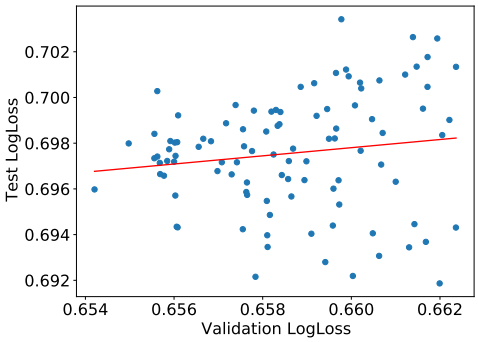
(d) 2019.

(e) 2020.

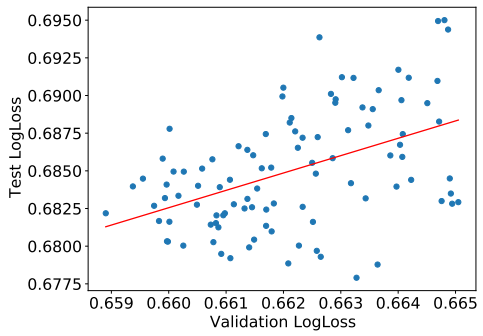
Figure 3.6: Standard temporal cross-validation: test cross-entropy versus validation cross-entropy of the 100 best models of the random hyperparameters search.



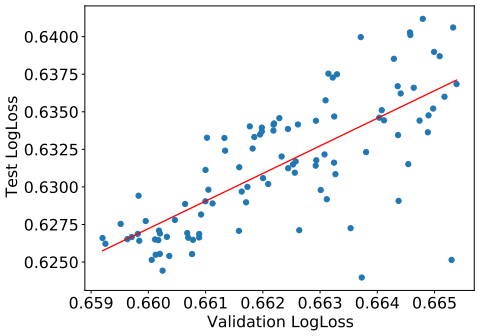
(a) 2016.



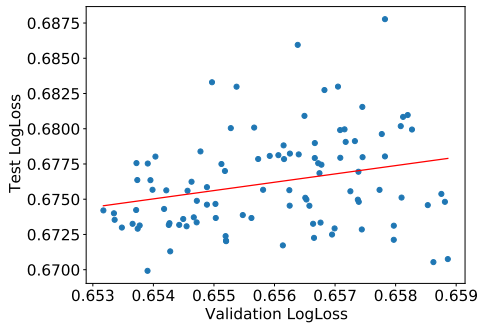
(b) 2017.



(c) 2018.



(d) 2019.



(e) 2020.

Figure 3.7: Company-wise cross-validation: test set cross-entropy versus validation cross-entropy of the 100 best models of the random hyperparameters search.

5-times repeated random sub-sampling company-wise cross-validation				
Year	Pearson correlation	R^2	Kendall tau	p-value of Kendall tau
2016	-0.54	0.29	-0.36	$8.0\mathrm{e}^{-8}$
2017	0.14	0.021	0.12	$6.7\mathrm{e}^{-2}$
2018	0.47	0.22	0.30	$1.1\mathrm{e}^{-5}$
2019	0.73	0.54	0.58	$1.5\mathrm{e}^{-17}$
2020	0.27	0.071	0.19	$5.4\mathrm{e}^{-3}$
Standard temporal split				
Year	Pearson correlation	R^2	Kendall tau	p-value of Kendall tau
2016	-0.43	0.18	-0.29	$1.6\mathrm{e}^{-5}$
2017	0.46	0.21	0.33	$9.2\mathrm{e}^{-7}$
2018	0.46	0.21	0.34	$7.7\mathrm{e}^{-7}$
2019	0.47	0.22	0.33	$1.3\mathrm{e}^{-6}$
2020	0.47	0.22	0.39	$7.6\mathrm{e}^{-9}$

Table 3.1: Dependence measures between the cross-entropies in the validation and test sets, for the 100 best models of the random hyper-parameters search.

models. Table 3.2 provides results on the absolute performance of the models for each test period for both the company-wise and the standard temporal splits. Both splitting methods have a decreasing cross-entropy as a function of time, except for 2020, which shows once again the special nature of this year in our dataset. This shows that the relevance of ESG features in price return formation increases as a function of time. Balanced accuracy displays a similar improvement before 2020. However, this time, yields of company-wise splits are increasingly better than temporal splits, which we believe is an encouraging sign of its ability to better leverage the latest and best data.

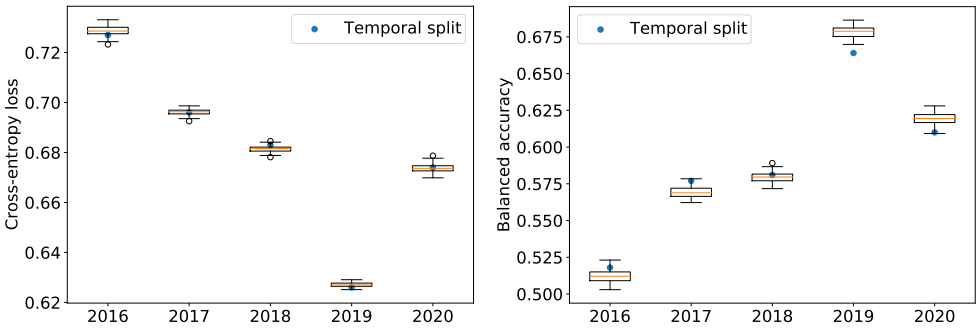
Figure 3.8 displays the time evolution of the cross-entropy and the balanced accuracy in the test sets. The boxplots are computed for the company-wise splits from the 100 associated predictions: the orange lines are the median of these performance measures, the rectangle delimits the first and third quartiles, and extreme limits are situated before the first quartile minus 1.5 times the interquartile range and after the third quartile plus 1.5 times the interquartile range. Any point outside of this range is considered an outlier.

The 5-times repeated random sub-sampling company-wise cross-validation outperforms the standard time-splitting scheme, which supports our claim that the not fully mature nature of ESG data can be partly alleviated by a suitable validation scheme.

Figure 3.9 shows the difference in performance between the models trained on ESG and benchmark features and the models trained only on benchmark features for the 5-times repeated random sub-sampling company-wise cross-validation. ESG features contain more relevant information as time goes on. Two explanations spring to mind: long positions are more and more driven by ESG-conscious

5-times repeated random sub-sampling company-wise cross-validation				
	Only Benchmark features		Benchmark and ESG features	
Year	Balanced Accuracy	Cross-entropy loss	Balanced Accuracy	Cross-entropy loss
2016	52.6	70.6	51.2	72.8
2017	57.4	69.2	56.9	69.6
2018	57.5	68.1	57.9	68.2
2019	65.6	63.1	67.9	62.7
2020	59.6	69.3	61.9	67.4
Standard temporal split				
	Only Benchmark features		Benchmark and ESG features	
Year	Balanced Accuracy	Cross-entropy loss	Balanced Accuracy	Cross-entropy loss
2016	53.2	68.8	51.8	70.3
2017	56.1	68.2	57.7	68.0
2018	56.2	67.5	58.1	67.4
2019	64.3	64.5	66.4	63.8
2020	58.5	70.5	61.0	69.6

Table 3.2: Performance measures in percent on the test set for both types of validation splits. The numbers for the company-wise splits are the median values of the performance of 100 random samplings of 5 models among 100 random company-wise validation splits.



(a) Cross-entropy.

(b) Balanced accuracy.

Figure 3.8: Performance measures on the test sets of the two train and validation schemes. The boxplots show the performance of 100 random samplings of 5 models among 100 random company-wise validation splits.

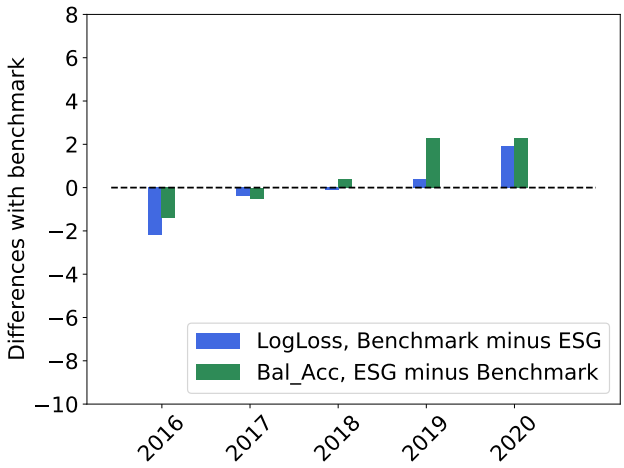


Figure 3.9: Performance measures with respect to benchmark for the 5-times repeated random sub-sampling company-wise cross-validation.

investors, or the quality of data increases as a function of time, which makes the relevance of ESG scores more apparent.

3.6 . Interpretability

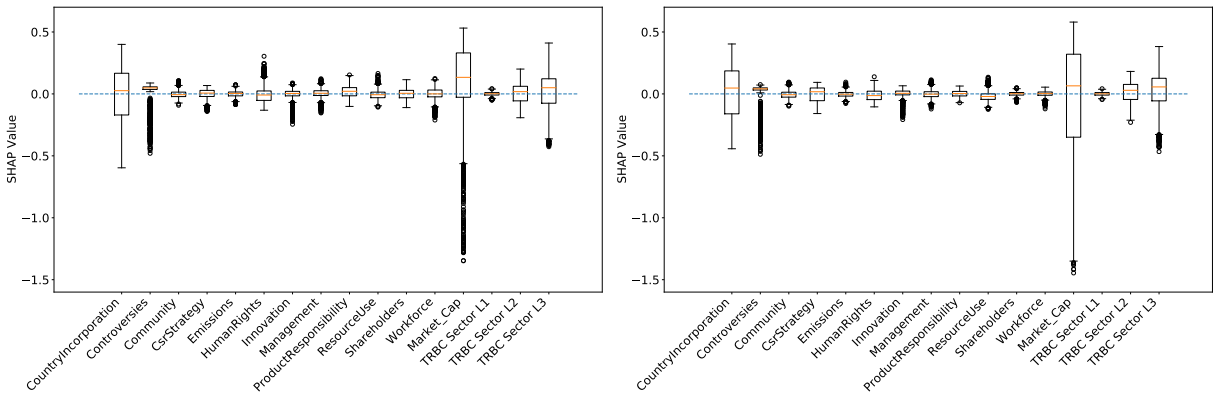
We now provide a breakdown of the impact of the different ESG features on the predicted probability of having positive idiosyncratic returns in the CAPM model. Because of the superior performance of the K-times repeated random sub-sampling company-wise cross-validation, we use this method in the following.

3.6.1 . Shapley values

Shapley values, first introduced in the context of game theory (Shapley, 1953), provide a way to characterize how each feature contributes to the formation of the final predictions. We provide in section 2.3.1 a mathematical explanation of the Shapley values and its SHAP applications in the field of machine learning, following Lundberg and Lee (2017) and Lundberg et al. (2018) research.

Let us note that, as we are using a LightGBM model in a classification setting, the prediction is not directly the probability of belonging to class 1, but rather the logit associated with this probability. Probability is an increasing function of the logit, and thus, SHAP values obtained for the logit can easily be transformed for the probability. Indeed, for a sample x_i, the predicted probability of belonging to class 1 p_i is linked to the logit logit_i according to

$$p_i = \frac{1}{1 + e^{-\text{logit}_i}}. \tag{3.3}$$



(a) Test year 2017.

(b) Test year 2020.

Figure 3.10: SHAP values distribution according to selected test year.

Evolution of ESG features contribution from 2017 to 2020

In Fig. 3.10, we plot the distribution of SHAP values for each feature and all test samples for models trained from 2002 to 2016 (Fig. 3.10a) and trained from 2002 to 2019 (Fig. 3.10b). The first teaching of this plot is that the contribution of ESG features to the predicted probability of having a positive return has not dramatically increased with the additional, more recent and more complete data. Benchmark features are the ones that have the most important impact impact on the prediction. However, we observe an important number of outliers for some SHAP values associated with some features, demonstrating that these ESG features have more impact on the prediction for these particular samples. It would be interesting to study these outliers to understand more why ESG features are more important in explaining price returns for some samples rather than others.

For instance, we observe in Fig. 3.11 the scores distributions for the outliers of the Controversy SHAP values. All Controversy scores are below 0.9, suggesting that the Controversy score is more informative when a company has indeed suffered controversies during the year and was then not able to reach a perfect score of 1. Observing outliers of SHAP values and their associated scores, we can make the hypothesis that ESG features are important and have a strong impact on the explanations of past returns if their score is extreme. This would mean that ESG information would lie in extreme scores, with more standard scores bringing much less information. Checking this hypothesis is beyond the scope of this work and is left for future investigations.

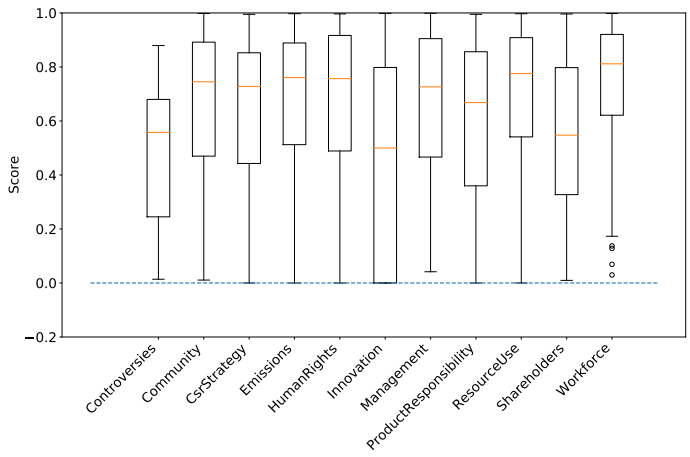


Figure 3.11: Distribution of data for lowest outliers of SHAP values for the 2020 test year and the Controversy score.

3.6.2 . Partial dependence plots: marginal effect of ESG features

A partial dependence plot shows the marginal effect of features on the prediction made by the model. It is a way of understanding the links the model made from features to the target and that it had understood from the data. Partial dependence plots are detailed in section 2.3.2. All partial dependence plots in this section are made with the most recent model, trained with data from 2002 to 2019, on a subsample of recent ESG data.

Marginal effect of ESG features

Using partial dependence plots, we first compute the marginal effect of each ESG feature on the probability of having a positive return during the year of publication of the ESG features (Fig. 3.12). Figure 3.13 reports the sector-by-sector probability of having a positive predicted return.

Figure 3.12 shows that ESG features are mostly not related in a monotonic way with the probability of having a positive return. A clear exception would be the Controversy score, on the top left, which shows a strong monotonic relation and strongly implies that being subject to controversies during a year leads to a lower probability of having a positive return. For the 10 pillar scores, one sees a much weaker dependence. For example, the probability of positive price return increases by 1 to 2% when the Product Responsibility and Shareholders scores increase from 0 to 1. Still, a trend is present for most of these ESG features: partial dependence plots for scores such as Resource Use, Innovation, Community or Management seem to be decreasing, suggesting that obtaining better ESG scores and practices comes at the price of slightly degraded financial performance.

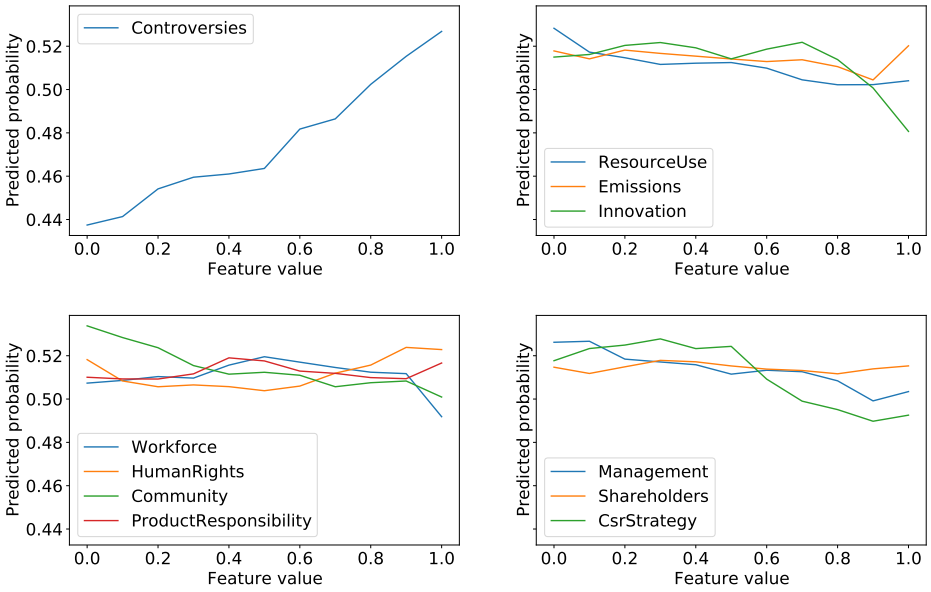


Figure 3.12: Marginal effect of each ESG feature on the predicted probability of having a positive return.

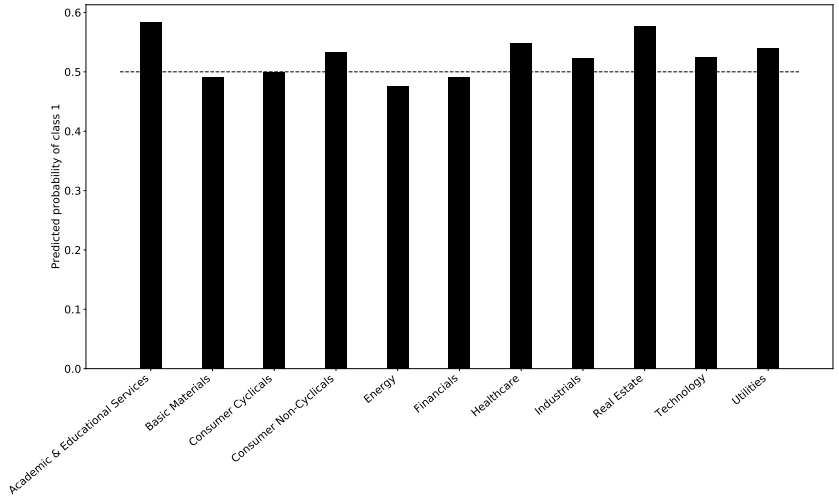


Figure 3.13: Marginal effect of the sector (TRBC sector at level 1) feature on the predicted probability of having a positive return.

Marginal effect of ESG features sector by sector: materiality matrices

Adding the sector dimension to partial dependence plots yields so-called materiality matrices. In our setting, it is a table whose rows represent ESG features and whose columns are economic sectors. A cell of this matrix shows how much the probability, expressed in percentage, of having a positive return is increased by going from a low score (between 0 and 0.2) to a high one (by 0.8 to 1). This quantity is easily obtained using partial dependence plots: for a specific selected economic sector, we can plot the evolution of the predicted probability against the feature value. Making the strong hypothesis of a monotonic and close-to-linear relationship, we can compute the value in the cell as the slope of the trend line of the precedent plot.

The obtained materiality matrix is presented in Figure 3.14. All the TRBC sectors at level 1 are included. Results for Academic and Educational Services should be handled with care as they are not based on as many samples as the ones for other sectors as shown in Fig. 3.4. Some ESG scores have a strong impact on the probability of having positive returns. The Controversy score especially has a similar impact for all sectors: not suffering controversies during the year increases the probability of having a positive return. On the contrary, the CSR Strategy row shows that working towards the integration of social and environmental dimensions into the day-to-day decision-making processes, in addition to economic and financial ones, leads to a loss of financial performance. It is also the case for Resource Use, Environmental Innovation, Community, and Management scores, each with a different magnitude.

Furthermore, we bucket the companies that serve to build this materiality matrix by market capitalization. We choose three buckets, with small market capitalization being below 2 billion euros, mid ones between 2 and 10 billion euros and large ones above 10 billion euros, which correspond to the buckets used by Refinitiv when calculating the Controversy score. The three obtained materiality matrices are presented in Fig. 3.15. The marginal effect of the Controversy score remains the same, even if it is slightly smaller for the small capitalizations. However, companies with a large market capitalization benefit from a better impact of ESG: for some features, working toward better ESG scores can preserve or even boost financial performance, whereas it would be the opposite for small capitalizations. For instance, large capitalization companies have an average materiality of 0.8 for the Resource Use score and 1.5 for the Emissions score, whereas small caps ones have respectively average scores of -4.6 and -1.1, denoting a clear difference.

To obtain a statistically meaningful interpretation of these results, we need to account for the fact that each cell corresponds to coefficients of a linear fit with associated p-values, i.e. one makes one null hypothesis per cell. We thus need to use multiple hypothesis correction to check globally which cells show statistically significant results. Here, we choose to control the False Discovery Rate (FDR) with the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). We set

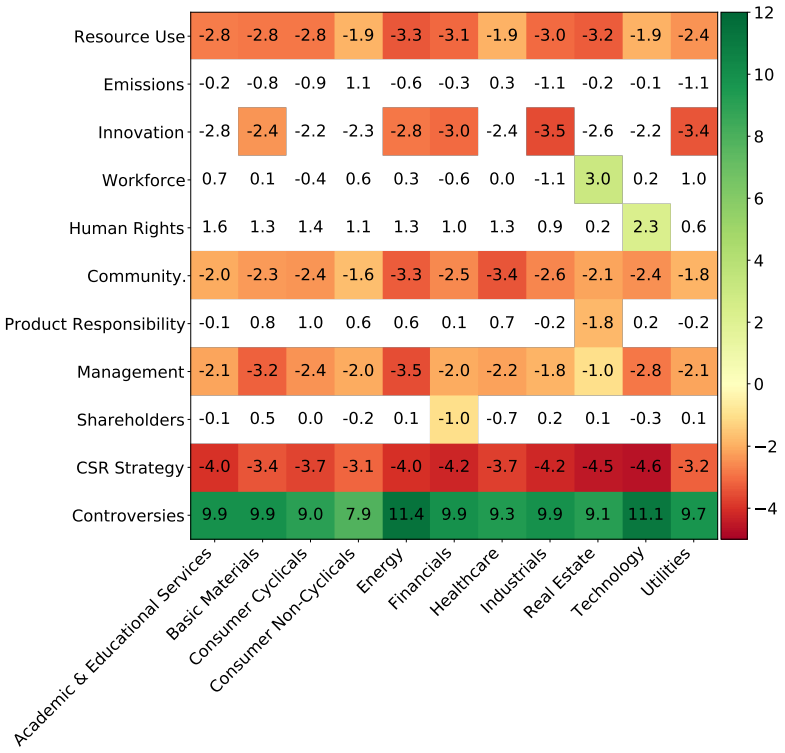


Figure 3.14: Materiality matrix: marginal effects of the combination ESG feature/sector feature on the predicted probability of having a positive return. Blank cells are those which were not found statistically significant by the Benjamini–Hochberg procedure.

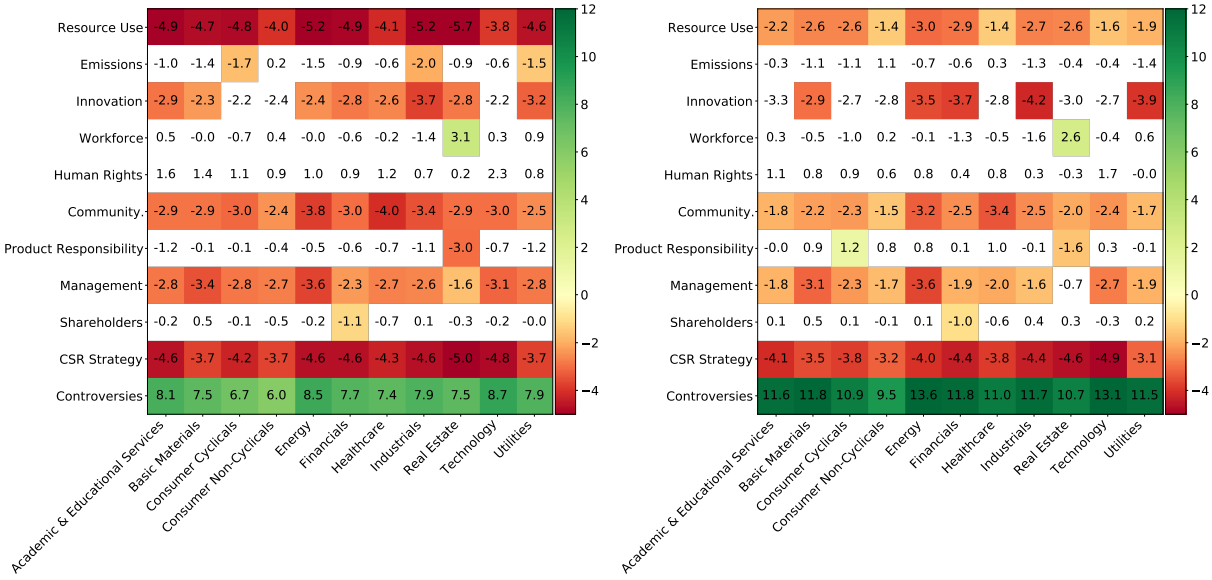
the FDR to 5%, which means that there are only about three false discoveries in each of the reported tables.

3.7 . Additional experiments

3.7.1 . Results using the target derived from the Fama-French 3-factor model

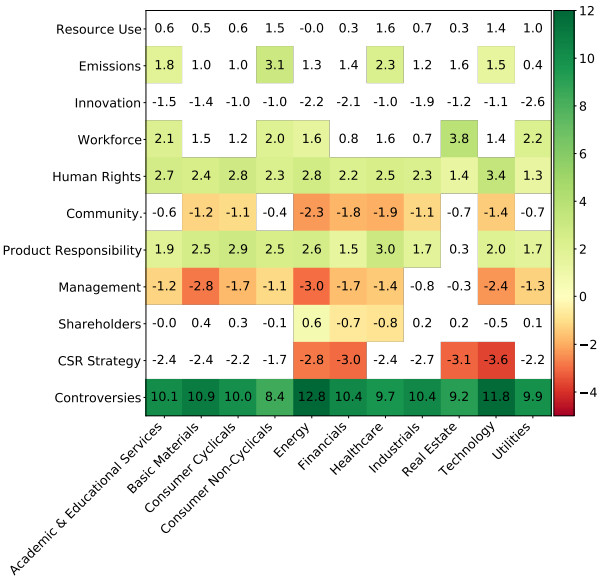
The following results were obtained with a target derived from the Fama-French 3-factor model following section 3.4.1. The target is composed of 47.72% of class 0 and 52.28% of class 1. This target was not selected as the results were not as satisfactory as those obtained with the target derived from the CAPM model. How to interpret results with this target, especially in terms of materiality matrices, was also less clear.

We present in Tab. 3.3 and in Fig. 3.16 results of the study of the relationship $\mathcal{L}_m^{\text{test}}$ versus $\mathcal{L}_m^{\text{validation}}$ for each model m within the top 100 validation cross-entropy losses, using a 5-times repeated random sub-sampling company-wise



(a) Small market capitalization (<2B€).

(b) Mid market capitalization (>2B€, <10B€).



(c) Large market capitalization (>10B€).

Figure 3.15: Materiality matrices: marginal effects of the combination ESG feature/sector feature on the predicted probability of having a positive return, bucketed by market capitalization. Blank cells are those which were not found statistically significant by the Benjamini–Hochberg procedure.

5-times repeated random sub-sampling company-wise cross-validation

Year	Pearson correlation	R^2	Kendall tau	p-value of Kendall tau
2016	-0.23	0.052	-0.14	4.4^{-2}
2017	-0.054	0.0030	0.010	8.8^{-1}
2018	0.29	0.085	0.19	4.2^{-3}
2019	0.67	0.44	0.49	7.1^{-13}
2020	0.053	0.0028	0.017	8.0^{-1}

Table 3.3: Dependence measures between the cross-entropy losses in the validation and test sets, for the 100 best models of the random hyperparameters search, for a target computed using the Fama-French 3-factor model; Refinitiv ESG dataset.

5-times repeated random sub-sampling company-wise cross-validation

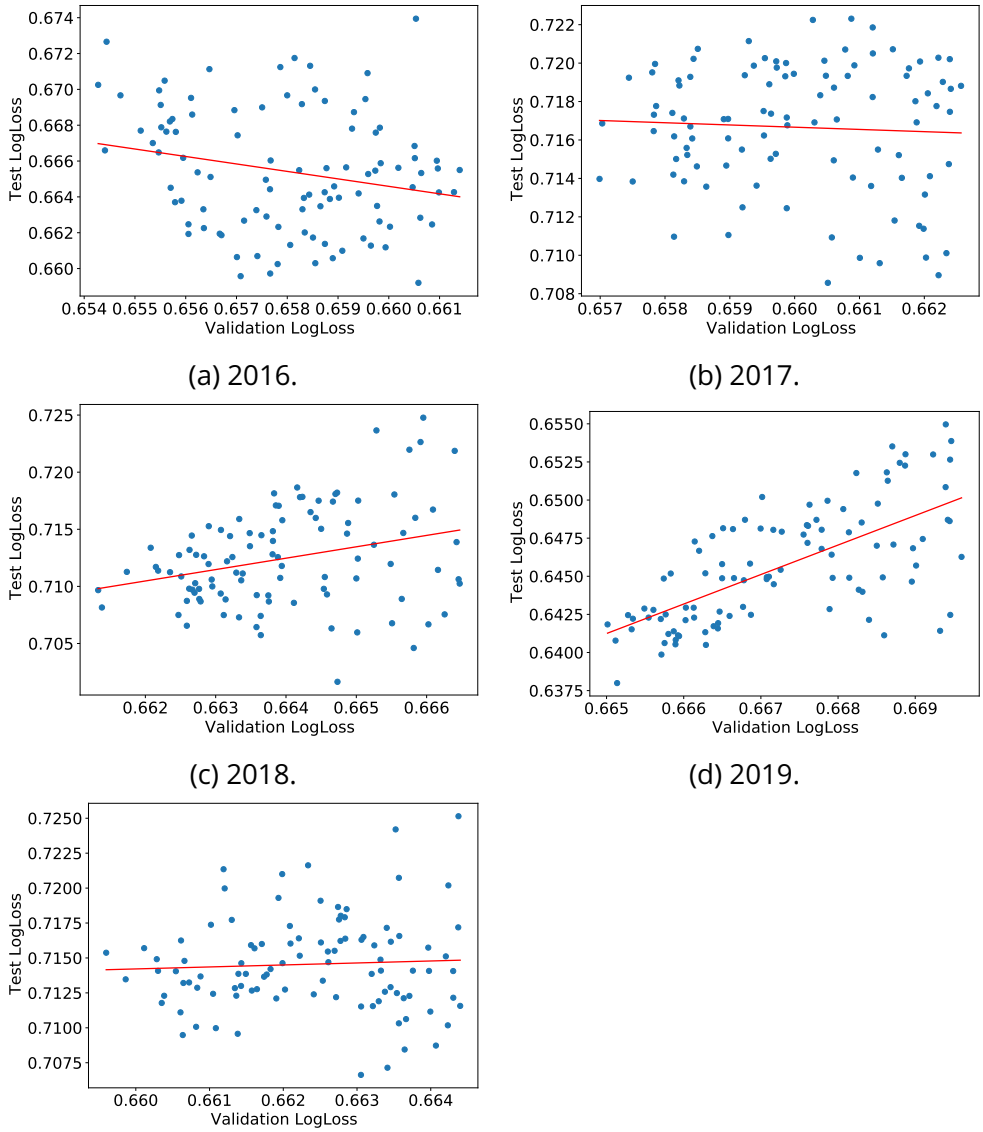
	Only Benchmark features		Benchmark and ESG features	
Year	Balanced Accuracy	Cross-entropy loss	Balanced Accuracy	Cross-entropy loss
2016	57.9	65.8	56.0	66.7
2017	55.0	70.6	55.2	71.6
2018	56.0	70.4	56.0	71.1
2019	62.4	64.6	64.7	64.1
2020	56.1	72.2	55.3	71.3

Table 3.4: Performance measures in percent on the test set, for a target computed using the Fama-French 3-factor model; Refinitiv ESG dataset.

splitting strategy. Similarly to what we observed for the target derived from the CAPM, the dependence measures increase significantly from 2017 to 2019 showing that the models start learning persistent relationships over these years. However, clear conclusions are harder to reach when analyzing the differences in performance showed in Tab. 3.4, between a model trained only on benchmark features and a model trained on both ESG and benchmark features.

3.7.2 . Application to MSCI data

The results obtained in the first part of this chapter are strongly dependent on the chosen dataset. Different ESG scores providers use different methodologies that capture different aspects of the ESG profile of a company (see section 1.3.2). Changing the ESG dataset used in this study leads to different results. In this experiment, we check this hypothesis by using the MSCI ESG scores from 2007 to 2020, derived from the MSCI ESG Ratings methodology (MSCI ESG Research, 2020), for a European universe of 2403 companies corresponding to 27,243 samples. The dataset is built using the same methodology described in section 3.4.1. The target is composed of 48.66% of class 0 and 51.34% of class 1. Figure 3.17 displays the evolution of the number of samples per year for the processed dataset. Models are trained using the 5-times repeated random sub-sampling company-wise



(a) 2016.

(b) 2017.

(c) 2018.

(d) 2019.

(e) 2020.

Figure 3.16: Company-wise cross-validation: test cross-entropy versus validation cross-entropy of the 100 best models of the random hyper-parameters search, for a target computed using the Fama-French 3-factor model; Refinitiv ESG dataset.

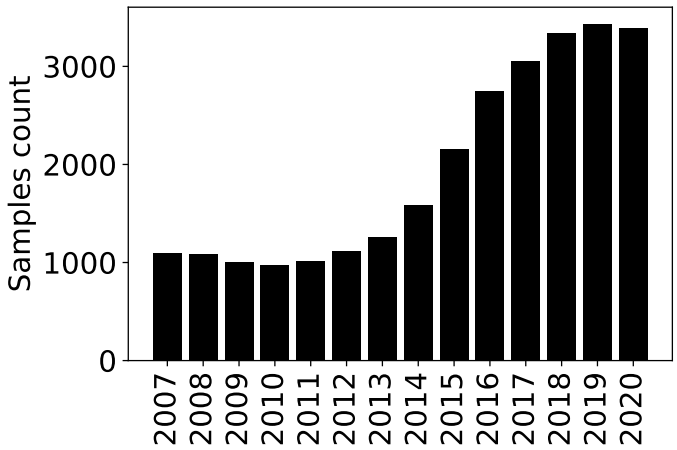
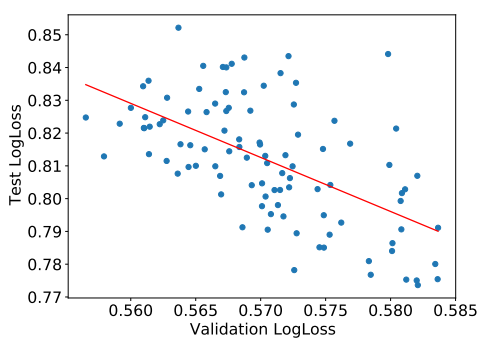


Figure 3.17: Time evolution of the number of samples in the MSCI ESG dataset used for explanation of price returns.

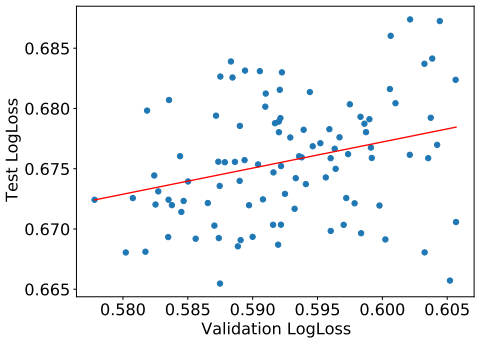
cross-validation methodology.

Figure 3.18 and the associated metrics available in Tab. 3.5 show the relationship between the cross-entropy losses in the validation and test set for the 100 models trained with different sets of hyperparameters and which have the best, i.e. the lowest, validation cross-entropy loss. Similarly to the results obtained in section 3.5, we observe that for years 2017 and 2019, both validation and test cross-entropy losses are positively correlated. The correlation for the year 2016 is negative, which we believe is due to the lower quality of data before 2016. Correlation is also not good for the year 2020 and correlation for the year 2018 shows a drop in comparison to 2017 and 2019 while remaining positive. This may be due to the procedure used by MSCI to compute its scores in comparison to the Refinitiv one: Refinitiv scores are built using a systematic and quantitative methodology, the same being applied each year, while MSCI relies more on analysts that may differ from one year to another, or one company to another, explaining the lower correlations we observe between validation and test cross-entropy losses for some years.

