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

Jeremi Assael, Thibaut Heurtebize, Laurent Carlier, François Soupé

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

Jérémi Assael, [1,2], Thibaut Heurtebize[3], Laurent Carlier[1], and François Soupé[3]

[1]BNP Paribas Corporate & Institutional Banking, Global Markets Data & Artificial Intelligence Lab, Paris, France
[2]Chair of Quantitative Finance, MICS Laboratory, CentraleSupélec, Université Paris-Saclay, Gif-sur-Yvette, France
[3]BNP Paribas Asset Management, Quantitative Research Group, Research Lab, Paris, France

Abstract

As of 2022, greenhouse gases (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 scope 1 and scope 2 GHG emissions of companies not reporting them yet. Our model, designed to be transparent and completely adapted to this use case, is able to estimate emissions for a large universe of companies. It shows good out-of-sample global performances as well as good out-of-sample granular performances when evaluating it by sectors, countries, or revenue buckets. We also compare the model results to those of other providers and find our estimates to be more accurate. Explainability tools based on Shapley values allow the constructed model to be fully interpretable, the user being able to understand which factors split explains the GHG emissions for each particular company.

Keywords sustainability; disclosure; greenhouse gas emissions; machine learning; interpretability; carbon emissions; scope 1; scope 2;

JEL Classification C51; C52; C55; G17; G18; Q51; Q52; Q54;

1 Introduction

Past human activities or "footprints" are now 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, according to Wackernagel and Rees (1998). When it comes to an air emissions footprint, the greenhouse gases (GHG) emissions are the most widely analyzed as they allow the calculation of radiative forcing. When this radiative forcing is positive, the Earth system captures more energy than it radiates to space: it is a common measure for the global warming of the Earth (Hansen et al., 2005). The calculation of this carbon

footprint tends to account for all GHG emissions caused by an individual, event, organization, service, place, or product, and is expressed in units of carbon dioxide equivalent (CO_2-eq).

The annual meetings of the United Nations Climate Change Conference at the World Conferences of the Parties (COP) allow for a review of 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 the country's emissions are the sum of emissions coming from its inhabitants and its companies.

As such, 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 like 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 Corporate Sustainability Reporting Directive (CSRD) coming into force from 2024 for the largest companies to 2026 for Small and Medium-sized Enterprises (SME) in the European Union (EU). This directive will also apply to non-European companies, making over 150 million euros of turnover in Europe, according to the Council of the European Union. Companies will need to report audited GHG emissions as well as a quantitative pathway and remediation plan to cancel their net emissions.

Overall, these GHG emissions assessments measure exposure to transition risk 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. Nevertheless, as soon as financial institutions aggregate GHG emissions at the portfolio level for several companies, they need homogeneous methodologies. At this stage, company reporting of GHG emissions is either voluntary or mandatory depending on location and is linked to defined nomenclatures (mostly activity types and size of companies). As explained previously, 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 create biases. Moreover, not only may calculation methodologies vary, but they are also mainly not documented in the reports.

In the Global Warming Potential (GWP) framework, for any gas, CO_2-eq is calculated as the mass of CO_2, which would warm the earth as much as the mass of that gas: it provides a common scale for measuring the climate effects and global warming impacts of different gases. In practice, measuring the GHG emissions of a stakeholder requires much more information depending on how the GWP is released. To standardize these methodologies of calculation, the GHG Protocol, first published in 2001 (Ranganathan et al., 2015), is used by large companies, by 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, it has become the most widely used methodology in the world when it comes to assessing GHG emissions. The carbon 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 companies' total GHG emissions for some business industries like automakers or financial institutions. In practice, making the methods for calculating emissions in a given industry converge makes it easier not only to model but also to compare the emissions of each company with those of its peers.

To guarantee data quality of companies' reported GHG emissions, 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 Corporate Social Responsibility (CSR) reports, are more and more involved increasing convergence of methodologies and controls.

In this study, the methodology is limited to scope 1 and scope 2. Regarding scope 3 emissions, some framework like the Partnership for Carbon Accounting Financials (PCAF), officially recognized by the GHG protocol, allows measuring scope 1 and scope 2 emissions of a financial institution using reported emissions of investments sources but also estimates of scope 1 and 2, as stated in PCAF (2022).

Overall GHG emissions from large firms in developed countries follow a common methodology for calculating scope 1 and 2 emissions: results are either published, validated, or both by independent bodies, such as external auditors, the CDP, or both. In 2021, this was the case for more than 4000 companies worldwide, as observed in this study. For a typical investment universe of 15,000 companies, this means that about 11,000 companies (73%) were not reporting their scope 1 and 2 GHG emissions. This breadth 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. At the same time, even some recent studies like Bolton and Kacperczyk (2021) analyze GHG emissions of 14,468 companies, including 98% of publicly listed companies, without mentioning or analyzing that 80% of the data used is coming from GHG modeled estimates from the data provider Trucost. They even construct a regression model to fit all the scopes 1, 2, and 3 data and draw conclusions on global carbon premiums in the market. On the opposite, some studies using the same Trucost dataset specify and analyze more deeply the underlying quality of the GHG emissions data used (Aswani et al., 2022).

That is why it is so important to analyze in detail the corporate GHG emissions data: operational scopes (accounting consolidation scope, some of biggest factories), standards of calculations (GHG protocol or others), calculation basis (scope 2 Market-based versus Location-based) and this, even if the data is modeled (simple derivation from a previous year to more complex non-linear models). When it comes to comparing corporations across geographies and sectors or drawing conclusions at the global level for anthropogenic GHG emissions, we need a fair assessment of the GHG emissions at country, corporation, factory, and personal levels.

This study focuses narrowly on the unreported estimated emissions of companies. The model framework focuses on estimating the targeted high-quality GHG emissions, especially waiting for international regulatory bodies to bring a homogeneous framework for corporations to report their GHG emissions in extra financial statements. For financial use cases needing GHG emissions for portfolio construction, there is also a distinction to be made between point-in-time estimates using only information available at the date of the estimated emission and "as-of-today" estimates using all available information, including those posterior to the date of the estimated emissions. These two types of estimates answer different use cases and require different calibration strategies.

2 Literature Review—Hypothesis

Focusing on scopes 1 and 2, the available reported data is typically issued from voluntary reporting based on the CDP or on extra financial reports (CSR reports) from companies. With a few exceptions, like France with Article 173 of the French Energy Transition law, GHG emissions reporting is not yet mandatory, but the corporate regulatory framework to report GHG emissions was improved recently with the CSRD in the EU or the Securities and Exchange Commission (SEC) proposed rules in the United States (Securities and Exchange Commission , SEC).

Corporate GHG emissions models make the link between the industrial processes of each business model and the carbon 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, neither the information required to quantify companies' use of those processes nor their intensity in the overall annual production chain, is publicly available. Linking detailed industrial processes and technologies with an accounting of GHG emissions is a perilous task, even when it is handled by big corporate sustainability expert teams or by CDP experts.

To mitigate such a lack of data, financial data vendors rely on relatively simple models to estimate GHG emissions for some companies that do not currently report. These estimates are usually sector-level extrapolations based on indicators, such as the number of employees and income generated, or both. Sector averages or regression models constructed from the existing reported GHG emissions data from peer companies have the advantage of simplicity for explainability, but the number of regressors is usually limited, as are the sample sizes. Model validation tends to rely on the quality of the regression in-samples where data is available.

Data providers, such as Bloomberg (Quants, 2022), MSCI ESG (Shakdwipee and Lee, 2016; Andersson et al., 2016; De Jong and Nguyen, 2016), Refinitiv ESG - previously known as Thomson Reuters ESG - (Refinitiv, 2023; Paribas, 2016; Boermans et al., 2017), S&P Global Trucost and CDP, use models to estimate the GHG emissions of companies that fail to publish emissions data. Such models rely mainly on rules of proportionality between emissions and the size of the company operations or, more recently, on more complex approaches using non-linear models. The simple models tend to use historical data available for the industry as a basis for the calculation, and focus on predicting the logarithm of GHG emissions. Occasionally, they also use energy-specific metrics like GHG intensity per the company's energy consumption and production or 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 for most samples when evaluating the logarithm of the emissions. To be noted, these performance levels are tested in-the-sample, meaning the R^2 computed with the logarithm of the

GHG emissions is tested with the data used to calibrate the model. The performance of the model is calculated on the same companies used for the calculation of the regressions levels. On the other hand, out-of-sample performance tests require a completely new dataset to test the model on unseen companies with reported emissions. Good out-of-sample performance shows that the model avoids overfitting and is able to generalize well.

