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, specifically 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, by countries or by revenues buckets. We also compare our results to those of other providers and find our estimates to be more accurate. Thanks to the proposed explainability tools using Shapley values, our model is fully interpretable, the user being able to understand which factors split explain 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-being activities or "footprint" are now commonly held responsible for the current pollution of our environment. Our footprint is measures by how fast we consume resources and generate waste versus how fast our planet can absorb our waste and generate resources, according to Wackernagel and Rees (1998). When it comes to our 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 our planet (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 to review the objectives of the global effort to fight climate change. They assess GHG footprints at 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 country level creating a new framework applicable to companies, the underlying hypothesis being that the country 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 companies point of view, the assessment of their GHG footprint can be useful not only for regulatory or accounting disclosure, but also to implement strategies designed to mitigate and reduce their emissions. All frameworks like carbon pricing policies, measuring alignment to climate scenario with the Paris Agreement Capital Transition Assessment (PACTA) or engaging toward net zero GHG emissions via Net Zero Banking Alliances (NZBA) need correct GHG emissions baseline. This momentum will be emphasis by the new Corporate Sustainability Reporting Directive (CSRD) coming into force from 2024 for 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 of 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 financial fundamental analysis and slowly implemented in corporate valuation methodologies at least for most vulnerable sectors. Nevertheless as soon as financial institutions aggregate GHG emissions at portfolio level for several companies, they need homogeneous methodologies. At this stage, company reporting of GHG emissions are either voluntary or mandatory depending on location and link 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 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 this way, for a gas with a GWP of 10, two tons of it would have CO_2-eq of 20 tons. In practice, measuring GHG emissions of a stakeholder requires much more information depending on how the GWP is released. To standardize this 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 case, 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 makes 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 our study, we limit the methodology 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 to measure 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 4 000 companies worldwide, as we 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 requiring to take into account the GHG emissions of companies. At the same time, even some recent study like Bolton and Kacperczyk (2021) analyses 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 conclude 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, like (Aswani et al., 2022).

That is why it is so important to analyze in details 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 even if the data is modeled (simple derivation from a previous year to more complex non-linear models). When it comes to comparing corporate across geographies and sector or drawing conclusion at global level for anthropogenic GHG emissions we need fair assessment of the GHG emissions at countries, corporations, factories and personal levels. This study is narrowing to unreported estimated emissions of companies. Our model framework focus on estimating the targeted high quality reported emissions for every use cases, especially waiting for international regulatory bodies to bring a homogeneous framework for corporation to report their GHG emissions in extra financial statements.

2 Literature review - Hypothesis

Focusing on scopes 1 and 2, 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 the 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 improve 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), 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 accounting of GHG emission is a perilous task even when it is handle by big corporate sustainable expert teams or by CDP experts.

To mitigate such 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, 2017; Paribas, 2016; Boermans et al., 2017) and S&P Global Trucost and CDP use models to estimate the GHG emissions of companies that fail to publish emissions data. Such models relies mainly on rules of proportionality between emissions and the size of the company operations or more recently on more complex approach 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 metric like GHG intensity per the company's energy consumption and production or per tons of produce cements. However, these metrics are only available for the limited numbers of companies reporting them without reporting their GHG emissions. These models are calibrated on the samples of reported data. Performance is around 60% in terms of R^2 for most samples, when evaluating on 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 calculation of the regressions levels. On the other hand, out-of-sample performance test requires a completely new dataset to test the model on unseen companies with reported emissions. Good performance out-of-sample shows that the model avoided overfitting and is able to generalize well.

Some more advanced models described in Goldhammer et al. (2017); Griffin et al. (2017); 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 but relies more usually 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 efficacy of the model was found for predicting aggregate 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, if we consider the five dirtiest industries representing about 90% of total scope 1 emissions (Utilities, Materials, Energy, Transportation, Capital Goods), their average in-the-sample R^2 computed with the logarithm of the GHG emissions is only 51%. If we consider the five dirtiest industries accounting for about 70% of the total emissions in terms of scope 2 (Materials, Energy, Utilities, Capital Goods, Automobiles & Components), their average in-the-sample R^2 computed with the logarithm of the GHG emissions is only 52%. Moreover, their model fails for Insurance, both for scope 1 and scope 2, with R^2 of -378% and -151%, respectively.
- 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 Environment, Social and Governance (ESG) data, fundamentals, and industry segmentation data. The GBDT allows for non-linear patterns to be found even if not all featured data are available. Moreover, an important debiasing approach compares the features 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 go from 84% for firms with good disclosures (lots of features available) to 41% for companies with average or poor features disclosures.

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 the

GHG emissions reporting and auditing is not compulsory, the only viable alternative is to predict non-reported company emissions relying on estimation models. From the above, we believe that 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 4,000 companies actually report. This model is inspired by the work of Heurtebize et al. (2022), and aims at achieving the following qualities:

- important final coverage, aiming at using the model for smaller and private companies in a scope of 50,000 companies.
- simplicity and interpretability, making sure the statistical explanations of the outputs are clear and exhaustive enough.
- implementation ease with no manual data correction.
- flexibility, making sure the model can be adapted to all types of usages within a set of different calibrations, and to be able to easily include new input data with the evolution of regulations, especially on disclosure.

To succeed in this, we made some significant choices in departing from existing approaches. First, our models are always tested on data samples never seen during the calibration, so that we can truly measure their generalization abilities. The second important decision was to keep model complexity to a minimum by relying only on the most accurate non-linear machine learning approaches. The last important decision was to keep the raw dataset from data providers with no manual corrections. Indeed, even if the use of incorrect features or emissions data can reduce the accuracy of models, their industrialization brings some operational constraints to produce automatic updates of the model. Thus, we introduce shortly an automatic data polishing process at the end of this study. In the remainder of the paper, we shall describe the data retained to calibrate and evaluate our model. This will be followed by sections in which we present in depth the designed methodology, insisting on our particular implementation choices, necessary to apply it to the use case of estimating GHG scope 1 and 2 emissions. We will then discuss the results associated to this methodology both by comparing our estimates to true reported by companies GHG emissions and by comparing our estimates to the ones from other providers. Finally, we will provide tools to get insights on how our model works and why it estimates such values of GHG emissions.

3 Datasets

We have at our disposal an important variety of data sources. 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

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.

company, we gather all indicators exhibit in Tab. 1a, selecting yearly data. Such indicators allows to get a sense at the company profitability, assets size, assets location and how they are used.

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

Data are extracted between 2010 and 2020 from 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

Our goal is to develop a data-driven model estimating scope 1 and scope 2 greenhouse gases emissions of a company which has not reported them. Using the vast amount of available indicators, whose selected ones have been exhibit in Section 3, we can build a high quality dataset and calibrate a machine

learning model which output the estimated emission of a company. This automatized method allows to estimate the emissions of any company as long data related to it can be gathered. Scope 1 and scope 2 emissions will be estimated through two separate models.