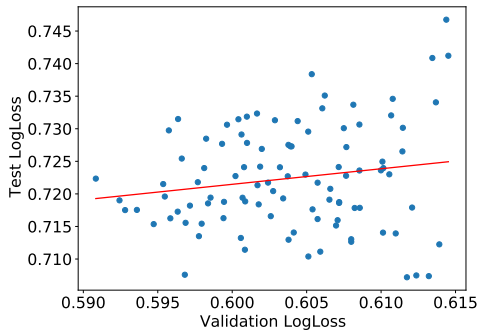
Table 3.6 and Fig. 3.19 and 3.20 show cross-entropy and balanced accuracy results on the test sets. In practice, we train 25 models using the company-wise cross-validation method and then compute the median performance on 100 random subsets of size $K = 5$ among these 25 models, similar to what is done in section 3.5. Results seem less clear-cut than the ones using the Refinitiv dataset. The underperformance of the model trained with ESG data in comparison to the benchmark is much higher in 2016 for the MSCI dataset than for the Refinintiv one. On the opposite, we find an overperformance for the year 2017 that was not shown by the model trained on the Refinintiv dataset. The observed trend on the Refinitiv dataset in which we found that ESG data increasingly participate over



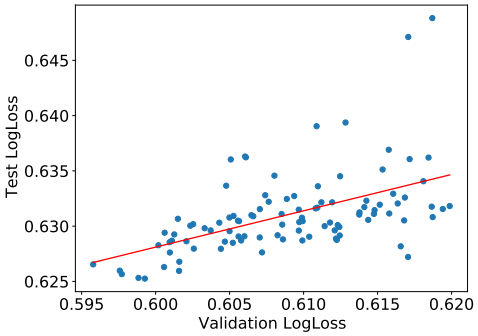
(a) 2016.



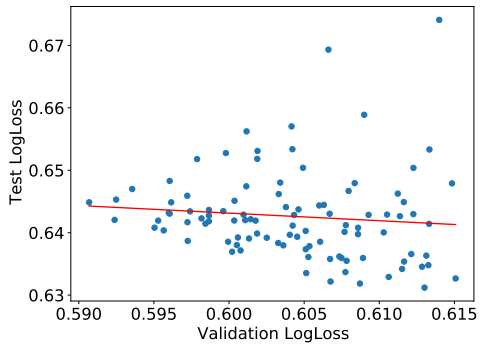
(b) 2017.



(c) 2018.



(d) 2019.



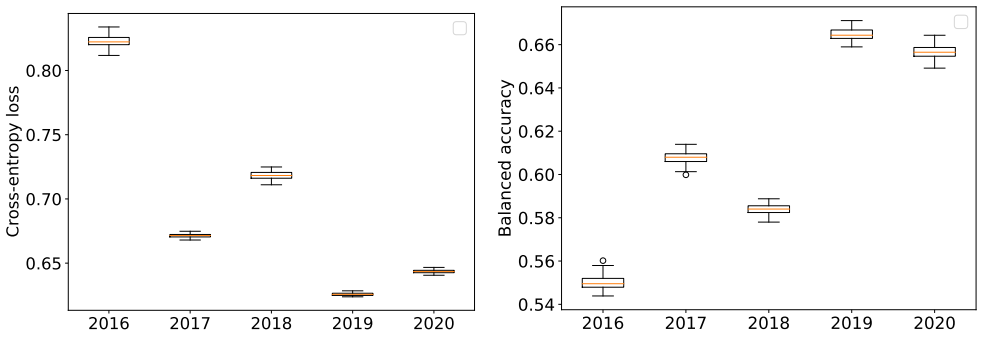
(e) 2020.

Figure 3.18: Company-wise cross-validation: test cross-entropy versus validation cross-entropy of the 100 best models of the random hyper-parameters search; MSCI ESG dataset.

5-times repeated random sub-sampling company-wise cross-validation

Year	Pearson correlation	R^2	Kendall tau	p-value of Kendall tau
2016	-0.59	0.34	-0.41	1.7e^{-9}
2017	0.29	0.086	0.20	2.8e^{-3}
2018	0.17	0.029	0.083	2.2e^{-1}
2019	0.52	0.27	0.43	3.6e^{-10}
2020	-0.010	0.0099	-0.17	1.2e^{-2}

Table 3.5: Dependence measures between the cross-entropies in the validation and test sets for the 100 best models of the random hyper-parameters search; MSCI ESG dataset.



(a) Cross-entropy. (b) Balanced accuracy.

Figure 3.19: Performance measures on the test sets of the 5-times repeated random sub-sampling cross-validation scheme; MSCI ESG dataset. The boxplots show the performance of 100 random samplings of 5 models among 25 random company-wise validation splits.

time in the formation of equity returns remains to be validated using the MSCI ESG scores when more years of data will be available.

3.7.3 . Explaining the full value of the idiosyncratic part of price returns

In section 3.4.1, we investigate whether ESG features help explain the sign of the part of price returns not accounted for by the market factor. We go further in this section by researching if the developed methodology can yield interesting results when explaining the full value of the idiosyncratic part of price returns.

In both experiments, we apply the defined methodology, optimizing a regression loss, the MSE, using a LightGBM model. Performance is evaluated using the RMSE, the MAE and the balanced accuracy computed on the sign of the estimated return from the model. These metrics are further detailed in section 2.1.4.

5-times repeated random sub-sampling company-wise cross-validation

	Only Benchmark features		Benchmark and ESG features	
Year	Balanced Accuracy	Cross-entropy loss	Balanced Accuracy	Cross-entropy loss
2016	56.4	69.4	55.0	82.2
2017	56.4	69.1	60.8	67.2
2018	57.8	69.3	58.4	71.8
2019	60.9	64.7	66.4	62.6
2020	63.2	64.8	65.6	64.3

Table 3.6: Performance measures in percent on the test set; MSCI ESG dataset. The numbers for the company-wise splits are the median values of the performance of 100 random samplings of 5 models among 25 random company-wise validation splits.

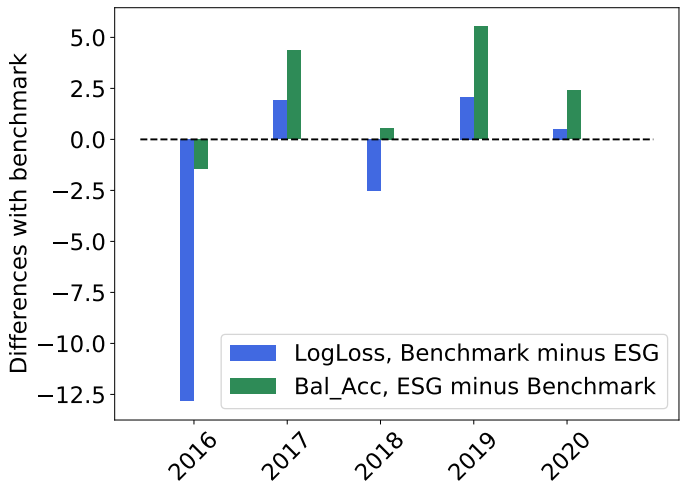
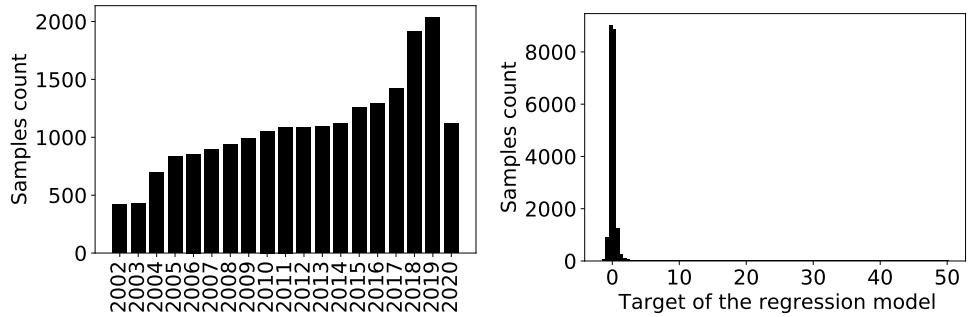


Figure 3.20: Performance measures with respect to benchmark, for the 5-times repeated random sub-sampling company-wise cross-validation; MSCI ESG dataset.



(a) Time evolution of the number of samples in the Refinitiv ESG dataset used for explanation in a regression setting.

(b) Histogram of the targets associated with the Refinitiv ESG dataset used for explanation in a regression setting.

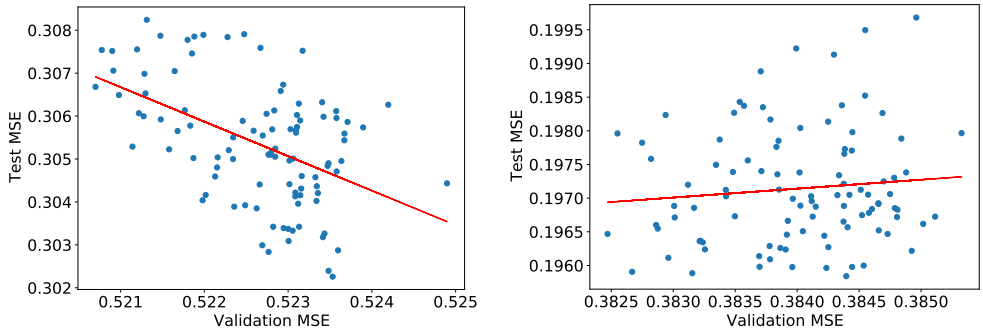
Figure 3.21: Description of the Refinitiv ESG dataset used in regression without target filtering.

Training of the full dataset without filtering the targets

In this experiment, we apply the proposed methodology and train the models using a 5-times repeated random sub-sampling company-wise cross-validation on the full dataset. Figure 3.21a shows the evolution of the number of available samples with time for the Refinitiv dataset, using the full data of 20,541 samples and 2433 companies. The distribution of target is shown in Fig. 3.21b: most of the targets lie between -1 and 1, which is logical for equity returns but we observe numbers of outliers. They are due to important overperformance or underperformance compared to the market for the associated stocks. There are 60 samples with a target below -1 and 500 samples with a target above 1. Because of the scale of the Y-axis of Fig. 3.21b, they cannot be detected in this figure. The scale of the X-axis gives an idea of how spread they are. In this section, we do not filter the dataset by removing these outliers.

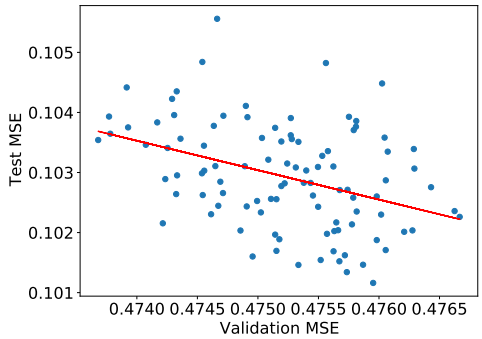
Figure 3.22 and Tab. 3.7 show the relationship between the validation and test MSE losses. Correlations are not as good as the ones for the model trained to explain the sign of the idiosyncratic part of price returns, especially for the year 2018. They greatly improve in 2019 and 2020.

Performance is displayed in Tab. 3.8 and in Fig. 3.23 and 3.24. In practice, we train 25 models using the company-wise cross-validation method and then compute the median performance on 100 random subsets of size $K = 5$ among these 25 models. Performance of the year 2016 shows that the model makes large mistakes in its estimations, as illustrated by the high RMSE in comparison to MAE. Figure 3.24 still displays the trend that we found in classification, with ESG features increasingly explaining the idiosyncratic part of price returns when considering only the market factor.

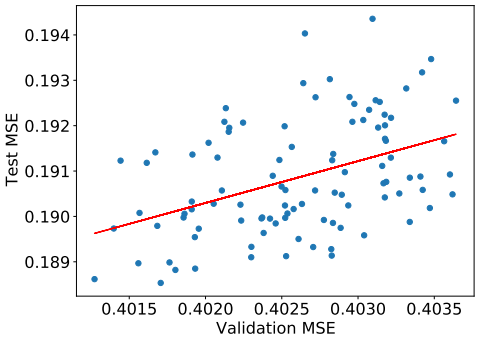


(a) 2016.

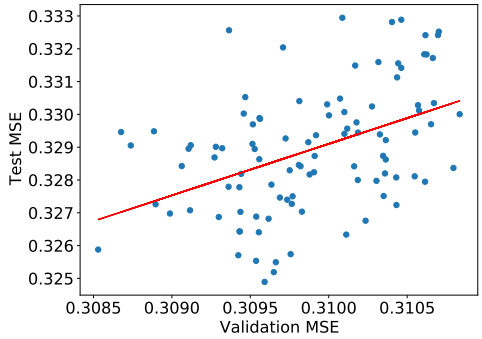
(b) 2017.



(c) 2018.



(d) 2019.



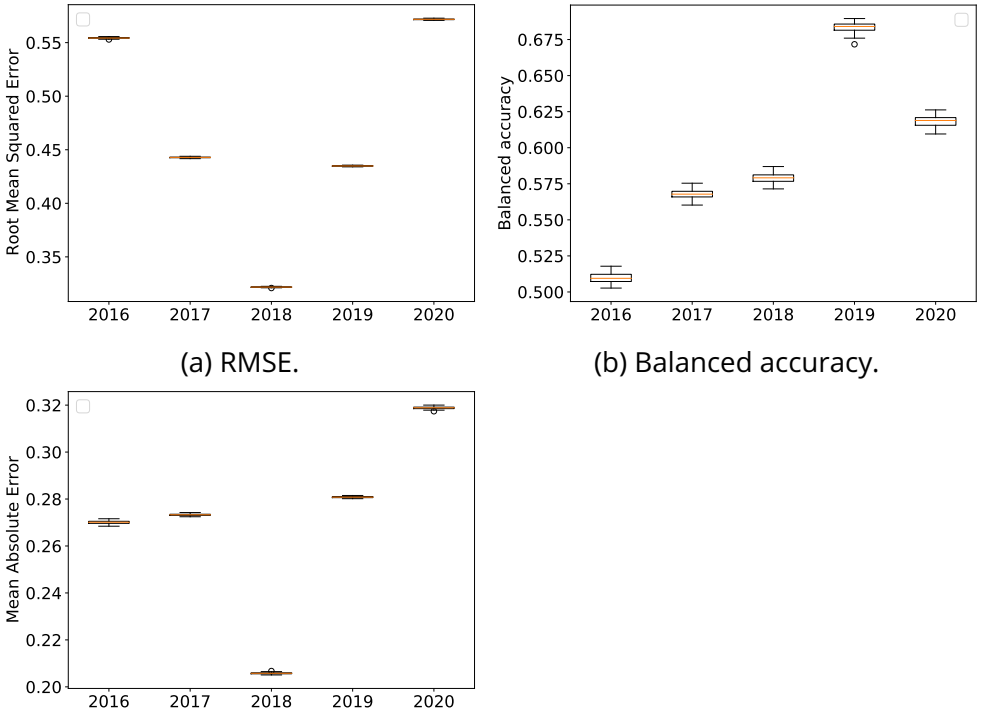
(e) 2020.

Figure 3.22: Company-wise cross-validation in a regression setting without target filtering: test MSE versus validation MSE of the 100 best models of the random hyperparameters search; Refinitiv ESG dataset.

5-times repeated random sub-sampling company-wise cross-validation

Year	Pearson correlation	R^2	Kendall tau	p-value of Kendall tau
2016	-0.48	0.23	-0.28	$4.5e^{-5}$
2017	0.10	0.010	0.048	$4.8e^{-1}$
2018	-0.38	0.15	-0.25	$1.8e^{-4}$
2019	0.43	0.18	0.29	$1.4e^{-5}$
2020	0.44	0.20	0.30	$9.9e^{-6}$

Table 3.7: Dependence measures between the MSE in the validation and test sets, for the 100 best models of the random hyperparameters search, in a regression setting without target filtering; Refinitiv ESG dataset.



(a) RMSE.

(b) Balanced accuracy.

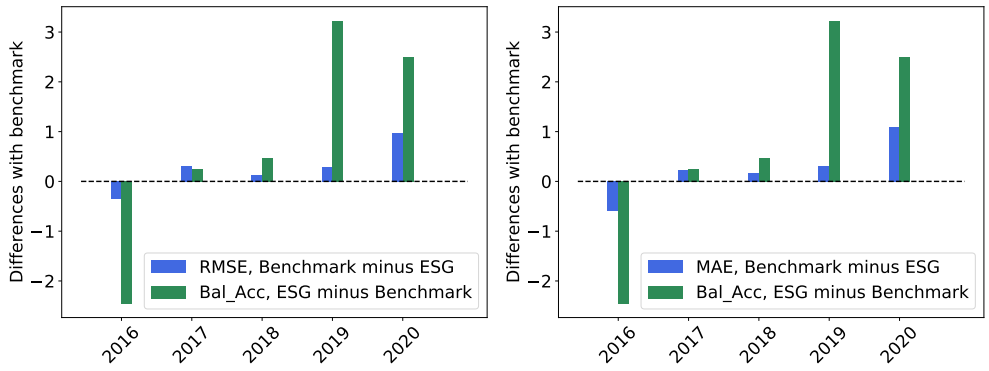
(c) MAE.

Figure 3.23: Performance measures on the test sets of the 5-times repeated random sub-sampling company-wise cross-validation, in a regression setting without target filtering; Refinitiv ESG dataset. The boxplots show the performance of 100 random samplings of 5 models among 25 random company-wise validation splits.

5-times repeated random sub-sampling company-wise cross-validation

	Only Benchmark features			Benchmark and ESG features		
Year	Balanced Accuracy	RMSE	MAE	Balanced Accuracy	RMSE	MAE
2016	53.4	55.1	26.4	51.0	55.4	27.0
2017	56.5	44.6	27.6	56.8	44.3	27.3
2018	57.4	32.3	20.8	57.9	32.2	20.6
2019	65.2	43.8	28.4	68.4	43.5	28.1
2020	59.4	58.1	33.0	61.9	57.2	31.9

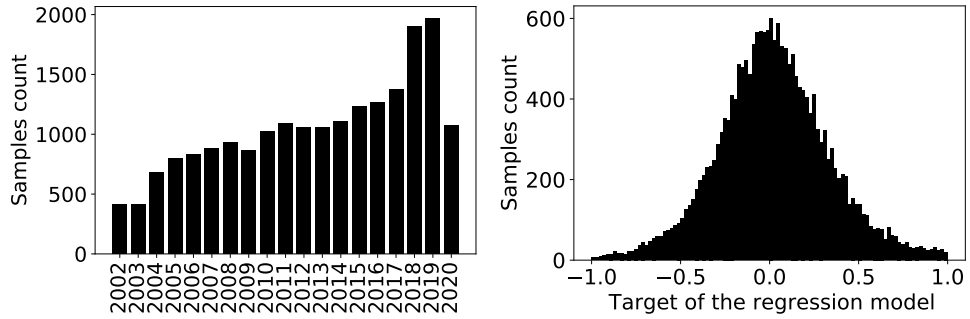
Table 3.8: Performance measures in percent on the test set, in a regression setting without target filtering; Refinitiv ESG dataset. The numbers for the company-wise splits are the median values of the performance of 100 random samplings of 5 models among 25 random company-wise validation splits.



(a) RMSE and balanced accuracy.

(b) MAE and balanced accuracy.

Figure 3.24: Performance measures with respect to benchmark, for the 5-times repeated random sub-sampling company-wise cross-validation, in a regression setting without target filtering; Refinitiv ESG dataset.



(a) Time evolution of the number of samples in the Refinitiv ESG dataset used for explanation in a regression setting.

(b) Histogram of the targets associated with the Refinitiv ESG dataset used for explanation in a regression setting.

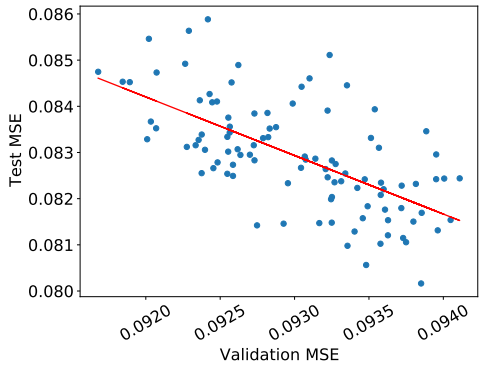
Figure 3.25: Description of the Refinitiv ESG dataset used in regression, with target filtering.

Training of the full dataset with a target filter

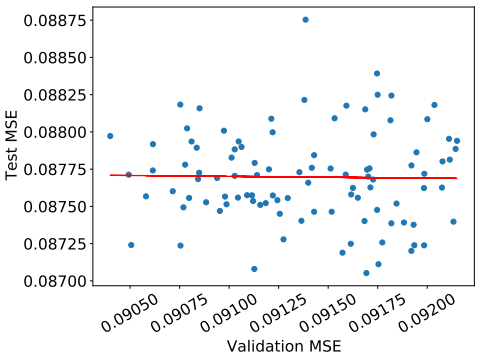
In this experiment, we apply the same methodology and train the models using a 5-times repeated random sub-sampling company-wise cross-validation on a filtered version of the dataset, removing the 560 samples with an outlier target below -1 or above 1. Figure 3.25a shows the evolution of the number of available samples with time for the processed Refinitiv dataset, composed of 19,981 samples and 2407 companies. The distribution of the target is shown in Fig. 3.25b: thanks to filtering, it is much closer to a Gaussian law, which should help the machine learning model perform better.

Figure 3.26 and Tab. 3.9 show the relationship between the validation and test MSE losses. Correlations are improving when training models on a dataset without outliers, except for the year 2016 which was already not good. Filtering the dataset to remove outliers leads to a better correlation between performance on the validation set and performance on the test set.

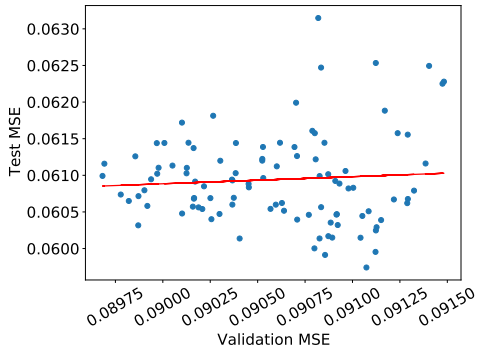
Performance is displayed in Tab. 3.10 and in Fig. 3.27 and 3.28. In practice, we train 25 models using the company-wise cross-validation method and then compute the median performance on 100 random subsets of size $K = 5$ among these 25 models. Removal of the outliers allows for much better performance in terms of RMSE and MAE, while the performance in terms of balanced accuracy on the sign of the estimated output stays similar, suggesting that filtering the outliers did not help the estimation of the sign. Figure 3.24 shows the same trend that we found in classification and in regression using the full dataset, with ESG features increasingly explaining the idiosyncratic part of price returns when considering only the market factor.



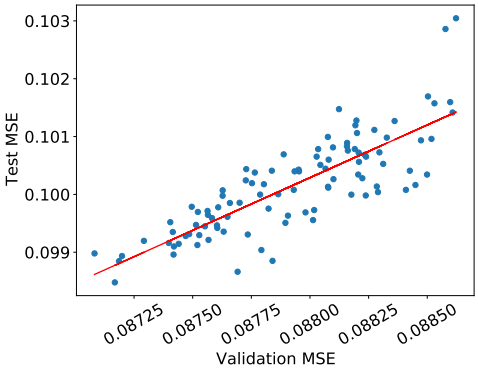
(a) 2016.



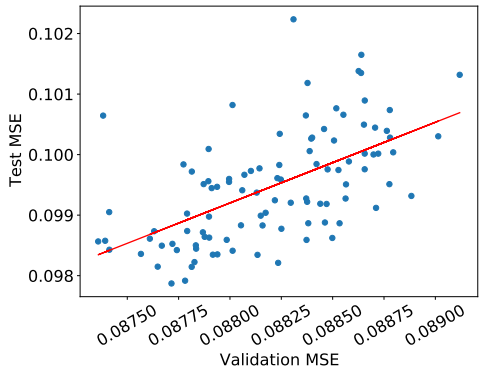
(b) 2017.



(c) 2018.



(d) 2019.



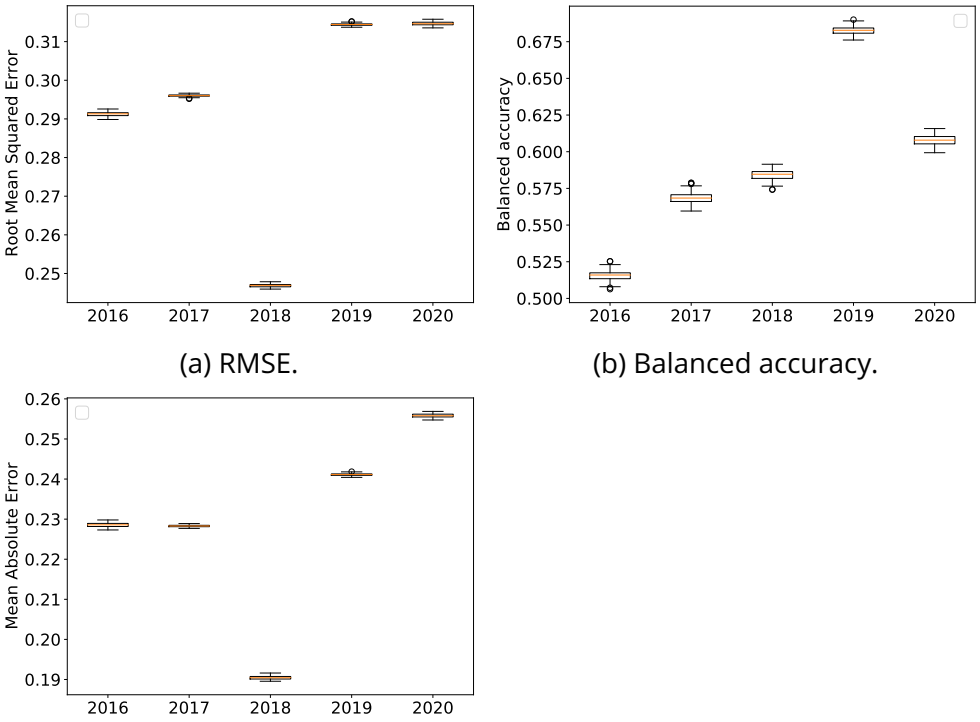
(e) 2020.

Figure 3.26: Company-wise cross-validation: test MSE versus validation MSE of the 100 best models of the random hyperparameters search, in a regression setting with target filtering; Refinitiv ESG dataset.

5-times repeated random sub-sampling company-wise cross-validation

Year	Pearson correlation	R^2	Kendall tau	p-value of Kendall tau
2016	-0.63	0.41	-0.46	7.1e^{-12}
2017	-0.018	0.00032	-0.022	7.5e^{-1}
2018	0.073	0.0053	-0.050	4.6e^{-1}
2019	0.82	0.67	0.63	1.3e^{-20}
2020	0.59	0.34	0.44	1.2e^{-10}

Table 3.9: Dependence measures between the MSE in the validation and test sets, for the 100 best models of the random hyperparameters search, in a regression setting with target filtering; Refinitiv ESG dataset.



(a) RMSE.

(b) Balanced accuracy.

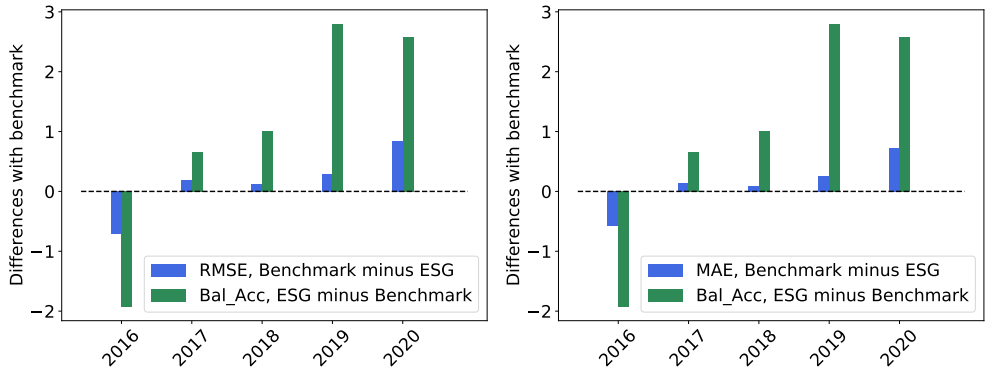
(c) MAE.

Figure 3.27: Performance measures on the test sets of the 5-times repeated random sub-sampling company-wise cross-validation, in a regression setting with target filtering; Refinitiv ESG dataset. The boxplots show the performance of 100 random samplings of 5 models among 25 random company-wise validation splits.

5-times repeated random sub-sampling company-wise cross-validation

	Only Benchmark features			Benchmark and ESG features		
Year	Balanced Accuracy	RMSE	MAE	Balanced Accuracy	RMSE	MAE
2016	53.5	28.4	22.3	51.6	29.1	22.9
2017	56.2	29.8	23.0	56.8	29.6	22.8
2018	57.5	24.8	19.1	58.5	24.7	19.0
2019	65.5	31.7	24.4	68.3	31.4	24.1
2020	58.2	32.3	26.3	60.8	31.5	25.6

Table 3.10: Performance measures in percent on the test set, in a regression setting with target filtering; Refinitiv ESG dataset. The numbers for the company-wise splits are the median values of the performance of 100 random samplings of 5 models among 25 random company-wise validation splits.



(a) RMSE and balanced accuracy. (b) MAE and balanced accuracy.

Figure 3.28: Performance measures with respect to benchmark, for the 5-times repeated random sub-sampling company-wise cross-validation in a regression setting with target filtering; Refinitiv ESG dataset.

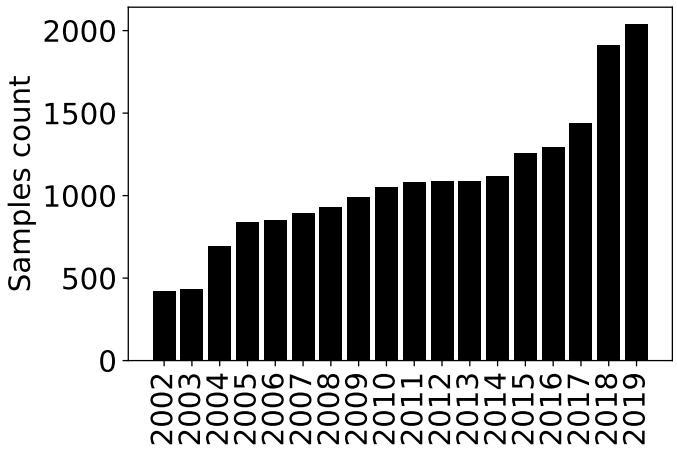


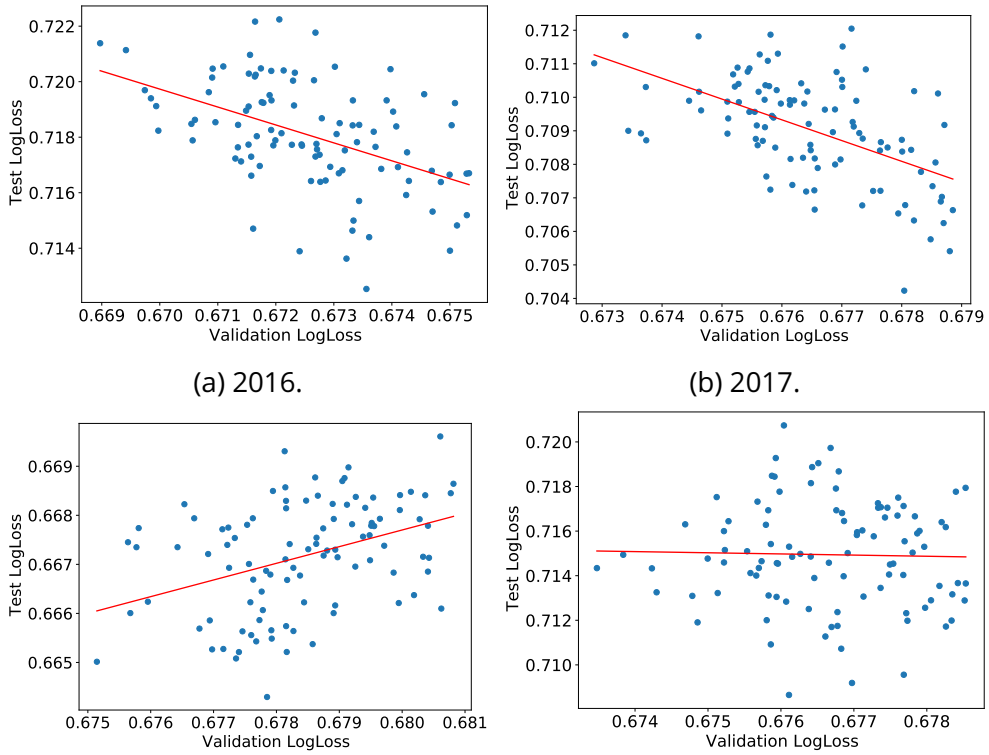
Figure 3.29: Time evolution of the number of samples in the Refinitiv ESG dataset used for prediction.

3.7.4 . From explanation to prediction using Refinitiv data

In section 3.4.1, the model focuses on explaining the idiosyncratic part of price returns using machine learning. In this experiment, the goal is to go from explanation to prediction, predicting the future idiosyncratic part of price returns one year ahead of the publication of the ESG scores. The same methodology as the one previously defined is applied, the only change is the target which is from now on the sign of the one-year in the future idiosyncratic part of price returns when only considering the market factor. We are then coming back to a classification setting. The used dataset now stops at 2019 to account for the one-year lag between the used ESG features and the chosen target. The target is composed of 49.36% of class 0 and 50.64% of class 1. Figure 3.29 displays the evolution of the number of available samples with time. The models are trained on a dataset of 19,434 samples and 2418 companies.

Figure 3.30 and Tab. 3.11 show the relationship between the validation and test cross-entropy losses. Correlations are not very good, with a slightly better correlation for the year 2018. Performance on the validation set does not bear a strong relationship with performance on the test set in a prediction setting using the Refinitiv ESG data.

Table 3.12 and Fig. 3.31 and 3.32 show the performance of the models. In practice, we train 25 models using the company-wise cross-validation method and then compute the median performance on 100 random subsets of size $K = 5$ among these 25 models. The model struggles to beat the benchmark: performance of a model trained on benchmark features is better than the one of a model trained on both ESG and benchmark features. It is only for the year 2019 that the model trained on both ESG and benchmark features manage to slightly beat the one trained on the benchmark only, and only for the balanced accuracy metric.



(a) 2016. (b) 2017.

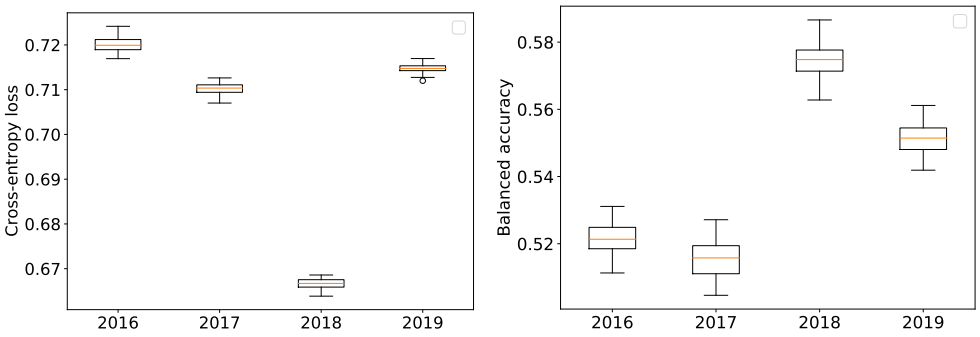
(c) 2018. (d) 2019.

Figure 3.30: Company-wise cross-validation: test cross-entropy versus validation cross-entropy of the 100 best models of the random hyper-parameters search, in a prediction setting; Refinitiv ESG dataset.

5-times repeated random sub-sampling company-wise cross-validation

Year	Pearson correlation	R^2	Kendall tau	p-value of Kendall tau
2016	-0.47	0.22	-0.33	1.0e^{-6}
2017	-0.54	0.29	-0.37	4.0e^{-8}
2018	0.39	0.15	0.26	8.5e^{-5}
2019	-0.025	0.00063	-0.021	7.6e^{-1}

Table 3.11: Dependence measures between the cross-entropies in the validation and test sets, for the 100 best models of the random hyper-parameters search, in a prediction setting; Refinitiv ESG dataset.



(a) Cross-entropy. (b) Balanced accuracy.

Figure 3.31: Performance measures on the test sets of the 5-times repeated random sub-sampling company-wise cross-validation, in a prediction setting; Refinitiv ESG dataset. The boxplots show the performance of 100 random samplings of 5 models among 25 random company-wise validation splits.