Some more advanced models described in Goldhammer et al. (2017), Griffin et al. (2017) and CDP (2020) proposed the use of Ordinary Least Squares (OLS) and Gamma Generalized Linear Regression (GGLR) with a broader dataset of publicly available company data for the construction of models. Such models go beyond using just simple factors and rely more 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-the-sample R^2 computed with the logarithm of the GHG emissions around 80%.

More recently, two studies proposed the use of statistical learning techniques to develop models for predicting corporate GHG emissions from publicly available data. These machine-learning approaches take the form of:

- In Nguyen et al. (2021), a meta-learner relying on the optimal set of predictors combining OLS, Ridge regression, Lasso 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 predict reported emissions that were not used to construct the model. Nevertheless, the strongest predictive efficiency of the model was found for predicting aggregated direct and indirect emission scopes as opposed to predicting each of them 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. Moreover, extrapolating these results to a typical wider investment universe is difficult since their used GHG emissions dataset is small, with around 2300 reporting firms against 4300 in this study. The paper also lacks discussions on the achievable coverage of GHG emissions estimates and on the interpretability of the model, with no explanation of why it outputs such estimates.

- In Quants (2022), amortized inference with Gradient Boosted Decision Trees (GBDT) models (Friedman, 2001), re-calibrated using Conditional Mixture of Gammas and Mean Maximum Discrepancy (MMD)-based patterned dropout for regularization. The model is trained on hundreds of features, including Environment, Social, and Governance (ESG) data, fundamental data, and industry segmentation data. The GBDT allows for non-linear patterns to be found even if not all data 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. That paper especially lacks transparency, with several implementation elements like the choice of features not explained, making it not reproducible. That paper also lacks a discussion on the interpretability

of the designed model.

Understanding the risks and opportunities arising from the GHG emissions of companies requires good financial and non-financial data. In some countries and for some companies, as long as GHG emission reporting and auditing is not compulsory, the only viable alternative is to predict non-reported company emissions relying on estimation models. From the above, the current state-of-the-art does not yet provide good enough models for the task at hand. In our view, the quality of data made available by the specialized data vendors is not yet sufficient. Understanding the reasons behind this problem and being able to propose alternative approaches that can lead to better models and more accurate predictions of unreported data is thus of great importance.

The recently proposed approaches by Nguyen et al. (2021) and Quants (2022) based on statistical learning, offer a promising starting point. The central challenge with such statistical learning approaches is to strike the right balance between increasing both the model complexity and accuracy while limiting the risk of overfitting. In this paper, we propose a statistical learning model to predict unreported scope 1 and scope 2 company emissions in an investment universe of about 50,000 companies, of which only about 4000 companies actually report. This model is inspired by the work of Heurtebize et al. (2022) and aims at achieving the following qualities:

- accuracy, globally and by granular sub-sectors, with good and balanced performances on each sub-sector's on point-in-time estimates.
- operability, transparency of the methodology, and reproducibility of results, keeping the complexity of the model to a minimum while achieving good global and granular performances. For example, all data preprocessing steps must be fully automated with no manual corrections. The model of this study is flexible and easily allows 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, including small ones.
- interpretability, a regulatory requirement as highlighted by Heurtebize et al. (2022). The model of this study provides clear and exhaustive statistical explanations of the outputs.

To succeed, we made some significant choices in departing from existing approaches. First, models are always tested on data samples never seen during the calibration, so that their generalization abilities can truly be measured, which was not done in Heurtebize et al. (2022). The second important decision was to always evaluate the model globally and by granular sub-sectors, countries, and buckets of revenues. Obtained estimates are also compared to the ones from other providers through a detailed methodology. To our knowledge, this paper is the first to propose evaluating the model at such levels of granularity. The third important decision was to keep the raw dataset from data providers with a totally documented and fully automated data preprocessing and no manual corrections. Even if the use of incorrect data can reduce the accuracy of models, it allows for full reproducibility and industrialization, as this latter brings some operational constraints to produce automated updates of the model. We also introduce shortly an automated data polishing process at the end of this study. The fourth important decision was to keep model complexity to a minimum by relying only on a fixed small set of predictive features and the most accurate non-linear machine learning approaches without losing interpretability. As a matter of fact, the last important decision was to make a fully interpretable model using a model-agnostic method so that interpretability does not come at the expense of performance.

Indeed, for instance, the implementation from Heurtebize et al. (2022) required keeping a linear layer in the model, degrading accuracy. To our knowledge, such an extensive part on the interpretability of a machine learning model estimating GHG emissions has not been done in the current literature and allows us to understand why the model produced such emissions values.

In the remainder of the paper, we describe the data retained to calibrate and evaluate our model and present the designed methodology in-depth, insisting on the particular implementation choices made, necessary to apply it to the use case of estimating GHG scope 1 and 2 emissions. We then discuss the results associated with this methodology both by comparing our estimates to true GHG emissions reported by companies and by comparing our estimates to the ones from other providers. Finally, we provide tools to understand how the constructed model works and why it estimates such values of GHG emissions.

3 Datasets

An important variety of data sources is 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 Table 1a, selecting yearly data. Such indicators allow one to get a sense of the company profitability, assets size, assets location, and how assets are used.

The second set of indicators is the regional ones, also selected each year, and presented in Table 1b. They provide information on the environment the company is incorporated in.

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 for 2020 alone.

4 Methods

4.1 Problem settings

The goal of this study is to develop a data-driven model estimating scope 1 and scope 2 greenhouse gas emissions of companies that have never reported them. Using the vast amount of available indicators, whose selected ones have been exhibited in Section 3, we build a high-quality dataset and calibrate a machine learning model that outputs the estimated emission of a company. This automated method allows the estimation of the emissions of any company as long as enough financial and non-financial data is available. Scope 1 and scope 2 emissions are estimated through two separate models.

This is a regression setting: the model learns for each possible couple (i, t) the reported emission $Y_{i,t}$ from a set of P potentially explanatory factors called features. Here, i represents a company, and t is the year of sampling of the features and emissions. Let us relabel all the couples (i, t) by the index $n \in \{1, \cdots, N\}$. This regression problem consists in estimating Y_n, the reported emission, with a vector X_n with P components, or equivalently, to explain the vector $Y \in \mathbb{R}^N$ from the lines of matrix $X \in \mathbb{R}^{N \times P}$. Y is called the target and X the features matrix.

Optimizing a machine learning model supposes the division of the full dataset into three parts called the training, validation, and test sets:

- The training set is used to optimize the parameters of the machine learning model on a set of features associated with their GHG emission. Practically, it learns the mapping between the lines of X and the components of the vector Y.

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

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

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

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

Table 1: Data sources and indicators used in the model.

- The validation set allows optimizing the model hyperparameters.
- The test set is used to evaluate the generalization capacities of the model on data samples never seen in training or validation. In inference, the model takes as input a vector of features and outputs the estimated GHG emission.

The state of the art for regression problems on tabular data like this one is provided by Gradient Boosting models (Friedman, 2001), as shown for instance in Shwartz-Ziv and Armon (2022). Gradient boosting consists in using a sequence of weak learners, making wrong predictions, that iteratively correct the mistakes of the previous ones, eventually yielding a strong learner, making good predictions. We use here decision trees as weak learners: we are using the GBDT algorithm. Different implementations of the GBDT method have been proposed, e.g., XGBoost (Chen and Guestrin, 2016), LightGBM (Ke et al., 2017), and CatBoost (Prokhorenkova et al., 2018). We use LightGBM. The advantage of such methods with respect to linear regression is that they are able to learn more generic functional forms.

The model is trained to minimize the mean-squared error (cost function), also referred to in this paper as MSE, defined as

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|^2, \tag{1}$$

where $\hat{y}_i$ is the predicted output from the model (decimal logarithm of the GHG estimation, as explained in Section 4.2.2) and y_i is the ground truth (decimal logarithm of the reported emission).

4.2 Target computation

4.2.1 Raw target obtention

The explained variable, 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 quality level. 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 company's extra-financial communication.

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. Their reported emissions are expressed in tCO_2-eq.

4.2.2 Target cleaning procedure

GHG emissions are reported on different dates during the year. To unify samples and preserve meaning with the used training features, 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 per year remains.

Variability is an important characteristic of GHG emissions data, leading sometimes to inconsistencies, with important changes in emissions for a company over the years: this could be due to changes in the reporting methodology, to a corporate action like the acquisition of a subsidiary or mergers. The chosen cleaning procedures mitigated these issues: a fully automated jump-cleaning methodology was developed.

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

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 value. 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. 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 goes further in terms of performances but keeps this interpretability idea.