We are in 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 the year of sampling of the features and emission. Let us relabel all the couples (i, t) by the index $n \in \{1, \cdots, N\}$. This regression problem consists in explaining Y_n, the reported emission, by 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. The problem is then to train a machine learning method to learn the mapping between the lines of X and the components of the vector Y. Once the training is complete, such a model takes as input a vector of features and outputs the estimated greenhouse gas 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, to eventually yields 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 has been proposed, e.g. XGBoost (Chen and Guestrin, 2016), LightGBM (Ke et al., 2017), CatBoost (Prokhorenkova et al., 2018). We use here LightGBM. The advantage of such methods with respect to linear regression is that they are able to learn more generic functional forms.

The models are trained to minimize the mean-squared error (cost function), also refers in this paper as MSE, defined as

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

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. We are 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. We are using the reported GHG emissions gathered by Bloomberg, sourced from the companies extra-financial communication.

As the CDP data is reviewed, with direct contact with the considered company or following the company answers to CDP questionnaires, we choose to prioritize the CDP reported emissions when both data sources report a emission for this company. There reported emissions are expressed in tCO_2-eq.

4.2.2 Target cleaning procedure

GHG emissions are reported at different dates during the year. To unify samples and keep meaning with the used training features, we attribute the GHG emissions reported between January and June of the year y to the year $y-1$ and the GHG emissions reported between July and December of the year y to the same year y. For both scopes, we remain with only one reported GHG emission per company and per year.

Variability is an important characteristic of GHG emissions data, leading to sometimes 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 mitigates these issues: we develop for this purpose a jump cleaning methodology.

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 serie of GHG emissions of a company, all data points before the jump and the jump are removed. A jump is unexplained if we cannot find a concomitant and large enough corporate action to justify it. In practice, a jump is said *explained* if we find, using a Bloomberg corporate action dataset, at least one corporate action amounting for at least 20% of the company revenues during the year before or after the considered jump. The different thresholds were determined by trials and errors.

To reduce the negative impact of the skewed nature of the GHG emissions distribution, we are training our model 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. A error of one unit in the decimal logarithm estimation mean an error of one order of magnitude in the raw GHG emission (multiplied/divided by 10 in comparison to the true value).

4.3 Training features

For each of the obtained target, we fetch a vector of features using the data sources exposed in Tab. 1. As two models are trained for the scopes 1 and 2, we obtain two feature matrices, representing the training features for each of the scope. The scope 1 training set has 16234 samples and the scope 2 16925. We summarize in Tab. 2 and 3 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 let as such as the LightGBM model is able to handle them while preserving performance.

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 Revenues features are obtained annually for each companies for which we have a target, meaning a reported GHG emission for scope 1 and/or scope 2. These values are converted from the reporting currency to dollars using the foreign exchange rate from the 31st December of the considered year.

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.

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:

$$LifeExpectancy = \frac{GPPE}{DepreciationExpense}$$

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 in average older and may emit more GHG.

As we do not have the Depreciation Expense indicator in our dataset and the Gross property Plant & Equipment feature have a lot of missing values, we use the equivalent following formula:

$$LifeExpectancy = \frac{NPPE - CapitalExpenditure + AccumulatedDepreciation}{Depreciation, Depletion\&Amortization}$$

The numerator is modified by decomposing the Gross Property Plant & Equipment term. We make the approximation that, 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 a significant impact of these approximations on the final GHG emission estimation.

4.3.2 Industry classification

Industry classification features allows the model to grasp the nature of the assets of a company by understanding in which fields they are used. The GHG emission profile of companies operating in different sectors are different. There exists numerous industry classifications, grouping companies differently: this feature is critical for our model, as we do not want to rely on a classification which would never, at any level, make the difference between companies operating in the Oil & Gas, Renewable Energy or Nuclear fields. We rely on the industry classification feature which gave us the best results and is detailed enough so that the model is about to grasp granular differences in the companies profiles.

For each company, we get its main industry classification using the Bloomberg Industry Classification Standard (BICS). The industry in which a company is classified corresponds to the one in which it is making the biggest faction of its revenues. The BICS classification is a detailed one, with a good level of granularity with seven hierarchical levels. It makes granular distinctions between the different sector, going as far as distinguishing companies both 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, we do not have enough density for the deeper levels in our training dataset: not all companies have a 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 make it more prompt to overfitting, the model having more difficulties to generalize to other samples. In our preprocessing, we remove all occurrences of industries that are present less than 10 times in the training set. They are replaced by a NaN value, missing values being directly handled by the LightGBM model.

As precise as the BICS classification is, we complement it with the New Energy Exposure Rating from Bloomberg. It is a categorical feature which 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 the GHG emissions and allows 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 can be explained as 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 Tab. 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 emissions targets, we attribute the values reported between January and June of the year y to the year $y-1$ and those reported between July and December of the year y 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 in. 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 in. Data are 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 serie for the considered country is extended using the last known value.

The model also relies on a categorical feature describing if a system of carbon taxes or an Emission Trading System (ETS) have 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 on in the USA for instance).

4.4 High Quality Dataset

Using the features and target cleaning procedures, we obtain the final training datasets to estimate scope 1 and scope 2 GHG emissions. 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

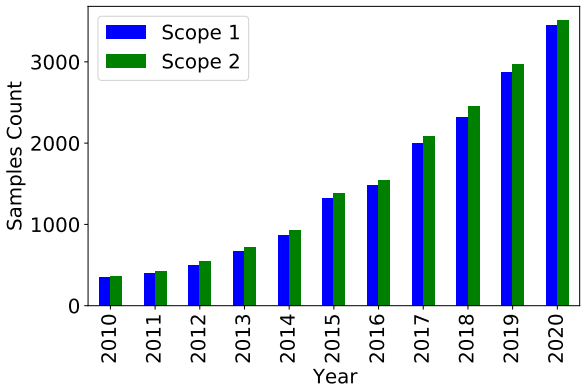


Figure 1: GHG emissions: number of companies with a reported emission per year for scopes 1 and 2

reported GHG emission per year was obtained. We observe an important increase of data quantity through the years, which illustrates the growing importance of GHG emissions 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 causal consecutive train, validation and test data sets. The model learns the mapping between features and target on the training set, determined its parameters on the validation one and is finally tested on the 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 which has not reported them during the last year. 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. 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. We want to rely on the entire time span of data we have.
- similarly, GHG emissions data are non-stationary, leading to inaccurate results using this standard splitting scheme.