5-times repeated random sub-sampling company-wise cross-validation

	Only Benchmark features		Benchmark and ESG features	
Year	Balanced Accuracy	Cross-entropy loss	Balanced Accuracy	Cross-entropy loss
2016	52.1	70.6	52.1	72.0
2017	53.1	70.3	51.6	71.0
2018	58.1	66.6	57.5	66.7
2019	55.0	71.0	55.1	71.5

Table 3.12: Performance measures in percent on the test set, in a prediction setting; Refinitiv ESG dataset. The numbers for the company-wise splits are the median values of the performance of 100 random samplings of 5 models among 25 random company-wise validation splits.

The obtained results are all the more not satisfactory as the used Refinitiv ESG dataset is not point-in-time: data are adjusted after their publication, leaking a part of the future, as explained in section 3.3.2.

3.8 . Conclusion

While ESG data are not yet fully mature and lack long enough quality records to be amenable to easy conclusions, powerful machine learning and cross-validation techniques make it already possible to show that they do influence yearly price returns, and increasingly so: ESG features successfully explain the part of annual price returns not accounted for by the market factor. By breaking down their influence sector-by-sector, subscore-wise and according to market capitalization, we have demonstrated that ESG scores are informative. Our findings indicate that

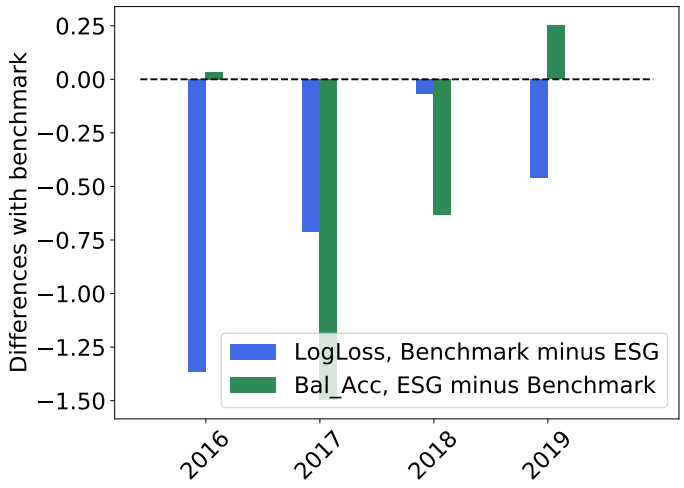


Figure 3.32: Performance measures with respect to benchmark, for the 5-times repeated random sub-sampling company-wise cross-validation, in a prediction setting; Refinitiv ESG dataset.

the relationship between controversies and price return is the most robust one. The average influence of all the other ESG scores significantly depends on the market capitalization of a company: strikingly, most of the statistically significantly influential ESG scores weigh negatively on the price returns of small or mid-size companies. Large-capitalization companies, on the other hand, have significantly advantageous ESG score types. Our findings are specific to the Refinitiv ESG dataset for the European market, and caution should be exercised in generalizing them to other ESG datasets. Indeed, as methodologies to build ESG scores vary across providers, the resulting ESG scores do not necessarily capture the same information. Outcomes using the MSCI ESG dataset, in the European market, show a less conspicuous indication of an increasing explanatory capability of ESG scores on price returns.

The research in this chapter demonstrates the capacity of certain ESG features to provide supplementary information in explaining the fraction of annual price returns not accounted for by the market factor compared to a predetermined set of benchmark features, capturing the size, activity and country of the considered companies. The benchmark features selection was tailored to the purpose of this study: alternative choices of benchmark features could have uncovered other types of additional information embodied in ESG features.

While this work focuses on explaining the sign of excess price returns derived from the CAPM model, those derived from the Fama-French 3-factor model lead to results that are less clear-cut for the time being. However, correlations between validation and test set errors increased in both 2018 and 2019, indicating the potential increasing information value of ESG data in explaining price returns dis-

counted by the market, size and value factors. Future investigations will use data for years 2020, 2021 and 2022 to verify these initial findings. Moreover, extending this research to the study of the explanatory power of ESG data with respect to more equity factors, such as quality, would enhance its comprehensiveness.

We also give a first result regarding the application of the methodology in a regression setting: we still find using the same Refinitiv dataset that ESG scores increasingly participate in the formation of price returns. However, a regression setting necessitates more preprocessing steps and results should be handled with care as the RMSE and MAE remain high. We also propose an extension of the framework on the same Refinitiv ESG dataset to evaluate the predictive power of these ESG features by estimating the sign of the future idiosyncratic part of price returns when considering only the market factor. The results were not conclusive. Future research could focus on deepening these additional experiments. Additional ESG datasets could be tried, training a model for each of the providers or mixing providers with additional feature selection steps.

Future work will also include studying outliers of the SHAP values distribution and testing the hypothesis that extreme scores in the ESG field are more informative. Moreover, a study of the links between ESG and equity returns is comprehensive only if the systematic and idiosyncratic aspects of risks and returns are studied together (Giese and Lee, 2019): indeed, it may be that having better ESG scores not only decreases price returns but also reduces risk. Future research will investigate the information content of ESG datasets to evaluate risk measures concerning a company's stock, such as volatility or drawdown. This would provide a broader understanding of the interplay between ESG factors, risk, and equity returns.

We propose in Tab. A.1 and A.2 in the Appendices two summary tables exhibiting the main results obtained in this chapter for the different discussed settings: Refinitiv or MSCI data, target derived from the CAPM model or the Fama-French 3-factor model, classification or regression, explanation or prediction.

4 - Greenhouse gas emissions: estimating corporate non-reported emissions using interpretable machine learning

4.1 . Context

Scope 1 and scope 2 GHG emissions reporting from large firms in developed countries generally follow a common methodology and results are either published and/or validated by independent bodies, such as external auditors and the CDP. In 2021, this was the case for more than 4000 companies worldwide. For a typical investment universe of 15,000 companies, this means that 11,000 companies (73%) did not report their scope 1 and 2 GHG emissions in 2021. This lack of reporting is not sustainable even in the short term, knowing the increasing number of regulatory bodies and investors who either want or are required to take into account the GHG emissions of companies. This begs for models that estimate these GHG emissions. These models are particularly useful to fill the gaps but the chosen methodologies are often undisclosed and can largely vary from one model to another, from simple derivation from previous year data to more complex non-linear methods.

When it comes to comparing corporations across geographies and sectors or drawing conclusions at the global level for anthropogenic GHG emissions, precise assessments of GHG emissions at country, corporation, factory and personal levels are needed, whether these emissions are modeled on not. Operational scopes (accounting consolidation scopes), standards of calculations (GHG protocol or others) and calculation basis must be analyzed in details. Omitting this assessment can lead to biased results and a lack of transparency. For instance, Bolton and Kacperczyk (2021) analyze GHG emissions of 14,468 companies, including 98% of publicly listed companies, without mentioning that 80% of the data used is coming from GHG estimates from the data provider Trucost. They construct a regression model to fit the scopes 1, 2, and 3 data and draw conclusions on global carbon premiums in the market. On the other hand, Aswani et al. (2022) use the same Trucost dataset and analyze more deeply the underlying quality of the GHG emissions data used.

The study in this chapter focuses on the unreported scope 1 and scope 2 emissions of companies. We propose a machine learning model to estimate unreported scope 1 and scope 2 company emissions in an investment universe of about 50,000 companies, out of which only around 4000 entities have reported their emissions data. This model is built to be used in financial applications to estimate GHG emissions at portfolio levels. To this end, the model needs to produce so-called "point-in-time" estimates using only information available at the date of the estimated emission. This model has the following aims:

- accuracy, globally and by granular sub-sectors, with good and balanced performance on each sub-sector.
- transparency of the methodology and reproducibility of results, keeping the complexity of the model to a minimum while achieving good performance. For example, all data preprocessing steps must be fully automated with no manual corrections. The proposed model should be flexible and easily allow the inclusion of new input data with the evolution of regulations, especially on GHG disclosure.
- large final coverage, aiming at using the model for a scope of 50,000 companies, both public and private and including small ones.
- interpretability, a regulatory requirement as highlighted by Heurtebize et al. (2022), with clear and exhaustive statistical explanations of the outputs.

We make crucial decisions that deviate from existing approaches to achieve the desired attributes.

1. Models are always tested on data samples never seen during calibration so that their generalization abilities can truly be measured, as required in any machine learning setting. As the availability of labeled data is limited in the context of GHG emissions, we propose a methodology to keep enough data in the training set while having test sets allowing a fair evaluation of performance.
2. GHG reported emissions data are quite recent, their number and quality improving with time. We rely on the company-wise cross-validation scheme introduced in chapter 3, that makes it possible to train and validate models with the most recent (and thus most reliable) data.
3. Models are always evaluated globally and by sub-sectors. Estimates are compared to the ones from other providers through a proposed methodology.
4. Models are reproducible. Having fully automated data preprocessing steps and no manual correction is a requirement.
5. Models are fully interpretable, using model-agnostic methods so that interpretability does not come at the expense of performance.

In the remainder of this chapter, we present three iterations of the model implemented during this thesis. We start by describing the data retained to calibrate and evaluate our model and present the proposed methodology in-depth, mostly common to the three model iterations. We then discuss the results associated with this methodology for the first model iteration both by comparing our estimates to the actual reported GHG emissions and by comparing our estimates to

the ones from other providers. We provide interpretability elements to understand how the constructed model works and to what extent each feature participates in the formation of the GHG emissions estimations. The second and third iterations of the model are then discussed, exposing the changes made and the associated results. This chapter ends with a review of additional experiments realized along the construction of the different iterations of the model.

4.2 . Literature review

Corporate GHG emissions models link the industrial processes of each business model and the emissions associated with each stage of those processes. The Environmental Input-Output Analysis (EIO) and the Process Analysis (PA) models give precise results for a given industrial process (Wiedmann, 2009). However, the information required to quantify these processes and their intensity in the overall annual production chain is not publicly available. Linking detailed industrial processes and technologies with an accounting of GHG emissions is a perilous task, even when it is handled by large corporate sustainability expert teams or by CDP experts.

To mitigate this lack of data, financial data vendors, such as Bloomberg (Bloomberg Enterprise Quants, 2022), MSCI (Shakdwipee and Lee, 2016; Andersson et al., 2016; De Jong and Nguyen, 2016), Refinitiv - previously known as Thomson Reuters - (Refinitiv, 2023; BNP Paribas, 2016; Boermans et al., 2017), S&P Global Trucost, and CDP use models to estimate the GHG emissions of companies that do not disclose them. Such models rely mainly on rules of proportionality between emissions and the size of the company operations with sectorial adjustments or, recently, on more complex approaches using non-linear models. Sector averages and other regression models constructed using the existing reported GHG emissions data from peer companies have the advantage of simplicity for explainability but the number of regressors and samples is usually limited. The simple models tend to use historical data available for the industry as a basis for the calculation and focus on estimating the logarithm of GHG emissions. Occasionally, they also use energy-specific metrics such as the GHG intensity per the considered company's energy consumption or energy production or even per ton of produced cement. However, these metrics are only available for the limited number of companies reporting them without reporting their GHG emissions. These models are calibrated on samples of reported data. Performance is around 60% in terms of R^2 when evaluating estimations of the logarithm of the GHG emissions. To be noted, these performance levels are obtained in-sample, meaning the R^2 is computed with the data used to calibrate the models.

Some more advanced models described in Goldhammer et al. (2017), Griffin et al. (2017) and CDP (2020) propose the use of Ordinary Least Squares (OLS) regression and Gamma Generalized Linear Regression (GGLR) with a broader dataset

of publicly available company data to calibrate models. Such models go beyond using just simple factors and rely on data correction processes or smaller sub-samples of industries where the models work correctly. These models are more effective than the previous ones, with in-sample R^2 computed with the logarithm of the GHG emissions around 80%.

More recently, two studies proposed the use of more complex statistical learning techniques to develop models for estimating corporate GHG emissions from publicly available data.

In Nguyen et al. (2021), a meta-learner relies on an optimal set of predictors and combines OLS regression, ElasticNet, multilayer perceptron, K-nearest neighbors, random forest and extreme gradient boosting as base learners. Their approach generates more accurate predictions than previous models even in out-of-sample situations, i.e. when used to estimate reported emissions that were not used to calibrate the model. Nevertheless, the highest predictive accuracy of the model was found for estimating aggregated scope 1 and 2 emissions as opposed to predicting each of the scopes separately. Furthermore, despite the improvement over existing approaches, the authors also noted that relatively high prediction errors were still found, even in their best model. Indeed, the five dirtiest industries representing about 90% of total scope 1 emissions (Utilities, Materials, Energy, Transportation, Capital Goods) have an average in-the-sample R^2 computed with the logarithm of the GHG emissions of only 51%. The five dirtiest industries accounting for about 70% of the total emissions in terms of scope 2 (Materials, Energy, Utilities, Capital Goods, Automobiles & Components) have an average in-the-sample R^2 computed with the logarithm of the GHG emissions of only 52%. In addition, their model fails for Insurance, both for scope 1 and scope 2, with R^2 of -378% and -151%, respectively. The paper also lacks discussions on the achievable coverage of GHG emissions estimates and the interpretability of the model.

In Bloomberg Enterprise Quants (2022), amortized inference with GBDT models calibrated using a conditional mixture of Gammas and Maximum Mean Discrepancy (MMD) regularization is used. The model is trained on hundreds of features, including ESG data, fundamental data and industry segmentation data. The GBDT allows for non-linear patterns to be found even if not all features are available. Moreover, an important debiasing approach compares the feature distributions for the reporting companies and non-reporting companies by trying to match missing features between labeled data and unlabeled data using MMD. In this model, the R^2 computed directly with the GHG emissions goes from 84% for firms with good disclosures (lots of features available) to 41% for companies with average or poor features disclosures. However, this paper lacks transparency with several implementation elements, including details on the selected set of features, making it not reproducible. It also lacks a discussion on the interpretability of the designed model.

The current state-of-the-art does not yet seem to provide good enough and

transparent models to estimate scope 1 and scope 2 GHG emissions, encompassing all the desired qualities. The approaches recently proposed based on statistical learning are promising. The central challenge is to strike the right balance between increasing both the model complexity and accuracy while limiting the risk of overfitting, especially when used training data is non-stationary and of variable quality.

4.3 . Datasets

An important variety of data sources are available. Following Heurtebize et al. (2022), we rely on two sets of indicators. The first set refers to data retrieved at the company level. For a given company, we gather all indicators exhibited in Tab. 4.1a, selecting yearly data. Such indicators give indications on the company profitability, asset size, asset location, and how they are used.

The second set of indicators is the regional ones, also selected each year, and presented in Tab. 4.1b. They provide information on the environment the company is incorporated.

Company data are extracted between 2010 and 2020 from the Refinitiv Worldscope database, for a total of 531,408 samples. It represents 65,673 companies between 2010 and 2020 incorporated in 115 countries, with 48,429 companies incorporated in 112 countries in 2020 alone.

4.4 . Methods

4.4.1 . Problem settings

Using the vast amount of available indicators, whose selected ones have been exhibited in section 4.3, we build a high-quality dataset and calibrate a machine learning model on the reported emissions of companies, for the subset of companies disclosing them. Scope 1 and scope 2 emissions are estimated through two separate models.

Following section 2.1, this is a regression setting applied to tabular data. The goal is to estimate the reported emissions using a vector of features consisting in financial and extra-financial data on sample companies for different years. The state-of-the-art for machine learning with tabular data is gradient boosting models. Gradient boosting models and specifically GBDT are developed in section 2.2. We use here the GBDT algorithm and its LightGBM implementation.

The model is trained to minimize the MSE between the predicted output from the model and the ground truth. This loss is detailed in section 2.1.4.

Type of indicator	Data Provider	Name of indicator
General	Refinitiv	Country of Incorporation
		Employees
Industry Classification	Bloomberg	BICS Classification Levels 1 to 7
		New Energy Exposure Rating
Financial	Refinitiv	Accumulated Depreciation
		Capital Expenditure
		Depreciation, Depletion & Amortization
		Enterprise Value
		Revenues
		Property, Plant & Equipment - Gross
		Property, Plant & Equipment - Net
Energy	Bloomberg	Corporate Actions
		Energy Consumption
Greenhouse Gas Emissions		Total Power Generated
		Reported GHG Emission - Scope 1
		Reported GHG Emission - Scope 2
	Carbon Disclosure Project	Reported GHG Emission - Level 7 quality - Scope 1
		Reported GHG Emission - Level 7 quality - Scope 2

(a) Indicators retrieved at the company level. BICS refers to the Bloomberg Industry Classification Standard and is a business classification.

Type of indicator	Data Provider	Name of indicator
Regional	International Energy Agency	Country Energy Mix Carbon Intensity
	WorldBank	Existence of an Emission Trading System
		Existence of carbon taxes

(b) Indicators retrieved at the regional level for each country or sub-region in which a company is incorporated.

Table 4.1: Data sources and indicators used in the GHG estimation model.

4.4.2 . Target computation

Raw target obtention

The reported GHG emissions for scopes 1 and 2 are sourced using two databases:

- CDP data, using the non-modeled and audited emissions from CDP which are at level 7, the highest level of quality. Details on CDP methodology and quality review are available in their documentation (CDP, 2020).

- Bloomberg data, using the reported GHG emissions gathered by Bloomberg, sourced from the companies' extra-financial communications.

When both data sources are available for a company and year, CDP data is prioritized over Bloomberg. Indeed, Bloomberg GHG data is directly sourced from companies' extra-financial communications. Norms and audit processes for these data may differ per country, whereas CDP used a uniform and audited process, based on the GHG Protocol (Ranganathan et al., 2015) for all companies in the world.

Reported emissions are expressed in tCO_2-eq.

Target cleaning procedure

GHG emissions are reported at different dates during the year. To unify samples, GHG emissions reported between January and June of the year y are attributed to the year $y-1$ and the GHG emissions reported between July and December of the year y are attributed to the same year y. For both scopes, only one reported GHG emission per company and year remains.

Variability is inherent to GHG emissions data, leading sometimes to inconsistencies with important changes in emissions for the same company over the years: this can either be due to changes in the reporting methodology or to a corporate action such as the acquisition of a subsidiary or mergers. Cleaning procedures mitigate these issues. We propose a fully automated jump-cleaning methodology.

We call *jump* a year-to-year variation in the GHG emission reported value of a company larger than 50%. This jump processing procedure aims at detecting jumps inside the dataset, removing all inconsistent points unless they can be explained by a significant corporate action. We make the hypothesis that the more recent data is of highest quality: if one or more unexplained jumps are detected in the time series of GHG emissions of a company, all the data points before the more recent jump and this jump are removed. In practice, a jump is said to be *explained* if, using the Bloomberg Corporate Actions dataset, there exists at least one corporate action amounting to at least 20% of the company's revenues during the year before or after the considered jump. A jump is *unexplained* if a concomitant and large enough corporate action to justify it cannot be found. The different thresholds were determined by trial-and-error.

Type of feature	Name	Values	Coverage
General	Year	2010 to 2020	100%
	Country of Incorporation	Country code (ISO 3166, alpha-3 code)	100%
Industry Classification	BICS Classification Levels 1 to 7	Industry Name	100%
	New Energy Exposure Rating	A1 Main driver: 50 to 100% A2 Considerable: 25 to 49% A3 Moderate: 10 to 24% A4 Minor: less than 10% NaN if missing	54.1%
Regional	CO_2 Law: Existence of an ETS or carbon taxes	National Implemented Subnational Implemented No CO_2 Law	100%

Table 4.2: Categorical features used to train the GHG emissions estimation model.

To reduce the negative impact of the skewed nature of the GHG emissions distribution, the model is trained to estimate the decimal logarithm of the GHG emissions instead of the raw values. Another advantage of using the decimal logarithm resides in the interpretation of the estimated value: an error of one unit in the decimal logarithm estimation means an error of one order of magnitude (power of 10) in the raw GHG emission. For some use cases in the financial world and depending on the practitioner, having estimated the right order of magnitude for the GHG emissions can be enough. This study seeks to go further in terms of performance but keeps this interpretability idea.

4.4.3 . Training features

For each of the obtained targets, we build a vector of features using the data sources exposed in Tab. 4.1. As a different model for each scope is trained, two feature matrices are obtained, representing the training features for each of the scopes. The scope 1 training set is composed of 16,234 samples, and the scope 2 one has 16,925 samples. In Tab. 4.2 and 4.3, we summarize the 21 features used to train the model as well as their distribution and average coverage in the two training sets. Let us note that missing values are usually left as such: in addition to the capacities of the LightGBM implementation to handle them, it is the setting for which the best performance was obtained as opposed to the data imputation methods used in Nguyen et al. (2021) and Heurtebize et al. (2022). The only exception is for the Carbon Intensity of Energy Mix feature, if for a considered country, there is a missing value only for the most recent year.

In the remainder of this section, we provide details on these different features.

Type of feature	Name	1st percentile	Median	99th percentile	Unit	Coverage
General	Employees	73	11,810	330,000	/	87.3%
Financial	Capital Expenditure	0	204	118,374	Million $	99.8%
	Enterprise Value	11.4	7,578	2,609,476	Million $	99.5%
	Revenues	56.3	4,167	1,939,292	Million $	100%
	Property, Plant & Equipment Gross	28.6	3,291	1,896,412	Million $	87.2%
	Property, Plant & Equipment Net	8.4	1,542	966,459	Million $	99.6%
	Life Expectancy of Assets	0.42	13.42	50	Year	99.2%
Energy	Energy Consumption	1.7	731	207,784	GWh	74.1%
	Total Power Generated	0.1	20,900	564,436	GWh	3.3%
Regional	Country Energy Mix Carbon Intensity	17.7	53.0	76.9	t CO_2/TJ	99.8%

Table 4.3: Numerical features used to train the GHG emissions estimation model.

Financial features

The model relies on financial features, allowing a better understanding of the size of a company and its assets. The Capital Expenditure, Enterprise Value, Gross Property Plant & Equipment (GPPE), Net Property Plant & Equipment (NPPE), and Revenues features are obtained annually for each reporting company. Both GPPE and NPPE are included as they both give elements on the tangible assets of a company that are physically responsible for its emissions: the difference between the two is accounting elements linked to the age of the assets, which provide interesting information to the model.

The feature values are converted from the reporting currency to dollars using the foreign exchange rate from the 31st December of the considered year. Apart from this conversion, financial data are used as reported from the companies' financial communications with no additional manual re-treatment.

In the first iteration of the model, the training set is filtered on the Revenues feature so that it is never missing, negative or null.

The last financial feature, the Life Expectancy of Assets, is obtained following Griffin et al. (2017) and Nguyen et al. (2021), using the following formula:

$$\text{Life Expectancy of Assets} = \frac{\text{GPPE}}{\text{Depreciation Expense}}. \tag{4.1}$$

The idea behind this proxy is to estimate the average life expectancy of the assets of a company by dividing the total amount of tangible assets of a company by the depreciation expense the company reported for the considered year. We make the hypothesis that a company associated with a higher value for the life expectancy feature has assets that are, on average, older and may emit more GHG.

As the Depreciation Expense indicator is not available and the GPPE feature has many missing values, the quasi-equivalent following formula is used:

$$\text{Life Expectancy of Assets} = \frac{\text{NPPE} - \text{Capital Expenditure} + \text{Accumulated Depreciation}}{\text{Depreciation, Depletion \& Amortization}}. \quad (4.2)$$

The numerator is modified by decomposing the GPPE term. If the Capital Expenditure or Accumulated Depreciation indicators are missing values, they are ignored and their values are set to 0. The denominator is modified by adding the depletion and amortization expenses. We did not measure any significant impact of these approximations on the final GHG emissions estimations.

Industry classification

Industry classification features allow the model to gain insights into the business model of a company. It is one of the most judgmental features used in GHG estimation models, truly distinguishing between companies by the nature of their activities according to their sectors. Indeed, the GHG emission profiles of companies operating in different sectors are not the same. For instance, sustainable energy companies are specifically tagged as such in some classifications and are not in others. There exist numerous industry classifications, grouping companies differently. This is critical for the model as it must not rely on a classification that would, for instance, never make the difference between companies operating in the Oil & Gas, Renewable Energy, or Nuclear fields. As a preliminary work, four typical business classifications were identified: The Refinitiv Business Classification (TRBC), the Standard Industrial Classification (SIC), the Global Industry Classification Standard (GICS), and the Bloomberg Industry Classification Standard (BICS). The BICS classification was selected for most of the studies done in this chapter. The industry in which a company is classified corresponds to the one in which it is making the largest fraction of its revenues. In this first iteration of the model, missing values for the first level of the BICS classification were removed.

Details and comparisons of the different business classifications are available in section 4.10.4. The results obtained from using different classification methods mostly did not yield very significant changes in model performance. The choice of the BICS classification was made for the following reasons:

- Ease of access.
- Maintenance of a database with sectorial evolutions with time. This is particularly interesting for companies that have changed activities in the past, deriving now most of their revenues from a different sector.
- Granular distinctions at deep levels, with seven hierarchical levels. For instance, it makes the distinction between companies working in the oil sector that derived most of their revenues either from oil marketing or oil extraction. This is particularly interesting for future analysis of the model, to group emissions and draw conclusions at granular levels.

- For future research, integrating features such as the revenues derived from each activity a company is involved in could improve the performance of the models. This feature is available per BICS sectors. The choice of the BICS classification would then simplify the integration of these novel features.

With the important level of details of the BICS classification, the deeper levels are not dense enough in the training dataset: not all companies have data for levels 5, 6, or 7 in the classification. As a result, having just a few instances of a particular industry at a deep level is only adding noise to the model and making it more prone to overfitting. In the preprocessing steps, all occurrences of industries that are present less than 10 times in the training set are removed. They are replaced with a NaN value, missing values being directly handled using the LightGBM model. The value of this parameter was determined by trial-and-error.

As precise as the BICS classification is, it is complemented by the New Energy Exposure Rating from Bloomberg. It is a categorical feature that estimates the percentage of an organization's value that is attributable to its activities in renewable energy, energy smart technologies, Carbon Capture and Storage (CCS), and carbon markets. This categorical data can take five values:

- A1 Main driver: 50 to 100% of the organization's value is estimated to derive from these activities.
- A2 Considerable: 25 to 49% of the organization's value is estimated to derive from these activities.
- A3 Moderate: 10 to 24% of the organization's value is estimated to derive from these activities.
- A4 Minor: less than 10% of the organization's value is estimated to derive from these activities.
- NaN if missing.

Energy data

Energy features, expressed in GWh, are often directly correlated to GHG emissions and allow the model to have a better understanding of how a company is using its assets. Energy Consumption is the amount of energy consumed by a company during a year. Total Power Generated is the energy produced in a year by a company, and therefore, it is only relevant for companies in some specific industries, explaining the low coverage displayed in Tab. 4.3.

The reporting period may differ between companies: similarly to the GHG emission targets, values reported between January and June of the year y are attributed to year $y-1$, and those reported between July and December of year y are attributed to the same year y.

Regional data

Regional data allows the model to get a sense of the environment the company is operating in, for the country in which it is incorporated. This country of incorporation is the first regional categorical feature used. Similarly to what is done for the BICS classification features, we choose in the first model iteration not to allow any missing data for this feature and so we filter the training set to remove them.

The Carbon Intensity of Energy Mix refers to the CO_2 Emissions from fuel combustion for the country in which the considered company is incorporated. Data is gathered from the International Energy Agency (IEA). Depending on when these data are obtained, there may be missing data for the most recent years. In this case, the time series for the considered country is extended using the last known value.

The model also relies on a categorical feature describing whether a system of carbon taxes or an Emission Trading System (ETS) has been put in place at a national or sub-national level. This feature, called CO_2 Law, can take three values:

- No CO_2 law: no carbon tax or ETS has been put in place for the considered country.
- National Implemented: one or both of these systems are implemented in the whole considered country.
- Sub-national Implemented: one or both of these systems are implemented in part of the considered country (a state in Canada or the USA for instance).

4.4.4 . High quality dataset

Following the described preprocessing steps, the final training datasets to estimate scope 1 and scope 2 GHG emissions are built. All preprocessing steps are fully automated with no manual retreatment for the sake of reproducibility.

Figure 4.1 shows for scope 1 and scope 2 the number of companies for which a reported GHG emission per year was obtained in the produced dataset. There has been an important increase in data quantity through the years, which illustrates the growing importance of GHG emissions reporting.

4.4.5 . Cross-validation and hyperparameter tuning - Out-of-sample performance evaluation

The usual strategy in machine learning for time series consists of a single data split into causal consecutive train, validation and test datasets, as explained in section 2.1.2. This usual strategy is not appropriate for the current problem. Indeed:

- the usual splitting scheme does not comply with the use case: the goal is not to predict future GHG emissions but to estimate unreported ones during the last available year for which samples are present in the training set.

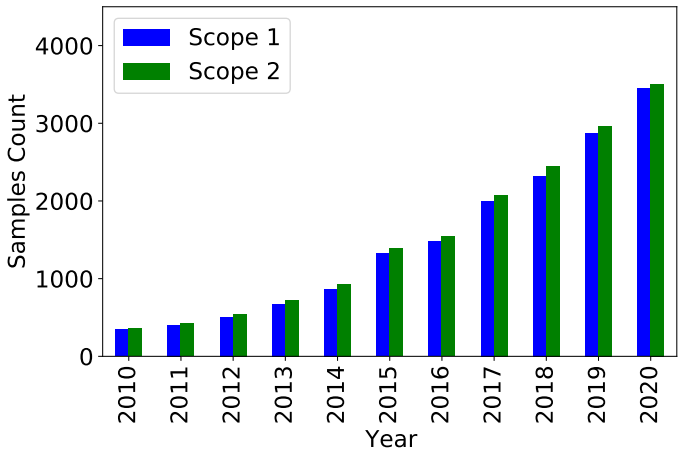


Figure 4.1: Number of companies with a reported GHG emission per year for scopes 1 and 2.

- the amount of GHG emissions data grows both in terms of quantity and quality. GHG emissions data are non-stationary. The oldest data are not exploitable alone: using this splitting scheme would lead to unreliable results as only old data would be in the training set. To get relevant results, we need to rely on the entire time span of available data.

To address these issues, a specific testing methodology and cross-validation scheme are proposed. To estimate unreported emissions of the last available year, the test set built to evaluate the models should only include companies that are never in the training or validation sets, even in a different year: the goal of the model is to estimate unreported emissions of companies which, most of the time, never reported their emissions before. Moreover, it enables avoiding a potential bias. Because of the persistent high year-over-year correlation of GHG emissions of a company, having the same company both in the training/validation set and in the test set during different years would lead to an overfitted model. In practice, the test set is built by selecting 30% of the companies for which there is a reported value during the last available year. These companies may have other reported emissions for other years: all these companies are removed from the training and validation sets. As shown in Fig. 4.1, there is not a great number of samples to train the model: this leads to small test sets with around 800 data points. As a result, the evaluation of the test set may be subject to a high variance: a few single wrongly estimated points could lead to an important deterioration of performance. We mitigate this issue by creating five different test sets and evaluating the model performance on these five test sets.

For training and validation, we use a K-times repeated random sub-sampling company-wise cross-validation, as proposed in chapter 3. Here, 80% of companies are randomly assigned to the training set and the remaining 20% to the validation

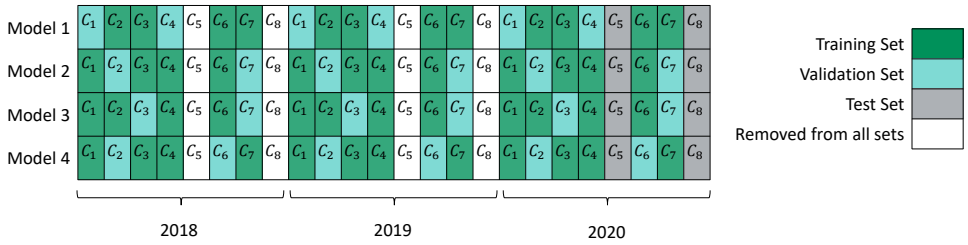


Figure 4.2: Company-wise cross-validation in the context of estimation of GHG emissions: the validation sets consists of randomly selected companies, which allows training to account for most of the most recent data.

one. We train 180 models on each of the K training sets varying the hyperparameters of the LightGBM algorithm with a random search and select the best model based on its average performance, measured using the MSE, on the respective validation sets. We use early stopping with a patience of 50 iterations. The current framework is respected, not having any company both in training and in validation, and models are trained with a large part of the most recent and more relevant data, while also being validated with the most recent and more relevant data. We take $K = 4$.

Figure 4.2 illustrates in a three-year and eight-company dataset the procedure used to build the training, validation and test sets.

4.5 . Results: evaluating the performance of the model

We first assess the quality and performance of the model on the designed testing sets, built as explained in section 4.4.5.

4.5.1 . Selected metrics

We seek to design a model with both good global performance on the test sets and good performance for each business sector at different levels of granularity, for each country and each decile of revenues.

To evaluate performance on the test sets, the selected metric is the RMSE between the GHG emission estimation (log-transformed) from the model and the ground truth (log-transformed). We also show global results using the MAE and the R^2. These metrics are defined in section 2.1.4.

RMSE and MAE are easier to interpret than R^2 in the context of GHG emissions as they are expressed in the same unit as the log-transformed GHG emission. RMSE penalizes more large errors than MAE: large errors are undesirable in the context of estimating GHG emissions, justifying the choice of the RMSE metric as the main used metric in the remainder of this chapter.

	Scope 1		Scope 2	
Metric	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
R^2	0.832	0.007	0.746	0.017
RMSE	0.578	0.007	0.522	0.031
MAE	0.401	0.006	0.341	0.010

Table 4.4: Results of the model on the five different test sets: mean and standard deviation of the R^2, RMSE and MAE metrics.

4.5.2 . Global performance

Table 4.4 displays the average global results of the scope 1 and scope 2 models for the RMSE, MAE, and R^2 metrics on each test set. These metrics are computed using the decimal logarithm of the predicted emission and the decimal logarithm of the reported emission.

4.5.3 . Breakdown of performance by sectors, countries and revenues

Besides assessing the global performance of the model, we consider a breakdown of the model performance per sector, per country and per bucket of revenues: it allows for a transparent review of the performance of the model and to better understand its strengths and weaknesses.

Figure 4.3 shows the RMSE distribution across the five test sets for BICS Sectors L1 and L2. The green boxplots correspond to the L2 sectors results across the five test sets and the pink ones in the background correspond to the associated L1 sectors results across the five test sets. Results are ordered from the highest to the lowest emissivity of the BICS Sector L2, computed on the full set of reported data. These figures highlight that the model has rather stable performance across all sectors, with good performance in the most emissive ones. These plots also highlight the importance of the chosen sectorization methodology when evaluating a GHG model: sectors should regroup similar companies in terms of emissions. Knowing that some sectors gather sub-industries with heterogeneous GHG emissions schemes could explain why the model currently has a bit more difficulty in estimating emissions for some sectors. Indeed, for instance in the mining sector, depending on the chosen technique, one ton of aluminum production can create around 10 times more emissions than one ton of steel production. We also provide in Fig. 4.4 the breakdown of the out-of-sample performance of the model across the five test sets for the different BICS Sector L3, ranked from high to low emissivity and for sectors accounting for at least 1% of the total GHG emissions of reporting companies.

Figure 4.5 takes a similar approach by proposing the RMSE distributions per country across the five test sets, for both scopes. Results are ordered by how emissive a country is in regard to the set of reported data.

We also show in Fig. 4.6 the RMSE performance across the five test sets per decile of revenues. The 9^{th} decile of revenues corresponds to the one with the highest revenues and the 0^{th} is the one with the lowest. These graphs show that, on average, it is easier for the model to estimate the GHG emissions of companies with higher revenues. This may be because the training sets have more samples coming from large companies than ones coming from SMEs, as shown in Tab. 4.3. Gathering more data from SMEs is a source of improvement for future versions of the model.

4.6 . Results: comparison of estimates with other providers

The quality of the estimates from our model called the GHGv1 model is now assessed in comparison to other data providers, comparing both coverage and accuracy. To this end, we propose a specific comparison methodology. Comparisons are made with data from providers as of August 2022.