4.3 Training features

For each of the obtained targets, a vector of features using the data sources exposed in Table 1 is fetched. We train a different model for each scope: two feature matrices are obtained, representing the training features for each of the scopes. The scope 1 training set has 16,234 samples, and scope 2 has 16,925. In Tables 2 and 3, we summarize the 21 features used to train the model as well as their distribution and average coverage in the two training sets. In the remainder of this section, we provide details on these different features. Missing values are left as such: in addition to the capacities of the LightGBM implementation to handle them, it is the setting for which the best performances were obtained as opposed to the data imputation used in Nguyen et al. (2021) and Heurtebize et al. (2022).

4.3.1 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 Revenue features are obtained annually for each company for which there is a target, meaning a reported GHG emission for scope 1 and/or scope 2. 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 (scope 1 and 2): the difference between the two is accounting elements linked to the age of the assets, that provide interesting information to the model. These values are converted from the reporting currency to dollars using the foreign exchange rate from the

Type of feature	Name	Values	Coverage
General	Year	2010 to 2020	100%
General	Country of Incorporation	Country code (ISO 3166, alpha-3 code)	100%
Industry Classification	BICS Classification Levels 1 to 7	Industry Name	100%
Industry Classification	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	54.1%
Regional	CO_2 Law: Existence of an ETS or carbon taxes	National Implemented Subnational Implemented No CO_2 Law	100%

Table 2: Categorical features used to train the GHG emissions estimation model.

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%
Financial	Enterprise Value	11.4	7 578	2 609 476	Million $	99.5%
Financial	Revenues	56.3	4 167	1 939 292	Million $	100%
Financial	Property, Plant & Equipment Gross	28.6	3 291	1 896 412	Million $	87.2%
Financial	Property, Plant & Equipment Net	8.4	1 542	966 459	Million $	99.6%
Financial	Life Expectancy of Assets	0.42	13.42	50	Year	99.2%
Energy	Energy Consumption	1.7	731	207 784	GWh	74.1%
Energy	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 3: Numerical features used to train the GHG emissions estimation model.

31^{st} December of the considered year. Apart from this conversion, financial data are used as reported from the company's financial communication with no additional manual re-treatment, guaranteeing reproducibility.

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:

$$\text{Life Expectancy of Assets} = \frac{\text{GPPE}}{\text{Depreciation Expense}} \tag{2}$$

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 whose assets have a longer life expectancy 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 equivalent following formula is used:

$$\text{Life Expectancy of Assets} = \frac{\text{NPPE} - \text{Capital Expenditure} + \text{Accumulated Depreciation}}{\text{Depreciation, Depletion \& Amortization}} \tag{3}$$

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 expense. We did not measure any significant impact of these approximations on the final GHG emission estimation.

4.3.2 Industry classification

Industry classification and sectorization features allow the model to grasp 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, which 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). Testing these different classifications and different combinations of them, we retain the one for which the model gave the best performances in terms of MSE by subsectors, the BICS. No manual retreatment is done for reproducibility purposes.

For each company, its main industry classification is obtained using the BICS. The industry in which a company is classified corresponds to the one in which it is making the biggest fraction of its revenues. The BICS classification is a detailed and granular one, with seven hierarchical levels. It makes granular distinctions between the different sectors, going as far as distinguishing companies operating in the Oil & Gas Production field but either working on Petroleum Marketing or focusing on Exploration & Production.

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, the model has more difficulties to generalize to other samples. 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 by the LightGBM model.

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.

4.3.3 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 shown in Table 3. To be noted, the distinction between renewable and non-renewable power generated is available in our dataset but was not used in this version of the model. 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.

4.3.4 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. 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 in the USA for instance).

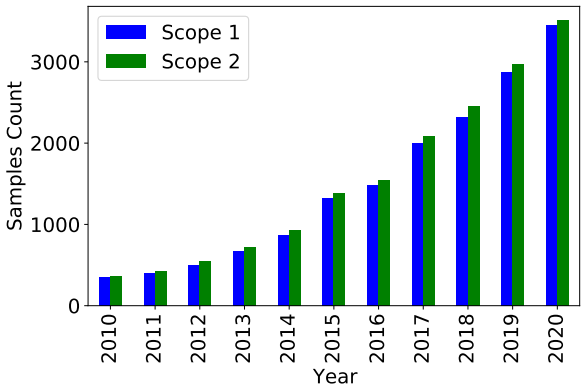


Figure 1: GHG emissions: number of companies with a reported emission per year for scopes 1 and 2

4.4 High quality dataset

Using the features and target cleaning procedures, the final training datasets to estimate scope 1 and scope 2 GHG emissions are obtained. All preprocessing steps are transparent and fully automated with no manual retreatment for the sake of reproducibility. These two high-quality datasets are used in all the remaining parts of this study. Figure 1 shows for scope 1 and scope 2 the number of companies for which a reported GHG emission per year was obtained. There has been an important increase in data quantity through the years, which illustrates the growing importance of GHG emission reporting.

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 a causal consecutive train, validation, and test data sets. The model learns the mapping between features and targets on the training set, determines its parameters on the validation one, and is finally tested in test one. To avoid overfitting and preserve the generalization capacity of the model, the test set should only be used at the end of the training, to evaluate the model. This usual strategy is not appropriate for the current problem, estimating the GHG emissions of companies that have never reported them. 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.
- the amount of data grows from a very low baseline both quantity- and quality-wise. 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 meaningful results, we need to rely on the entire time span of available data.
- similarly, GHG emissions data are non-stationary, leading to inaccurate results using this standard splitting scheme.

To address these issues, a specific testing methodology and cross-validation scheme were developed, inspired by Assael et al. (2022).

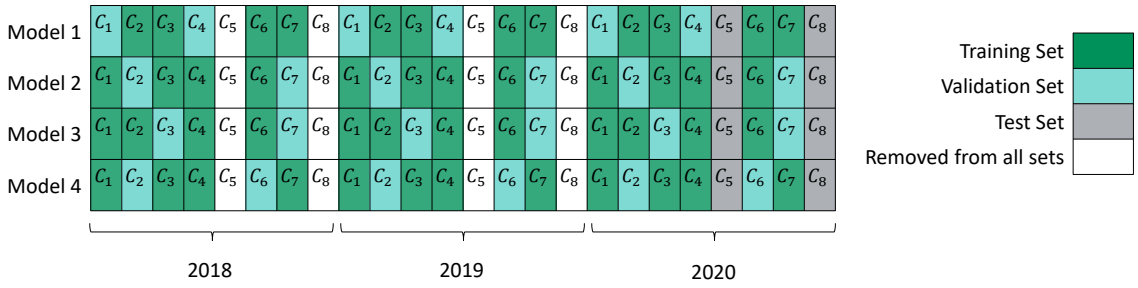


Figure 2: Company-wise cross-validation: the validation sets consists of randomly selected companies, which allows training to account for most of the most recent data.

To estimate unreported data during the last available year, the test set built to evaluate the models should only include companies that are not in the training or validation sets: the goal of the model is to estimate unreported emissions of companies which, most of the time, never reported their emissions before. Moreover, it avoids a potential bias: because of the huge year-over-year correlation of GHG emissions for the same company, having the same company both in training/validation and in tests 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 samples constitute the test set. These companies may have other reported emissions for other years: all these companies are removed from the training and validation sets.

For training and validation, a K-fold company-wise cross-validation is used: 80% of companies are randomly assigned to the training set and the remaining 20% to the validation one. We train 180 models on each of the K training sets varying the hyperparameters of the LightGBM algorithm and select the best one based on its average performances, measured using the MSE, on the respective validation sets. In this way, 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 validating them with the most recent and more relevant data. We take $K = 4$.

Figure 2 illustrates in a three-year and eight-company dataset the procedure used to build the training, validation, and test sets.

5 Results: evaluating the performances of the model

We first assess the quality and performances of the model on the designed high-quality testing set, built as explained in Section 4.5.

5.1 Selected metrics

An important contribution of this work is to design a model with both good global performances on the test set and good performances for each business sector, at different levels of granularity, for each country and for each decile of revenues.

To evaluate performances on the test set, the selected metric is the root-mean-squared error also referred to in this paper as *RMSE*, defined as

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{4}$$

where $\hat{y}_i$ is the GHG estimation (log-transformed) from the model and y_i is the ground truth (log-transformed).

Another measure of performance is the mean-absolute error, referred to in this paper as *MAE*, defined as

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{5}$$

where $\hat{y}_i$ is the GHG estimation (log-transformed) from the model, and y_i is the ground truth (log-transformed).

The R^2 metric is also a common measure of performance. It is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2} \tag{6}$$

where $\hat{y}_i$ is the GHG estimation (log-transformed) from the model, y_i is the ground truth (log-transformed), and $\bar{y}_i$ is the average of the ground truth emissions (log-transformed).