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

As we want to estimate unreported data during the last available year, the test set built to evaluate our models should only include companies which are not in the training or validation sets: we want to estimate unreported emissions of companies which, most of the time, never reported their emissions before. Moreover, it allows us to avoid a potential bias: because of the huge year-over-year correlation of GHG emissions for a same company, having the same company both in training/validation and in

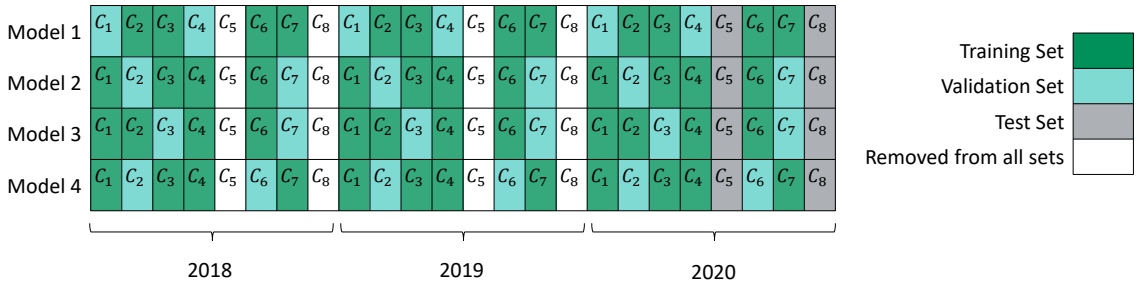


Figure 2: Company-wise cross-validation: the validation sets consists of randomly selected companies, which allows training to account for most of the most recent data.

test during different years would lead to an overfitted model. In practice, we build the test set by selecting 30% of the companies for which we have 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, we use a K-fold company-wise cross-validation: 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, we respect the framework we are in, 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 validating the model also with the most recent and and more relevant data. We take $K = 4$.

Figure 2 illustrates on a three years and eight companies dataset the used procedure to build our 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, build as explained in section 4.5.

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 each decile of revenues.

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

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} ||y_i - \hat{y}_i||^2}$$

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, refers in this paper as MAE, defined as

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

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:

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

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 different 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. Regarding the R^2 metric, it 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 than MAE larger errors: large errors are undesirable in the context of estimating GHG emissions, that's why we rely mainly on the RMSE metric in the remainder of this study.

Multiple test sets As shown in Fig. 1, we do not have a great quantity of samples to train our model: this leads to small test sets with around 800 data points. As a result, evaluation on the test set may be subject to a high variability: a few single wrongly estimated point could lead to an important deterioration of performance. We mitigate this issue by creating five different test sets and evaluate the model performances on these five test sets.

5.1 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 sets. In comparison to the literature like Goldhammer et al. (2017) or Griffin et al.

		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

(2017), we display only out-of-sample results as they are the ones that really 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. It means for instance that a RMSE of 1 corresponds to an error of one order of magnitude in terms of raw greenhouse gas emissions. 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.

5.2 Breakdown of performances by sectors, countries and revenues

Beside assessing the global performance of the models, we consider a breakdown of the models performances per sectors, per countries and per revenues: it allows for a transparent review of the performances of the model and to understand better its strengths and weaknesses.

Results are presented in Fig. 3 and 4 respectively for the 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 sector. Results are ordered from the highest to the lowest emissivity of the BICS Sector L2, computed on the full set of reported data. This figures highlights that the model has rather stable performances across all sectors, with in particular good performances on the most emissive sectors. This plots also highlights 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 subindutries with heterogenous GHG emissions schemes, this could explain why our model currently have a bit more difficulties in estimating emissions for some sectors. For instance, in the mining sector, depending on the chosen technique, one ton of aluminium production can create around 10 times more emission than one ton of steel production. The model performances for the most emissive BICS Sector L3 (Level 3 of granularity) is proposed in the Appendix A.

Figures 3b and 4b takes 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 Fig. 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, the 0^{th} is the one with the lowest. These graphs shows that in average, it is easier for the model to estimate the GHG emissions of companies with higher revenues. This may due to the fact that we have in our

training sets more samples coming from big companies, as shown in Tab. 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

To be more thorough, we now assess the quality of the estimates from our model, called the GHG-2022 model, in comparison to other data providers. These comparisons are done as of August 2022.

Retraining the model on the full dataset In section 5, the models performances are evaluated using test sets. The samples in those test set 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 models on the whole data, without test sets. Validation sets are still required to find the best models hyperparameters.

We consider the universe of 48429 companies extract from the Worldscope Refinitiv Database in 2020 to evaluate the prediction of our models and to compare them to other providers ones.

6.1 Comparison of coverage

Figure 5 displays for the scope 1 and scope 2 the number of reported GHG emissions and estimated GHG emissions each providers can provide for the year 2020. The test was conducted on the full universe of 48429 companies: for instance, our GHG-2022 model can provide for the scope 1 4360 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: we do not want our model to extrapolate on categories unseen during calibration.

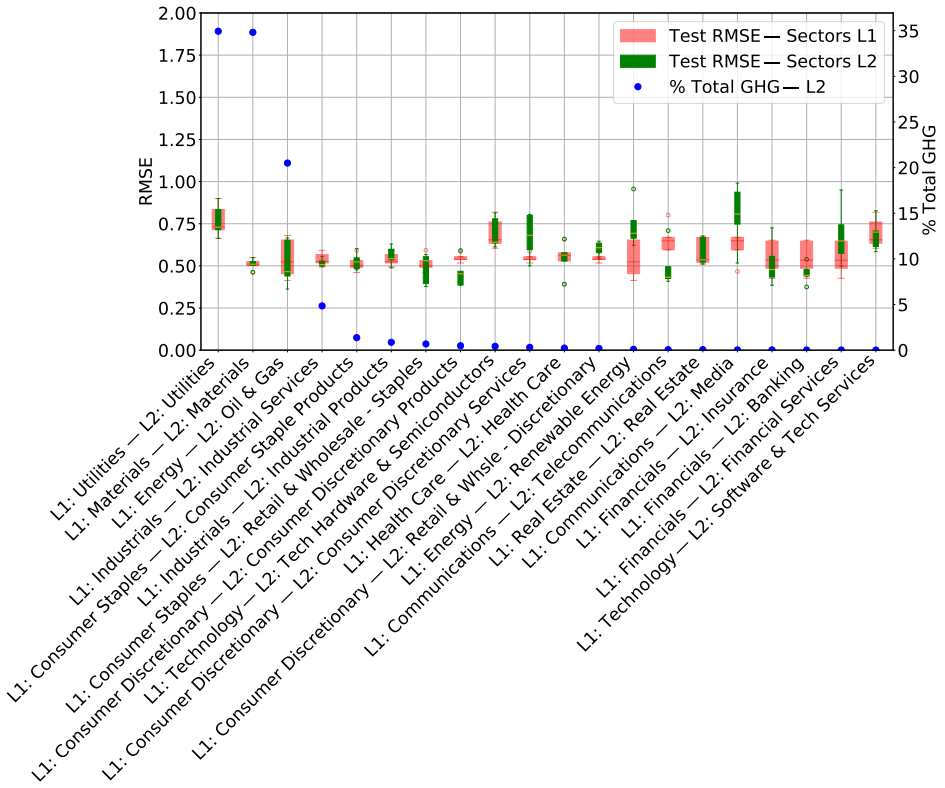
Coverages for Bloomberg, Trucost, Sustainalytics, MSCI and CDP are rounded as we may have slightly different results depending on the moment the datasets were obtained. Results provided in Fig. 5 were obtained using the elements at our disposal in August 2022.