4.6.1 . Retraining the model on the full dataset

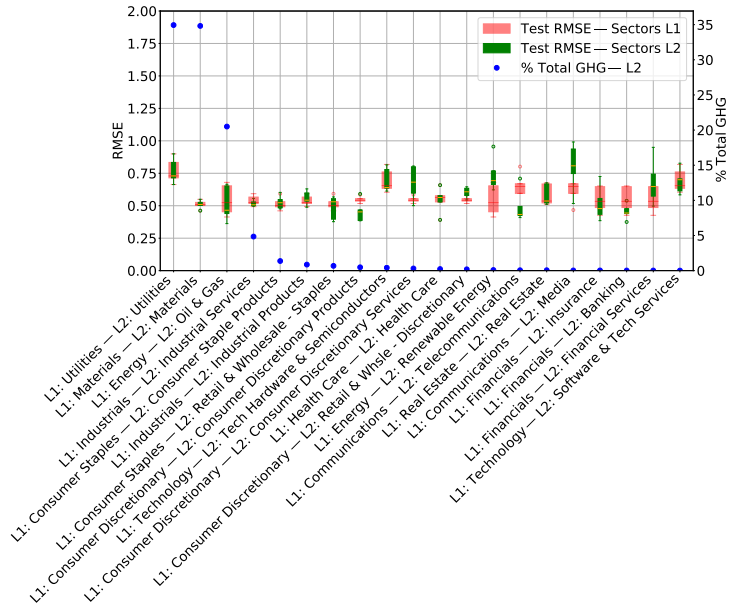
In section 4.5, the model performance is evaluated using test sets. The samples in those test sets could bring precious additional information to the model and should not be left aside in the final calibration of the model. Thus, to obtain the final models on which predictions will be made, we follow the procedure previously validated by the results in section 4.5 and train the models on the full dataset, without test sets. Validation sets are still required to find the best hyperparameters of the model. In the section 4.10.2, we show additional experiments validating this choice.

We consider the universe of 48,429 companies extracted from the Worldscope Refinitiv database for the year 2020 to evaluate the predictions of our GHG estimation model and to compare them to those of other providers.

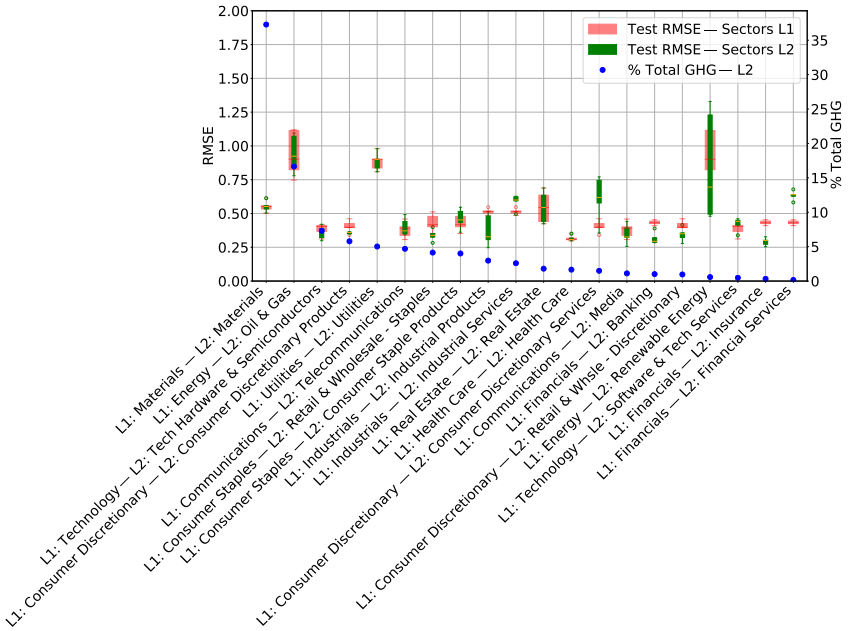
4.6.2 . Comparison of coverage

Figure 4.7 displays, for scope 1 and scope 2, the number of reported and estimated GHG emissions for each provider for the year 2020. The test was conducted on the full universe of 48,429 companies: for instance, the GHGv1 model can provide for scope 1 4360 reported data (sampled from CDP and Bloomberg and used for training as explained in section 4.4.2) and 32,261 estimates. For the remaining samples, the model was not able to provide an estimate mainly because of missing information for the company or because the considered values for categorical features were never seen during training; the model does not extrapolate on categories unseen during calibration. In future iterations of the model, we seek to enlarge this coverage by training the model on missing samples for sectors, countries or revenues (see sections 4.8 and 4.9).

Coverages for Bloomberg, Trucost, Sustainalytics, MSCI and CDP are rounded

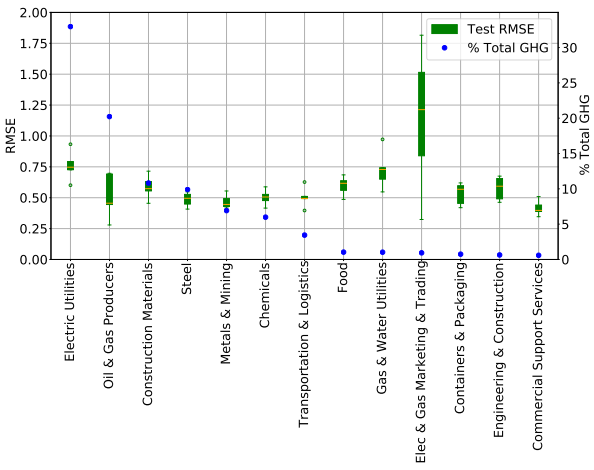


(a) Scope 1.

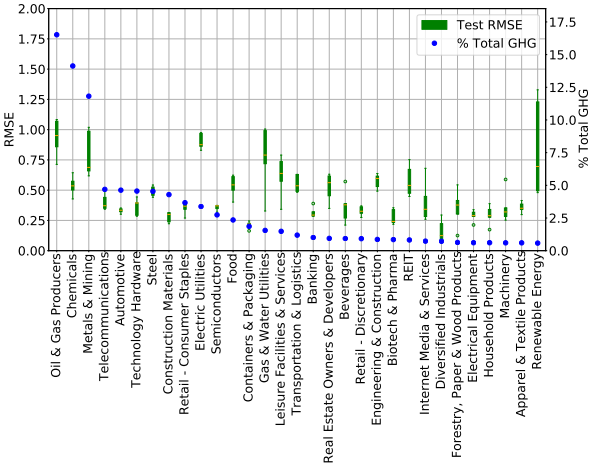


(b) Scope 2.

Figure 4.3: Distribution of performance of the model on five test sets per BICS sector levels 1 and 2 and ordered by level 2 sectors emissions. The green boxplots correspond to the L2 sectors results across the five test sets and the pink ones in the background correspond to the associated L1 sectors results across the five test sets. The percentage of total GHG represents the percentage of total GHG emissions the sector level 2 represents in the dataset of reporting companies.

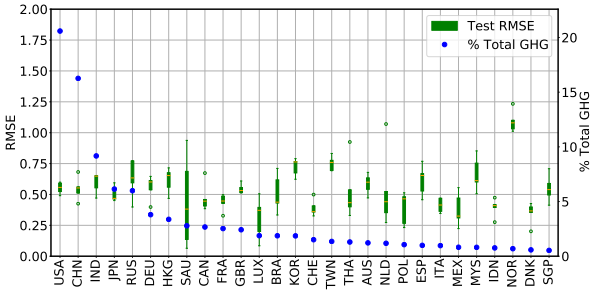


(a) Scope 1.

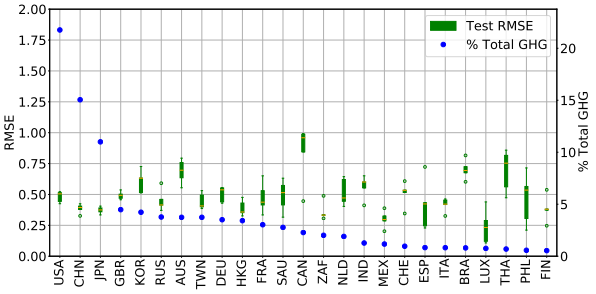


(b) Scope 2.

Figure 4.4: Distribution of performance of the model on five test sets per BICS sector level 3 and ordered by level 3 sectors emissions. The percentage of total GHG represents the percentage of total GHG emissions the sector level 3 represents in the dataset of reporting companies.

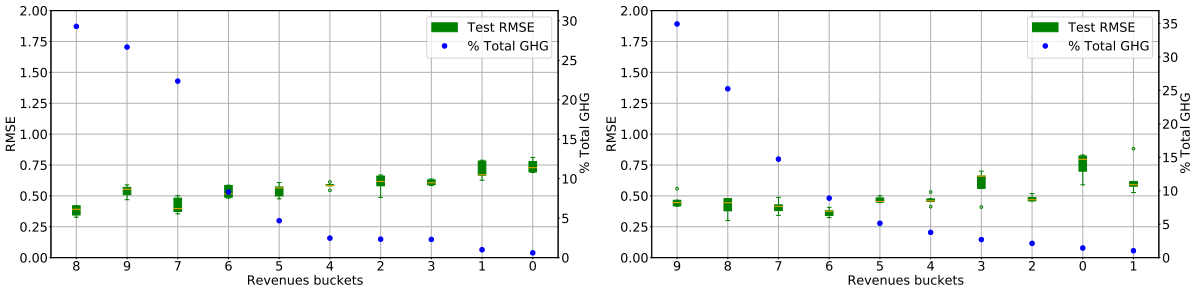


(a) Scope 1.



(b) Scope 2.

Figure 4.5: Distribution of performance of the model on five test sets per country and ordered by countries emissions. The percentage of total GHG represents the percentage of total GHG emissions the country represents in the dataset of reporting companies.



(a) Scope 1.

(b) Scope 2.

Figure 4.6: Distribution of performance of the model on five test sets per decile of revenues and ordered by deciles of revenues emissions. The percentage of total GHG represents the percentage of total GHG emissions the decile of revenues represents in the dataset of reporting companies.

as there may be slightly different results depending on the moment the datasets were obtained. Results provided in Fig. 4.7 were obtained using available elements in August 2022.

Figure 4.7 demonstrates that using a machine learning model, fully automated and with a systematic methodology, allows an important coverage, greater than any other provider, while preserving good performance.

4.6.3 . Comparison of estimates accuracy

To assess estimates quality from one provider to another, we propose a methodology relying on the high year-over-year correlation of reported GHG emissions. The methodology is as follows:

- Using the same procedure, two models are trained: one relying only on 2010 to 2018 data and a second one relying only on 2010 to 2019 data. These models, when used for predictions on 2018 and 2019 data, give respectively 2018 and 2019 point-in-time estimates.

- We consider the reported values in 2020 for companies that started reporting in 2020 and thus have never reported in 2018 or 2019. These 2020 reported values are called the ground truth.

- By comparing the 2019 estimates (or 2018 estimates if 2019 estimates are not available) from the GHGv1 model and the ones from the other providers' models to the 2020 ground truth, we determine which provider is the closest to the ground truth and thus which provider seems to have the most accurate model. Comparisons are done by computing the RMSE on the decimal logarithm of the estimations and ground truth.

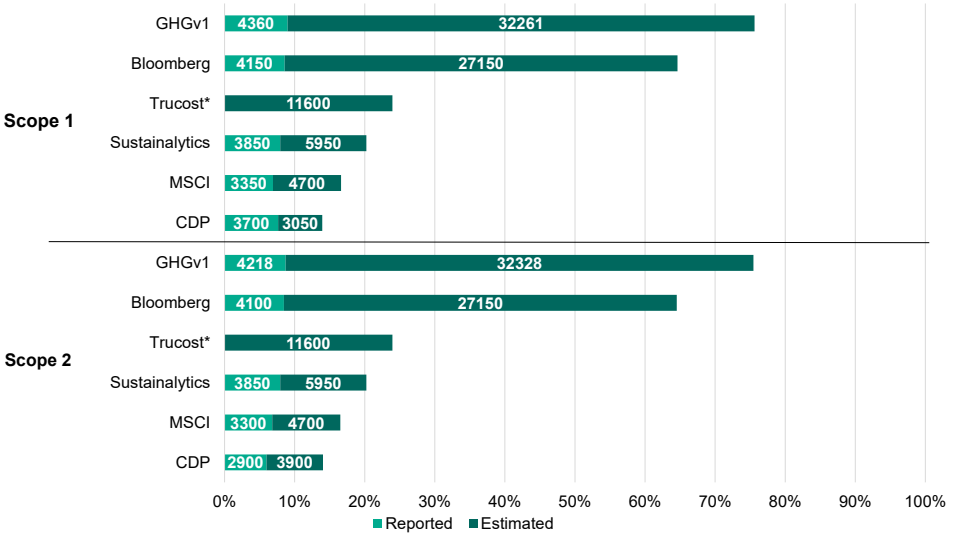


Figure 4.7: GHG emissions coverage, as of August 2022: number of reported data and estimates provided by each model. For the providers marked with an asterisk, the split between reported and estimated data was unclear, so all data points are marked as estimates.

We propose in section 4.10.3 a justification of this methodology by illustrating the high year-over-year correlation of reported GHG emissions.

Considering this methodology, we propose two solutions to evaluate the providers:

- First, we evaluate each of them separately. Tables 4.5a and 4.6a summarized these results for scope 1 and scope 2. The number of samples may greatly differ according to the coverage of the provider in estimates for companies that started reporting in 2020. The GHGv1 model has the best, i.e. lowest, RMSE in comparison to the other considered providers but comparability is not guaranteed as the evaluation sets are not the same between providers.
- Second, we consider each provider against the GHGv1 model. Results are available in Tab. 4.5b and 4.6b for scope 1 and scope 2. This time, the same samples for GHGv1 and the considered provider are used, insuring comparability. In each case, the GHGv1 is systematically more accurate than the considered provider.

Provider	RMSE	$N_{samples}$
Bloomberg	0.948	1119
CDP	1.222	546
GHGv1	0.828	1079
MSCI	0.882	509
Trucost	1.033	980

(a) All companies are considered.

Provider	RMSE: provider	RMSE: GHGv1	$N_{samples}$
MSCI	0.884	0.864	494
Bloomberg	0.956	0.828	1063
Trucost	1.039	0.812	952
CDP	1.228	0.849	530

(b) Only common companies between providers are considered.

Table 4.5: Last scope 1 GHG estimates from providers (2019 – 2018) compared to 2020 ground truth for companies that started reporting in 2020.

Provider	RMSE	$N_{samples}$
Bloomberg	0.809	1089
CDP	0.970	577
GHGv1	0.709	1042
MSCI	0.808	522
Trucost	0.822	955

(a) All companies are considered.

Provider	RMSE: provider	RMSE: GHGv1	$N_{samples}$
MSCI	0.780	0.707	502
Bloomberg	0.774	0.700	1029
Trucost	0.803	0.645	925
CDP	0.950	0.661	561

(b) Only common companies between providers are considered.

Table 4.6: Last scope 2 GHG estimates from providers (2019 – 2018) compared to 2020 ground truth for companies that started reporting in 2020.

Point-in-time data

Models are trained using only 2018 and 2019 data to avoid any leakage of the future in the 2018 and 2019 estimations respectively. It may not be the case for the estimations from the other providers, which can bias the evaluation towards better performance for the other providers. The only provider for which estimates are done point-in-time with certitude is CDP. Even considering this, the proposed model still has better performance than the considered providers.

Breakdown of performance per sector

The methodology developed to compare the GHGv1 model to providers can be extended per sector. This section only focuses on the provider CDP as it is the only one for which estimates are done point-in-time with certitude, even if the coverage of CDP is relatively small compared to any other provider.

For each sector of BICS level 1, we plot the distribution of the difference between the decimal logarithm of the ground truth and of the 2019 estimate (or 2018 if the 2019 one is not available) from the considered model. Results are displayed in Fig. 4.8: the green boxplots represent the distributions of differences between CDP estimates and the ground truth; the pink boxplots correspond to the distributions of differences between the GHGv1 estimates and the ground truth. Distributions from the GHGv1 model are more concentrated around 0, meaning better accuracy than CDP. However, CDP estimates are more conservative than the ones from the GHGv1 model: when CDP estimates are not exact, they tend to overestimate, whereas the GHGv1 model is rather balanced between overestimation and underestimation. Both behaviors and calibrations can have their strengths and weaknesses depending on the use case. Let us note however that regulatory frameworks seem to favor overestimations in comparison to underestimations.

4.7 . Interpretability

The interpretability of machine learning models producing GHG emissions is becoming a regulatory requirement. In this part, we provide tools to interpret how the model works and why it estimates such values of GHG emissions. Breakdowns of the impact of the different training features on the estimated emissions are computed.

A common criticism of GBDT is that, despite their superior performance in tabular settings, they remain difficult to interpret. A tool recently applied to the machine learning field and called Shapley values solves this issue. Shapley values, first introduced in the context of game theory (Shapley, 1953), provide a way in machine learning to characterize how each feature contributes to the formation of the final predictions. We provide in section 2.3.1 a mathematical explanation of the Shapley values and its SHAP applications in the field of machine learning, following Lundberg and Lee (2017) and Lundberg et al. (2018) research.

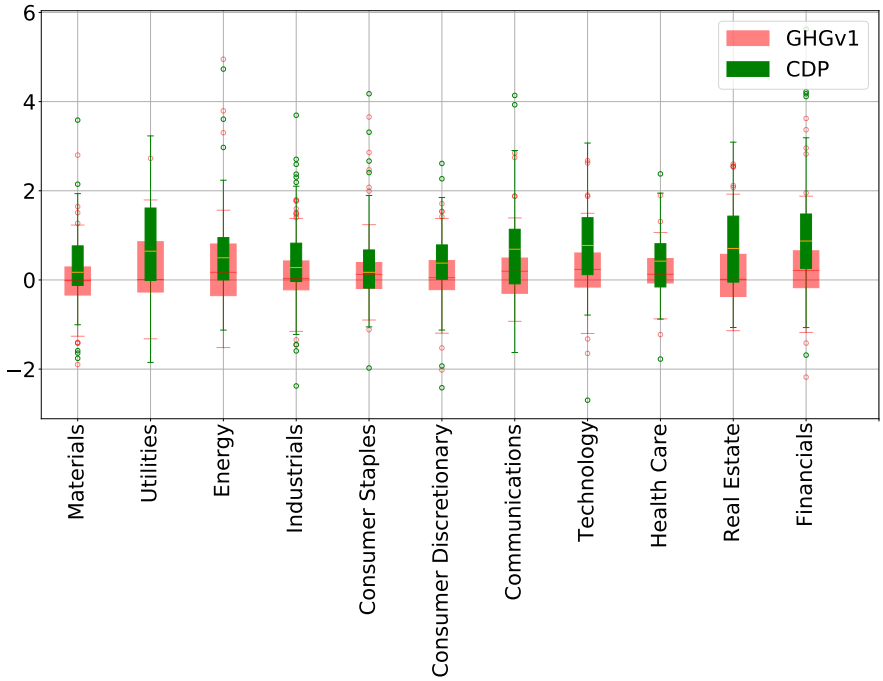


Figure 4.8: Distribution of the differences of estimated emissions from GHGv1 and from CDP (2019-2018) with 2020 ground truth for scopes 1 and 2, for companies that started reporting in 2020. The green boxplots represent the distributions of differences between CDP estimates and the ground truth; the pink boxplots correspond to the distributions of differences between the GHGv1 estimates and the ground truth.

4.7.1 . SHAP feature importance

Figure 4.9 shows the breakdown of SHAP values per feature for scope 1 and scope 2 GHG emissions, ordered by feature importance. For each feature, these graphs plot the distribution of SHAP values computed for each sample in the training set. They are key elements in the constructed model as they make it interpretable: they can be done for any set of samples, including those within a particular sector. This allows us to understand further why the model makes a specific decision and outputs these predicted estimates.

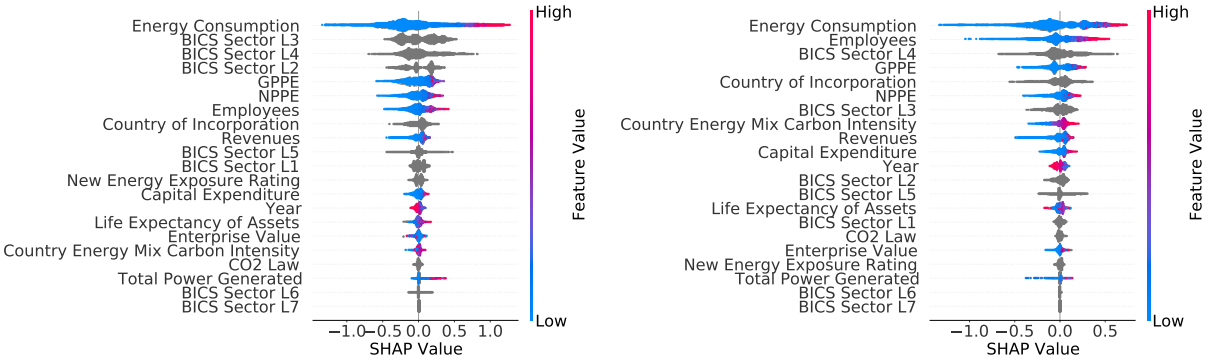
The Energy Consumption feature is the most important one used by the model for both scope 1 and scope 2. As expected from the definition of scope 2, the Employees, Country of Incorporation and Country Energy Mix Carbon Intensity features are more important for the estimation of scope 2 than the estimation of scope 1. The plots also highlight that the business classification features are paramount in GHG estimation models, with high importance for several levels of the BICS classification, both for scopes 1 and 2. It is important to choose a granular classification as features up to the level 6 of the classification are used. However, the too-deep level 7 of the BICS is not used by the model: as this level is too sparse, it does not bring additional information. The plots also show that the addition of the New Energy Exposure Rating complements well the BICS classification and contributes to the formation of the estimates.

Knowing these SHAP values not only allows us to better understand the estimates of the model but also to evaluate the reliability of the estimates based on the presence or the absence of a feature: if the Energy Consumption feature is not given for a sample, it would lead, for certain sectors, to a less reliable estimate. This can be evaluated further by comparing the distribution of SHAP values for a set of companies that reported this feature and another set of companies that did not. Such analyses are conducted on the third iteration of the model in section 4.9.4.

4.7.2 . Relationship between feature values and GHG estimates

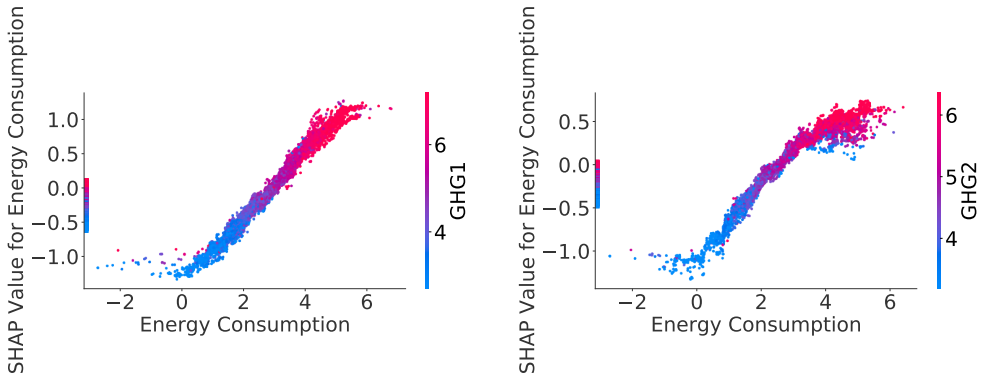
Numerical features

SHAP values can be computed for each feature on each sample, showing the relationship captured by the model between a feature and the estimated GHG emission. For numerical features, we can plot the SHAP values for a specific feature against this feature values in the dataset. For instance, Fig. 4.10 shows the relationship between SHAP values of the Energy Consumption feature and the decimal logarithm of the Energy Consumption feature values. Apart from the points which are on the Y-axis and which represent missing values for the Energy Consumption feature, there is for both scopes a near-linear increasing relationship between the SHAP values of the Energy Consumption feature and the decimal logarithm of this feature values.



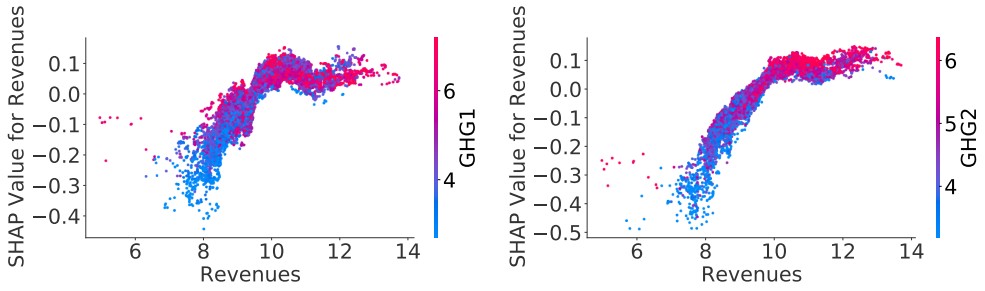
(a) Scope 1. (b) Scope 2.

Figure 4.9: SHAP values: impact of each feature on the predicted GHG emission, ordered by importance. The colorbars represent the distribution of each feature.



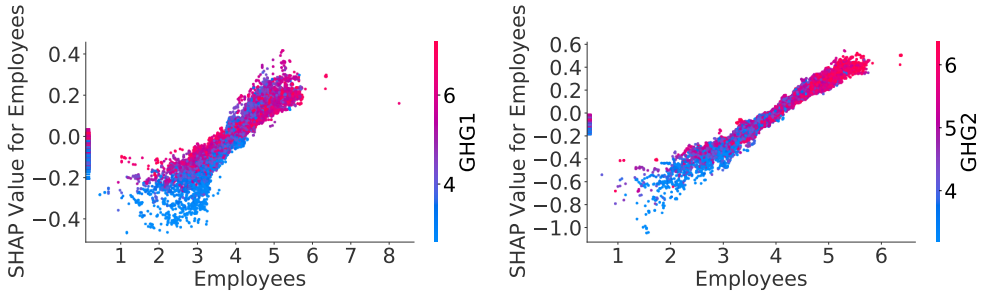
(a) Scope 1. (b) Scope 2.

Figure 4.10: Relationship between SHAP values of the Energy Consumption feature and the decimal logarithm of the Energy Consumption feature values. The colorbars represent the distribution of GHG scope 1 reported emissions (GHG1) and GHG scope 2 reported emissions (GHG2).



(a) Scope 1. (b) Scope 2.

Figure 4.11: Relationship between SHAP values of the Revenues feature and the decimal logarithm of the Revenues feature values. The colorbars represent the distribution of GHG scope 1 reported emissions (GHG1) and GHG scope 2 reported emissions (GHG2).

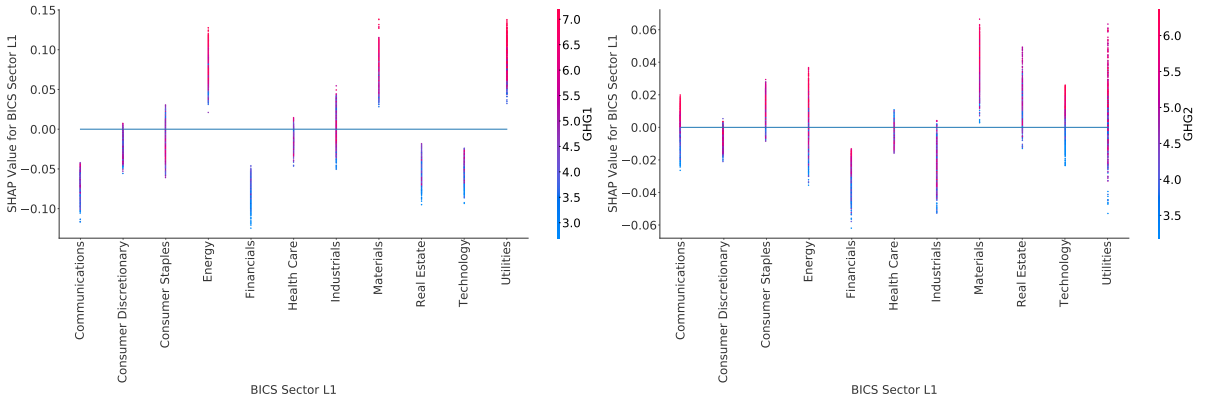


(a) Scope 1. (b) Scope 2.

Figure 4.12: Relationship between SHAP values of the Employees feature and the decimal logarithm of the Employees feature values. The colorbars represent the distribution of GHG scope 1 reported emissions (GHG1) and GHG scope 2 reported emissions (GHG2).

Figure 4.11 shows a near-linear relationship between the SHAP values of the Revenues feature and the decimal logarithm of the Revenues feature values until a sort of cap: beyond a certain level of revenues, the SHAP values are almost constant.

Figure 4.12 also exhibits a near-linear relationship between the SHAP values of the Employees feature and the decimal logarithm of the Employees feature values, apart from the few points on the Y-axis referring to missing data.



(a) Scope 1.

(b) Scope 2.

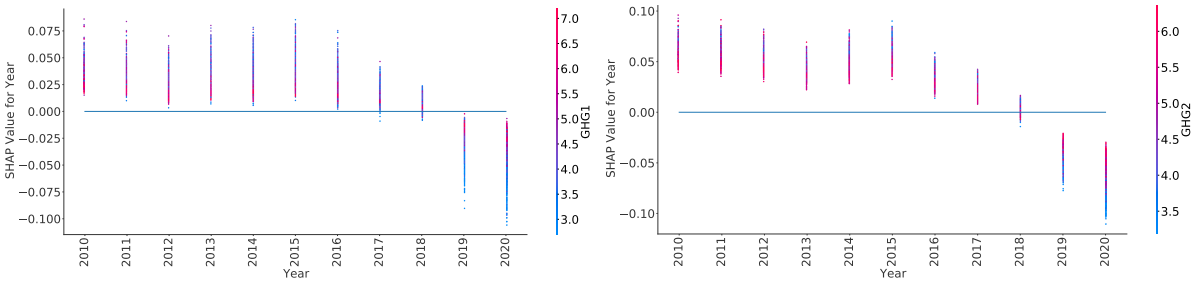
Figure 4.13: SHAP values: impact of the BICS Sector L1 feature on the predicted GHG emissions. The colorbars represent the distribution of GHG scope 1 reported emissions (GHG1) and GHG scope 2 reported emissions (GHG2).

Categorical features

SHAP values can also be used on categorical features to study their distribution for each possible value of the categorical feature. Figure 4.13 displays the distribution of SHAP values for the BICS Sector L1 feature, for each of the BICS sector of level 1. This plot highlights in what sectors companies are more likely to have higher GHG emissions. For instance, for scope 1, SHAP values for all companies in the Energy and Materials sectors show that belonging to these sectors generally leads to an increase in the estimated emission (positive SHAP values). On the contrary, samples in the Financial sector have negative SHAP values, belonging to this sector generally leads to a decrease in the estimated emission.

Figure 4.14 shows the distribution of SHAP values for the Year feature, for each year in the training set. It is interesting to see that for both scope 1 and scope 2, the model captures a tendency to have lower GHG estimates as time passes.

These plots can be done for all categorical features as they capture the distributions of SHAP values according to each category and provide elements for a better interpretation of the model.



(a) Scope 1.

(b) Scope 2.

Figure 4.14: SHAP values: impact of the Year feature on the predicted GHG emissions. The colorbars represent the distribution of GHG scope 1 reported emissions (GHG1) and GHG scope 2 reported emissions (GHG2).

4.8 . Second model iteration

We propose in this section new adjustments to the models, explain why we needed them and analyze the associated results.

4.8.1 . Changes in the second model iteration

In the second version of the model, the following changes were made:

- Preprocessing steps regarding dates are changed. Thanks to access to more documentation on Refinitiv Worldscope data, we found that Refinitiv has specific rules to attribute a year to a reported data point. For non-US companies, data for a fiscal year ending on or before 15th January is classified as the previous year's results. The cutoff date is 10th February for US companies. In our models, this same rule is now applied for all reported data including reported GHG emissions or reported Energy Consumption.

- We now include the released reported Bloomberg data for 2021, allowing running some first tests for this year. We did not have access yet to the CDP reported data for 2021. Addition of this data enables assessment of the stability of results between models trained using data until 2020 and data until 2021, knowing that 2021 data are not exhaustive as CDP labeled data samples are missing.

- The first iteration of the model highlights that the BICS Sector L7 attribute is too granular, with small coverage. It is not used much by the model and can be noisy. We remove this feature in this second version.

- In the first iteration of the model, the used labeled data is processed by removing all missing values for sectors and countries, leading to the inability

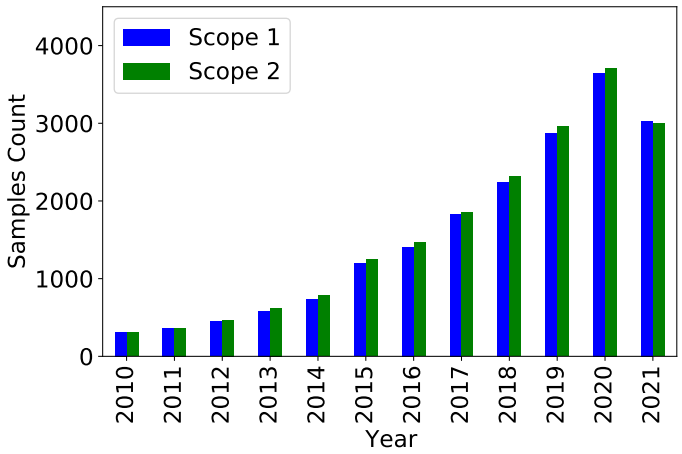


Figure 4.15: Number of companies with a reported GHG emission per year for scopes 1 and 2, for the dataset used in the second iteration of the model.

of the model to do estimations for companies for which these samples are unknown. This filtering step is removed in the second iteration to increase the coverage of the model, even if the resulting predictions can be of lower quality.

- In the first version of the model, the training data is processed by removing all sectors which do not appear in at least 10 samples. This parameter was determined by trial-and-error. In the second iteration, this number is managed as a hyperparameter of the model and is selected, using a grid search, to optimize the training loss on the validation sets. Possible values are 0, 10, 20 and 50. Each model, depending on the scope and the training years, can use a different value for this hyperparameter.

Figure 4.15 displays the evolution with time of the number of companies used in training from 2010 to 2021, with a reported GHG emission. The drop we observe in 2021 is due to the absence in our dataset of the reported CDP data for this year. We propose in Tab. 4.7 a description of the used labeled datasets, highlighting the number of added samples without sectorial or country information. Because of the kept filter on missing revenues, this iteration of the model is not trained on any sample with the country feature missing.

Considering the universe of 48,429 companies extracted from the Worldscope Refinitiv database for the year 2020, this version of the model is able to propose 35,058 estimates for scope 1 and 35,343 for scope 2.

	Scope 1		Scope 2	
Last year of training	**2020**	**2021**	**2020**	**2021**
$\mathbf{N}_{\text{samples}}$ total	15641	18668	16103	19100
$\mathbf{N}_{\text{samples}}$ with missing BICS	651	833	658	838
$\mathbf{N}_{\text{samples}}$ with missing country	0	0	0	0
$\mathbf{N}_{\text{samples}}$ with missing BICS and country	651	833	658	838

Table 4.7: Number of missing values for the BICS sector L1 and country features in the dataset used in the second iteration of the model.

4.8.2 . Results: evaluating the performance of the model

Table 4.8 displays the global results obtained for models trained between 2010-2020 and models trained between 2010-2021. Table 4.9 exhibits global results for the same trained models evaluated on test sets where samples with missing BICS features have been filtered out.

- When evaluating the models on samples with missing sectorial information, we observe a slight decrease in average performance in comparison to the evaluation on samples without missing BICS features. This decrease in performance is higher for models trained until 2021 than for models trained until 2020. The average performance of the models evaluated on samples without missing BICS features are comparable to the one obtained with the first iteration of the model (see section 4.5). It suggests that the models have more difficulties outputting good estimates for samples with missing information but that the inclusion of these samples in the training set does not impair performance on the other samples.
- Performance of the models trained until 2021 seems slightly lower than the one of the models trained until 2020. Comparing these two models is not straightforward: the test sets are not the same and CDP samples are missing in 2021.

Figures 4.16 and 4.17 shows the breakdown of performance expressed in RMSE per BICS Sector L1, for models evaluated on the full test sets and the filtered test sets. Performance remains balanced between the different sectors. Distributions of performance on samples with and without BICS missing values are very similar, reinforcing the argument that the addition of samples with missing sectorial information in training did not impair the performance of the model on complete samples.

	Scope 1		Scope 2	
Metric	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
R^2	0.827	0.007	0.750	0.027
RMSE	0.588	0.023	0.548	0.038
MAE	0.396	0.016	0.345	0.016

(a) Test 2020.

	Scope 1		Scope 2	
Metric	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
R^2	0.824	0.022	0.762	0.037
RMSE	0.608	0.034	0.543	0.034
MAE	0.412	0.023	0.343	0.025

(b) Test 2021.

Table 4.8: Results of the second iteration of the model on the five different test sets, including samples with missing sectorial information: mean and standard deviation of the R^2, RMSE and MAE metrics.