We provide global results using these three metrics for comparability purposes across the literature on GHG emissions models. *RMSE* and *MAE* metrics can vary between 0 and infinity, a value of 0 meaning that the model is perfectly accurate. The R^2 metric varies between 0 and 1, a value of 1 meaning that the model is perfectly accurate. *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 in the remainder of this study.

5.2 Multiple test sets

As shown in Figure 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 variability: 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 performances on these five test sets.

		Scope 1		Scope 2	
Range	**Metric**	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
$[0, 1]$	R^2	0.832	0.007	0.746	0.017
$[0, +\inf[$	RMSE	0.578	0.007	0.522	0.031
$[0, +\inf[$	MAE	0.401	0.006	0.341	0.010

Table 4: Results of the model on the five different test sets: mean and standard deviation of the R^2, RMSE and MAE metrics. The three metrics, computed on the decimal logarithm of the emissions, are given for comparability purposes across the literature and should not be compared to each other

5.3 Global performances

Table 4 displays the mean global results of the scope 1 and scope 2 models for the RMSE, MAE, and R^2 metrics on each test set. In comparison to the literature like Goldhammer et al. (2017) or Griffin et al. (2017), we display only out-of-sample results as they are the ones that show the performance of the model and its capacity to generalize beyond the training data.

These metrics are computed using the decimal logarithm of the predicted emission and the decimal logarithm of the reported emission. As stated as an introduction to this section, *RMSE*, *MAE*, and R^2 metrics are displayed for comparability purposes across the literature. As they differ in their definition, they should not be compared to each other.

5.4 Breakdown of performances by sectors, countries and revenues

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

Results are presented in Figures 3 and 4, respectively, for scope 1 and scope 2.

Figures 3a and 4a show the *RMSE* distribution across the five test sets for BICS Sectors L1 (Level 1 of granularity) and L2 (Level 2 of granularity); the green box-plots correspond to the L2 sectors results and the pink ones in the background corresponds to the associated L1 sectors. 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 performances across all sectors, with particularly good performances in the most emissive sectors. 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 some sectors, like mining, 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. 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. The model performance for the most emissive BICS Sector L3 (Level 3 of granularity) is proposed in the Appendix A.

Figures 3b and 4b take a similar approach by proposing the *RMSE* distribution across the five test sets per countries, for both scopes. Results are ordered by how emissive a country is in regard to the set of reported data.

Finally, we show in Figures 3c and 4c the *RMSE* performances across the five test sets per deciles 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 due to the fact that there are in the training sets more samples coming from big companies, as shown in Table 3, than ones coming from SMEs. Gathering more data from SMEs is a source of improvement for future versions of the model.

6 Results: comparison of estimates with other providers

The quality of the estimates from our model, called the GHG-2022 model, is now assessed in comparison to other data providers, comparing both coverage and accuracy. An innovative methodology was developed to achieve this goal. Comparisons are done as of August 2022.

6.1 Retraining the model on the full dataset

In Section 5, the model performances are 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 model on which predictions will be made, we follow the procedure previously validated by the results in Section 5 and train the model on all the data, without test sets. Validation sets are still required to find the best model hyperparameters.

We consider the universe of 48,429 companies extracted from the Worldscope Refinitiv Database in 2020 to evaluate the prediction of the GHG-2022 model and to compare them to other providers.

6.2 Comparison of coverage

Figure 5 displays, for scope 1 and scope 2, the number of reported GHG emissions and estimated GHG emissions each provider can provide for the year 2020. The test was conducted on the full universe of 48,429 companies: for instance, the GHG-2022 model can provide for scope 1 4,360 reported data (sampled from CDP and Bloomberg and used for training as explained in Section 4.2.1), 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.

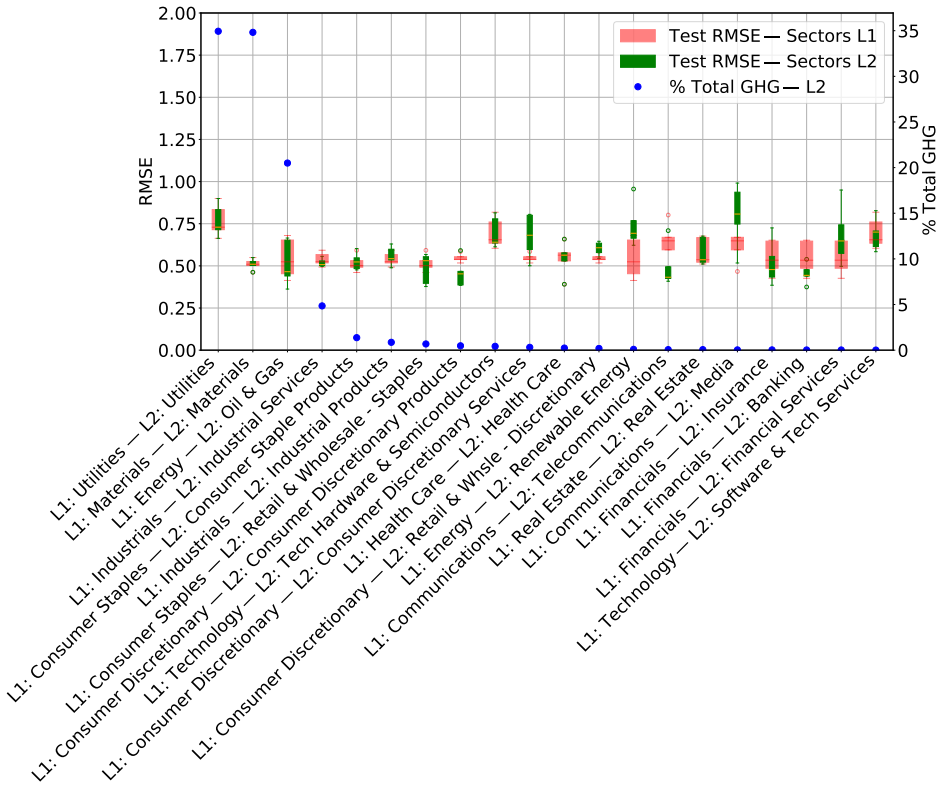
Coverages for Bloomberg, Trucost, Sustainalytics, MSCI, and CDP are rounded as there may be slightly different results depending on the moment the datasets were obtained. Results provided in Figure 5 were obtained using available elements in August 2022.

Figure 5 clearly demonstrates that using a machine learning model, fully automated and with a systematic methodology, allows achieving an important coverage, greater than any other provider, while preserving good performances.

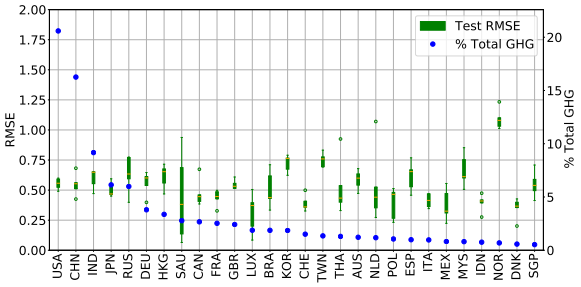
6.3 Comparison of estimates accuracy

To assess how good estimates are from one provider to another, we developed a methodology relying on the high year-over-year correlation of GHG estimates. The methodology is as follows:

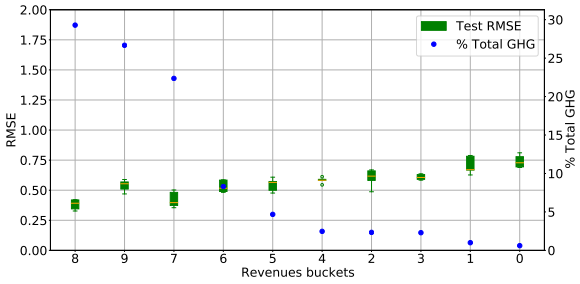
- Using the same procedure, two models are trained: one relying only on 2010 to 2018 data and a second one relying only on 2010 to 2019 data. These models, when used for predictions on 2018 and 2019 data, respectively, give 2018 and 2019 point-in-time estimates.
- We consider the reported values in 2020 for companies that started reporting in 2020 and thus have never reported in 2018 or in 2019. This 2020 reported value is called the ground truth.



(a) Boxplot of test RMSE on the five different test sets per BICS sectors levels 1 and 2, ordered by level 2 sectors emissions.