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

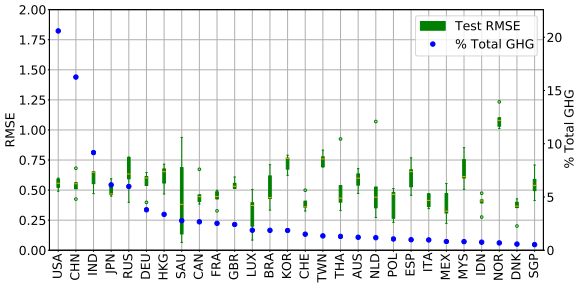
6.2 Comparison of estimates accuracy

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:

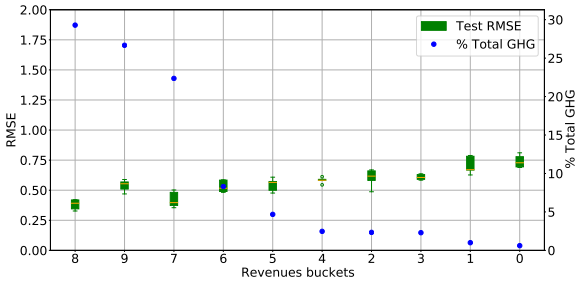
- Using the same procedure, we train a model relying only on 2010 to 2018 data and a second one relying only on 2010 to 2019 data. These models, when used for predictions on 2018 and 2019 data respectively give 2018 and 2019 point-in-time estimates.
- We consider the reported values in 2020 for companies which started reporting in 2020 and thus have never reported in 2018 nor 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.

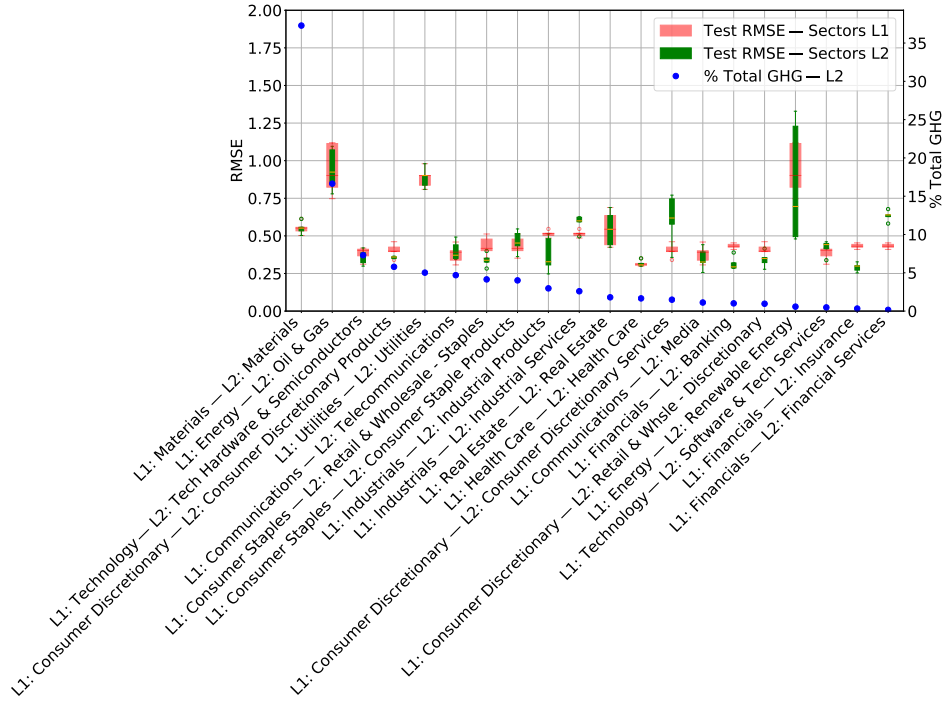


(b) Boxplot of test RMSE on the five different test sets per countries, ordered by countries emissions.

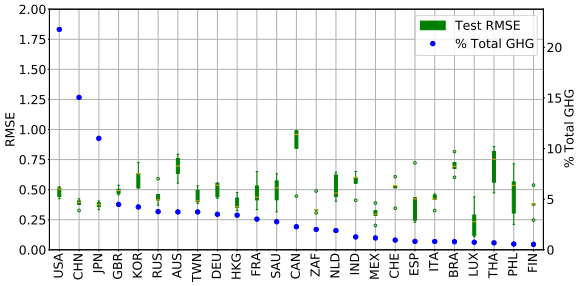


(c) Boxplot of test RMSE on the five different test sets per deciles of revenues, ordered by revenues deciles emissions.

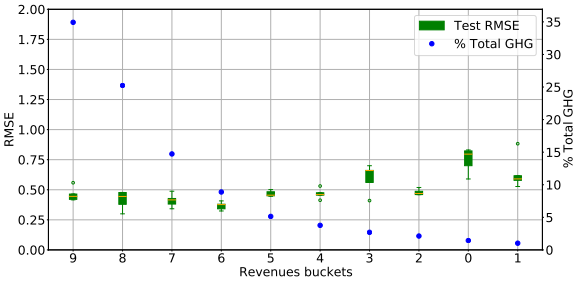
Figure 3: GHG emissions scope 1: distribution of performances of the model on five test sets according to different characteristics of companies.



(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.