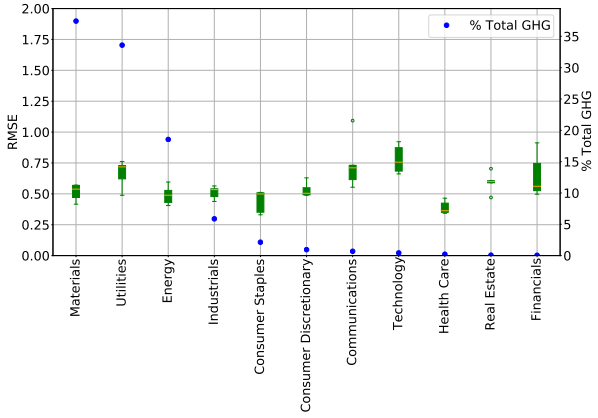
	Scope 1		Scope 2	
Metric	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
R^2	0.833	0.007	0.752	0.027
RMSE	0.579	0.022	0.539	0.035
MAE	0.389	0.015	0.338	0.015

(a) Test 2020.

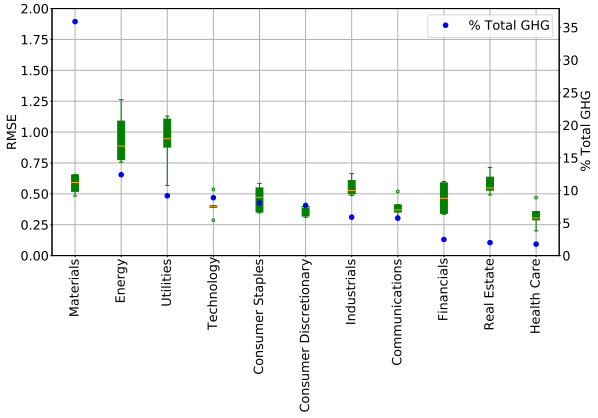
	Scope 1		Scope 2	
Metric	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
R^2	0.839	0.022	0.768	0.048
RMSE	0.576	0.034	0.525	0.047
MAE	0.392	0.021	0.333	0.025

(b) Test 2021.

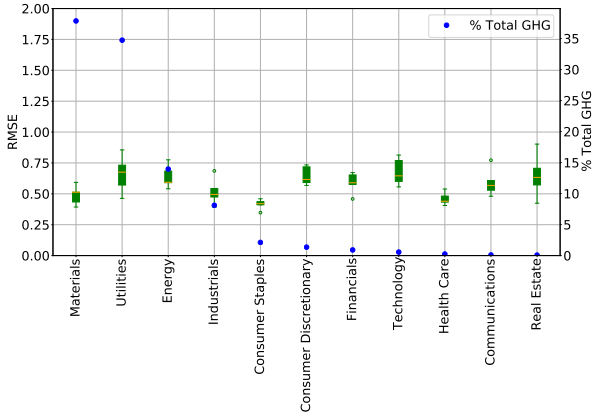
Table 4.9: Results of the second iteration of the model on the five different test sets, excluding samples with missing sectorial information: mean and standard deviation of the R^2, RMSE and MAE metrics.



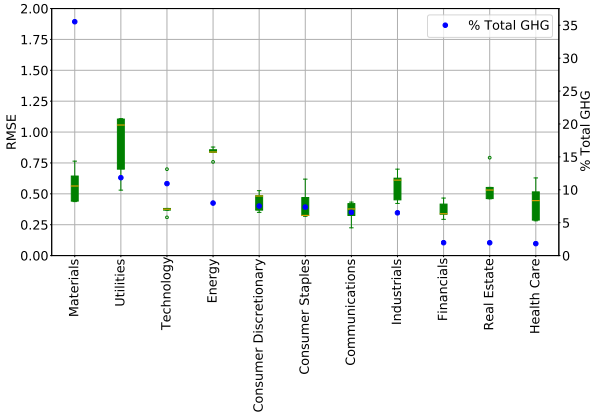
(a) Scope 1 - Test 2020.



(b) Scope 2 - Test 2020.

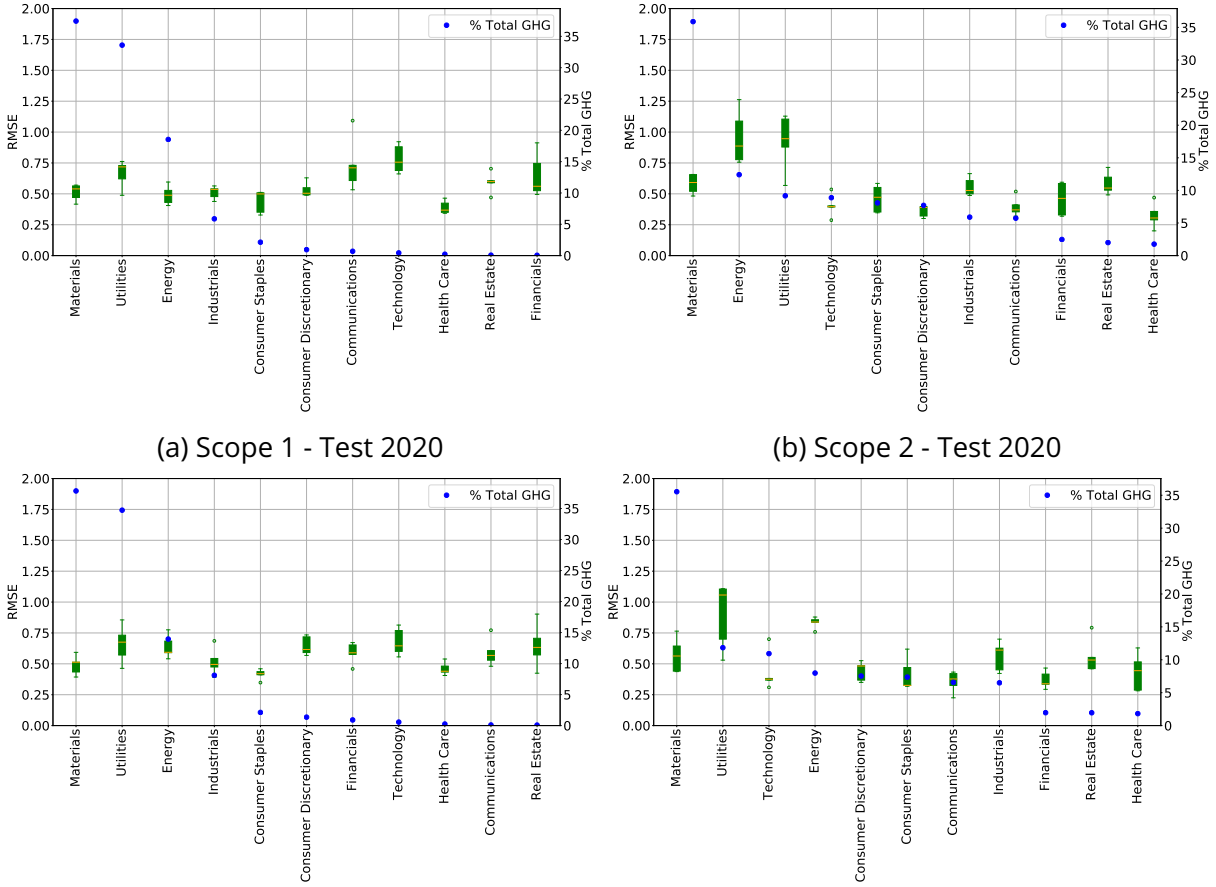


(c) Scope 1 - Test 2021.



(d) Scope 2 - Test 2021.

Figure 4.16: Distribution of performance of the second iteration of the model on five test sets including samples with missing sectorial information, per BICS sector level 1 and ordered by level 1 sectors emissions. The percentage of total GHG represents the percentage of total GHG emissions the sector level 1 represents in the dataset of reporting companies.



(a) Scope 1 - Test 2020

(b) Scope 2 - Test 2020

(c) Scope 1 - Test 2021

(d) Scope 2 - Test 2021

Figure 4.17: Distribution of performance of the second iteration of the model on five test sets excluding samples with missing sectorial information, per BICS sector level 1 and ordered by level 1 sectors emissions. The percentage of total GHG represents the percentage of total GHG emissions the sector level 1 represents in the dataset of reporting companies.

4.8.3 . Results: comparison of estimates with other providers

Following the methodology developed in section 4.6, we compare the quality of estimations from the second iteration of the model called GHGv2 to the ones from external providers. Comparisons are done with data from providers as of August 2022.

For models trained until 2019, estimates from 2019 (or 2018 if the 2019 one is missing) are compared to the 2020 ground truth, for companies that started reporting in 2020. Similarly, for models trained until 2020, estimates from 2020 (or 2019 if the 2020 one is missing) are compared to the 2021 ground truth, for companies that started reporting in 2021.

Results are presented in Tab. 4.10, 4.11, 4.12 and 4.13 for scopes 1 and 2 and training sets until 2019 and 2020. This analysis is proposed using all available samples and using only samples with sectorial information.

We observe that the second iteration of the GHG model remains better than external providers, whether it is evaluated on samples with or without sectorial information. However, by reducing the test set to samples with this sectorial information, the average performance of our models is better, suggesting the more difficult nature of estimating GHG emissions on incomplete samples.

Let us also note that the performance of models, external or not, greatly depends on the test set they are evaluated on. This is illustrated by the apparent good performance of the Bloomberg model in comparison to the GHGv2 one in Tab. 4.13a. The test set the Bloomberg model is tested on is almost four times smaller than the one the GHGv2 model is tested on. When restricting both models to the same test set in Tab. 4.13b, the superior performance of the GHGv2 model is highlighted.

4.9 . Third model iteration

Adjustments proposed in section 4.8.1 allowed to increase the coverage of the model but are not sufficient to analyze the impact of any missing values as the model remains trained on samples always including revenues information.

4.9.1 . Changes in the third model iteration

Two main objectives seek to be achieved with the third iteration of the GHG model.

- Increase the scope of the model, enabling it to make relevant estimations for a larger universe of companies.
- Build additional interpretation elements aiming at understanding the impact of any missing feature in inference.

To fulfill these objectives, the preprocessing steps consisting of filtering the dataset to remove any sample with missing revenues are abandoned.

	With missing BICS		Without missing BICS	
Provider	**RMSE**	$\mathrm{N}_{\mathrm{samples}}$	**RMSE**	$\mathrm{N}_{\mathrm{samples}}$
Bloomberg	0.906	771	0.911	738
CDP	1.145	603	1.148	592
GHGv2	0.877	1422	0.854	1309
MSCI	0.857	591	0.854	544
Trucost	1.016	1105	1.014	1071

(a) All companies are considered.

	With missing BICS			Without missing BICS		
Provider	**RMSE: provider**	**RMSE: GHGv2**	$\mathrm{N}_{\mathrm{samples}}$	**RMSE: provider**	**RMSE: GHGv2**	$\mathrm{N}_{\mathrm{samples}}$
MSCI	0.857	0.853	591	0.854	0.820	544
Bloomberg	0.906	0.830	771	0.911	0.830	738
Trucost	1.016	0.837	1105	1.014	0.823	1071
CDP	1.145	0.805	603	1.148	0.806	592

(b) Only common companies between providers are considered.

Table 4.10: Last scope 1 GHG estimates from providers (2019 – 2018) compared to 2020 ground truth for companies that started reporting in 2020, for the second iteration of the model.

	With missing BICS		Without missing BICS	
Provider	**RMSE**	$\mathrm{N}_{\mathrm{samples}}$	**RMSE**	$\mathrm{N}_{\mathrm{samples}}$
Bloomberg	0.803	777	0.783	744
CDP	1.025	635	1.011	623
GHGv2	0.734	1416	0.730	1303
MSCI	0.866	612	0.853	560
Trucost	0.832	1092	0.823	1061

(a) All companies are considered.

	With missing BICS			Without missing BICS		
Provider	**RMSE: provider**	**RMSE: GHGv2**	$\mathrm{N}_{\mathrm{samples}}$	**RMSE: provider**	**RMSE: GHGv2**	$\mathrm{N}_{\mathrm{samples}}$
Bloomberg	0.803	0.710	777	0.783	0.703	744
MSCI	0.866	0.733	612	0.853	0.719	560
Trucost	0.832	0.682	1092	0.823	0.678	1061
CDP	1.025	0.689	635	1.011	0.680	623

(b) Only common companies between providers are considered.

Table 4.11: Last scope 2 GHG estimates from providers (2019 – 2018) compared to 2020 ground truth for companies that started reporting in 2020, for the second iteration of the model.

	With missing BICS		Without missing BICS	
Provider	**RMSE**	$\mathrm{N}_{\mathrm{samples}}$	**RMSE**	$\mathrm{N}_{\mathrm{samples}}$
Bloomberg	0.825	262	0.828	252
CDP	1.170	327	1.178	314
GHGv2	0.880	843	0.844	714
MSCI	0.936	376	0.947	320
Trucost	1.061	558	1.055	536

(a) All companies are considered.

	With missing BICS			Without missing BICS		
Provider	**RMSE: provider**	**RMSE: GHGv2**	$\mathrm{N}_{\mathrm{samples}}$	**RMSE: provider**	**RMSE: GHGv2**	$\mathrm{N}_{\mathrm{samples}}$
MSCI	0.936	0.861	376	0.947	0.848	320
Bloomberg	0.825	0.698	262	0.828	0.695	252
Trucost	1.061	0.811	558	1.055	0.803	536
CDP	1.170	0.818	327	1.178	0.812	314

(b) Only common companies between providers are considered.

Table 4.12: Last scope 1 GHG estimates from providers (2020 – 2019) compared to 2021 ground truth for companies that started reporting in 2021, for the second iteration of the model.

	With missing BICS		Without missing BICS	
Provider	**RMSE**	$\mathrm{N}_{\mathrm{samples}}$	**RMSE**	$\mathrm{N}_{\mathrm{samples}}$
Bloomberg	0.610	274	0.608	266
CDP	0.960	379	0.969	369
GHGv2	0.742	889	0.714	758
MSCI	0.859	395	0.864	338
Trucost	0.831	591	0.838	571

(a) All companies are considered.

	With missing BICS			Without missing BICS		
Provider	**RMSE: provider**	**RMSE: GHGv2**	$\mathrm{N}_{\mathrm{samples}}$	**RMSE: provider**	**RMSE: GHGv2**	$\mathrm{N}_{\mathrm{samples}}$
Bloomberg	0.610	0.598	274	0.608	0.587	266
MSCI	0.859	0.697	395	0.864	0.693	338
Trucost	0.831	0.688	591	0.838	0.686	571
CDP	0.960	0.705	379	0.969	0.704	369

(b) Only common companies between providers are considered.

Table 4.13: Last scope 2 GHG estimates from providers (2020 – 2019) compared to 2021 ground truth for companies that started reporting in 2021, for the second iteration of the model.

	Scope 1		Scope 2	
Last year of training	**2020**	**2021**	**2020**	**2021**
$N_{samples}$ total	16013	19119	16512	19595
$N_{samples}$ with missing revenues	421	497	458	543
$N_{samples}$ with missing BICS	1071	1331	1113	1377
$N_{samples}$ with missing country	420	495	456	539
$N_{samples}$ with missing revenues and non-missing BICS	1	2	2	4
$N_{samples}$ with missing revenues and non-missing country	1	2	2	4

Table 4.14: Number of missing values for the revenues, BICS sector L1 and country features in the dataset used in the third iteration of the model.

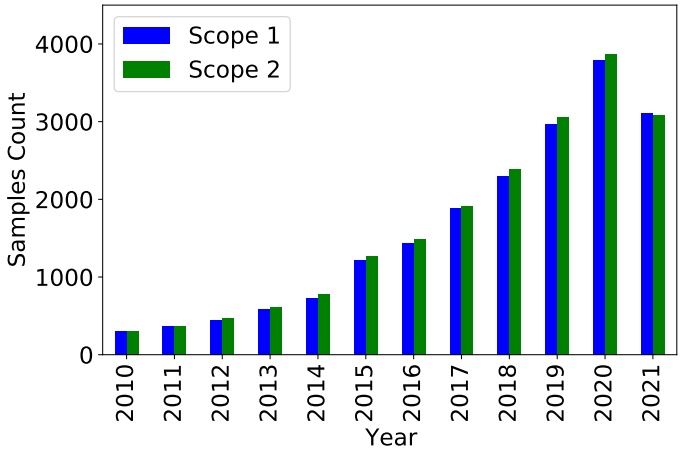


Figure 4.18: Number of companies with a reported GHG emission per year for scopes 1 and 2, for the dataset used in the third iteration of the model.

Figure 4.18 exhibits the evolution from 2010 to 2021 of the number of companies reporting their GHG emissions for the dataset used in training. Similarly to what is explained in 4.8.1, the 2021 CDP reported samples were not available: the drop observed in 2021 is due to their absence. Table 4.14 describes the added samples to the used datasets. We note that this version of the model is not only trained on samples with missing revenues but also on samples with missing country, which was not the case before (see section 4.8.1): indeed, in the gathered data, only samples with missing revenues also have sometimes missing country information. Moreover, when revenues are missing, likely, many fundamental and sectorial information is also missing.

Considering the universe of 48,429 companies extracted from the Worldscope Refinitiv database for the year 2020, this version of the model can propose 40,898 estimates for scope 1 and 41,218 for scope 2.

	Scope 1		Scope 2	
Metric	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
R^2	0.803	0.015	0.735	0.017
RMSE	0.631	0.027	0.573	0.037
MAE	0.418	0.012	0.366	0.017

(a) Test 2020.

	Scope 1		Scope 2	
Metric	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
R^2	0.816	0.019	0.755	0.023
RMSE	0.615	0.035	0.559	0.038
MAE	0.418	0.021	0.352	0.017

(b) Test 2021.

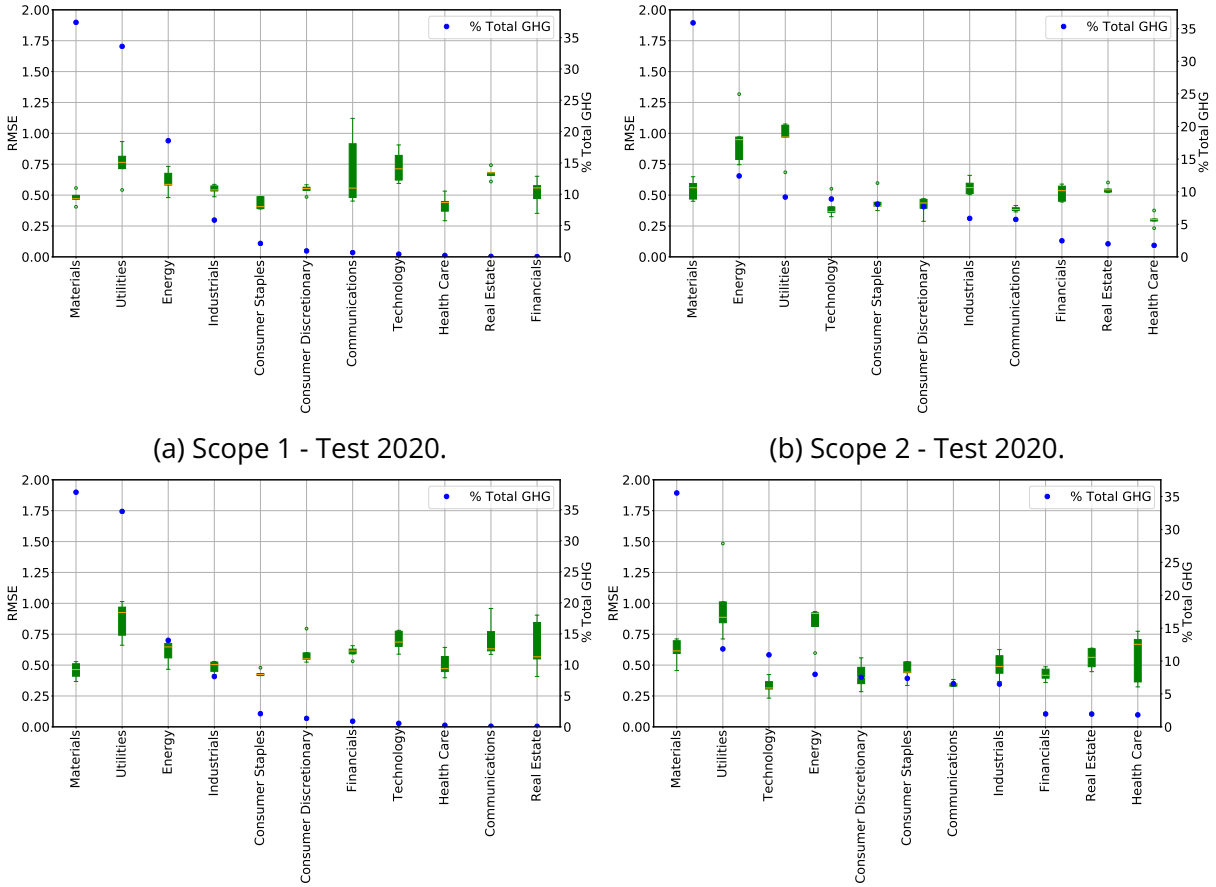
Table 4.15: Results of the third iteration of the model on the five different test sets, including samples with missing revenues: mean and standard deviation of the R^2, RMSE and MAE metrics.

4.9.2 . Results: evaluating the performance of the model

Global results of models trained on years 2010 to 2020 and years 2010 to 2021 are exhibited in Tab. 4.15. Table 4.16 displays global results when evaluating the models on test sets excluding any sample with missing revenues.

Accentuating what is observed in section 4.8.2 with the BICS features, the average performance of the model is much better when evaluated on samples without missing revenues. The drop in performance following the inclusion of these samples is due to the double effect of including samples with missing revenues which in reality have many different features missing. What is interesting is that the inclusion of these samples in the training set seems to not have impaired the performance of the model on the more complete samples, preserving a good accuracy. This is most certainly due to the tree structure we used to train models: the GBDT algorithm has, for samples with missing revenues, very little information on which to base its estimates and, as a result, samples follow the same paths in the different trees leading to a same estimate with poor accuracy. For complete samples, the algorithm builds efficient decision trees. We note that the average performance of this model is close to those of the first and second model iterations when evaluating it on samples without missing revenues.

Distributions of performance expressed in RMSE per BICS Sector L1 are exhibited in Fig. 4.19 and 4.20, for models evaluated on the full test sets and the filtered test sets. Similarly to the results obtained in section 4.8.2, performance of models evaluated including or excluding samples with missing revenues is very similar: the inclusion of incomplete samples in the training set did not impair the quality of the estimates based on more complete samples.



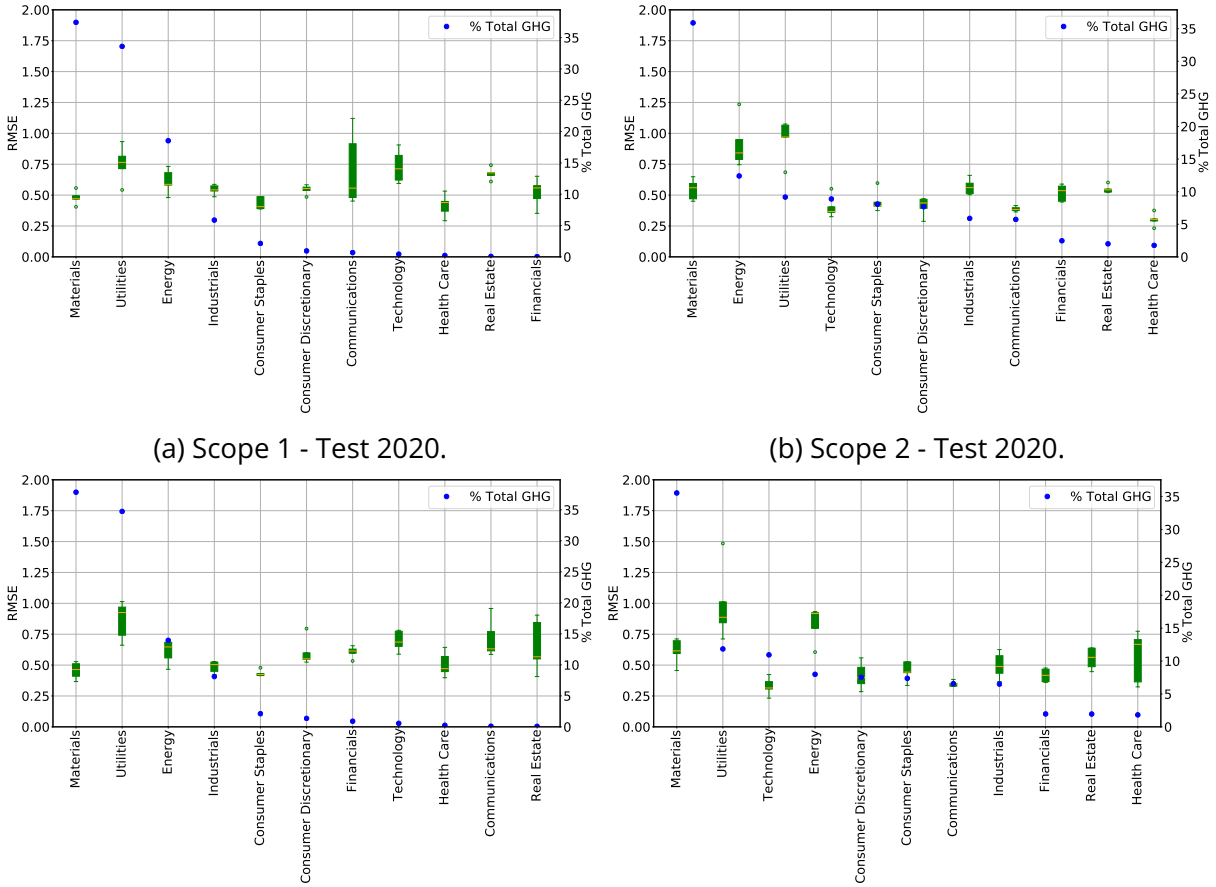
(a) Scope 1 - Test 2020.

(b) Scope 2 - Test 2020.

(c) Scope 1 - Test 2021.

(d) Scope 2 - Test 2021.

Figure 4.19: Distribution of performance of the third iteration of the model on five test sets including samples with missing revenues, per BICS sector level 1 and ordered by level 1 sectors emissions. The percentage of total GHG represents the percentage of total GHG emissions the sector level 1 represents in the dataset of reporting companies.



(a) Scope 1 - Test 2020.

(b) Scope 2 - Test 2020.

(c) Scope 1 - Test 2021.

(d) Scope 2 - Test 2021.

Figure 4.20: Distribution of performance of the third iteration of the model on five test sets excluding samples with missing revenues, per BICS sector level 1 and ordered by level 1 sectors emissions. The percentage of total GHG represents the percentage of total GHG emissions the sector level 1 represents in the dataset of reporting companies.

	Scope 1		Scope 2	
Metric	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
R^2	0.825	0.016	0.756	0.020
RMSE	0.593	0.022	0.548	0.038
MAE	0.396	0.012	0.350	0.017

(a) Test 2020.

	Scope 1		Scope 2	
Metric	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
R^2	0.819	0.019	0.756	0.023
RMSE	0.605	0.035	0.556	0.037
MAE	0.411	0.020	0.346	0.015

(b) Test 2021.

Table 4.16: Results of the third iteration of the model on the five different test sets, excluding samples with missing revenues: mean and standard deviation of the R^2, RMSE and MAE metrics.

4.9.3 . Results: comparison of estimates with other providers

The quality of estimations of the third iteration of the model, called GHGv3, is compared to the quality of estimations from external providers using the methodology proposed in section 4.6. Comparisons are done with data from providers as of August 2022.

Similarly, we compare both results from estimations of 2019-2018 to the 2020 ground truth and from estimations of 2020-2019 to the 2021 ground truth.

Results are exhibited in Tab. 4.17, 4.18, 4.19 and 4.20 for each of the trained model. We propose this analysis on both the full dataset and a filtered dataset excluding samples with missing revenues.

When conducting this analysis on samples with revenues information and selecting the same universe of companies for both the GHGv3 model and the external provider, the proposed GHGv3 model consistently outputs estimates that are of higher quality than the ones from other providers. The performance of the GHGv3 model is however greatly impaired by the inclusion of samples with missing revenues in the evaluation sets. For scope 1 and comparing the 2019 estimates to the 2020 ground truth, the MSCI model even slightly performs better in this specific case. Nonetheless, in all other cases, and despite degraded average performance, the GHGv3 estimates remain more accurate than those of other providers: despite the lower accuracy of estimates on incomplete samples, it seems that they remain on average of better quality than those of the other providers.

	With missing revenues		Without missing revenues	
Provider	**RMSE**	$\mathrm{N}_{\mathrm{samples}}$	**RMSE**	$\mathrm{N}_{\mathrm{samples}}$
Bloomberg	0.913	781	0.915	773
CDP	1.154	651	1.145	603
GHGv3	0.918	1508	0.864	1430
MSCI	0.855	603	0.856	592
Trucost	1.022	1157	1.016	1106

(a) All companies are considered.

	With missing revenues			Without missing revenues		
Provider	**RMSE: provider**	**RMSE: GHGv3**	$\mathrm{N}_{\mathrm{samples}}$	**RMSE: provider**	**RMSE: GHGv3**	$\mathrm{N}_{\mathrm{samples}}$
MSCI	0.855	0.867	603	0.856	0.837	592
Bloomberg	0.913	0.827	781	0.915	0.818	773
Trucost	1.022	0.879	1157	1.016	0.826	1106
CDP	1.154	0.875	651	1.145	0.800	603

(b) Only common companies between providers are considered.

Table 4.17: Last scope 1 GHG estimates from providers (2019 – 2018) compared to 2020 ground truth for companies that started reporting in 2020, for the third iteration of the model.

	With missing revenues		Without missing revenues	
Provider	**RMSE**	$\mathrm{N}_{\mathrm{samples}}$	**RMSE**	$\mathrm{N}_{\mathrm{samples}}$
Bloomberg	0.803	782	0.803	778
CDP	1.026	677	1.025	635
GHGv3	0.768	1491	0.739	1423
MSCI	0.864	626	0.867	614
Trucost	0.834	1135	0.833	1093

(a) All companies are considered.

	With missing revenues			Without missing revenues		
Provider	**RMSE: provider**	**RMSE: GHGv3**	$\mathrm{N}_{\mathrm{samples}}$	**RMSE: provider**	**RMSE: GHGv3**	$\mathrm{N}_{\mathrm{samples}}$
Bloomberg	0.803	0.720	782	0.803	0.716	778
MSCI	0.864	0.753	626	0.867	0.738	614
Trucost	0.834	0.707	1135	0.833	0.684	1093
CDP	1.026	0.731	677	1.025	0.689	635

(b) Only common companies between providers are considered.

Table 4.18: Last scope 2 GHG estimates from providers (2019 – 2018) compared to 2020 ground truth for companies that started reporting in 2020, for the third iteration of the model.

	With missing revenues		Without missing revenues	
Provider	**RMSE**	$\mathrm{N}_{\mathrm{samples}}$	**RMSE**	$\mathrm{N}_{\mathrm{samples}}$
Bloomberg	0.821	265	0.825	262
CDP	1.153	341	1.170	327
GHGv3	0.887	878	0.882	847
MSCI	0.932	379	0.936	376
Trucost	1.050	581	1.060	557

(a) All companies are considered.

	With missing revenues			Without missing revenues		
Provider	**RMSE: provider**	**RMSE: GHGv3**	$\mathrm{N}_{\mathrm{samples}}$	**RMSE: provider**	**RMSE: GHGv3**	$\mathrm{N}_{\mathrm{samples}}$
MSCI	0.932	0.860	379	0.936	0.853	376
Bloomberg	0.821	0.703	265	0.825	0.685	262
Trucost	1.050	0.817	581	1.060	0.808	557
CDP	1.153	0.825	341	1.170	0.815	327

(b) Only common companies between providers are considered.

Table 4.19: Last scope 1 GHG estimates from providers (2020 – 2019) compared to 2021 ground truth for companies that started reporting in 2021, for the third iteration of the model.

	With missing revenues		Without missing revenues	
Provider	**RMSE**	$\mathrm{N}_{\mathrm{samples}}$	**RMSE**	$\mathrm{N}_{\mathrm{samples}}$
Bloomberg	0.609	276	0.610	274
CDP	0.948	395	0.956	376
GHGv3	0.758	928	0.747	892
MSCI	0.857	399	0.861	393
Trucost	0.825	611	0.831	589

(a) All companies are considered.

	With missing revenues			Without missing revenues		
Provider	**RMSE: provider**	**RMSE: GHGv3**	$\mathrm{N}_{\mathrm{samples}}$	**RMSE: provider**	**RMSE: GHGv3**	$\mathrm{N}_{\mathrm{samples}}$
Bloomberg	0.609	0.590	276	0.610	0.590	274
MSCI	0.857	0.693	399	0.861	0.685	393
Trucost	0.825	0.684	611	0.831	0.679	589
CDP	0.948	0.707	395	0.956	0.697	379

(b) Only common companies between providers are considered.

Table 4.20: Last scope 2 GHG estimates from providers (2020 – 2019) compared to 2021 ground truth for companies that started reporting in 2021, for the third iteration of the model.

4.9.4 . Further interpretability elements

This section proposes additional interpretability elements and in particular, an analysis of the impact of missing values in inference for a subset of selected features. We first assess the impact of each feature without missing values on the raw GHG emissions (without the log-transformation). Second, we propose to use the properties of SHAP values to compute the difference in accuracies between estimations made on samples without missing values and estimations made on the same samples with one feature artificially set as missing.

The experiments in this section are conducted using models trained until 2021.

Raw impact of features using SHAP values

We compute the SHAP values associated with the training set. They represent for each sample the extent of the participation of the considered feature to the formation of the GHG estimation.

Using notations from section 2.3.1, we note $\phi_{j,i}$ the SHAP value of feature j for a sample i and $h(\mathbf{x_i})$ the associated prediction for this sample i. Let us denote by P the number of features in the model, $f(\mathbf{x_i}) = 10^{h(\mathbf{x_i})}$ the raw value of the estimated emission for sample i and $\mathbb{E}_X[h(X)]$ the average of all predictions, called the raw base value. We have, rewriting equation (2.22)

$$\mathbb{E}_X[h(X)] + \sum_{j=1}^{P} \phi_{j,i} = h(\mathbf{x_i}). \tag{4.3}$$

The terms of this equation are expressed at the decimal logarithm scale. Transforming these decimal logarithms of emissions into their raw values, we obtain

$$10^{\mathbb{E}_X[h(X)]} \times \prod_{j=1}^{P} 10^{\phi_{j,i}} = f(\mathbf{x_i}). \tag{4.4}$$

We call the coefficient $10^{\phi_{j,i}}$ the raw importance of feature j for a sample i. For each feature, we select only samples without missing values and compute these coefficients on the resulting dataset. They can be interpreted as the impact the considered feature has on the raw base value: if they are below 1 (respectively above 1), they tend to make the estimated emission decrease (respectively increase). Selecting only samples without missing values for the considered feature allows for a better assessment of the impact of this latter, without additional noise. Analyzing the average of these coefficients per sector can yield meaningful results to understand their raw importance and potential differences in behavior per industry.

Figures 4.21a and 4.21b display the plots of these computed coefficients per BICS Sector L1 for a set of numeric features including Revenues, Energy Consumption, Employees, GPPE and NPPE. Figure 4.22 zooms in on Fig. 4.21a for better visibility. Considering scope 1, the Energy Consumption feature has a large weight, especially for the Energy, Materials and Utilities sectors. On average, for

these sectors, it leads to an increase of 600% to 1000% of the raw base value. It is only for the Financial industry (scopes 1 and 2) and the Real Estate industry (scope 2) that the Energy Consumption feature reduces on average the raw base value. In Fig. 4.21c and 4.21d, similar graphs are presented for the six levels of the BICS classification used in the models. The most granular levels have a higher impact on the raw base value, whether this impact is positive or negative.

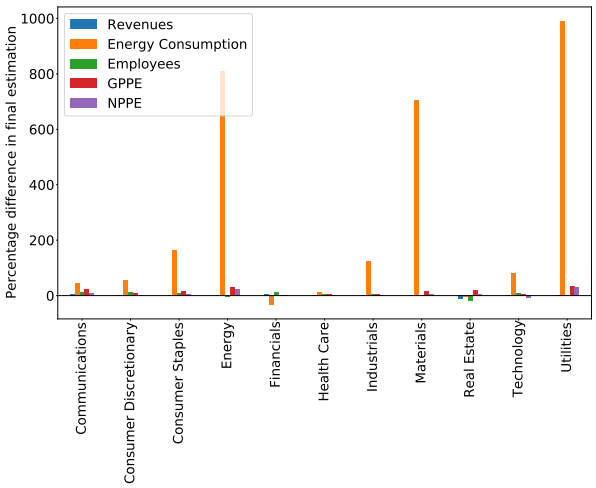
Analysis of the impact of missing values in inference on accuracy

Relying on equation (4.3), we propose a methodology to assess the impact of missing values in inference for a considered feature.