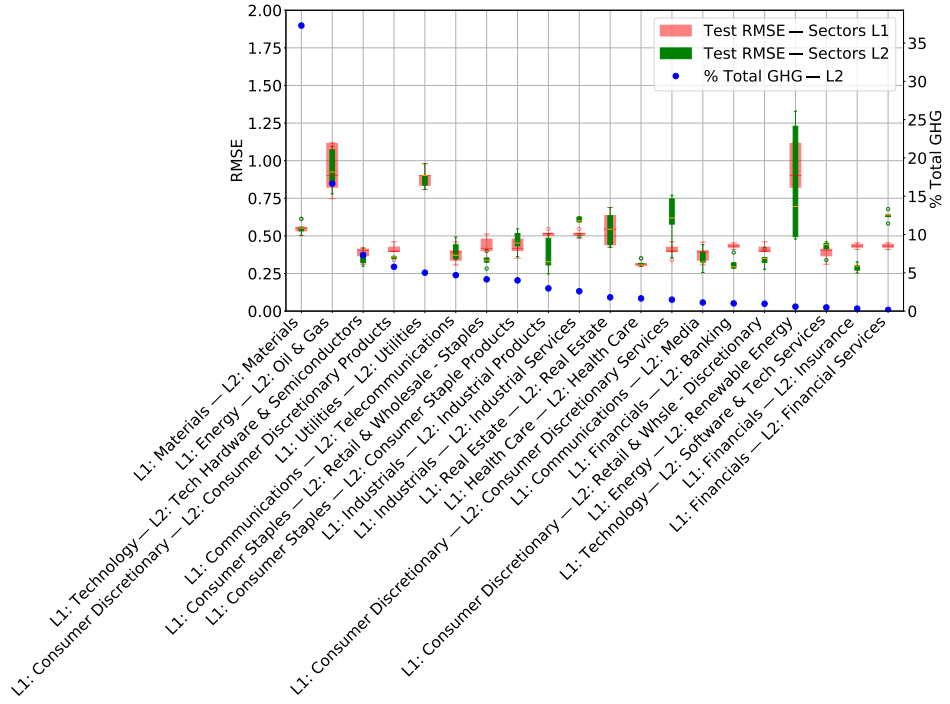


(b) Boxplot of test RMSE on the five different test sets per countries, ordered by countries emissions.

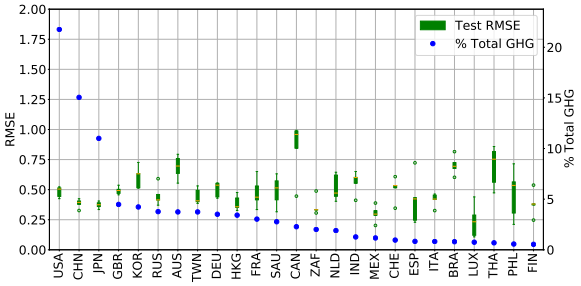


(c) Boxplot of test RMSE on the five different test sets per deciles of revenues, ordered by revenues deciles emissions.

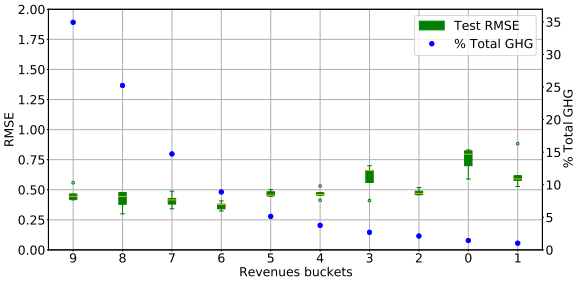
Figure 3: GHG emissions scope 1: distribution of performances of the model on five test sets according to different characteristics of companies.



(a) Boxplot of test RMSE on the five different test sets per BICS sectors levels 1 and 2, ordered by level 2 sectors emissions.



(b) Boxplot of test RMSE on the five different test sets per countries, ordered by countries emissions.



(c) Boxplot of test RMSE on the five different test sets per deciles of revenues, ordered by revenues deciles emissions.

Figure 4: GHG emissions scope 2: distribution of performances of the model on five test sets according to different characteristics of companies.

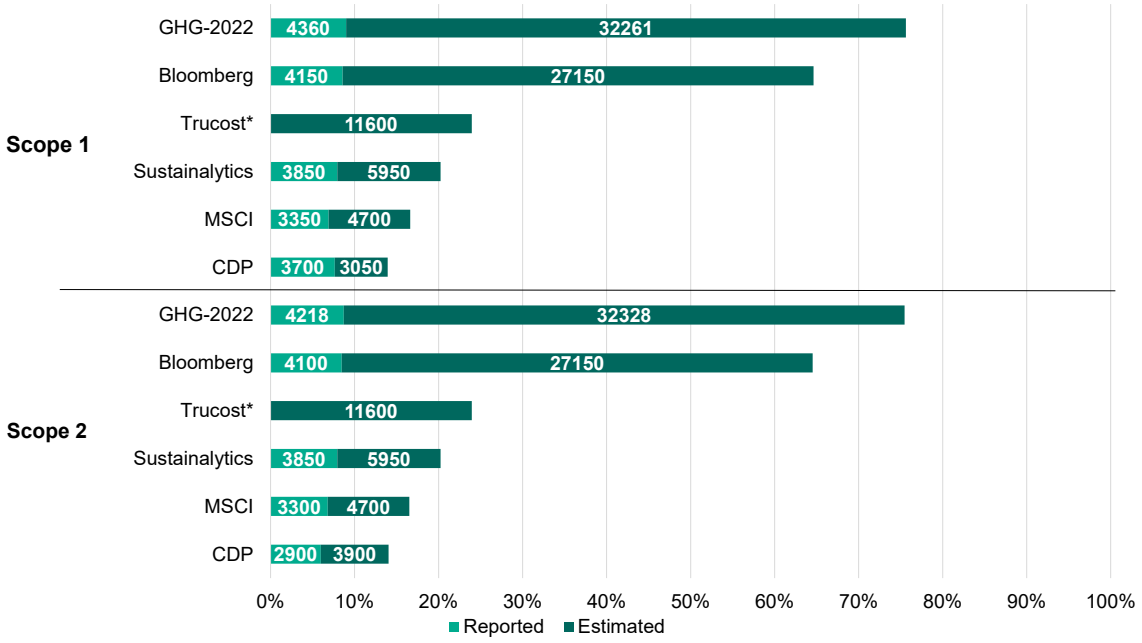


Figure 5: GHG emissions coverage, as of August 2022: number of reported data and estimates provided by each model. For the providers marked with an asterisk, the split between reported and estimated data was unclear, so all data points are marked as estimates.

Provider	RMSE	Number of samples
GHG-2022	0.828	1079
MSCI	0.882	509
Bloomberg	0.948	1119
Trucost	1.033	980
CDP	1.222	546

(a) All companies are considered.

Provider	RMSE: provider	RMSE: GHG-2022	Number of 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 5: Last scope 1 GHG estimates from providers (2019 – 2018) compare to 2020 ground truth, for companies which starts reporting in 2020.

- By comparing the 2019 estimates (or 2018 estimates if 2019 estimates are not available) from the GHG-2022 model and the ones from the other provider 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. Comparison is done by computing an RMSE on the decimal logarithm of the estimation and ground truth.

Considering this methodology, we propose two ways of evaluating the providers:

- First, we evaluate each of them separately. Tables 5a and 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 GHG-2022 model has the best, i.e. lowest, RMSE in comparison to the other considered providers.
- Second, we consider each provider against the GHG-2022 model. Results are available in tables 5b and 6b for scope 1 and scope 2. This time, the same samples for GHG-2022 and the considered provider are used, increasing the comparability. In each case, the GHG-2022 is systematically more accurate than the considered provider.

Point-in-time data Models are trained using only 2018 and 2019 data respectively, to avoid any leakage of the future in the 2018 and 2019 estimations. It may not be the case for the estimations of the other providers, which can bias the evaluation towards better performances of 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 performances than the considered providers.

Breakdown of performances per sectors The methodology developed to compare the GHG-2022 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.

Provider	RMSE	Number of samples
GHG-2022	0.709	1042
MSCI	0.808	522
Bloomberg	0.809	1089
Trucost	0.822	955
CDP	0.970	577

(a) All companies are considered.

Provider	RMSE: provider	RMSE: GHG-2022	Number of samples
Bloomberg	0.774	0.700	1029
MSCI	0.780	0.707	502
Trucost	0.803	0.645	925
CDP	0.950	0.661	561

(b) Only common companies between providers are considered.

Table 6: Last scope 2 GHG estimates from providers (2019 – 2018) compare to 2020 ground truth, for companies which starts reporting in 2020.