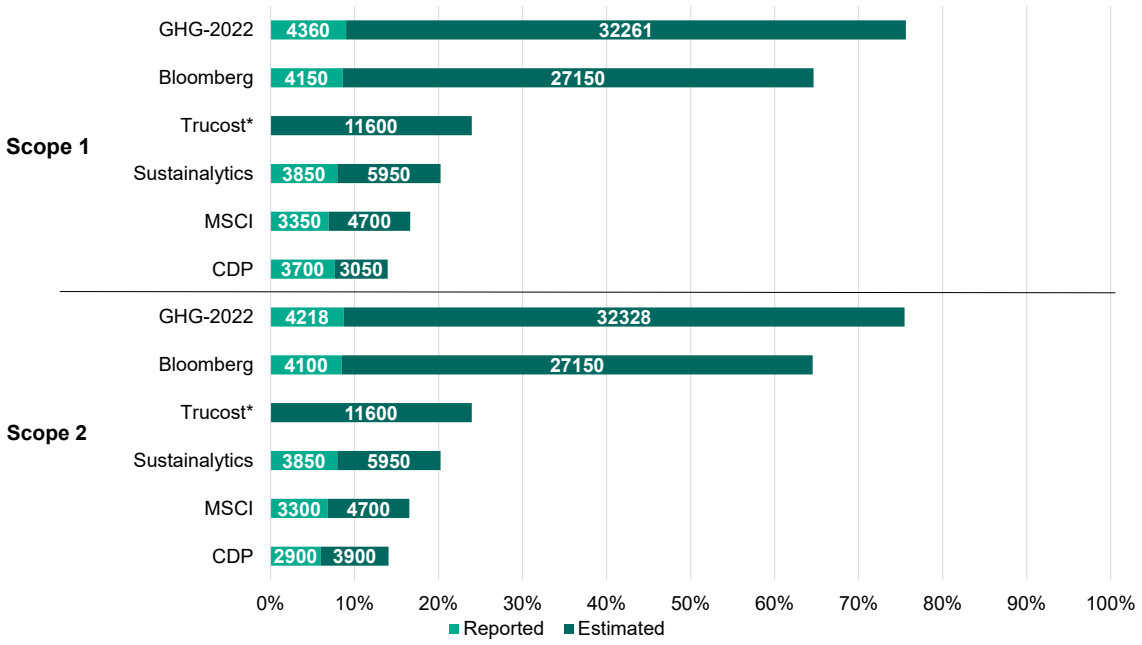


Figure 5: GHG emissions coverage, as of August 2022: number of reported data and estimates provided by each model. For the providers marked with an asterisk, the split between reported and estimated data was unclear, so all data points are marked as estimates.

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 the 2019 is not available) from our GHG-2022 model and the ones from the other providers models to the 2020 ground truth, we can 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 a 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 respectively. The number of samples may greatly differ according to the coverage of the provider in estimates for companies which started reporting in 2020. Our model has the best, i.e. lowest, RMSE in comparison to the other considered providers.
- Second, we consider each provider against our GHG-2022 model. Results are available in tables 5b and 6b for scope 1 and scope 2 respectively. We rely this time on the same samples for GHG-2022 and the considered provider, increasing the comparability. In each case, we are systematically more accurate than the considered provider.

Point-in-time data We train models using only 2018 and 2019 data respectively to avoid any leakage of the future in our 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 we are sure estimates are done point-in-time is CDP. Even considering this, our model still have better performances than the considered providers.

Breakdown of performances per sectors The methodology developed to compare our model to providers can be extended per sectors. In this section we only focus on the provider CDP as it is the only one for which we are sure estimates are done point-in-time, 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 Fig. 6: in green, we observe the distribution of differences between CDP estimates and the ground truth, in pink, we observe the distribution of differences between our GHG-2022 estimates and the ground truth. Distributions from our model are more centered around 0, meaning better accuracy for our model than CDP. We can however notice than CDP estimates are more conservative than ours: when CDP estimates are not exact, they have a tendency to overestimate whereas our GHG-2022 model is rather balance between overestimation and underestimations. Both behavior and calibration from our model can have their strengths and weaknesses depending on the use case.

7 Interpretability - Shapley values

In the remainder of this work, we will provide tools to get insights on how the model works and why it estimates such values of GHG emissions. We provide a breakdown of the impact of the different training features on the estimated emissions from the models. 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 combinations 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 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

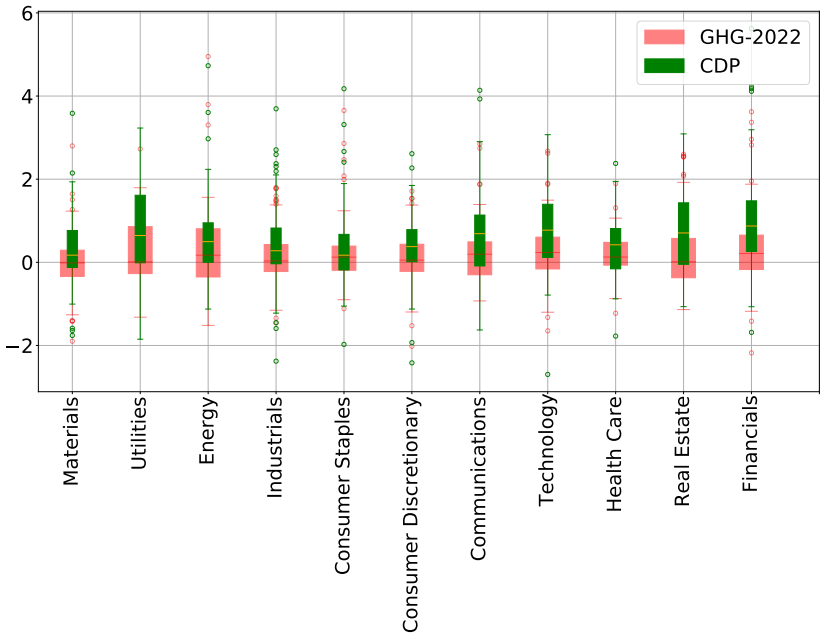


Figure 6: Differences of the emission from estimations from CDP and from GHG-2022 with ground truth for scopes 1 and 2.

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:

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

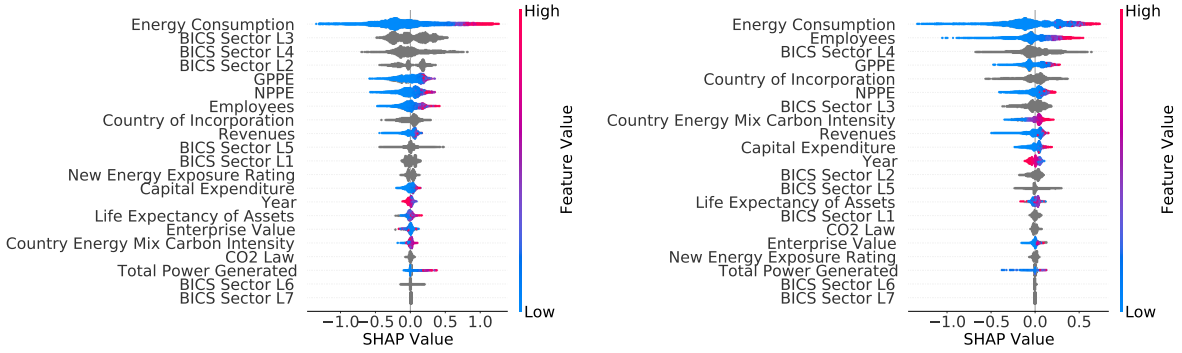
The dummy property states that the Shapley value of a feature which does not change the prediction, whatever combinations of features it is added to, should be 0.

