Marketing Communications - For Professional Investors

MACHINE LEARNING PREDICTION OF CORPORATE CARBON EMISSIONS

RAUL LEOTE DE CARVALHO – QRG
PARIS, 04/04/2023

BNP PARIBAS ASSET MANAGEMENT

The sustainable investor for a changing world

Motivation

Fund managers need to know about company’s GHG emissions

- Acknowledge the need to accelerate the transition towards net-zero
  - Net Zero Asset Managers initiative (NZAM): 301 signatories with USD 59 trillion in AUM
  - Glasgow Financial Alliance for Net Zero (GFANZ): 550+ members from 7 sector-specific net-zero alliances
- Companies with higher carbon emissions face higher regulatory and legal action risks
- Environmental factors increasingly important for investors
- Understanding emissions to evaluate environmental sustainability

Many companies do not yet report GHG emissions

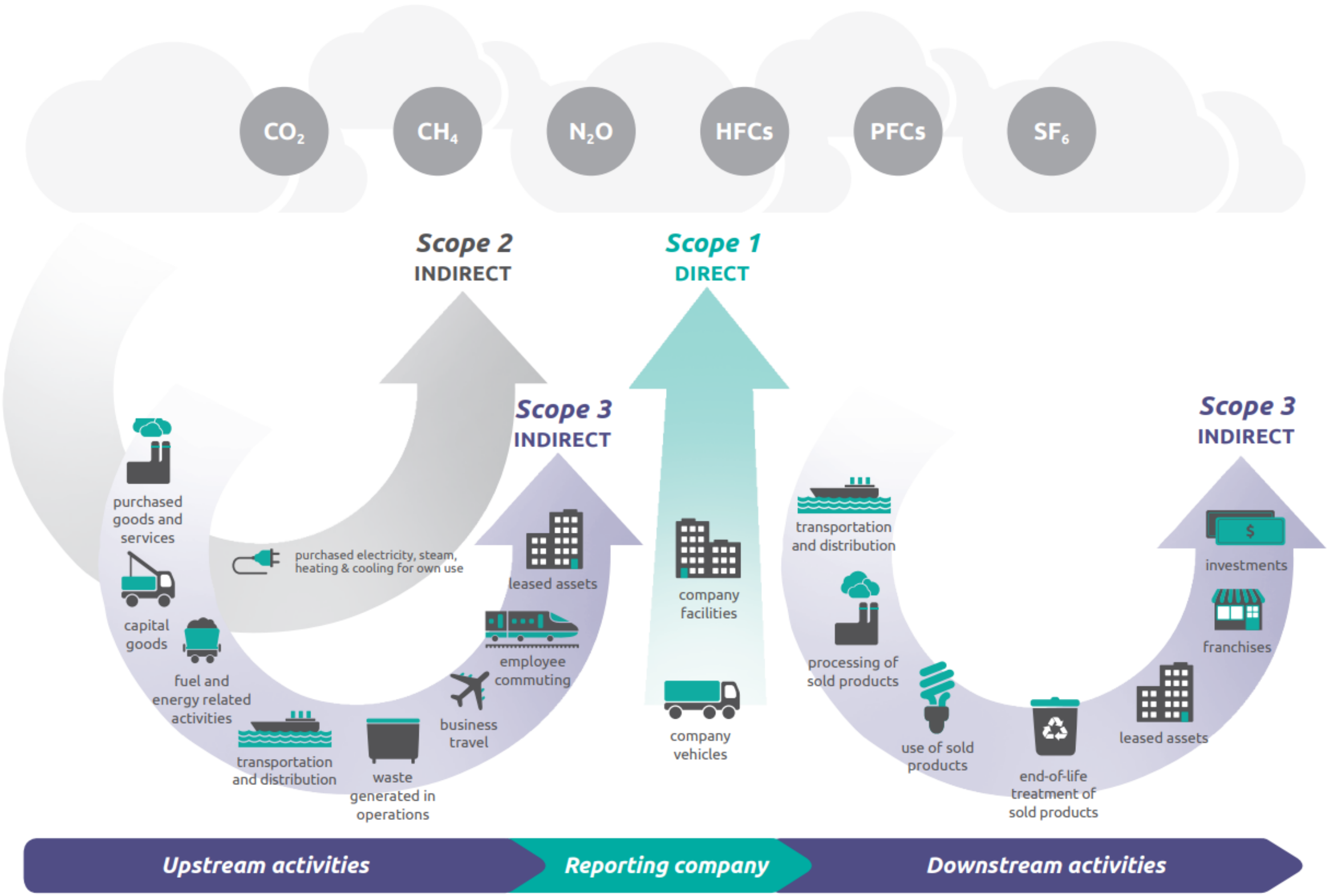
- Poor quality of existing models to predict unreported GHG