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 Figure 6: in green, the distribution of differences between CDP estimates and the ground truth is shown; in pink, it is the distribution of differences between the GHG-2022 estimates and the ground truth. Distributions from the GHG-2022 model are more centered around 0, meaning better accuracy than CDP. However, CDP estimates are more conservative than the ones from the GHG-2022 model: when CDP estimates are not exact, they have a tendency to overestimate, whereas the GHG-2022 model is rather balanced between overestimation and underestimation. Both behavior and calibration can have their strengths and weaknesses depending on the use case.

7 Interpretability - Shapley values

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

A common critic of GBDT is that, despite their superior performances 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. Shapley values and their uses in the context of machine learning are well-described in Molnar (2020).

The Shapley value of a feature can be obtained by averaging the difference of prediction between each combination of features containing and not containing the said feature. For each sample in the dataset, each feature possesses its own Shapley value representing the contribution of this feature to

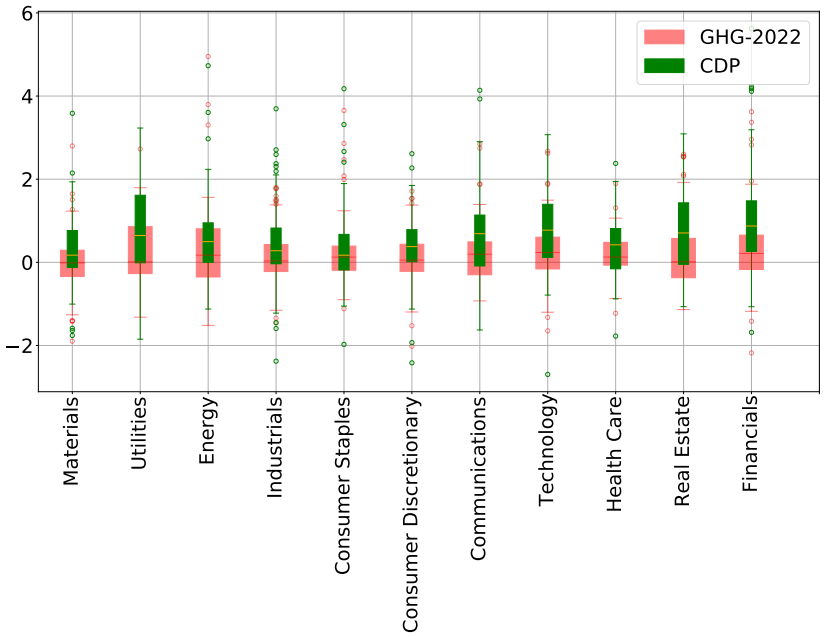


Figure 6: Differences of the emission from estimations from CDP and from GHG-2022 with ground truth for scopes 1 and 2.

the prediction for this particular sample. Shapley values have interesting properties, like the efficiency property. If we note $\phi_{j,i}$ the Shapley value of feature j for a sample x_i and $\hat{f}(x_i)$ the prediction for the sample x_i, Shapley values must add up to the difference between the prediction for the sample x_i and the average of all predictions $\mathbb{E}_X(\hat{f}(X))$ and then follow the following formula:

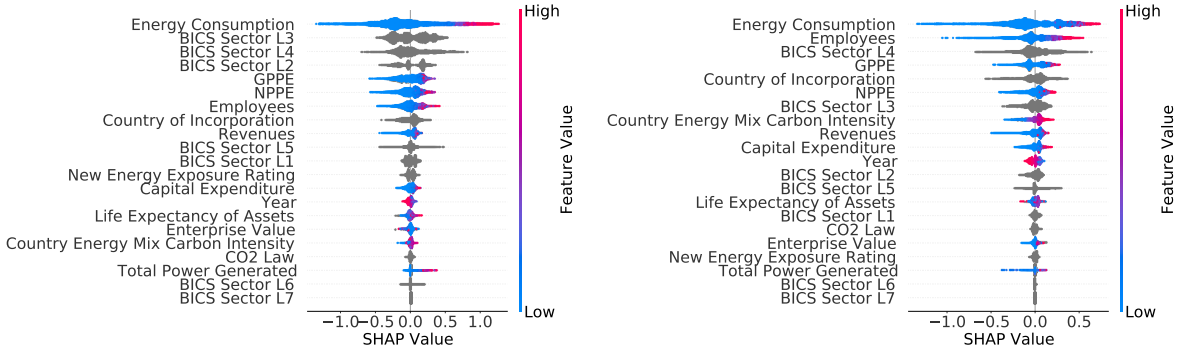
$$\sum_{j=1}^{p} \phi_j = \hat{f}(x) - \mathbb{E}_X(\hat{f}(X)) \tag{7}$$

The dummy property states that the Shapley value of a feature that does not change the prediction, whatever combinations of features it is added to, should be 0.

Shapley value calculation is quite time- and memory-intensive. Lundberg and Lee (2017) and later Lundberg et al. (2018) proposed an implementation of a fast algorithm called TreeSHAP, which allows approximating Shapley values for tree models like the LightGBM and which is used in the following. Shapley values computed with this algorithm are referred to as SHAP values.

7.1 SHAP feature importance

We provide in Figure 7 the breakdown of SHAP values per feature for the scope 1 and scope 2 GHG emissions, ordered by importance. For each feature, this graph shows the distribution of SHAP values across each sample in the training set. These graphs are key elements in the constructed model as they make it interpretable: they can be computed for any set of features, allowing us to understand why the model makes a specific decision and outputs this predicted estimate.



(a) Scope 1

(b) Scope 2

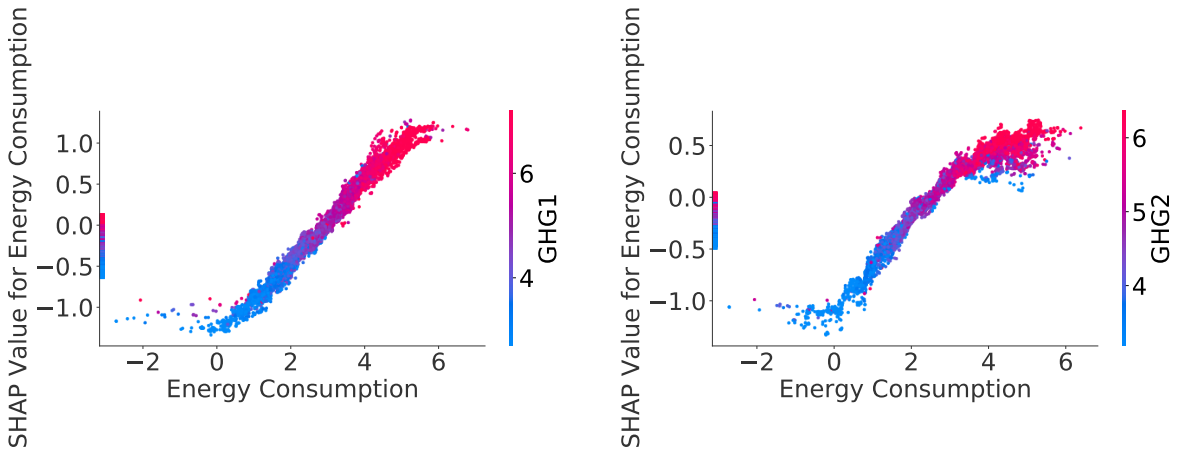
Figure 7: SHAP values: impact of each feature on the predicted GHG emission, order by importance.

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 plot also highlights 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 was important to choose a granular classification as features up to the classification Level 6 were used. However, the too deep Level 7 of the BICS was not used by the model: as this Level 7 was too sparse, it did 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 which did not.

7.2 Relationship between features values and GHG estimates

Numerical features SHAP values can be computed for each feature on each sample, allowing us to understand the relationship captured by the model between a feature and the estimated GHG emission. Indeed, for numerical features, we can plot the SHAP values for a specific feature against this feature value in the dataset. For instance, Figure 8 shows the relation between SHAP values of the Energy Consumption feature and the decimal logarithm of the Energy Consumption feature value. 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 value. Appendix B.1 provides some other examples of SHAP values plots for numerical features, allowing for a better interpretation of what the model is learning.