Shapley values 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 to approximate Shapley values for trees models like the LightGBM, which we use in the following and refer to as SHAP values.

7.1 SHAP feature importance

We provide in Fig. 7 the breakdown of SHAP values per features 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 samples in our training set. These graphs are key elements in our model as they make it interpretable: they can be computed for any set of features, allowing to understand why the model make a specific decision and output this predicted estimate.

The Energy Consumption feature is the most important one used by the model for both scope 1 and scope 2, which seems logic. Industry classification features are also very important for both



(a) Scope 1

(b) Scope 2

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

scopes. As we could have expected from the definition of scope 2, the features Employees, Country of Incorporation and Country Energy Mix Carbon Intensity are more important for the estimation of scope 2 than the estimation of scope 1.

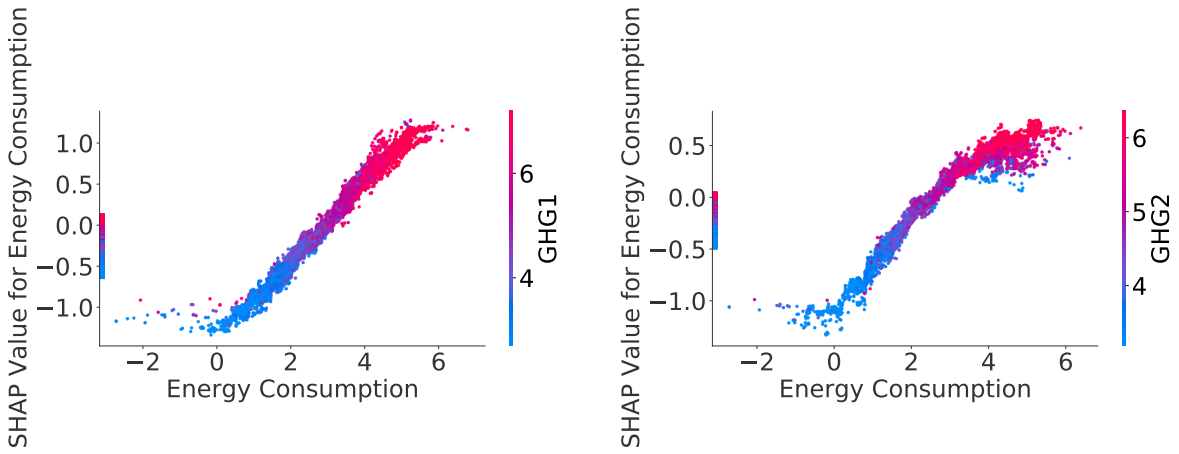
Knowing these SHAP values not only allows to better understand an estimate the model output but also to evaluate the reliability of the estimate based on the presence or the absence of a feature: if the Energy Consumption feature is not given for a sample, for certain sectors, it would lead to a less reliable estimate.

7.2 Relationship between features values and GHG estimates

Numerical features SHAP values can be computed for each feature on each sample. This tool allows to better 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, Fig. 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 represents 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. Appendices B.1 provides some others examples of SHAP values plots for numerical features, allowing for a better interpretation of what the model is learning.

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, Fig. 9 shows the distribution of SHAP values for the BICS Sector L1 feature, for each of the BICS Sector L1 sectors. Thanks to this plot, we can understand 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 to the estimated emission (positive SHAP values). On the contrary, samples in the Financial sector have negative SHAP values, showing a decrease to the estimated emission.

This plot can be done for all categorical features, allowing to understand the distribution of SHAP values according to each category, and then having a better interpretation of the model. In Appendix



(a) Scope 1

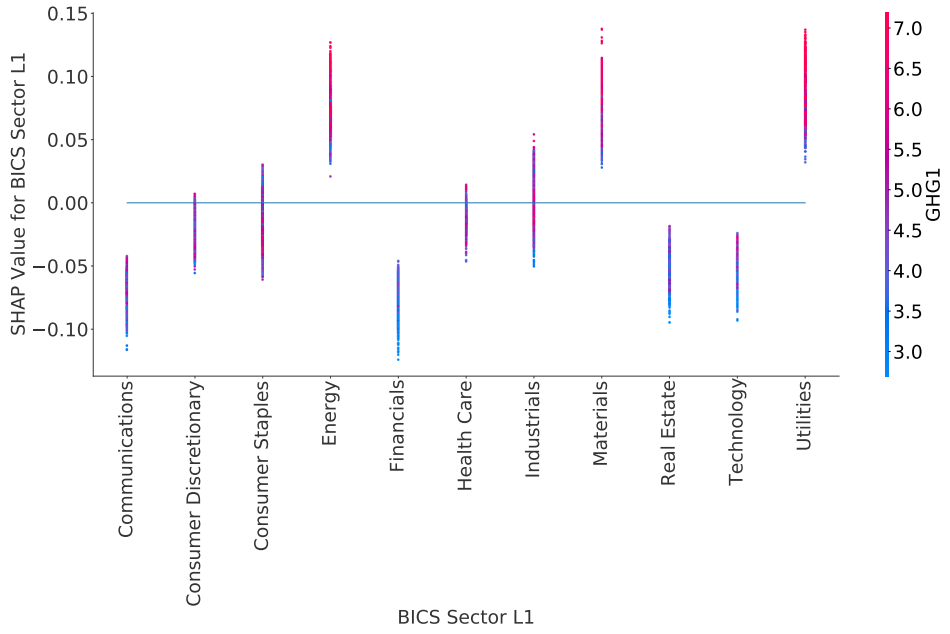
(b) Scope 2

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

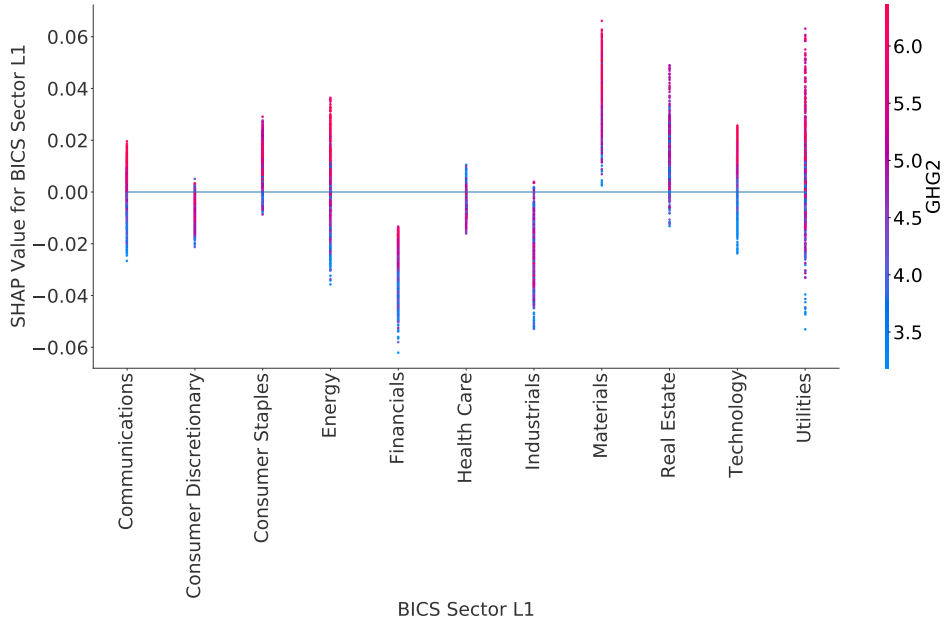
B.2, we provide an additional SHAP values plot for categorical features for further interpretation elements.