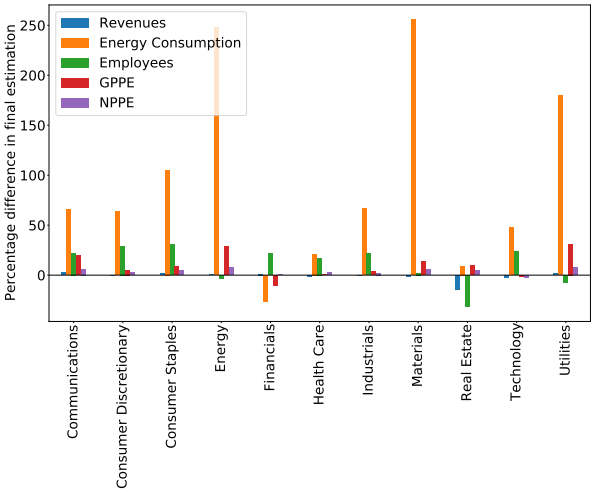
The idea is to filter the used dataset, removing all samples for which this feature is missing. SHAP values associated with remaining samples are computed and by applying equation (4.3), we obtain the estimates computed on the complete set of features. In a second phase, using the same dataset and replacing the considered feature with a NaN value for all samples, we reiterate the procedure and obtain new estimates computed on a partial set of features. Each of the two obtained sets of estimates is compared to the reported ground truth using the RMSE metric. Assessing the deterioration of the RMSE between estimates computed on the complete set of features and estimates computed on the partial set of features highlights the impact of missing values on accuracy for the considered feature. This methodology can be applied by grouping samples per sector, yielding differences in the effects of missing values for each of them.

Figures 4.23a and 4.23b display the loss of RMSE per BICS Sector L1 when applying this methodology for a set of numeric features including Revenues, Energy Consumption, Employees, GPPE and NPPE. These plots highlight the importance of the Energy Consumption feature for both scopes, as missing values lead to an important loss of accuracy. As expected, filling in the Employees feature increases the quality of the estimates, particularly for scope 2. Figures 4.23c and 4.23d exhibit similar plots for the six levels of the BICS classification used in the model. Depending on the sectors, knowledge of levels 3 and 4 is especially important to make qualitative estimates. Filling levels 5 and 6 is less important: this could be due to their frequent absence in the training set, as they are not mandatory. The model has learned how to treat these missing values.

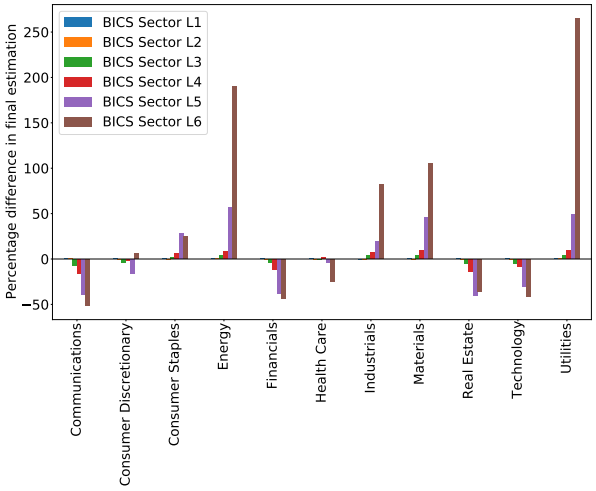
Similar results can be obtained by directly using the model to make estimations on the complete and partial sets of features. Using SHAP values enables assessment of the weight of each feature when set as present or as missing. Figures 4.24 and 4.25 propose such assessment with the examples of the BICS Sector L4 and Energy Consumption as reference features, for the scope 1 model. The average SHAP values per BICS Sector L1 for a subset of features are compared when the reference feature is set as present and missing. These figures highlight that setting a feature as missing does not only impact its associated SHAP value but can change the distribution of SHAP values across all features.



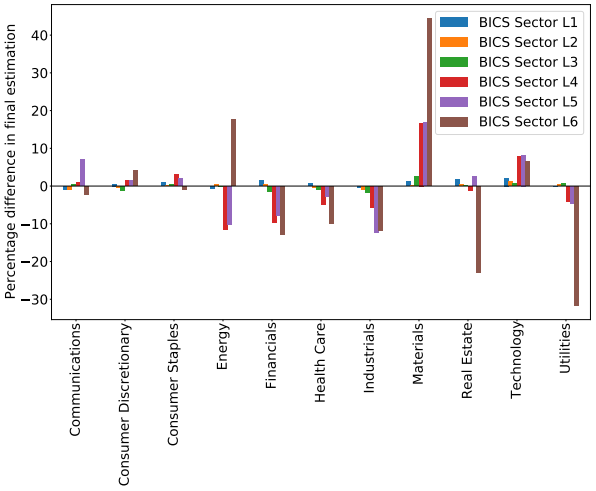
(a) Scope 1 - Subset of numerical features.



(b) Scope 2 - Subset of numerical features.



(c) Scope 1 - Business classification features.



(d) Scope 2 - Business classification features.

Figure 4.21: For a subset of numerical features and the set of business classification features, plot of the average raw importance of the considered feature per BICS Sector L1. Results are obtained using models trained until 2021.

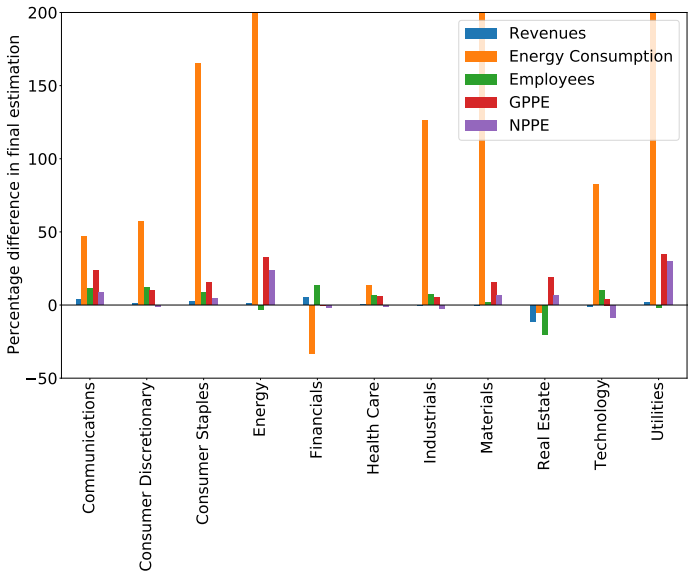


Figure 4.22: Plot of the average raw importance of the considered feature per BICS Sector L1: enlargement of the Y-axis. Scope 1 - Subset of numerical features.

Limitations

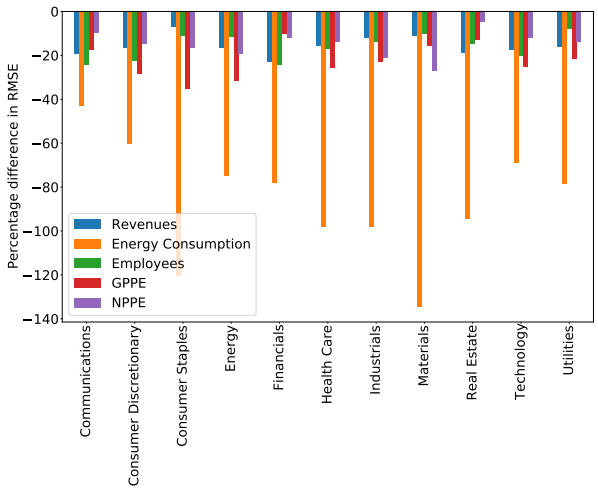
This methodology gives insights into the importance of each feature and the impact of missing values in inference. However, the choice of replacing independently each feature one after the other by missing values can lead to a combination of features values never seen by the model during calibration: this is for instance the case for the sectorial features, the presence of a deep level always meaning the presence of the more general ones. The model can thus produce less reliable estimates.

4.10 . Additional experiments

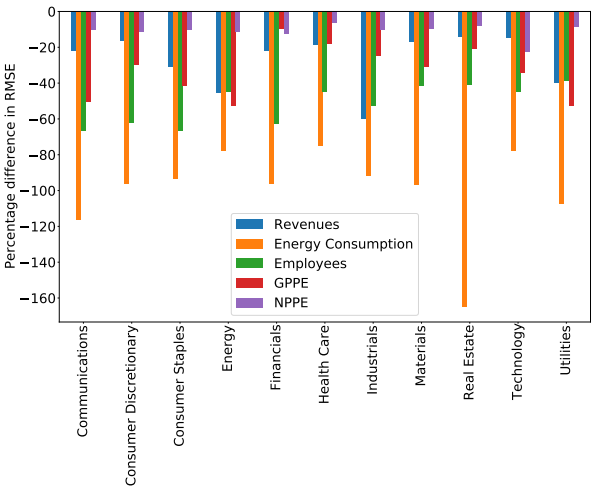
4.10.1 . Benchmark model: estimating GHG emissions using sectorial means

We propose in this section a benchmark model, a reference model to evaluate the performance of the models described in sections 4.5, 4.8.2 and 4.9.2. The benchmark model consists of computing the sectorial means of GHG emissions on selected training sets processed as in section 4.4.4. In inference, a company belonging to a specific sector is associated with the mean emissions of the considered sector.

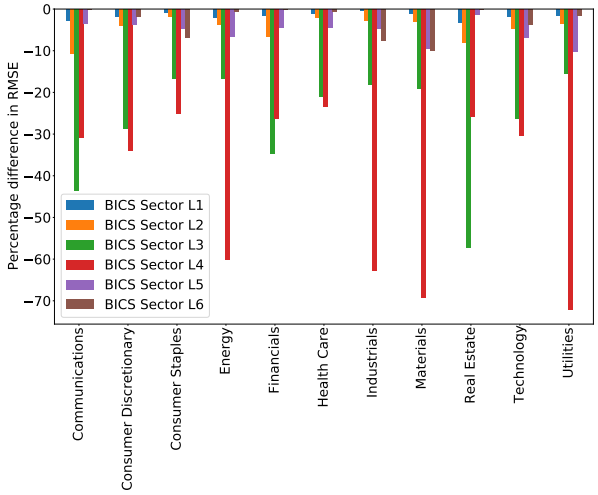
Using a grid search on the different levels of the BICS classification and a cross-validation procedure, we find that using the BICS Sector L4 to compute the



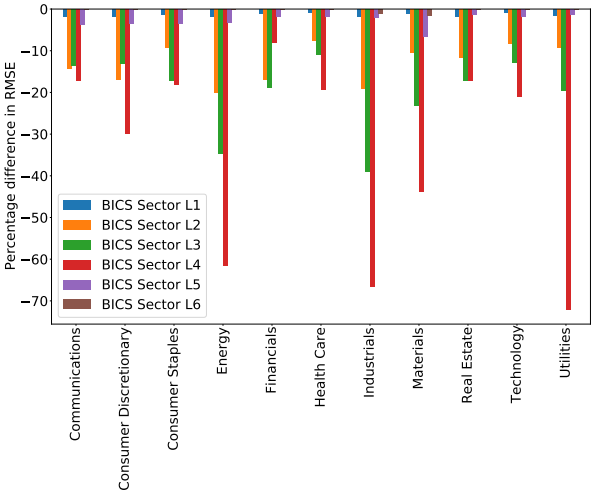
(a) Scope 1 - Subset of numerical features.



(b) Scope 2 - Subset of numerical features.

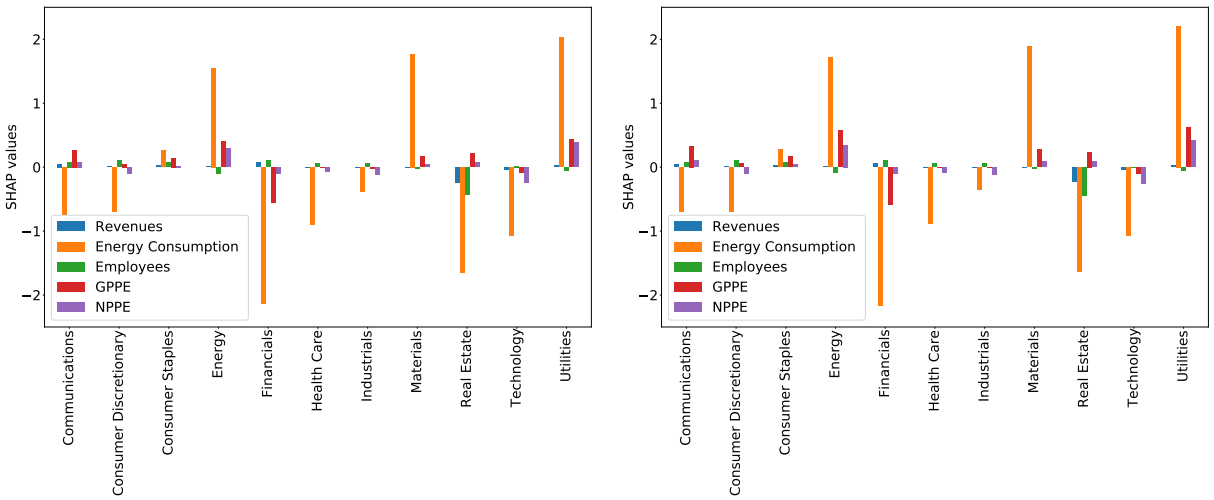


(c) Scope 1 - Business classification features.



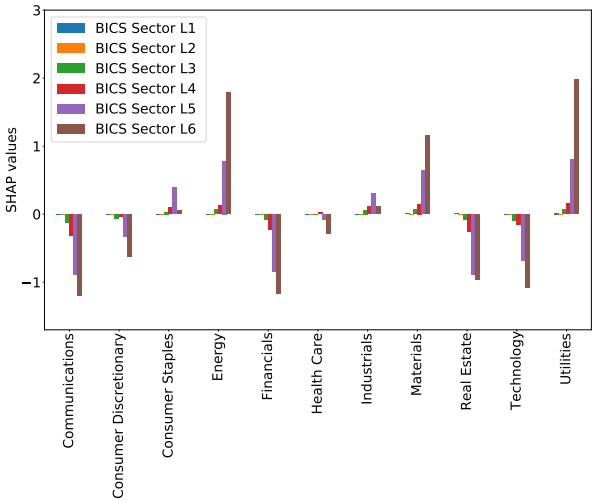
(d) Scope 2 - Business classification features.

Figure 4.23: For a subset of numerical features and the set of business classification features, average percentage difference in RMSE per BICS Sector L1, when the considered feature is fully present or fully missing. Results are obtained using models trained until 2021.

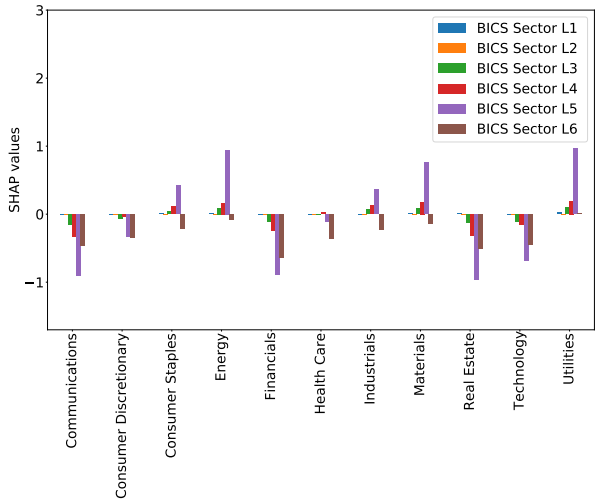


(a) BICS Sector L4 present for all samples.

(b) BICS Sector L4 missing for all samples.

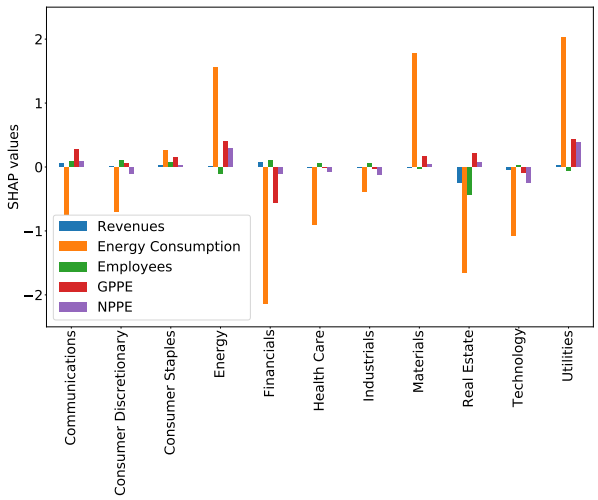


(c) BICS Sector L4 present for all samples.

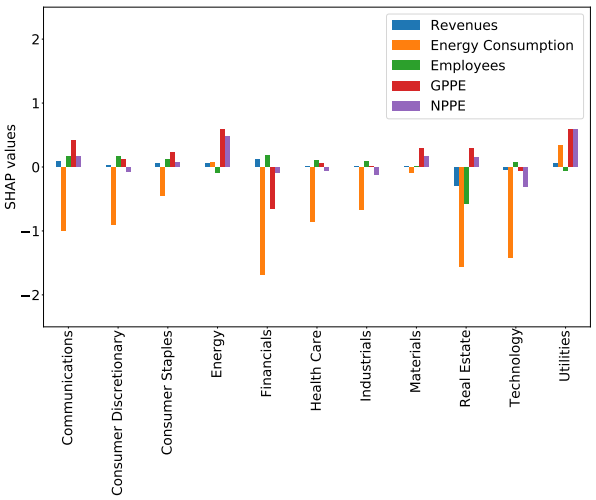


(d) BICS Sector L4 missing for all samples.

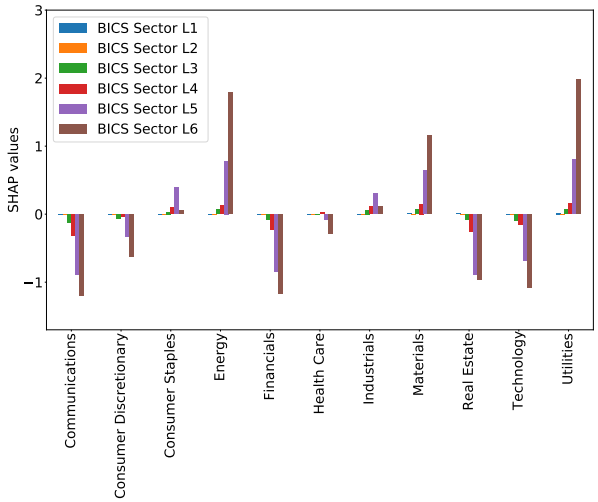
Figure 4.24: Comparison between SHAP values per BICS Sector L1 for a subset of features when the BICS Sector L4 feature is present or missing for all samples. Results are obtained using the scope 1 model trained until 2021.



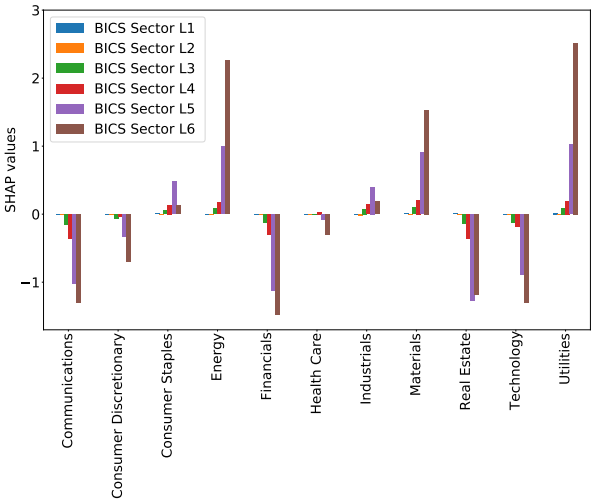
(a) Energy Consumption present for all samples.



(b) Energy Consumption missing for all samples.



(c) Energy Consumption present for all samples.



(d) Energy Consumption missing for all samples.

Figure 4.25: Comparison between SHAP values per BICS Sector L1 for a subset of features when the Energy Consumption feature is present or missing for all samples. Results are obtained using the scope 1 model trained until 2021.

	Scope 1		Scope 2	
Metric	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
R^2	0.491	0.032	0.164	0.033
RMSE	1.02	0.031	1.00	0.032
MAE	0.767	0.017	0.750	0.022

Table 4.21: Results of the benchmark model on the five different test sets for test year 2020: mean and standard deviation of the R^2, RMSE and MAE metrics.

sectorial means leads to the best validation performance for the benchmark model.

The benchmark model is evaluated using five test sets built as described in section 4.4.5, for test year 2020. The performance measures of this model are displayed in Tab. 4.21. They are significantly lower than those of models built using boosting algorithms. This low performance is also observed when evaluating the distributions of performance expressed in RMSE per BICS Sector L1 and BICS Sector L2 displayed in Fig. 4.26.

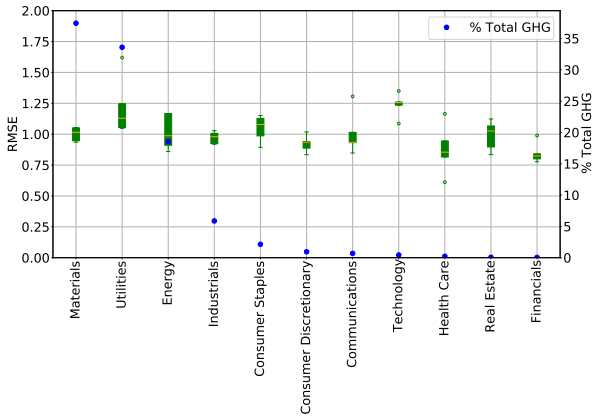
4.10.2 . Training the models on the full dataset

The final models used in this thesis are trained on the whole dataset, without relying on any test sets (see section 4.6.1). In particular, estimates compared to those of external providers are obtained using such models (see sections 4.6, 4.8.3 and 4.9.3).

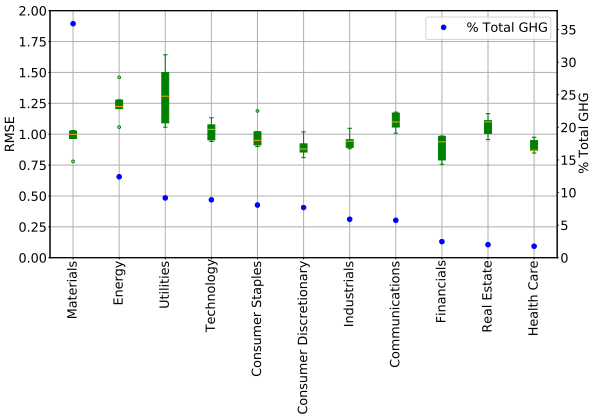
This procedure allows using all available and recent data, an important advantage when working with small non-stationary data. However, the drawback is that we do not have a precise number for the out-of-sample performance of the model. Trained models remain meaningful as the methodology has previously been validated with several test sets. This section brings additional elements to justify this procedure. We show in Fig. 4.27a and 4.27b that by removing at random more and more data points from the training sets, the validation losses increase, leading to lesser performance. This suggests, in the context of GHG emissions, that adding more data to the training set allows the model to improve its generalization capabilities and perform better.

The associated performance on the test sets is presented in Fig. 4.27c and 4.27d and seem rather stable once 50% of the full dataset is used for training. These figures are provided for information as performance on test sets should not be used to choose a training methodology.

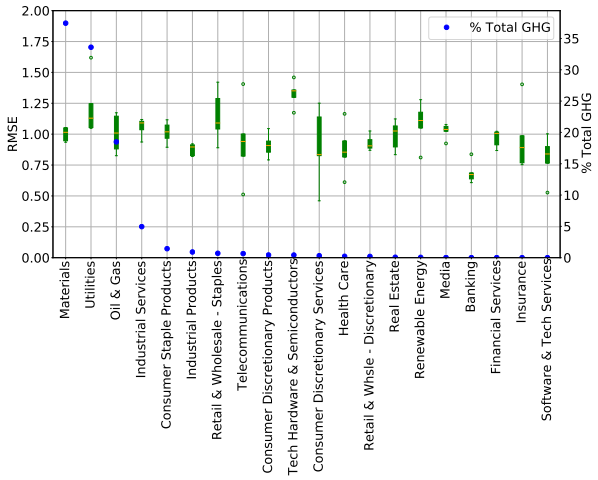
Tests were done using the third iteration of the model trained until 2020.



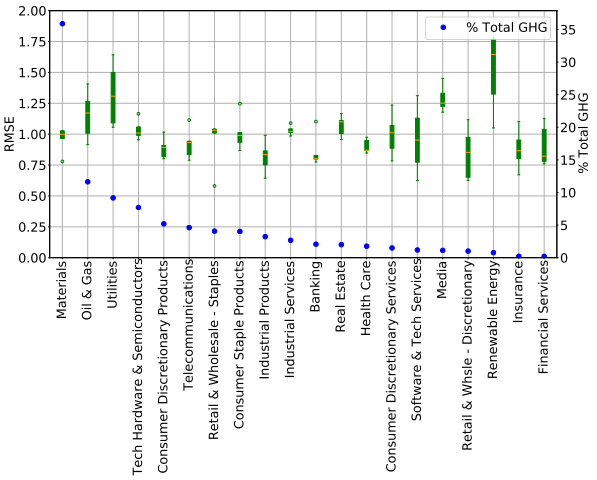
(a) Scope 1 - Per BICS Sector L1.



(b) Scope 2 - Per BICS Sector L1.

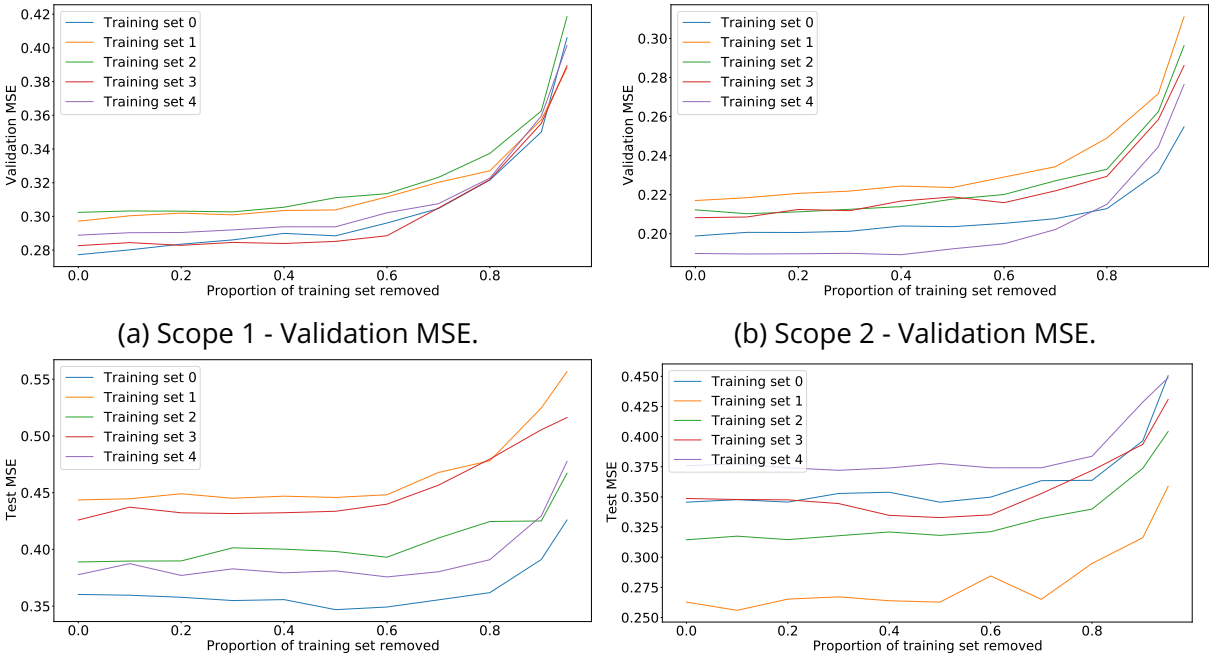


(c) Scope 1 - Per BICS Sector L2.



(d) Scope 2 - Per BICS Sector L2.

Figure 4.26: Distribution of performance of of the benchmark model on five test sets of 2020 and ordered by sectors emissions. The percentage of total GHG represents the percentage of total GHG emissions the sector level represents in the dataset of reporting companies.



(a) Scope 1 - Validation MSE.

(b) Scope 2 - Validation MSE.

(c) Scope 1 - Test MSE.

(d) Scope 2 - Test MSE.

Figure 4.27: Evolution of validation and test losses (MSE) when training models on a degraded training set built by increasingly and randomly removing a proportion of its samples.

	Scope 1		Scope 2	
Model	**RMSE**	$\mathrm{N}_{\mathrm{samples}}$	**RMSE**	$\mathrm{N}_{\mathrm{samples}}$
Estimating 2020 with 2019	0.00499	2807	0.00488	2894
Estimating 2021 with 2020	0.00468	2271	0.00467	2284

Table 4.22: Performance of a model taking the last reported GHG emissions of year $y-1$ to estimate GHG emissions of year y.

4.10.3 . The high year-over-year correlation of GHG emissions

Section 4.6.3 introduces a methodology to compare estimates from our models to those of external providers. It relies on the high year-over-year correlation of reported GHG emissions: we assess a provider by comparing its estimates from year $y-1$ to reported emissions of year y.

The goal of the following experiment is to illustrate the high year-over-year correlation of reported GHG emissions. Let us consider a model taking the reported value of year $y-1$ to estimate emissions of year y, the ground truth, and its associated performance measured by the RMSE. We use the training datasets described in 4.9.1 for the third iteration of the model. Obtained RMSE are exhibited in Tab. 4.22. The RMSE are very low when considering both the reported data of 2019 to estimate 2020 GHG emissions and the reported data of 2020 to estimate 2021 GHG emissions, highlighting the high year-over-year correlation of reported GHG emissions. We find this result for both scopes 1 and 2. It proves the consistency of the proposed methodology to compare providers between them: if the considered model for the year $y-1$ is accurate, the RMSE between estimates of year $y-1$ and the reported ground truth of year y should be low.

4.10.4 . Model performance and chosen business classification

The presented models are built using the BICS classification. We propose here an analysis of the performance of the models when trained with other available business classifications. The used test sets are the same for all models to allow comparisons of performance.

We experiment with five different sets of business classification features:

- Levels 1 to 6 of the BICS classification.
- Levels 1 to 4 of the GICS classification.
- Levels 1 to 4 of the primary SIC classification, classifying the primary activity of companies. The primary activity of a company is the one the company derives the majority of its revenues from.
- Levels 1 to 4 of the primary SIC classification and levels 1 to 4 of the secondary SIC classification. The secondary activity of a company is the one responsible for the second largest source of revenues of the company.

	Scope 1		Scope 2	
Last year of training	**2020**	**2021**	**2020**	**2021**
$\mathbf{N}_{\text{samples}}$ total	15641	18668	16103	19100
BICS	651	833	658	838
GICS	570	605	482	521
Primary SIC	87	95	82	88
Secondary SIC	1719	2104	1772	2149
TRBC	13	14	13	14

Table 4.23: Number of missing values for the first level of granularity of each business classification.

- Levels 1 to 5 of the TRBC classification.

The preprocessing steps described in sections 4.4.3 and 4.8.1 are applied for each of these business classification features. Tests are done using the second iteration of the GHG model.

Table 4.23 exhibits the number of missing values for the first level of granularity of each business classification. The BICS classification has the lowest coverage, with on average 4.3% of the samples without any sectorial information. The coverage of the secondary SIC classification is lower but this is compensated by the presence of the primary SIC classification.

Performance of the models trained with the different classifications is displayed in Tab. 4.24. The use of the different classifications does not show significant differences in performance when evaluating the models on the same test sets: for the different models, the best-performing classification is not always the same. Figures 4.28, 4.29, 4.30 and 4.31 show the associated distributions of performance on the five sets per BICS Sector L1. We observe few and marginal differences from the use of one classification to another, with similar distributions of performance across BICS sectors of level 1. The divergences of performance may be higher when evaluating the models on more granular subsectors.

This experiment shows the impact of using one classification rather than another. It should not be used to justify the choice of classification features: this choice should always be made using a proper validation set (or in cross-validation).

4.10.5 . Interpretability without SHAP values: experiments using linear models

We propose here to use directly interpretable models. Experiments are done using the training set defined for the first iteration of the GHG model.

Results using a linear regression model with L1 and L2 regularization, whose hyperparameters are tuned using company-wise cross-validation, are shown in Tab. 4.25. These models display very poor performance in comparison to GBDT and thus are not further analyzed.

	Scope 1		Scope 2	
Sectors	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
BICS	0.588	0.023	0.548	0.038
GICS	0.601	0.045	0.543	0.035
Primary SIC	0.588	0.016	0.541	0.038
Primary and secondary SIC	0.597	0.019	0.544	0.038
TRBC	0.589	0.026	0.549	0.033

(a) Test 2020.

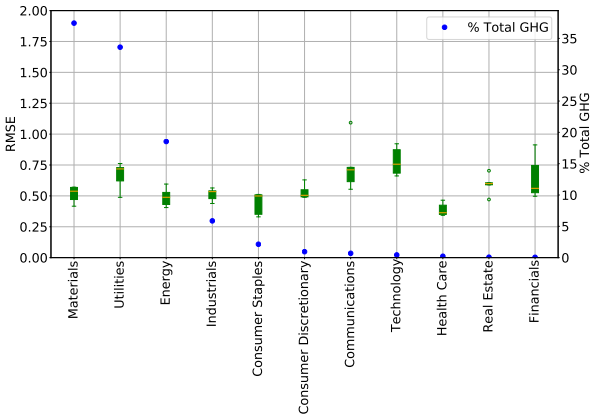
	Scope 1		Scope 2	
Sectors	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
BICS	0.608	0.034	0.543	0.034
GICS	0.592	0.044	0.533	0.036
Primary SIC	0.614	0.040	0.540	0.033
Primary and secondary SIC	0.615	0.038	0.538	0.038
TRBC	0.596	0.042	0.525	0.032

(b) Test 2021.

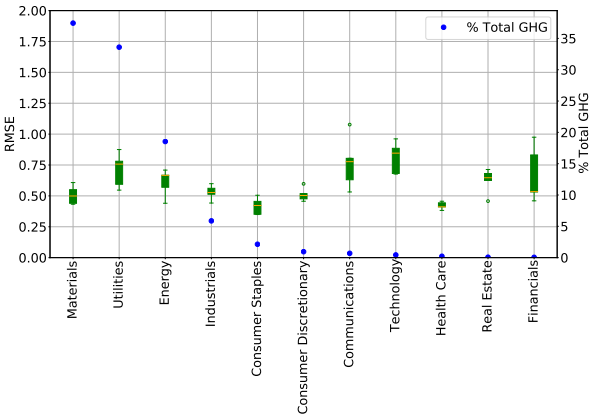
Table 4.24: Results of models trained with five different sets of business classification features, on the five different test sets: mean and standard deviation of RMSE.

	Scope 1		Scope 2	
Metric	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
R^2	0.613	0.010	0.382	0.021
RMSE	0.878	0.010	0.813	0.032
MAE	0.670	0.008	0.614	0.015

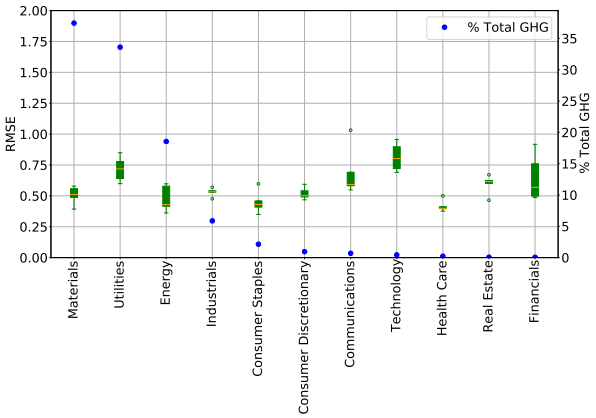
Table 4.25: Results of the linear model on the five different test sets: mean and standard deviation of the R^2, RMSE and MAE metrics.



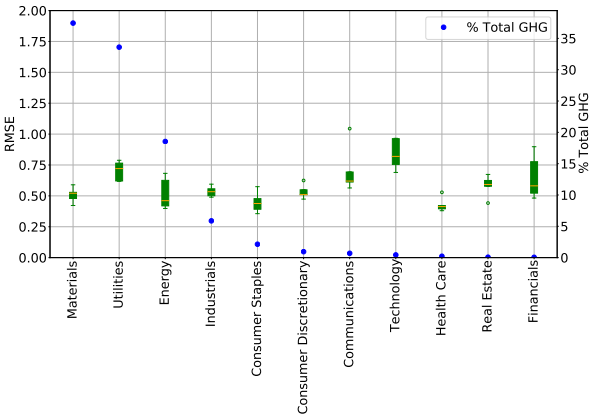
(a) BICS.