(a) Scope 1

(b) Scope 2

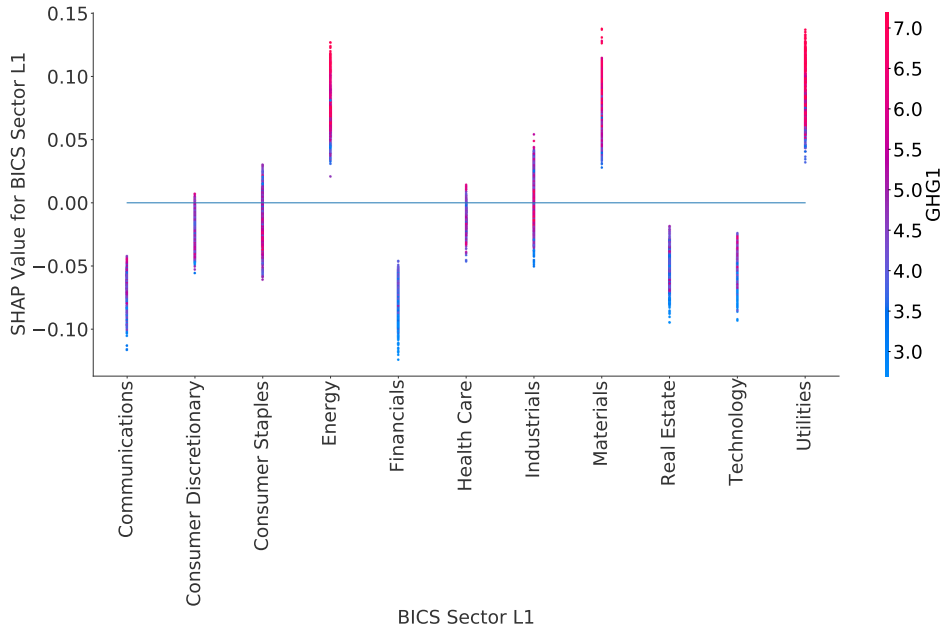
Figure 8: Relationship between SHAP values of the Energy Consumption feature and the decimal logarithm of the Energy Consumption feature value.

Categorical features SHAP values can also be used on categorical features to study their distribution, for each value a categorical feature can take. For instance, Figure 9 shows the distribution of SHAP values for the BICS Sector L1 feature, for each of the BICS Sector L1 sectors. 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 an increase in the estimated emission (positive SHAP values). On the contrary, samples in the Financial sector have negative SHAP values, showing a decrease in the estimated emission.

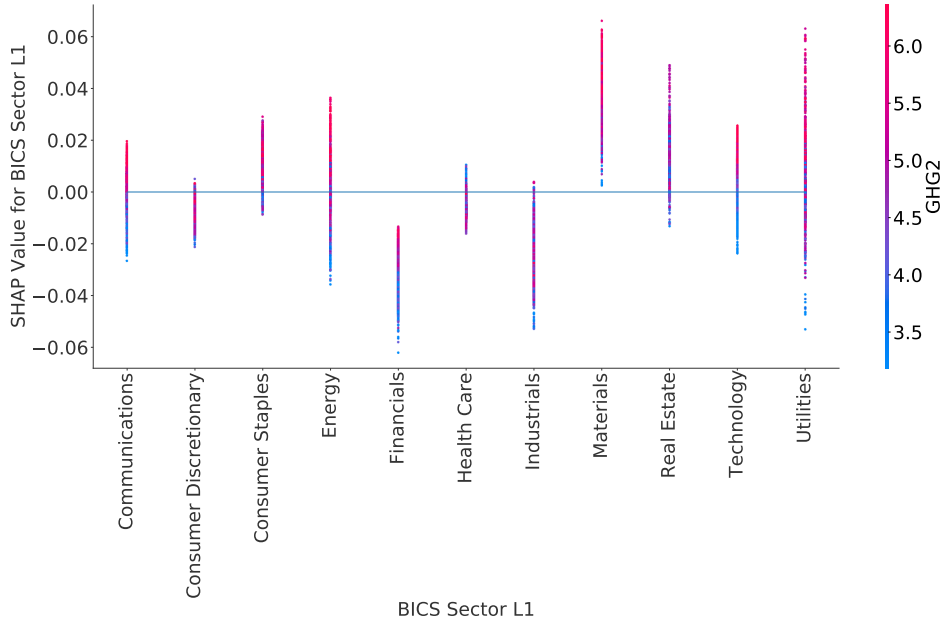
This plot can be done for all categorical features, allowing us to understand the distribution of SHAP values according to each category, and then to have a better interpretation of the model. In Appendix B.2, an additional SHAP value plot for categorical features is provided, for further interpretation elements.

Plots in Figure 9 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. Working on these clusters and removing the ones with too few samples could be a solution to improve the model by removing outliers and preventing overfitting. 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 from 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. The removal of this cluster with very few samples allows to improve the quality of the training data by removing outliers. Similar studies on other sectors lead to the same results: for the Materials sector, it can lead to the removal of the only years a company did not report its Energy Consumption, for instance.

Data polishing This methodology should, however, be automated and applied systematically. A first implementation using the SHAP distribution for each BICS Sector L4 (Level 4 of granularity) was done. For each L4 sector, a hierarchical clustering algorithm is applied, separating clusters if their



(a) Scope 1



(b) Scope 2

Figure 9: SHAP values: impact of belonging to a particular level 1 BICS sector on the predicted GHG emission.

		Without data polishing		With data polishing	
Range	**Metric**	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
$[0, 1]$	R^2	0.832	0.007	0.859	0.009
$[0, +\text{inf}[$	RMSE	0.578	0.007	0.501	0.020
$[0, +\text{inf}[$	MAE	0.401	0.006	0.347	0.013

(a) Scope 1

		Without data polishing		With data polishing	
Range	**Metric**	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
$[0, 1]$	R^2	0.746	0.017	0.778	0.017
$[0, +\text{inf}[$	RMSE	0.521	0.031	0.464	0.025
$[0, +\text{inf}[$	MAE	0.341	0.010	0.312	0.011

(b) Scope 2

Table 7: Results of the model on five different test sets, without and with the data polishing methodology applied: mean and standard deviation of the R^2, RMSE and MAE metrics. The three metrics, computed on the decimal logarithm of the emissions, are given for comparability purposes across the literature and should not be compared to each other.

distance is above 0.04 in the SHAP values space, and removing clusters of data with an insufficient number of samples, i.e., less than 10. All these parameters were found by trial-and-error. For both scope 1 and scope 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. Results are presented in Table 7, on average on the 5 different test sets: for both scopes, there is an average *RMSE* decrease between 11% and 13%. It may come at the price of an increase of variability between results in the different test sets, especially for scope 1. As future work, studying more the impact of this methodology on both global and granular performances may lead to a more accurate and robust model.

8 Conclusion

To mitigate the fact that GHG emissions reporting and auditing are not yet compulsory for all companies and that methodologies of measurement and estimations are not unified, we proposed a machine-learning model to estimate non-reported company GHG emissions for scopes 1 and 2. The resulting model showed good out-of-sample performances when assessing it globally as well as good and balanced out-of-sample performances when assessing it per sector, country and bucket of revenues. Comparing the obtained results to those of other providers, as of August 2022, we found our generated estimates to be available for a larger number of companies and more accurate.

In addition to its large coverage and accuracy, this model is also flexible, allowing for easy evolution of the input data as regulations evolve. It is also transparent, reproducible, and explainable: the methodology is described in this study extensively, and the implemented tools based on Shapley values allow us to understand the role played by each feature in the construction of the final output. We focused on some important interpretability elements, but many more interactions could have been studied: interaction between sectors, revenues, and the estimated GHG emission redoing the study done in this paper for each sector separately. These would give even more information on how the model is working and allow us to understand the specificities of GHG emissions per sector. Studying

all these SHAP values interactions is beyond the scope of this study and could be the object of an entire future publication.

Future work to improve the model will first focus on gathering and including more training data from SMEs so that the coverage of the model is further improved. As it was stressed in this analysis, the used industry classification is critical: sometimes, companies operating in very different sectors in terms of GHG emissions can be grouped together. Gathering data about all the activities a company reports being active in and working on including this new and more precise industry classification in the model will help improving its accuracy. The data polishing method introduced in the last section of this study will also be developed with the goal of obtaining a more robust model. Future work on the interpretability of the model will focus on the performance improvement linked to 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.

References

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

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. *Available at SSRN 3988318*, 2022.

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

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.

CDP. CDP full GHG emissions dataset – Technical annex III: Statistical framework., 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.

Marielle De Jong and Anne Nguyen. Weathered for climate risk: a bond investment proposition. *Financial Analysts Journal*, 72(3):34–39, 2016.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Bernhard Goldhammer, Christian Busse, and Timo Busch. Estimating corporate carbon footprints with externally available data. *Journal of Industrial Ecology*, 21(5):1165–1179, 2017.

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.

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.

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.

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.

Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

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.

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

PCAF. The global GHG accounting and reporting standard part A: Financed emissions. Second edition., 2022.

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, 2018.

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

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. Refinitiv ESG carbon data and estimate models, 2023. URL `https://www.refinitiv.com/content/dam/marketing/en_us/documents/fact-sheets/esg-carbon-data-estimate-models-fact-sheet.pdf`.