Plots in Fig. 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, working with the distribution of SHAP values for the Utilities sector in scope 1, we see that there is a cluster of SHAP values below 0.04 with few samples. These corresponds 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. The removal of this cluster with very few samples allows us to improve the quality of the training data by removing outliers. Similar studies on other sectors leads 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 automatized and applied systematically. We did a first implementation using the SHAP distribution for each BICS Sector L4 (Level 4 of granularity). For each L4 sector, we applied a hierarchical clustering algorithm, separating clusters if their distance is above 0.04 in the SHAP values space, and remove clusters of data with an insufficient number of sample, i.e. less than 10. All these parameters were found by trials-and-errors. 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 of the global performance of the model. Results are presented in Tab. 7, on average on the 5 different test sets: we observe for both scopes an average RMSE decrease between 11% and 13%. Let's however note that it may come at the price of an increase of variability between results in the different test set, especially for scope 1, as shown by the standard deviation of the RMSE metric in Tab. 7a. 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.



(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
$[0, +\inf[$	RMSE	0.578	0.007	0.501	0.020
$[0, +\inf[$	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
$[0, +\inf[$	RMSE	0.521	0.031	0.464	0.025
$[0, +\inf[$	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.

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 predict non-reported company emissions for scopes 1 and 2. Our 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, per country and per buckets of revenues. Comparing our results to those of other providers, we found, as of August 2022, our estimates to be available for a larger number of companies and more accurate.

In addition to the large coverage and the accuracy, this model is also flexible, allowing for easy evolution in the input data as regulations evolve. We built a transparent and explainable model: the methodology is described in this study extensively, and the implemented tools based on Shapley values allow to understand the role played by each feature in the construction of the final output. We focused on some interpretability elements we found particularly interesting. Many more interactions could have been studies: 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 to understand the specificities of GHG emissions per sectors. 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 our model will first focus on gathering and including more training data from SMEs, so that the coverage of our model can be further improve. As we stressed in our analysis, the used industry classification is critical. We will gather data about all the activities a company is reporting being active in, and work on including this new industry classification in our model: it will help in improving accuracy in industries where companies operating in very different sectors in terms of GHG emissions have been grouped. The data polishing method that we introduced in the last

section of this study will also be developed with the goal of obtaining a more robust model.

Funding

This work originates from a partnership between CentraleSupélec, Université Paris-Saclay and BNP Paribas. This research received no external funding.

Author Contributions

Conceptualization, J.A. and T.H.; Methodology, J.A., T.H. and L.C.; Software, J.A.; Validation, J.A. and T.H.; Formal Analysis, J.A. and T.H.; Investigation, J.A. and T.H.; Resources, T.H. and L.C.; Data Curation, J.A. and T.H.; Writing – Original Draft Preparation, J.A. and T.H.; Writing – Review & Editing: J.A., T.H., L.C. and F.S.; Visualization, J.A.; Supervision, L.C. and F.S.; Project Administration, J.A., T.H., L.C. and F.S.

Data Availability Statement

Restrictions apply to the availability of some data. Data obtained from Refinitiv, Bloomberg, MSCI, Trucost and CDP are available from the authors with the permission of respectively Refinitiv, Bloomberg, MSCI, Trucost and CDP.

Data from WorkldBank and the International Energy Agency are publicly accessible and can be found respectively here `https://carbonpricingdashboard.worldbank.org` and here `https://api.iea.org/stats/indicators/CO2Intensity`.

Conflicts of Interest

The authors declare no conflict of interest.

Abbreviations

BICS	Bloomberg Industry Classification Standard
BICS Sector L1	BICS Sector Level 1, first level of granularity
BICS Sector L2	BICS Sector Level 2, second level of granularity
BICS Sector L3	BICS Sector Level 3, third level of granularity
BICS Sector L4	BICS Sector Level 4, fourth level of granularity
CCS	Carbon Capture and Storage
CDP	Carbon Disclosure Project
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
GBDT	Gradient Boosted Decision Trees
GGLR	Gamma Generalized Linear Regression
GHG	Greenhouse Gases
GPPE	Gross Property Plant & Equipment
GWP	Global Warming Potential
IEA	International Energy Agency
MAE	Mean Absolute Error
MMD	Mean Maximum Discrepancy
MSE	Mean-Squared Error
NPPE	Net Property Plant & Equipment
NZBA	Net Zero Banking Alliances
OLS	Ordinary Least Squares
PA	Process Analysis
PACTA	Paris Agreement Capital Transition Assessment
PCAF	Partnership for Carbon Accounting Financials
RMSE	Root Mean-Squared Error
SEC	Securities and Exchange Commission
SME	Small and Medium-sized Enterprise
WBCSD	World Business Council for Sustainable Development
WRI	World Resources Institute

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, François Soupé, and Raul Leote de Carvalho. Corporate carbon footprint: A machine learning predictive model for unreported data. *Available at SSRN*, 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 an estimate models, 2017. URL `refinitiv.com/ESG`.

Securities and Exchange Commission (SEC). The enhancement and standardization of climate-related disclosures for investors, Proposed rules., 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 Fig. 10a and 10b, we provide for transparency purposes the breakdown of the out-of-sample performances of our model across the five test sets for the different BICS Sector L3, ranked from high to low emissivity, for sectors accounting for at east 1% of the total GHG emissions of the reporting companies.