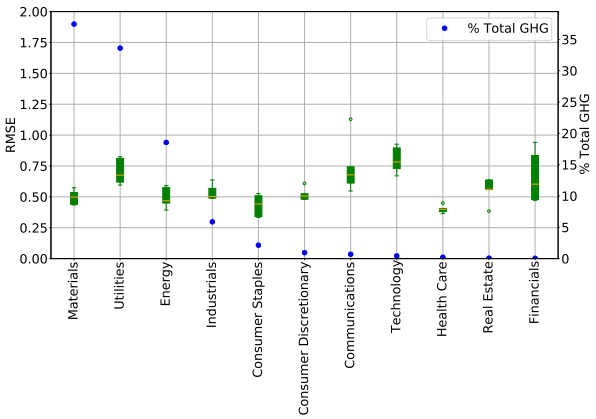
(b) GICS.



(c) Primary SIC.

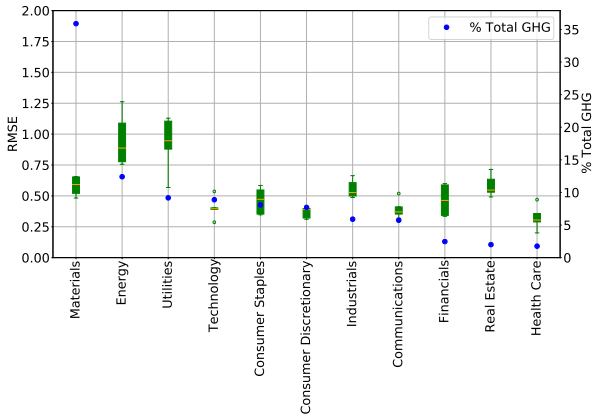


(d) Primary and secondary SIC.

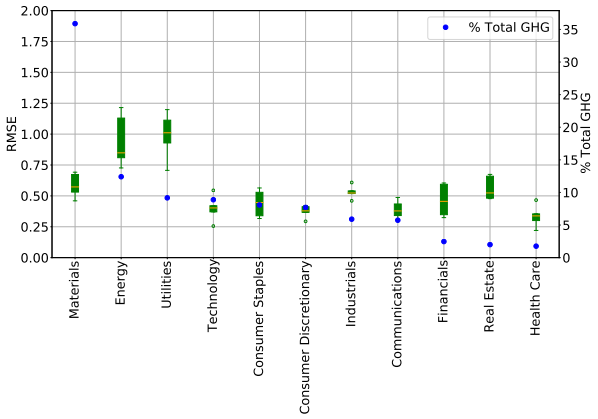


(e) TRBC.

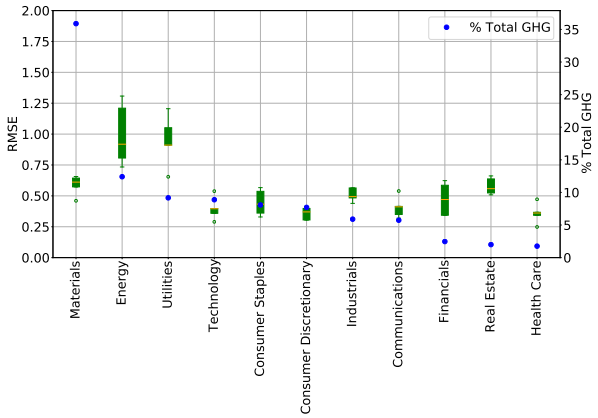
Figure 4.28: Distribution of performance of scope 1 models trained until 2020 with five different sets of business classification features, on five test sets, per BICS sector level 1 and ordered by level 1 sectors emissions. The percentage of total GHG represents the percentage of total GHG emissions the sector level 1 represents in the dataset of reporting companies.



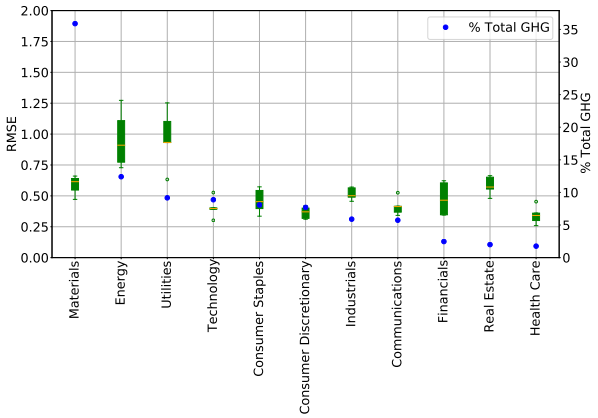
(a) BICS.



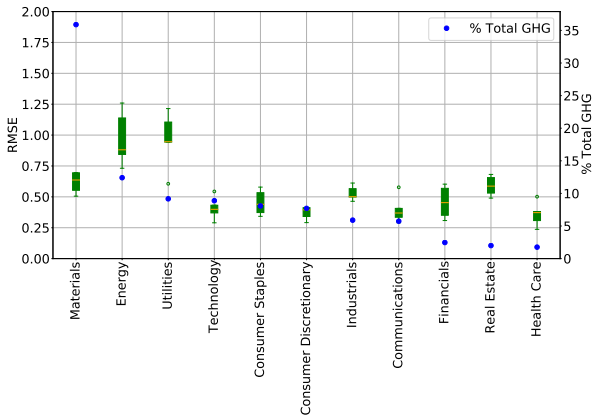
(b) GICS.



(c) Primary SIC.

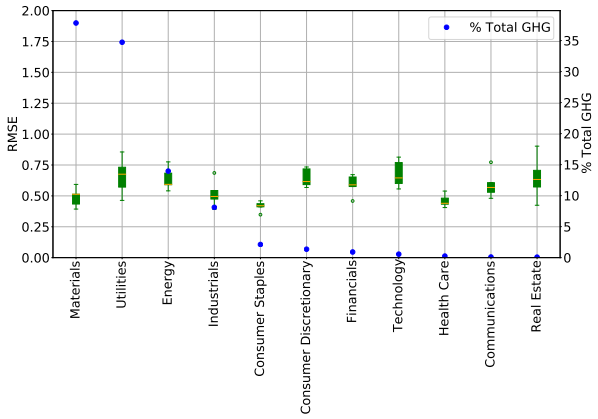


(d) Primary and secondary SIC.

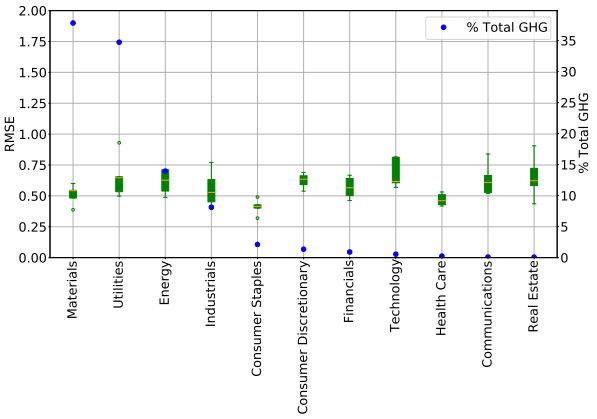


(e) TRBC.

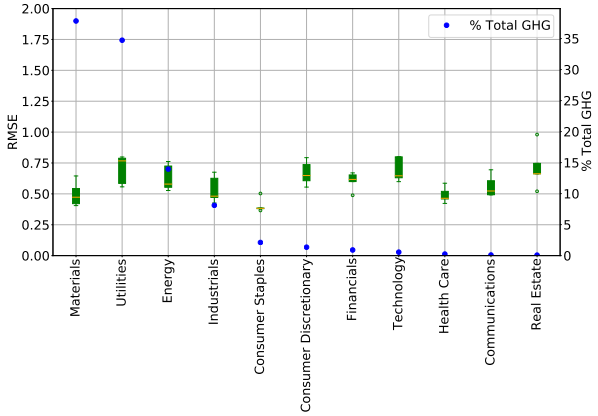
Figure 4.29: Distribution of performance of scope 2 models trained until 2020 with five different sets of business classification features, on five test sets, per BICS sector level 1 and ordered by level 1 sectors emissions. The percentage of total GHG represents the percentage of total GHG emissions the sector level 1 represents in the dataset of reporting companies.



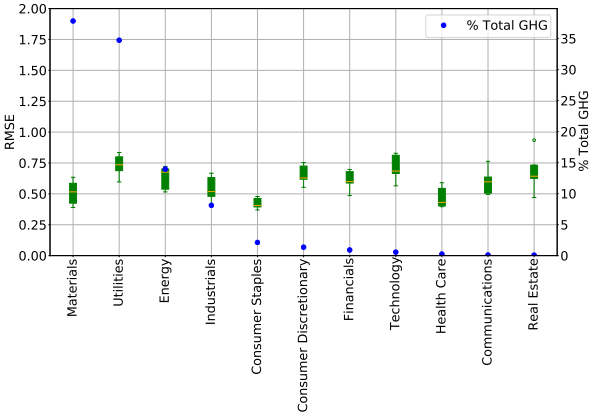
(a) BICS.



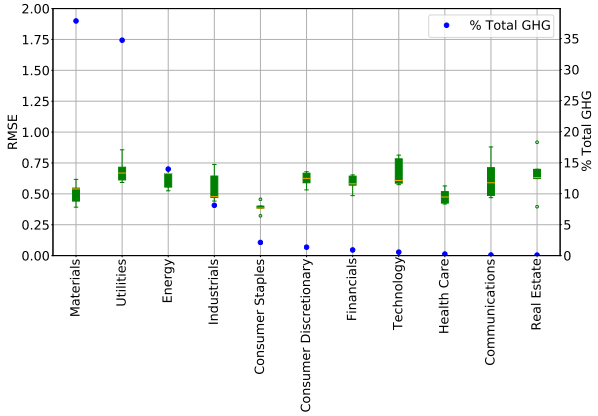
(b) GICS.



(c) Primary SIC.

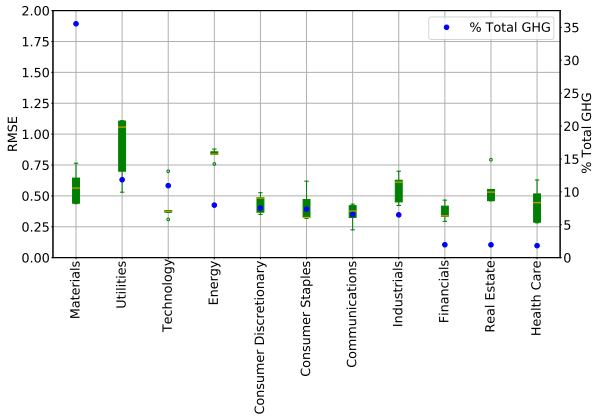


(d) Primary and secondary SIC.

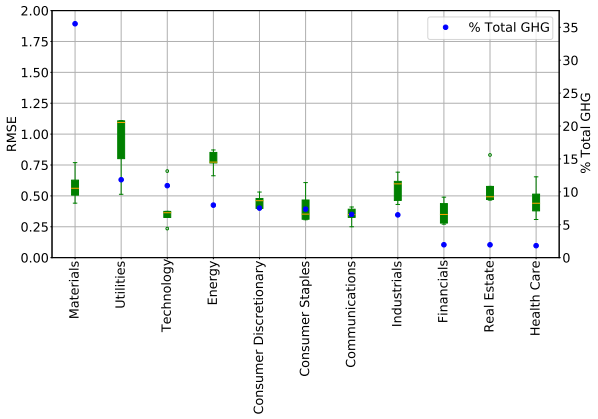


(e) TRBC.

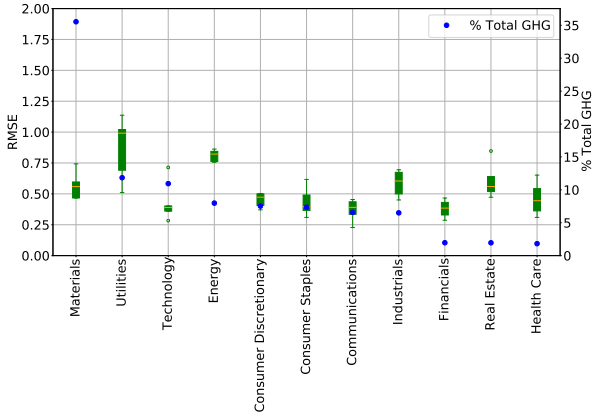
Figure 4.30: Distribution of performance of scope 1 models trained until 2021 with five different sets of business classification features, on five test sets, per BICS sector level 1 and ordered by level 1 sectors emissions. The percentage of total GHG represents the percentage of total GHG emissions the sector level 1 represents in the dataset of reporting companies.



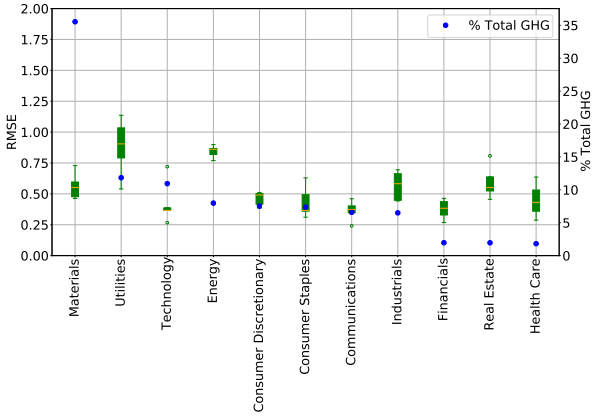
(a) BICS.



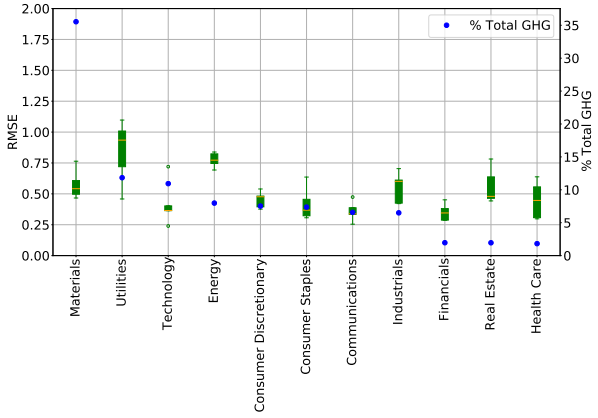
(b) GICS.



(c) Primary SIC.



(d) Primary and secondary SIC.



(e) TRBC.

Figure 4.31: Distribution of performance of scope 2 models trained until 2022 with five different sets of business classification features, on five test sets, per BICS sector level 1 and ordered by level 1 sectors emissions. The percentage of total GHG represents the percentage of total GHG emissions the sector level 1 represents in the dataset of reporting companies.

4.10.6 . Towards a more robust model: an outliers removal methodology

Principles of data polishing

Plots in Fig. 4.13 highlight some clusters of SHAP values inside the distribution of BICS Sector L1 SHAP values per BICS Sector L1. These clusters show differences in the distribution of the initial data. For instance, the distribution of SHAP values for the Utilities sector in scope 1 displays a cluster of SHAP values below 0.04 with few samples. These correspond to the years 2012 to 2014 of a specific company for which the reported Energy Consumption is around 19,000 GWh whereas the reported values for the same company from 2015 to 2020 are between 30 and 65 GWh. Removing this cluster with very few samples allows for the improvement of the quality of the training data by removing outliers. Similar studies on other sectors lead to the same results: for example, for the Materials sector, it can lead to the removal of the only years a company did not report its Energy Consumption. Working on these clusters and removing the ones with too few samples can be a solution to improve the model by removing outliers, noisy data and preventing overfitting.

This methodology should, however, be automated and applied systematically. A first implementation using the SHAP distribution for each BICS Sector L4 was done, based on the first iteration of the model.

For each BICS sector of level 4, a hierarchical clustering algorithm is applied, separating clusters if their distance is above 0.04 in the SHAP values space. Clusters of data with an insufficient number of samples, i.e. less than 10, are removed. These parameters were found by trial-and-error. For both scopes 1 and 2, it leads to the removal of about respectively 11.5% and 5% of the training data, enabling an improvement in the global performance of the model. The number of remaining data points per year is displayed in Fig. 4.32 for both scopes.

Results are presented in Tab. 4.26, on average on the five different test sets. For both scopes, we observe an average RMSE decrease between 11% and 13%. It may come at the price of an increased variance of results between the different test sets, especially for scope 1.

Sector-wise comparison

We propose in Fig. 4.33 a comparison of the performance of the model trained on the full dataset with the one trained on the dataset filtered for outliers, per BICS sector of level 1.

Regarding scope 1, we observe, comparing Fig. 4.33a and 4.33b that removing outliers leads to an improvement of performance for three out of the four most emissive sectors, namely Utilities, Materials and Industrials. Results are less clear-cut for the Energy sector and would require a more in-depth analysis. There is however no degradation in performance.

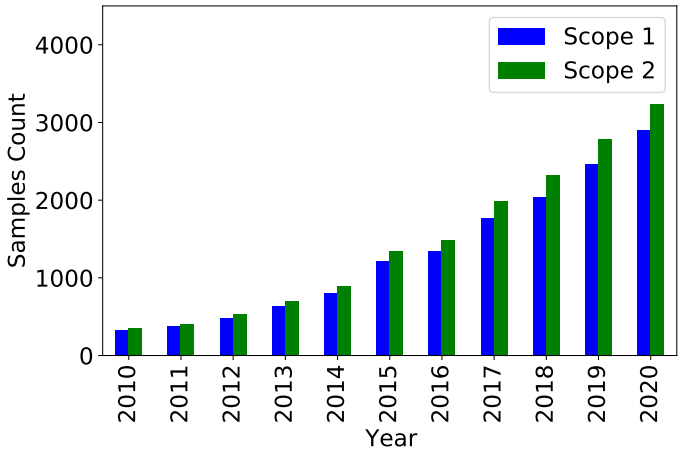


Figure 4.32: Number of companies with a reported GHG emission per year for scopes 1 and 2, for the dataset filtered for outliers.

	Without data polishing		With data polishing	
Metric	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
R^2	0.832	0.007	0.859	0.009
RMSE	0.578	0.007	0.501	0.020
MAE	0.401	0.006	0.347	0.013

(a) Scope 1

	Without data polishing		With data polishing	
Metric	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
R^2	0.746	0.017	0.778	0.017
RMSE	0.521	0.031	0.464	0.025
MAE	0.341	0.010	0.312	0.011

(b) Scope 2

Table 4.26: Results of the model trained on the full dataset and of the model trained on a dataset on which the outliers removal methodology was applied, on the five different test sets: mean and standard deviation of the R^2, RMSE and MAE metrics.

Similar results are observed for scope 2 in Fig. 4.33c and 4.33d. Sectors accounting for at least 5% of the total emissions of the reporting companies have seen their performance improved (Materials, Energy, Consumer Staples, Industrials, Utilities) or remained similar (Consumer Discretionary, Technology, Communications).

Limitations

The use of this cleaning methodology to remove outliers leads to an improvement in performance for most sectors. For some others, performance remains similar. The use of this methodology can however come at the cost of more variance of performance measures in the different test sets. Indeed, this methodology leads to the removal of some specific data points that are outliers in the original training set but that can correspond to real company situations. These situations may not be unique and the constructed test sets may include companies in similar ones. The model does not know anymore how to handle these situations, leading to lower quality estimates for the corresponding samples. The removal of some specific sectors or countries can lead to similar issues.

4.10.7 . Using sample weights

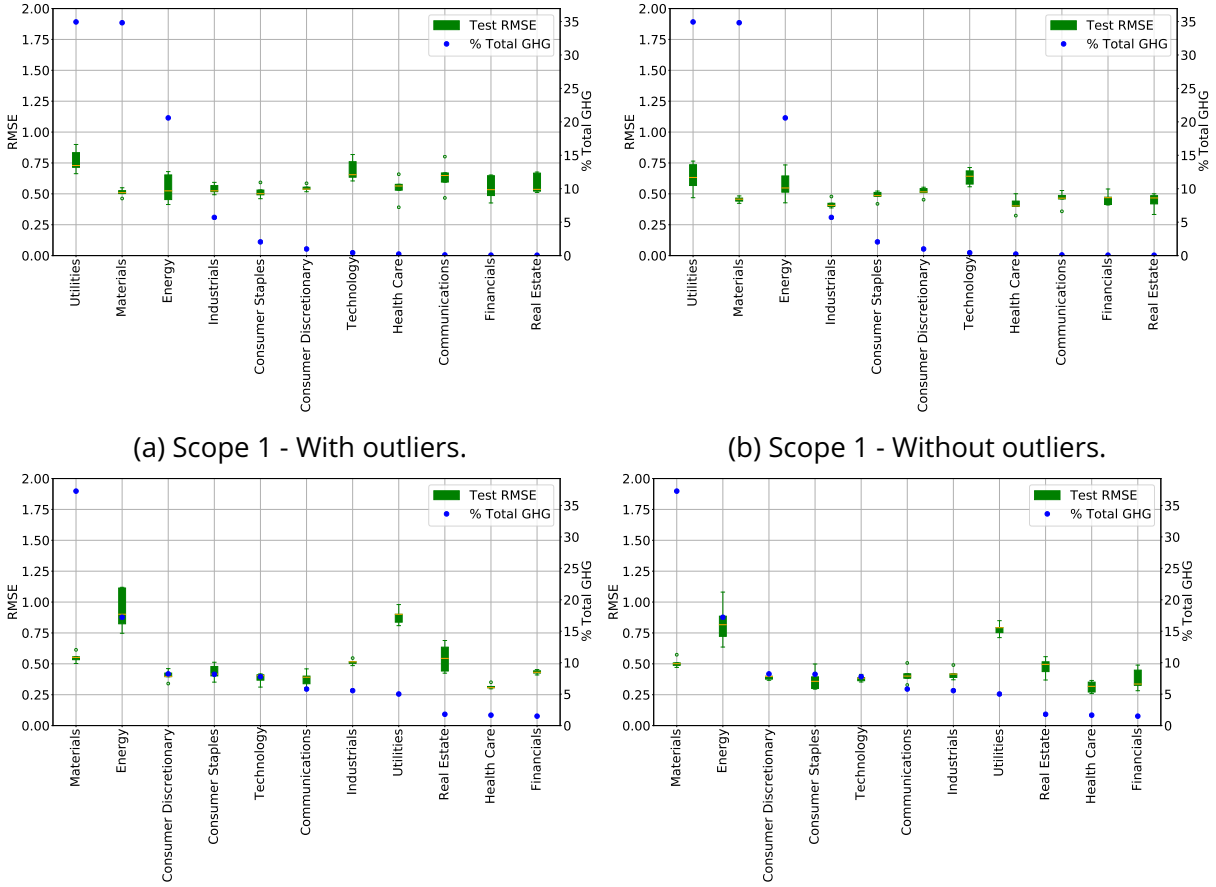
In the different proposed iterations of the models, sample weights are not used, each sample contributing equally to the loss. Here, we describe experiments done to overweight the contribution to the loss of samples with large emissions so that the model focuses more on estimating correctly high GHG emissions.

Experiments are conducted using the dataset introduced for the second iteration of the GHG model and the same test sets. They are done with the sample weights proposed in Fig. 4.34, corresponding to equation

$$w_i = 1.2^{y_i}, \tag{4.5}$$

where w_i is the weight associated with sample i and y_i the reported GHG emission associated with sample i.

Table 4.27 shows the obtained performance of models using sample weights. Results are slightly impaired in comparison to the ones obtained without sample weights in Tab. 4.8. Figure 4.35 displays the breakdown of performance per BICS Sector L1: similarly, performance measures are not improved by the use of sample weights. Specifically, it is not the case for the most emissive sectors for which the sample weights were introduced.



(a) Scope 1 - With outliers.

(b) Scope 1 - Without outliers.

(c) Scope 2 - With outliers.

(d) Scope 2 - Without outliers.

Figure 4.33: Distribution of performance of the model trained on the full dataset and of the model trained on a dataset on which the outliers removal methodology was applied, on five test sets, per BICS sector level 1 and ordered by level 1 sectors emissions. The percentage of total GHG represents the percentage of total GHG emissions the sector level 1 represents in the dataset of reporting companies.

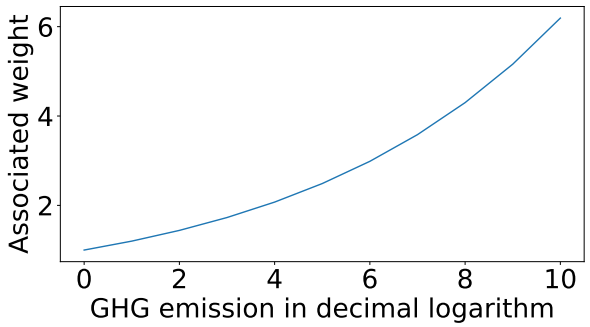


Figure 4.34: Sample weight against reported GHG emission in decimal logarithm.

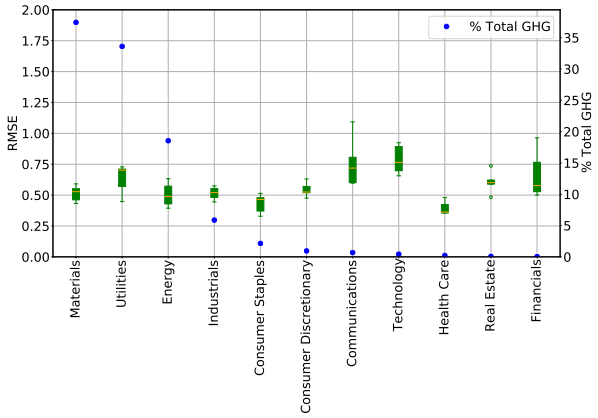
	Scope 1		Scope 2	
Metric	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
R^2	0.823	0.008	0.747	0.028
RMSE	0.595	0.025	0.551	0.039
MAE	0.395	0.016	0.344	0.016

(a) Test 2020.

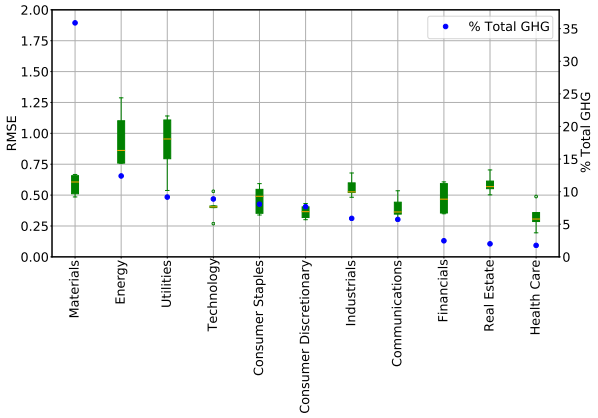
	Scope 1		Scope 2	
Metric	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
R^2	0.821	0.021	0.755	0.033
RMSE	0.613	0.033	0.551	0.028
MAE	0.410	0.019	0.345	0.015

(b) Test 2021.

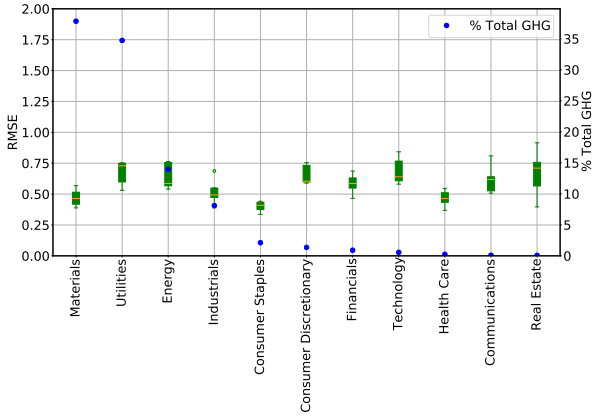
Table 4.27: Results of the model trained using sample weights on the five different test sets: mean and standard deviation of the R^2, RMSE and MAE metrics.



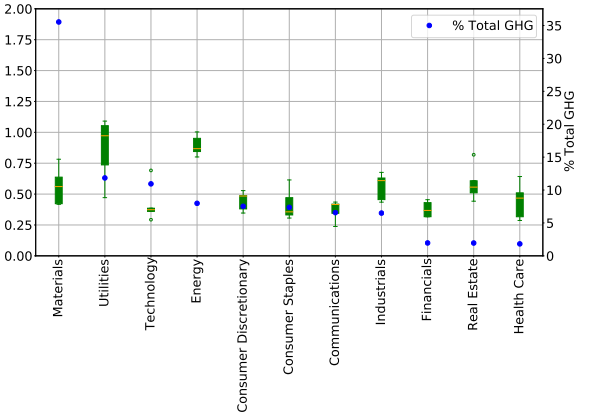
(a) Scope 1 - Test 2020.



(b) Scope 2 - Test 2020.



(c) Scope 1 - Test 2021.



(d) Scope 2 - Test 2021.

Figure 4.35: Distribution of performance of the model trained using sample weights on five test sets per BICS sector level 1 and ordered by level 1 sectors emissions. The percentage of total GHG represents the percentage of total GHG emissions the sector level 1 represents in the dataset of reporting companies.

4.10.8 . Using a custom loss

The GBDT model is trained on the MSE loss (see equation (2.8)). We propose here an experiment to make the model more conservative, meaning to force it to overestimate rather than underestimate GHG emissions. Experiments are conducted using the dataset introduced for the second iteration of the GHG model and the same test sets.

To do so, we train the model using the following loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} c_i, \tag{4.6}$$

with

$$c_i = \begin{cases} 10(y_i - h(\mathbf{x_i}))^2 & \text{if } y_i - h(\mathbf{x_i}) > 0, \\ (y_i - h(\mathbf{x_i}))^2 & \text{otherwise,} \end{cases} \tag{4.7}$$

where h is the calibrated model, $\mathbf{x_i}$ is the vector of features associated with sample i and y_i is the ground truth (decimal logarithm of the reported GHG emission).

This loss overweights samples that were underestimated. To optimize the loss, the algorithm has to learn in priority qualitative estimations for these samples.

Table 4.28 displays the obtained results. In comparison to the performance of the models trained using the classic MSE loss (see Tab. 4.8), the use of this custom loss leads to poor accuracy. Figure 4.36 shows the breakdown of performance per BICS Sector L1. Similar conclusions can be drawn from these plots.

Let us note that, despite deteriorated performance, the use of the custom loss indeed increases the tendency of the model to overestimate instead of underestimate GHG emissions. This is illustrated in Tab. 4.29 where the average number of overestimations and underestimations on the test sets are compared to the results obtained with the second iteration of the model.

4.10.9 . Extrapolation of GHG estimates and other comments

The model is designed to estimate scopes 1 and 2 GHG emissions for companies that do not report them for the same year as the last one in the training set. In this section, we assess the performance of the model in a different use case, when used in extrapolation, i.e. to estimate GHG emissions for the next year. Experiments are done using the third iteration of the model, filtering the dataset so that the revenues feature is never missing.

For each of the experiments, we do not show the results on the full considered sets which can be rather small. For each of them, we take five times 60% of their samples and compute the RMSE on these subsamples. The five obtained RMSE are then averaged to obtain the mean performance. Monitoring the mean performance and standard deviation of RMSE using this method mitigates the small size of the

	Scope 1		Scope 2	
Metric	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
R^2	0.803	0.011	0.714	0.029
RMSE	0.627	0.035	0.585	0.041
MAE	0.419	0.019	0.357	0.022

(a) Test 2020.

	Scope 1		Scope 2	
Metric	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
R^2	0.801	0.020	0.727	0.045
RMSE	0.646	0.029	0.582	0.036
MAE	0.430	0.019	0.360	0.016

(b) Test 2021.

Table 4.28: Results of the model trained using a custom loss on the five different test sets: mean and standard deviation of the R^2, RMSE and MAE metrics.

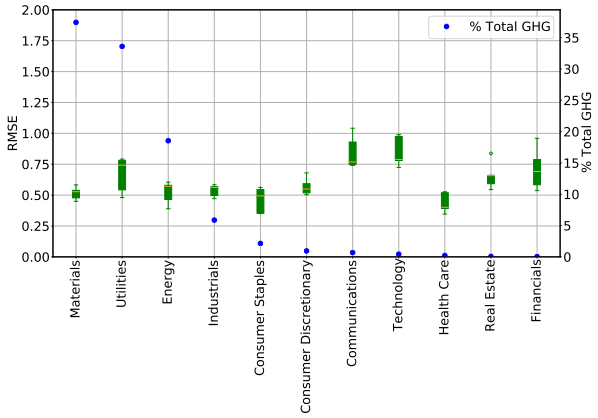
	Scope 1		Scope 2	
Metric	**Classic Loss**	**Custom Loss**	**Classic Loss**	**Custom Loss**
$\mathrm{N}_{\mathrm{samples}}$ total	729	729	742	742
Overestimation Mean	352.8	421.6	339.2	439.0
Underestimation	376.2	308.4	402.8	303.0

(a) Test 2020.

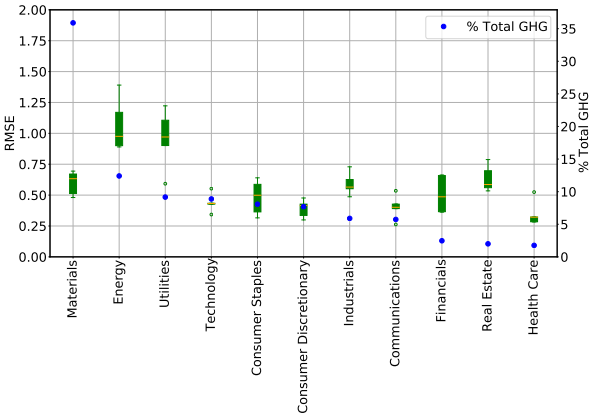
	Scope 1		Scope 2	
Metric	**Classic Loss**	**Custom Loss**	**Classic Loss**	**Custom Loss**
$\mathrm{N}_{\mathrm{samples}}$ total	605	605	599	599
Overestimation	289.8	349.6	276.8	361.2
Underestimation	315.2	255.4	322.2	237.8

(b) Test 2021.

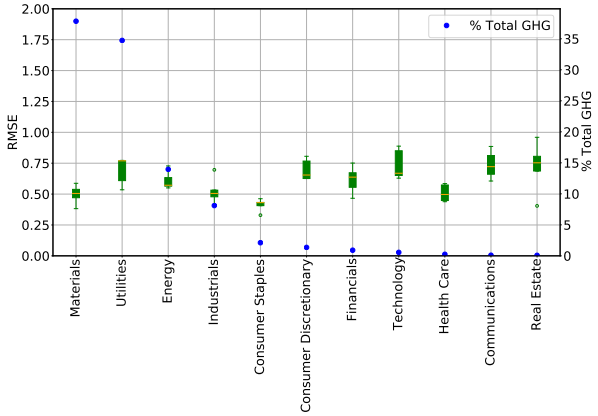
Table 4.29: Average number of overestimated and underestimated samples, across five different test sets, for models trained using the classic (MSE) or proposed custom loss.



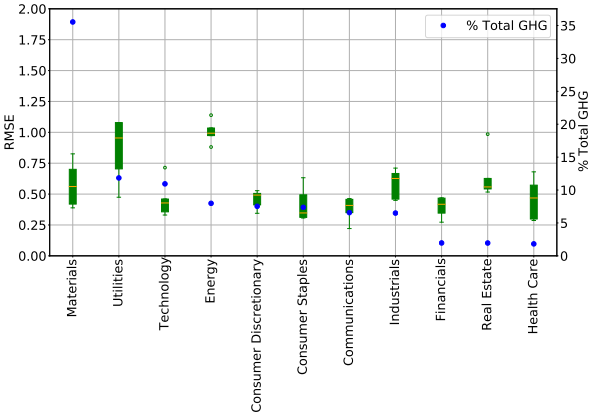
(a) Scope 1 - Test 2020.



(b) Scope 2 - Test 2020.



(c) Scope 1 - Test 2021.



(d) Scope 2 - Test 2021.

Figure 4.36: Distribution of performance of a model trained using a custom loss on five test sets per BICS sector level 1 and ordered by level 1 sectors emissions. The percentage of total GHG represents the percentage of total GHG emissions the sector level 1 represents in the dataset of reporting companies.

considered set for which performance can be driven too much by only one or two very wrong predictions.

Table 4.30 shows the performance of the model in an extrapolation use case. We analyze estimations for year y given by a model trained until year $y-1$. Such estimations are based on features released for the corresponding year y. We show the following results:

- RMSE for estimates on year y using a model trained until year $y-1$.
- RMSE for estimates on year y using a model trained until year $y-1$ for samples corresponding to reporting companies in year $y-1$.
- RMSE for estimates on year y using a model trained until year $y-1$ for samples corresponding to companies that did not report in year $y-1$.