Securities and Exchange Commission (SEC). The enhancement and standardization of climate-related disclosures for investors, Proposed rule, 2022.

Manish Shakdwipee and Linda-Eling Lee. Filling the blanks: Comparing carbon estimates again disclosures. *MSCI ESG Research Issue Brief*, 2016.

LS Shapley. A value for n-person games. *Contributions to the Theory of Games*, (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.

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.

A Model performances: BICS Sectors Level 3

In addition to the plots displayed in Figures 3a and 4a, we provide for transparency purposes the breakdown of the out-of-sample performances of the model across the five test sets for the different BICS Sector L3, ranked from high to low emissivity, for sectors accounting for at least 1% of the total GHG emissions of the reporting companies. Results are available in Figure 10.

B Relationship between features value and GHG estimates

In addition to the plots and interpretation elements provided in Section 7.2, some additional results are shown for a better understanding of the learned relationships in the model.

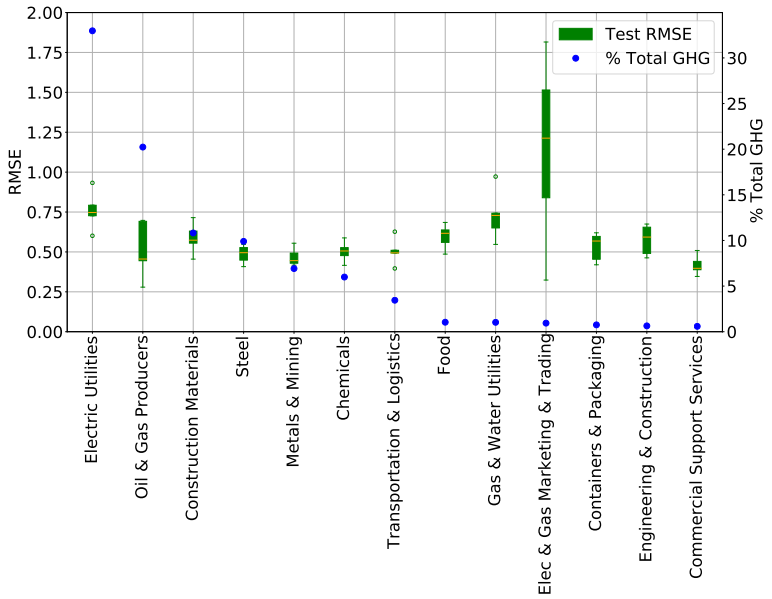
B.1 Numerical Data

Figure 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 revenue level, the SHAP values are almost constant.

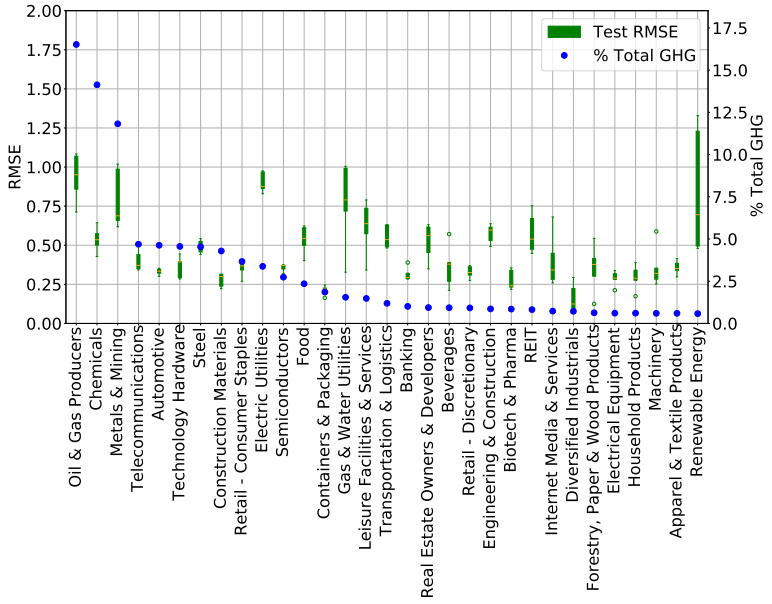
Figure 12 shows a near-linear relationship between the SHAP values of the Employees feature and the decimal logarithm of the Employees feature values, apart from the few points on the Y-axis referring to missing data.

B.2 Categorical Data

Figure 13 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.

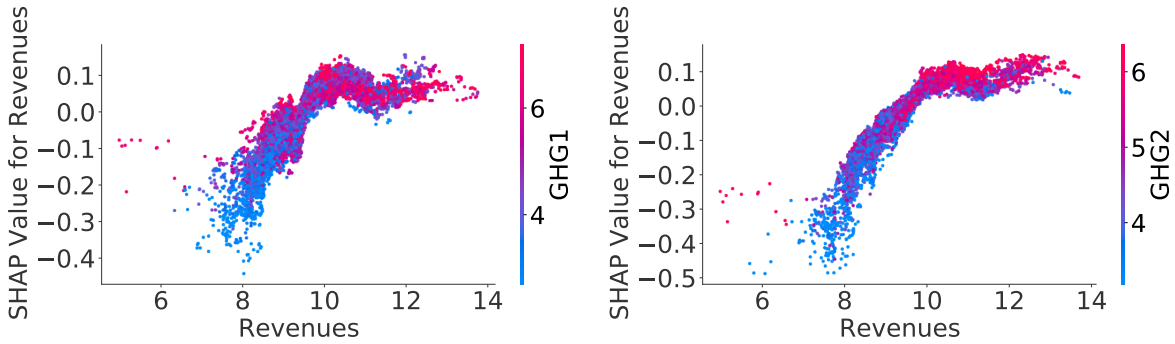


(a) Scope 1



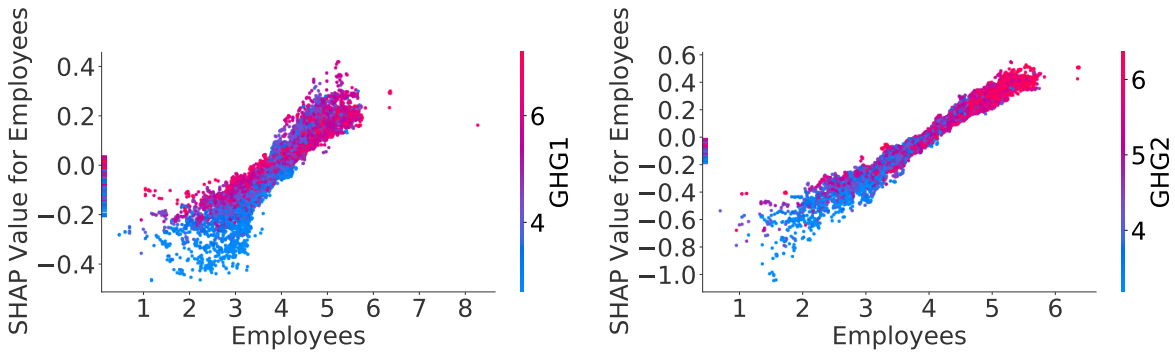
(b) Scope 2

Figure 10: GHG emissions: distribution of performances of the model on five test sets according to BICS sectors level 3.



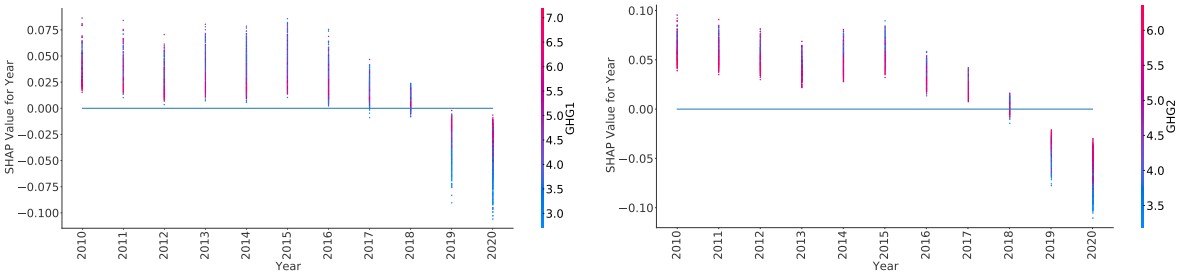
(a) Scope 1 (b) Scope 2

Figure 11: Relationship between SHAP values of the Revenues feature and the decimal logarithm of the Revenues feature value.



(a) Scope 1 (b) Scope 2

Figure 12: Relationship between SHAP values of the Employees feature and the decimal logarithm of the Employees feature value.



(a) Scope 1 (b) Scope 2

Figure 13: SHAP values: impact of the year on the predicted GHG emission.