B Relationship between features value and GHG estimates

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

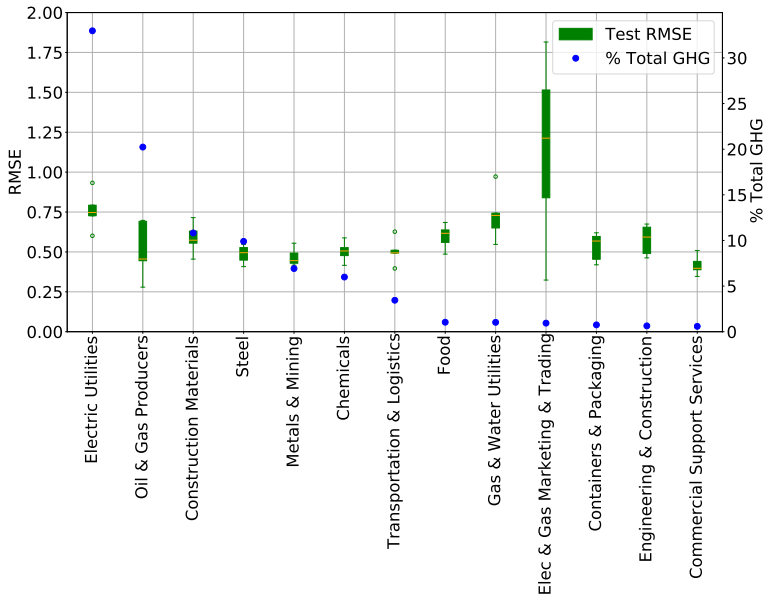
B.1 Numerical Data

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.

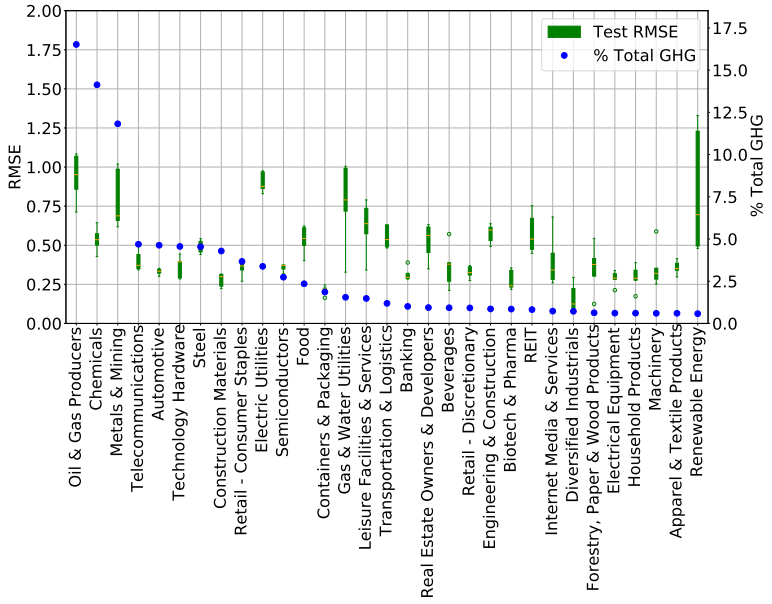
Figure 12 shows a near-linear relationship between the SHAP values of the Employees feature and the decimal logarithm of the Employees feature values, apart from the few points on the Y-axis referring to missing data.

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.

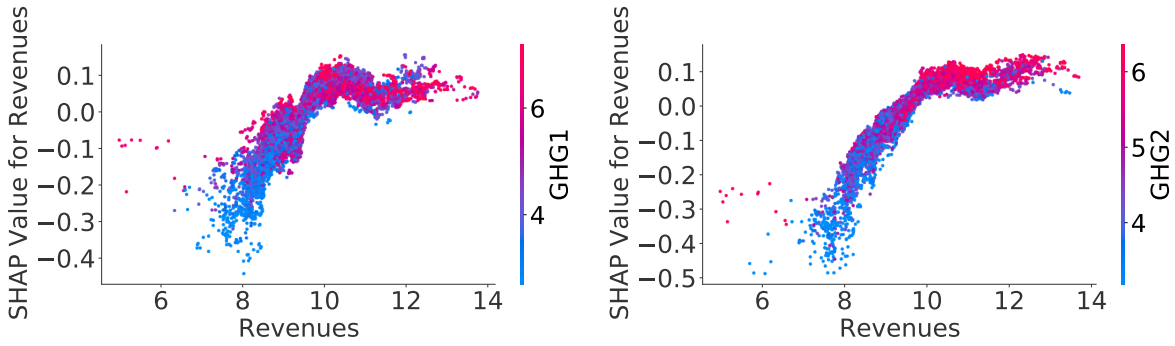


(a) Scope 1



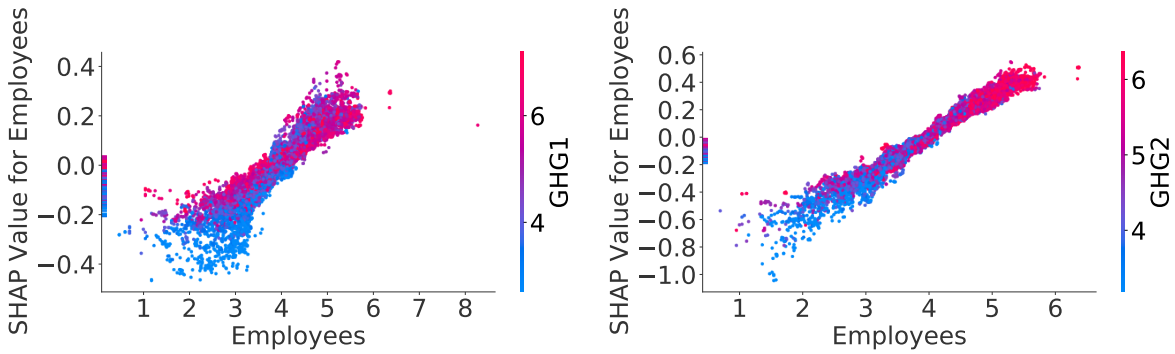
(b) Scope 2

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



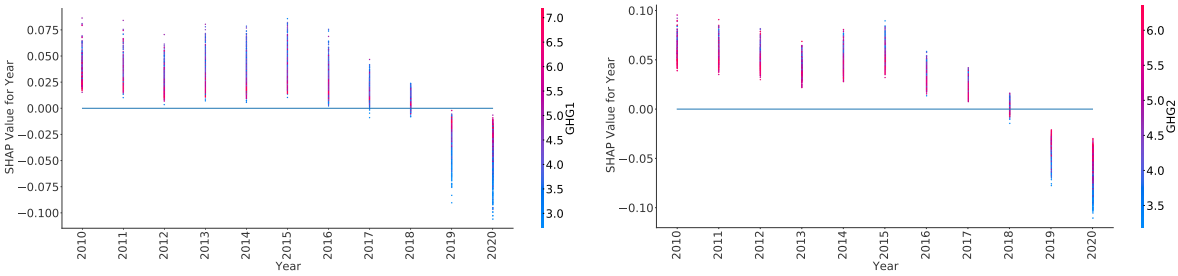
(a) Scope 1 (b) Scope 2

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.