Statistical learning techniques generate accurate predictions of Scope 1 and Scope 2

- Good description of data across all industries

Scopes of Greenhouse Gas (GHG) corporate emissions

Overview of GHG Protocol scopes and emissions across the value chain

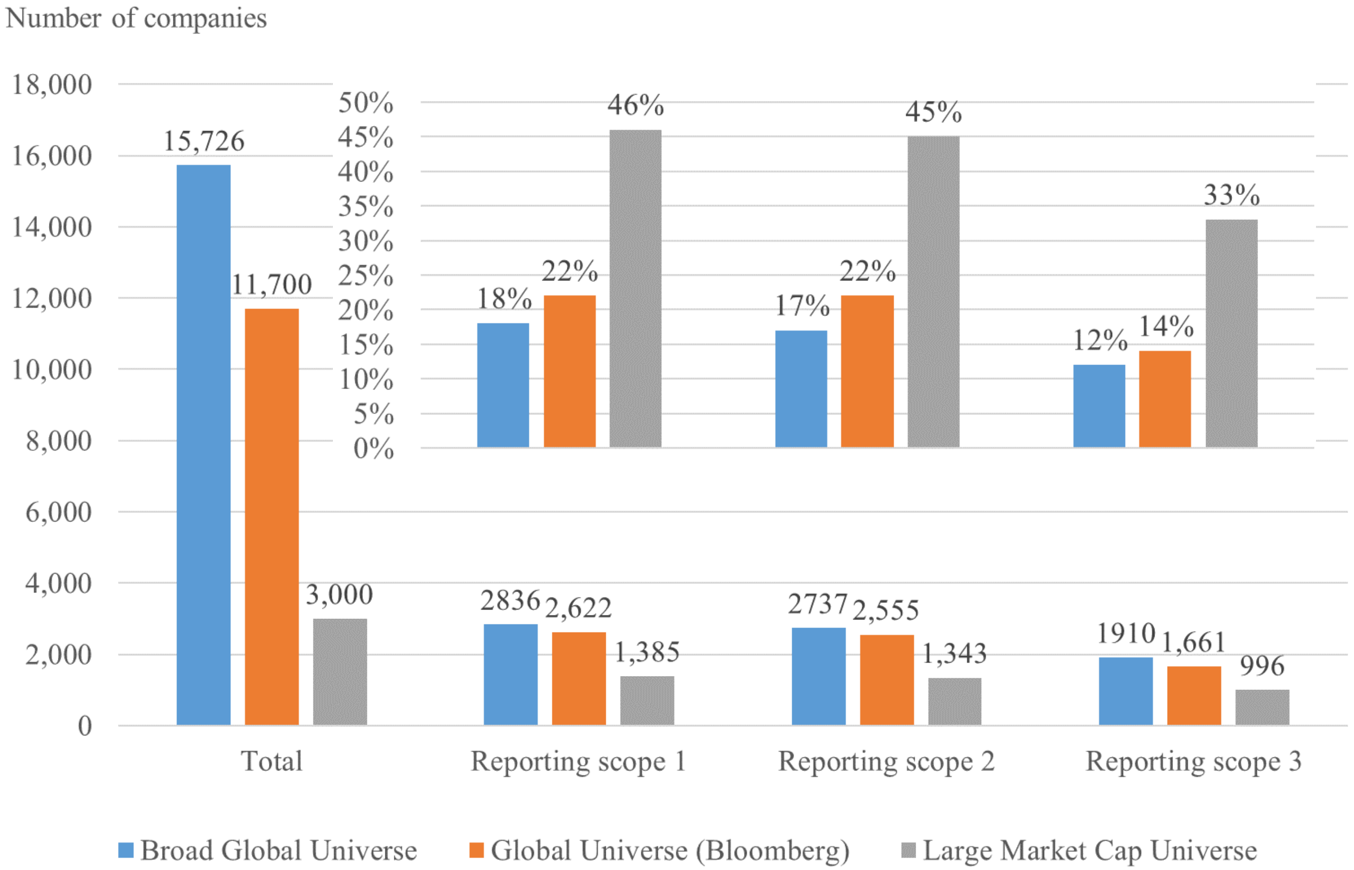


Source: Ranganathan, J., L. Corbier, P. Bhatia, S. Schmitz, P. Gage, and K. Oren. 2015. "The Greenhouse Gas Protocol: A Corporate Accounting and Reporting Standard, Revised Edition." World Business Council for Sustainable Development and World Resources Institute.