Global performance is similar when using the model in an extrapolation setting one year in the future in comparison to its main use case. However, we observe a large gap in performance between estimations for samples with history included in the training set and completely new samples for which accuracy is quite poor.

This is not restricted to the extrapolation use case. Using the model on samples that do not have a large history in the training set usually leads to smaller performance. Table 4.31 displays the performance of the model trained until year y on the full universe of reporting companies for year y. These estimations are here a mix of estimations obtained with in-sample and out-of-sample data, as a large part of the reported GHG emissions for year y are also present in the training set of the model trained until year y. We show:

- RMSE for estimates on year y using a model trained until year y.
- RMSE for estimates on year y using a model trained until year y for samples corresponding to reporting companies in year $y-1$ and y.
- RMSE for estimates on year y using a model trained until year y for samples corresponding to companies that did not report in year $y-1$ but did in year y.

Similarly, we observe lower accuracy for samples that do not have a history in the training set and only appear in its last year.

	Test set	Mean	Standard Deviation	$N_{samples}$
Scope 1	Full dataset	0.596	0.010	2881
	Historically reporting companies	0.534	0.007	2455
	Companies without history	0.827	0.020	425
Scope 2	Full dataset	0.520	0.009	2836
	Historically reporting companies	0.461	0.014	2393
	Companies without history	0.742	0.033	442

(a) 2020

	Test set	Mean	Standard Deviation	$N_{samples}$
Scope 1	Full dataset	0.583	0.009	2196
	Historically reporting companies	0.532	0.013	1915
	Companies without history	0.852	0.048	281
Scope 2	Full dataset	0.535	0.008	2163
	Historically reporting companies	0.501	0.010	1863
	Companies without history	0.754	0.020	300

(b) 2021

Table 4.30: Performance in RMSE of the GHG estimation model in an extrapolation use case.

	Test set	Mean	Standard Deviation	$N_{samples}$
Scope 1	Full dataset	0.512	0.009	2881
	Historically reporting companies	0.468	0.008	2455
	Companies without history	0.662	0.015	425
Scope 2	Full dataset	0.464	0.004	2836
	Historically reporting companies	0.427	0.010	2393
	Companies without history	0.611	0.024	442

(a) 2020

	Test set	Mean	Standard Deviation	$N_{samples}$
Scope 1	Full dataset	0.510	0.010	2196
	Historically reporting companies	0.489	0.014	1915
	Companies without history	0.659	0.035	281
Scope 2	Full dataset	0.414	0.009	2163
	Historically reporting companies	0.408	0.009	1863
	Companies without history	0.494	0.025	300

(b) 2021

Table 4.31: Performance in RMSE of the GHG estimation model: training and inference on the same year. Results are computed on a mix of in-sample and out-of-sample estimations.

4.11 . Conclusion

GHG emissions reporting and auditing are not yet compulsory for all companies and methodologies of measurement and estimation are not unified. As a result, we propose a machine-learning model to estimate non-reported corporate GHG emissions for scopes 1 and 2. GHG emissions for companies are non-stationary and the quality of reported data can dramatically change from one company to another. Thanks to suitable machine learning methods, the resulting models show good out-of-sample performance when assessed globally as well as good and balanced out-of-sample performance when evaluated per sector, country, and bucket of revenues. We propose a methodology to compare the estimated emissions from our models to those of external providers and find our estimates to be more qualitative. We also achieve better coverage with proposed estimations for a broad universe of companies. We exhibit in Tab. B.1 and B.2 in the Appendices two summary tables summing up these main results obtained in this chapter.

GHG emissions estimations methodologies are often black boxes or undisclosed, which goes against the very idea of transparency through ESG disclosure. The made methodological choices are described extensively in this chapter and the implemented tools based on Shapley values provide information on the role played by each feature in the construction of the final estimations. Many more interpretability elements could have been studied. For instance, the proposed analysis of the relationship between SHAP values and feature values can be done per sector, yielding meaningful results on the behavior of the model per business industry.

The research done in this chapter was iterative: we built successive versions of the model, enlarging its scope and trying to improve its performance and its interpretability. To be selected, an iteration of the model must always provide balanced test results across business sectors and be of higher quality than models from external providers. We proposed fully automated methods to evaluate models without the need for a human analyst to dissect each GHG estimation one by one.

We experimented with several algorithms, from the proposition of a methodology to remove outliers in the training set to an analysis of the impact of missing values in inference on accuracy. Experiments done with the proposed custom loss were not conclusive. However, working with new losses should deserve more attention in the future, for instance in the context of quantile regression (Koenker and Hallock, 2001). The idea of quantile regression goes beyond the work done in section 4.10.8 but bears some similarity. By optimizing a model using the MAE loss, we obtain an estimation of the median of the target. Then, by using different asymmetric losses, variants of the MAE with a weighting for overestimation and underestimation dependent on the quantile we seek to estimate, we obtain an estimation of the quantiles of the target. This is particularly interesting in the context of GHG emissions: our current models estimate the decimal logarithm of GHG emissions and we propose an exact value that is not perfectly accurate. By showing the quantiles instead of this exact value, the user has access to the distri-

bution of the GHG emissions for the considered samples. Another idea to propose such distributions is to use the tree structure of the GBDT algorithm: each of the successive trees is composed of leaves whose values are the average of the targets associated with samples falling inside these leaves. Instead of calculating this average, we can randomly sample values in the leaves at each iteration of the boosting algorithm. Applying the formula (2.20), we obtain a potential estimate. By doing this numerous times, we obtain a distribution of estimates for the considered sample.

The current model is biased toward large companies as they constitute our main source of data. Including more training data from SMEs would improve the coverage of the model and is left as future work. Moreover, as discussed in section 4.4.3, the used industry classification is critical: sometimes companies operating in very different sectors in terms of GHG emissions can be grouped by the used classification. Gathering data on all activities a company reports being active in and including this new and more precise information in the model can help improve its performance.

Finally, some used features for training do not have good coverage in the considered universe of companies. Future work will focus on performance improvement linked with the availability of the reported features. For instance, firms reporting energy consumption or production data without reporting their GHG emissions are the only beneficiaries of these particular features. This situation is not very common as often when a company reports its energy consumption, it also discloses its GHG emissions (at least scopes 1 and 2). We experimented by training models with the energy consumption feature masked for some samples but the results were not conclusive. Deepening this analysis remains as future work.

Conclusion and perspectives

This thesis proposes methodologies and solutions to leverage ESG data and yield meaningful results using machine learning methods. It focuses on the financial industry through two specific case studies.

In chapter 3, we systematically investigate the relationship between price returns and ESG scores in the European equity market. Using interpretable machine learning, we examine whether ESG scores can explain the part of price returns not accounted for by classic equity factors, especially the market one. To address the initial lack of quantity and quality of ESG data, a cross-validation scheme with repeated random sub-sampling company-wise validation is used and allows training and validating models with most of the latest and best data. We find that gradient boosting models successfully explain a part of annual price returns not accounted for by the market factor. Using benchmark features and an adapted ESG dataset, we show that ESG data explain significantly better price returns than basic fundamental features alone, and increasingly so over time. Lastly, we build materiality matrices, showing the most material ESG dimensions per industry and company size. We find that better ESG scores have opposite effects on the price returns of small and large capitalization companies, with better ESG scores generally associated with larger price returns for the latter and reversely for the former.

In chapter 4, we focus on particularly important and used ESG data points, the scope 1 and scope 2 GHG emissions. GHG emissions reporting and auditing are not yet compulsory for all companies, and methodologies of measurement and estimation are not unified. We propose a machine learning-based model to estimate scopes 1 and 2 GHG emissions of companies that have not yet reported them. Our model, thanks to adapted machine learning methods, can overcome the non-stationary nature of GHG emissions data. It proposes estimated emissions for a large universe of companies and shows good out-of-sample global performance, as well as good out-of-sample granular performance when evaluating it by sectors, countries, or buckets of revenues. Comparing the estimations of the model to those of external providers, we find our estimates to be more accurate. The constructed model is fully interpretable thanks to explainability tools based on Shapley values which provide a factor split explaining estimated GHG emissions for every particular company. We further discuss potential improvements, the use cases in which the model performs better and the impact of missing values on the final estimate during inference.

This dissertation shows that by carefully selecting machine learning algorithms and by adapting them to the specificities and challenges of ESG data, it is possible to produce meaningful results. Many ESG data sources are available and should always be carefully reviewed. Questioning the methodologies of ESG data providers

and of companies reporting ESG indicators, the consistency of the data, its external assessment, its scope and available history, are always necessary steps and should lead to the choice of the best possible source for the task at hand. When several qualitative sources are available, directly implementing feature selection processes could be a good idea. Classic cross-validation and testing methodologies are to be adapted as ESG data is quite recent and its quantity and quality have improved over time, leading to non-stationary time series for the different ESG indicators of the different companies. The very concept of ESG relies on a transparency pillar. Used methodologies should then always be reproducible and interpretable. Decisions made by the model should be assessed as well as the different potential use cases to prevent any misuse. In particular, when using models which support missing data, their impact should be evaluated.

This thesis was the opportunity to expose these challenges and to show how to overcome them through the two presented case studies. Limitations of our work for each of them are presented in the conclusion of their respective chapters. In particular, it could be interesting for both to go further in the interpretability of the built models using other tools such as SHAP interaction values, giving the common marginal contribution of two interacting features on the prediction. Testing different datasets and implementing feature selection processes could also enhance this work.

Chapter 1 highlights the vast potential of machine learning in harnessing ESG data and generating meaningful outcomes for the financial sector. Many other important and challenging case studies could have been addressed. Going further on the analysis of ESG data and alpha, we could rely on machine learning and deep learning methods to produce an ESG factor that could then be assessed for redundancy with other more classical equity factors. NLP methods could be leveraged to make ESG scores more reactive to immediate news. Our study of the GHG emissions scopes 1 and 2 could be extended to scope 3, which is much more challenging to estimate and constitutes the primary emission sources of most companies. Predictive models could be built, for instance, to estimate the future values of specific ESG indicators of some companies issuing sustainability-linked bonds, to measure the physical risk exposure of a company and assess the probability and magnitude of potential extreme weather events or even to assess the credibility of reported quantitative pathway of companies to cancel their net emissions.

Bibliography

Fabio Alessandrini and Eric Jondeau. ESG investing: From sin stocks to smart beta. *The Journal of Portfolio Management*, 46(3):75–94, 2020.

David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

Mats Andersson, Patrick Bolton, and Frédéric Samama. Hedging climate risk. *Financial Analysts Journal*, 72(3):13–32, 2016.

Mark Anson, Deborah Spalding, Kristofer Kwait, and John Delano. The sustainability conundrum. *The Journal of Portfolio Management*, 46(4):124–138, 2020.

Jérémi Assael, Laurent Carlier, and Damien Challet. Dissecting the explanatory power of ESG features on equity returns by sector, capitalization, and year with interpretable machine learning. *Journal of Risk and Financial Management*, 16 (3):159, 2023a.

Jérémi Assael, Thibaut Heurtebize, Laurent Carlier, and François Soupé. Greenhouse gases emissions: estimating corporate non-reported emissions using interpretable machine learning. *Sustainability*, 15(4):3391, 2023b.

Jitendra Aswani, Aneesh Raghunandan, and Shivaram Rajgopal. Are carbon emissions associated with stock returns? *Columbia Business School Research Paper Forthcoming*, 2022.

Autorité des Marchés Financiers. Finance durable: Comment donner du sens à son épargne?, 2020.
https://www.amf-france.org/sites/institutionnel/files/private/2020-10/guide-finance-durable-2020-bd-def_0.pdf
Accessed on 13 April 2023.

Steven Bacon and Arnfried Ossen. Smart ESG integration: Factoring in sustainability. *RobecoSam AG*, 2015.

Elizabeth Baldwin, Yongyang Cai, and Karlygash Kuralbayeva. To build or not to build? Capital stocks and climate policy. *Journal of Environmental Economics and Management*, 100:102235, 2020.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Florian Berg, Julian F Koelbel, and Roberto Rigobon. Aggregate confusion: The divergence of ESG ratings. *Review of Finance*, 26(6):1315–1344, 2022.

Christoph Bergmeir and José M Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 2012.

Bloomberg Enterprise Quants. Distributional Greenhouse gas emissions estimates: data challenges and modeling solutions, 2022.

BloombergNEF . Sustainable Debt Issuance Breezed Past $1.6 Trillion in 2021, 2022.
https://about.bnef.com/blog/sustainable-debt-issuance-breezed-past-1-6-trillion-in-2021
Accessed on 30 March 2023.

BNP Paribas. Stress-testing equity portfolios for climate change factors: The carbon factor, 2016.

MA Boermans, RJ Galema, et al. Pension funds carbon footprint and investment trade-offs. *DNB working papers*, (554), 2017.

Patrick Bolton and Marcin Kacperczyk. Global pricing of carbon-transition risk. Technical report, National Bureau of Economic Research, 2021.

Michael Branch, Lisa R Goldberg, and Pete Hand. A guide to ESG portfolio construction. *The Journal of Portfolio Management*, 45(4):61–66, 2019.

André Breedt, Stefano Ciliberti, Stanislao Gualdi, and Philip Seager. Is ESG an Equity Factor or Just an Investment Guide? *The Journal of Investing*, 28(2): 32–42, 2019.

Louison Cahen-Fourot, Emanuele Campiglio, Antoine Godin, Eric Kemp-Benedict, and Stefan Trsek. Capital stranding cascades: The impact of decarbonisation on productive asset utilisation. *Energy Economics*, 103:105581, 2021.

Cambridge Centre for Sustainable Finance. Environmental risk analysis by financial institutions: a review of global practice. *Cambridge Institute for Sustainability Leadership*, 2016.

Michael Cappucci. The ESG integration paradox. *Journal of Applied Corporate Finance*, 30(2):22–28, 2018.

CDP. CDP Full GHG Emissions Dataset – Technical annex III: Statistical Framework., 2020.

Ying Chan, Ked Hogan, Katharina Schwaiger, and Andrew Ang. ESG in Factors. *The Journal of Impact and ESG Investing*, 1(1):26–45, 2020.

Li Chen, Lipei Zhang, Jun Huang, Helu Xiao, and Zhongbao Zhou. Social responsibility portfolio optimization incorporating ESG criteria. *Journal of Management Science and Engineering*, 6(1):75–85, 2021.

Mike Chen and George Mussalli. An integrated approach to quantitative ESG investing. *The Journal of Portfolio Management*, 46(3):65–74, 2020.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Bradford Cornell and Aswath Damodaran. Valuing ESG: Doing good or sounding good? *NYU Stern School of Business*, 2020.

Marielle De Jong and Anne Nguyen. Weathered for climate risk: a bond investment proposition. *Financial Analysts Journal*, 72(3):34–39, 2016.

Angelo Drei, Théo Le Guenedal, Frédéric Lepetit, Vincent Mortier, Thierry Roncalli, and Takaya Sekine. ESG investing in recent years: New insights from old challenges. *Available at SSRN 3683469*, 2019.

Paola D'Orazio and Marco Valente. The role of finance in environmental innovation diffusion: An evolutionary modeling approach. *Journal of Economic Behavior & Organization*, 162:417–439, 2019.

European Commission. A European Green Deal, 2019.
https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_en
Accessed on 30 March 2023.

European Commission. EU Emissions Trading System, 2023a.
https://climate.ec.europa.eu/eu-action/eu-emissions-trading-system-eu-ets_en
Accessed on 30 March 2023.

European Commission. Sustainable Finance: Commission welcomes political agreement on European green bond standard, 2023b.
https://ec.europa.eu/commission/presscorner/detail/en/mex_23_1301
Accessed on 30 March 2023.

European Parliament. Renewed Sustainable Finance Strategy, 2021.
https://www.europarl.europa.eu/legislative-train/theme-a-european-green-deal/file-renewed-sustainable-finance-strategy
Accessed on 30 March 2023.

European Parliament and Council of the European Union. Regulation (EU) 2019/2088 of the European Parliament and of the Council on sustainability-related disclosures in the financial services sector, 2019.
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32019R2088&qid=1680357889022
Accessed on 30 March 2023.

European Parliament and Council of the European Union. Regulation (EU) 2020/852 of the European Parliament and of the Council on the establishment of a framework to facilitate sustainable investment, and amending Regulation (EU) 2019/2088, 2020.
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32020R0852
Accessed on 30 March 2023.

European Parliament and Council of the European Union. Directive (EU) 2022/2464 of the European Parliament and of the Council amending Regulation (EU) No 537/2014, Directive 2004/109/EC, Directive 2006/43/EC and Directive 2013/34/EU, as regards corporate sustainability reporting, 2022.
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022L2464&qid=1680357606687
Accessed on 30 March 2023.

Eugene F Fama and Kenneth R French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33:3–56, 1993.

Eugene F Fama and Kenneth R French. Fama and French Portfolios and Factors Data, 2021.
https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html
Accessed on 22 March 2021.

Gunnar Friede, Timo Busch, and Alexander Bassen. ESG and financial performance: aggregated evidence from more than 2000 empirical studies. *Journal of Sustainable Finance & Investment*, 5(4):210–233, 2015.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Stephan M Gasser, Margarethe Rammerstorfer, and Karl Weinmayer. Markowitz revisited: Social portfolio engineering. *European Journal of Operational Research*, 258(3):1181–1190, 2017.

Guido Giese and Linda-Eling Lee. Weighing the evidence: ESG and equity returns. *MSCI Research Insight*, 2019.

Bernhard Goldhammer, Christian Busse, and Timo Busch. Estimating Corporate Carbon Footprints with Externally Available Data. *Journal of Industrial Ecology*, 21(5):1165–1179, 2017.

Régis Gourdel, Irene Monasterolo, Nepomuk Dunz, Andrea Mazzocchetti, and Laura Parisi. The double materiality of climate physical and transition risks in the euro area. 2022.

Paul A Griffin, David H Lont, and Estelle Y Sun. The relevance to investors of greenhouse gas emission disclosures. *Contemporary Accounting Research*, 34 (2):1265–1297, 2017.

Douglas M Grim and Daniel B Berkowitz. ESG, SRI, and impact investing: A primer for decision-making. *The Journal of Impact and ESG Investing*, 1(1): 47–65, 2020.

Tian Guo, Nicolas Jamet, Valentin Betrix, Louis-Alexandre Piquet, and Emmanuel Hauptmann. ESG2risk: A deep learning framework from ESG news to stock volatility prediction. *arXiv preprint arXiv:2005.02527*, 2020.

James Hansen, Larissa Nazarenko, Reto Ruedy, Makiko Sato, Josh Willis, Anthony Del Genio, Dorothy Koch, Andrew Lacis, Ken Lo, Surabi Menon, et al. Earth's energy imbalance: Confirmation and implications. *science*, 308(5727):1431–1435, 2005.

Thibaut Heurtebize, Frederic Chen, François Soupé, and Raul Leote de Carvalho. Corporate Carbon Footprint: A Machine Learning Predictive Model for Unreported Data. *The Journal of Impact and ESG Investing*, 3(2):36–54, 2022.

Adolfo Hilario-Caballero, Ana Garcia-Bernabeu, Jose V Salcedo, and Marisa Vercher. Tri-criterion model for constructing low-carbon mutual fund portfolios: a preference-based multi-objective genetic algorithm approach. *International Journal of Environmental Research and Public Health*, 17(17):6324, 2020.

Maxence Jeunesse, Guillaume Chevalier, and Thomas Roulland. NLP-enabled impact. *Axa Investment Managers*, 2020.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154, 2017.

Roger Koenker and Kevin F Hallock. Quantile regression. *Journal of Economic Perspectives*, 15(4):143–156, 2001.

Marlene H Konqui, François Millet, and Serge Darolles. Why using ESG helps you build better porfolios. *Lyxor ETF Research Insights*, 2019.

Sakis Kotsantonis and George Serafeim. Four things no one will tell you about ESG data. *Journal of Applied Corporate Finance*, 31(2):50–58, 2019.

Ook Lee, Hanseon Joo, Hayoung Choi, and Minjong Cheon. Proposing an integrated approach to analyzing ESG data via machine learning and deep learning algorithms. *Sustainability*, 14(14):8745, 2022.

Stan Lipovetsky and Michael Conklin. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330, 2001.

Alexandra Luccioni, Emily Baylor, and Nicolas Duchene. Analyzing sustainability reports using natural language processing. *arXiv preprint arXiv:2011.08073*, 2020.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777, 2017.

Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

Vincent Margot, Christophe Geissler, Carmine de Franco, Bruno Monnier, et al. ESG Investments: Filtering versus Machine Learning Approaches. *Applied Economics and Finance*, 8(2):1–16, 2021.

Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

MSCI ESG Research. MSCI ESG Ratings Methodology, 2020.
https://www.msci.com/documents/1296102/0/MSCI+ESG+Ratings+Methodology+-+Exec+Summary+Dec+2020.pdf/9c54871f-361d-e1ff-adc7-dfdee299dfb3?t=1607501860114
Accessed on 24 January 2023.

Quyen Nguyen, Ivan Diaz-Rainey, and Duminda Kuruppuarachchi. Predicting corporate carbon footprints for climate finance risk analyses: a machine learning approach. *Energy Economics*, 95:105129, 2021.

Tim Nugent, Nicole Stelea, and Jochen L Leidner. Detecting ESG topics using domain-specific language models and data augmentation approaches. *arXiv preprint arXiv:2010.08319*, 2020.

OCDE. *OECD Guidelines for Multinational Enterprises, 2011 Edition*. OECD Publishing, 2011.

L'uboš Pástor, Robert F Stambaugh, and Lucian A Taylor. Dissecting green returns. *Journal of Financial Economics*, 146(2):403–424, 2022.

PCAF. The Global GHG Accounting and Reporting Standard Part A: Financed Emissions. Second Edition., 2022.

Ariel Pinchot, Lihuan Zhou, Giulia Christianson, Jack McClamrock, Ichiro Sato, et al. Assessing physical risks from climate change: do companies and financial organizations have sufficient guidance. *World Resources Institute*, 2021.

Jan-Carl Plagge and Douglas M Grim. Have investors paid a performance Price? Examining the behavior of ESG equity funds. *The Journal of Portfolio Management*, 46(3):123–140, 2020.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31:6638–6648, 2018.

Javier Lopez Prol and Kiwoong Kim. Risk-return performance of optimized ESG equity portfolios in the NYSE. *Finance Research Letters*, 50:103312, 2022.

H.-O. Pörtner, D.C. Roberts, H. Adams, I. Adelekan, C. Adler, R. Adrian, P. Aldunce, E. Ali, R. Ara Begum, B. Bednar Friedl, R. Bezner Kerr, R. Biesbroek, J. Birkmann, K. Bowen, M.A. Caretta, J. Carnicer, E. Castellanos, T.S. Cheong, W. Chow, G. Cissé G. Cissé, and Z. Zaiton Ibrahim. *Climate Change 2022: Impacts, Adaptation and Vulnerability*. Technical Summary. Cambridge University Press, Cambridge, UK and New York, USA, 2022. ISBN 9781009325844.

Janet Ranganathan, Laurent Corbier, Pankaj Bhatia, Simon Schmitz, Peter Gage, and Kjell Oren. *The Greenhouse Gas Protocol: a Corporate Accounting and Reporting Standard, Revised Edition*. World Business Council for Sustainable Development and World Resources Institute, 2015.

Refinitiv. Environmental, Social and Governance (ESG) Scores Methodology, 2020.
https://www.refinitiv.com/content/dam/marketing/en_us/documents/methodology/esg-scores-methodology.pdf
Accessed on 22 March 2021.

Refinitiv. Refinitiv ESG Carbon Data and Estimate Models, 2023.
https://www.refinitiv.com/content/dam/marketing/en_us/documents/fact-sheets/esg-carbon-data-estimate-models-fact-sheet.pdf
Accessed on 16 May 2022.

Anthony A Renshaw. ESG's evolving performance: First, do no harm. Frankfurt: Axioma, 2018.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Julie Rozenberg, Adrien Vogt-Schilb, and Stephane Hallegatte. Instrument choice and stranded assets in the transition to clean capital. *Journal of Environmental Economics and Management*, 100:102183, 2020.

Marc Schmitt. Deep Learning vs. Gradient Boosting: Benchmarking state-of-the-art machine learning algorithms for credit scoring. *arXiv preprint arXiv:2205.10535*, 2022.

David Schofield, Adam Craig, and Richard Yasenchak. What to look for on the road to ESG. *Intech - Janus Henderson*, 2019.

Securities and Exchange Commission. SEC Proposes Rules to Enhance and Standardize Climate-Related Disclosures for Investors, 2022.
https://www.sec.gov/news/press-release/2022-46
Accessed on 30 March 2023.

Manish Shakdwipee and Linda-Eling Lee. Filling The Blanks: Comparing Carbon Estimates Again Disclosures. *MSCI ESG Research Issue Brief*, 2016.

Llyod S Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41: 647–665, 2014.

Mukund Sundararajan and Amir Najmi. The many Shapley values for model explanation. In *International conference on machine learning*, pages 9269–9278. PMLR, 2020.

Robert Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Blaine Townsend. From SRI to ESG: The Origins of Socially Responsible and Sustainable Investing. *The Journal of Impact and ESG Investing*, 1(1):10–25, 2020.

Tim Verheyden, Robert G Eccles, and Andreas Feiner. ESG for all? The impact of ESG screening on return, risk, and diversification. *Journal of Applied Corporate Finance*, 28(2):47–55, 2016.

Mathis Wackernagel and William Rees. *Our ecological footprint: reducing human impact on the earth*, volume 9. New society publishers, 1998.

Thomas Wiedmann. Carbon footprint and input-output analysis - An introduction. *Economic Systems Research*, 21(3):175–186, 2009.

Thomas O Wiedmann, Manfred Lenzen, and John R Barrett. Companies on the scale: Comparing and benchmarking the sustainability performance of businesses. *Journal of Industrial Ecology*, 13(3):361–383, 2009.

List of Tables

List of Figures

Acronyms

APexJ	Asia-Pacific excluding Japan
Bal_Acc	Balanced Accuracy
BICS	Bloomberg Industry Classification Standard
BICS Sector L1	First level of granularity in BICS classification
BICS Sector L2	Second level of granularity in BICS classification
BICS Sector L3	Third level of granularity in BICS classification
BICS Sector L4	Fourth level of granularity in BICS classification
CAPM	Capital Asset Pricing Model
CCS	Carbon Capture and Storage
CDP	Carbon Disclosure Project
CO_2	Carbon Dioxide
CO_2-eq	Carbon Dioxide Equivalent
COP	Conferences Of the Parties
CSR	Corporate Social Responsibility
CSRD	Corporate Sustainability Reporting Directive
EIO	Environmental Input-Output
ESG	Environment, Social and Governance
ETS	Emission Trading System
EU	European Union
EuGBS	European Green Bond Standards
EUR	Europe
EVIC	Enterprise Value Including Cash
FDR	False Discovery Rate
FN	False Negatives
FP	False Positives
GBDT	Gradient Boosted Decision Trees
GGLR	Gamma Generalized Linear Regression
GHG	Greenhouse Gas
GHGv1	GHG emissions estimation model, first iteration
GHGv2	GHG emissions estimation model, second iteration
GHGv3	GHG emissions estimation model, third iteration
GICS	Global Industry Classification Standard
GPPE	Gross Property Plant & Equipment
GWP	Global Warming Potential
HML	High Minus Low, value Fama-French factor
IEA	International Energy Agency
IPCC	Intergovernmental Panel on Climate Change
JPN	Japan
KPI	Key Indicator of Performance
LIME	Local Interpretable Model-Agnostic Explanations

MAE	Mean-Absolute Error
MMD	Maximum Mean Discrepancy
MSE	Mean-Square Error
NAM	North America
NFRD	Non-Financial Reporting Directive
NLP	Natural Language Processing
NPPE	Net Property Plant & Equipment
NZBA	Net Zero Banking Alliances
OECD	Organisation for Economic Co-operation and Development
OLS	Ordinary Least Squares
PA	Process Analysis
PACTA	Paris Agreement Capital Transition Assessment
PAI	Principle Adverse Impact
PCAF	Partnership for Carbon Accounting Financials
RMSE	Root-Mean-Square Error
SEC	Securities and Exchange Commission
SFDR	Sustainable Finance Disclosure Regulation
SHAP	SHapley Additive exPlanations
SIC	Standard Industrial Classification
SLB	Sustainability Linked Bonds
SMB	Small Minus Big, size Fama-French factor
SMEs	Small and Medium-sized Enterprises
TN	True Negatives
TP	True Positives
TRBC	The Refinitiv Business Classification
VCS	Verified Carbon Standards
VCU	Verified Carbon Units
WBCSD	World Business Council for Sustainable Development
WRI	World Resources Institute

Appendices

A - Summary tables: Dissecting the explanatory power of ESG features on equity returns by sector, capitalization, and year with interpretable machine learning

A.1 . Results: dependence measures between the loss metrics in the validation and test sets

Table A.1 summarizes the obtained results in different settings: target derived from the CAPM model or the Fama-French 3-factor model, Refinitiv or MSCI data, classification or regression, explanation or prediction. In particular, we show the Person Correlation and Kendall Tau measures between the loss metrics in the validation and test sets for the 100 best models of the random hyperparameters search. The considered loss metric is the cross-entropy loss in classification settings and the MSE in regression settings.

A.2 . Results: performance measures

Table A.2 exhibits the performance on the test set in different settings: target derived from the CAPM model or the Fama-French 3-factor model, Refinitiv or MSCI data, classification or regression, explanation or prediction. We show the overperformance or underperformance of the model trained on benchmark and ESG features in comparison to the model trained only on benchmark features. The used benchmark features are defined in section 3.4.2.

Table A.2a shows the difference between the median cross-entropy (in a classification setting) or RMSE (in a regression setting) between a model trained only on benchmark features and a model trained on both ESG and benchmark features. Table A.2b shows the difference between the median balanced accuracy between a model trained on both ESG and benchmark features and a model trained only on benchmark features. The used notion of balanced accuracy in a regression setting is defined in section 2.1.4.

Medians are obtained by computing the median of the performance measures computed for different ensembles of models trained on different samples of the training data.

Year	**Refinitiv - Classification** Explanation - CAPM	**Refinitiv - Classification** Explanation - FF3	**MSCI - Classification** Explanation - CAPM	**Refinitiv - Regression Full** Explanation - CAPM	**Refinitiv - Regression Filtered** Explanation - CAPM	**Refinitiv - Classification** Prediction - CAPM
2016	-0.54	-0.23	-0.59	-0.48	-0.63	-0.47
2017	0.14	-0.054	0.29	0.10	-0.018	-0.54
2018	0.47	0.29	0.17	-0.38	0.073	0.39
2019	0.73	0.67	0.52	0.43	0.82	-0.025
2020	0.27	0.053	-0.010	0.44	0.59	/

(a) Pearson Correlation.

Year	**Refinitiv - Classification** Explanation - CAPM	**Refinitiv - Classification** Explanation - FF3	**MSCI - Classification** Explanation - CAPM	**Refinitiv - Regression Full** Explanation - CAPM	**Refinitiv - Regression Filtered** Explanation - CAPM	**Refinitiv - Classification** Prediction - CAPM
2016	-0.36	-0.14	-0.41	-0.28	-0.46	-0.33
2017	0.12	0.010	0.20	0.048	-0.022	-0.37
2018	0.30	0.19	0.083	-0.25	-0.050	0.26
2019	0.58	0.49	0.43	0.29	0.63	-0.021
2020	0.19	0.017	-0.17	0.30	0.44	/

(b) Kendall Tau.

Table A.1: Summary table. Dependence measures between the loss metrics (cross-entropy or MSE) in the validation and test sets, for the 100 best models of the random hyperparameters search, in different settings.

Year	**Refinitiv - Classification** Explanation - CAPM	**Refinitiv - Classification** Explanation - FF3	**MSCI - Classification** Explanation - CAPM	**Refinitiv - Regression Full** Explanation - CAPM	**Refinitiv - Regression Filtered** Explanation - CAPM	**Refinitiv - Classification** Prediction - CAPM
2016	-2.2	-0.92	-12.8	-0.35	-0.71	-1.4
2017	-0.38	-0.98	1.9	0.30	0.18	-0.71
2018	-0.098	-0.73	2.5	0.12	0.13	-0.067
2019	0.42	0.48	2.1	0.28	0.29	-0.46
2020	1.9	0.87	0.49	0.97	0.85	/

(a) Cross-entropy in classification and RMSE in regression.

Year	**Refinitiv - Classification** Explanation - CAPM	**Refinitiv - Classification** Explanation - FF3	**MSCI - Classification** Explanation - CAPM	**Refinitiv - Regression Full** Explanation - CAPM	**Refinitiv - Regression Filtered** Explanation - CAPM	**Refinitiv - Classification** Prediction - CAPM
2016	-1.4	-1.9	-1.4	-2.4	-1.9	0.034
2017	-0.52	0.18	4.4	0.26	0.66	-1.5
2018	0.42	0.012	0.58	0.47	1.0	-0.63
2019	2.3	2.3	5.5	3.2	2.8	0.25
2020	2.3	-0.78	2.4	2.5	2.6	/

(b) Balanced Accuracy.

Table A.2: Summary table. Performance in percent on the test set in different settings: overperformance or underperformance of the model trained on benchmark and ESG features in comparison to the model trained only on benchmark features.

B - Summary tables: Greenhouse gas emissions: estimating corporate non-reported emissions using interpretable machine learning

B.1 . Results: evaluating the performance of the models

Table B.1 displays the results of the benchmark model developed in section 4.10.1, as well as the results from the GHGv1, GHGv2 and GHGv3 models on five different test sets, built as proposed in 4.4.5. For the GHGv2 model, we propose results on test sets including samples without BICS information and on test sets excluding samples without BICS information. For the GHGv3 model, we exhibit results on test sets including samples without revenues information and on test sets excluding samples without revenues information.

B.2 . Results: comparison of estimates with other providers

Table B.2 summarizes the results of the comparison between the quality of estimates deriving from our designed models (GHGv1, GHGv2 and GHGv3) and estimates from external providers (Bloomberg, CDP, MSCI, Sustainalytics), using the methodology described in section 4.6.3. For the GHGv2 model, we propose results on sets including samples without BICS information and on sets excluding samples without BICS information. For the GHGv3 model, we exhibit results on sets including samples without revenues information and on sets excluding samples without revenues information.

These results are obtained using subsamples of estimates for companies covered by the two assessed models: our model and a model from an external provider. The exhibited results are the difference between the RMSE computed using the 2020 ground truth and 2019-2018 estimates for the considered external provider and the RMSE computed using the 2020 ground truth and 2019-2018 estimates for our selected GHG model. Thus, a positive number means that our model overperformed compared to the considered external provider.

Metric	**Benchmark**	**GHGv1**	**GHGv2** With missing BICS	**GHGv2** Without missing BICS	**GHGv3** With missing revenues	**GHGv3** Without missing revenues
R^2	0.491	0.832	0.827	0.833	0.803	0.825
RMSE	1.02	0.578	0.588	0.579	0.631	0.593
MAE	0.767	0.401	0.396	0.389	0.418	0.396

(a) Scope 1.

Metric	**Benchmark**	**GHGv1**	**GHGv2** With missing BICS	**GHGv2** Without missing BICS	**GHGv3** With missing revenues	**GHGv3** Without missing revenues
R^2	0.164	0.746	0.750	0.752	0.735	0.756
RMSE	1.00	0.522	0.548	0.539	0.573	0.548
MAE	0.750	0.341	0.345	0.338	0.366	0.350

(b) Scope 2.

Table B.1: Summary table. Results of the different models on the five different test sets of test year 2020: mean of the R^2, RMSE and MAE metrics.

Provider	**GHGv1**	**GHGv2** With missing BICS	**GHGv2** Without missing BICS	**GHGv3** With missing revenues	**GHGv3** Without missing revenues
Bloomberg	0.13	0.076	0.081	0.086	0.097
CDP	0.38	0.34	0.34	0.28	0.35
MSCI	0.020	0.0042	0.034	-0.012	0.019
Trucost	0.23	0.18	0.19	0.14	0.19

(a) Scope 1.

Metric	**GHGv1**	**GHGv2** With missing BICS	**GHGv2** Without missing BICS	**GHGv3** With missing revenues	**GHGv3** Without missing revenues
Bloomberg	0.074	0.093	0.080	0.083	0.087
CDP	0.29	0.34	0.33	0.30	0.34
MSCI	0.073	0.13	0.13	0.11	0.13
Trucost	0.16	0.15	0.15	0.13	0.15

(b) Scope 2.

Table B.2: Summary table. Last GHG estimates from providers (2019 – 2018) compared to 2020 ground truth for companies that started reporting in 2020: overperformance of our selected model in comparison to external providers, considering only companies covered by both our selected GHG model and the considered provider.