Reported GHG corporate emissions

Most GHG corporate emissions data is either estimated or unreported

Scope 3, with 17 items, is the least reported and most difficult to estimate

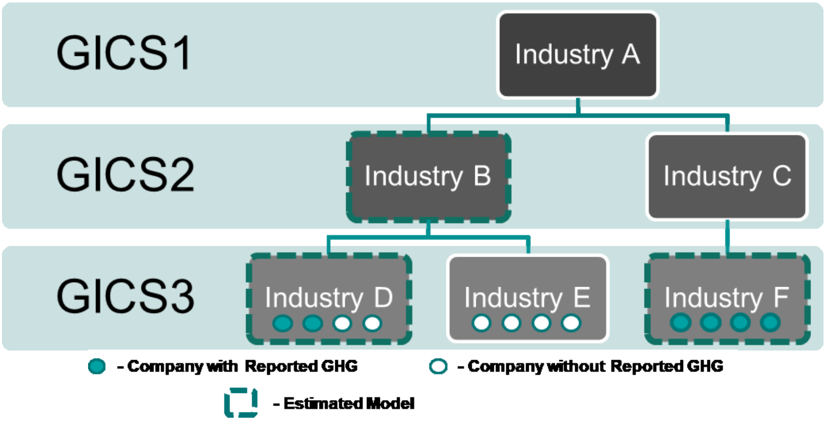


Source: BNPP AM, Bloomberg, Trucost, CDP. 31 Dec 2018.

Corporate GHG emission models

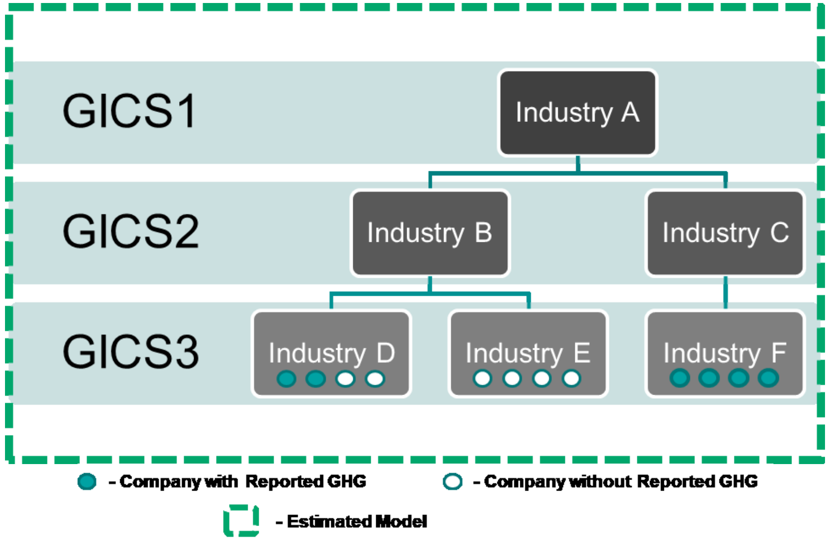
Other models

- *Create one model for each industry*
  - *Smaller data sets result in lower statistical significance*
  - *Poor models for industries with low reporting levels*



Our model

- *Create one single model with industry as a factor*
  - *Learn carbon predictive patterns from a larger data set*
  - *Can handle industries with lower reporting levels*



Other approaches to data quality

- *Unchecked errors in input data*
  - *Errors in data lead to less reliable models*

Our approach to data quality

- *Errors in input data checked and corrected iteratively*
  - *Outliers checked against company reported and corrected*

Source: Thibaut Heurtebize, Frederic Chen, François Soupé, and Raul Leote de Carvalho. The Journal of Impact and ESG Investing, Winter 2022, 3 (2) 36-54; DOI: https://doi.org/10.3905/jesg.2022.1.059. For illustration purposes only.

Predictors for scope 1 and 2 models

Categories	Sources	Description	Data range in period 2010 through 2018				Type
Industry Classification	Bloomberg	Sectors and industries	GICS 1, 2, 3 or 4 definitions with minor adjustments based on a more sustainability-oriented classification in particular for the utilities and energy sectors				Categorical
Regional information at country level	World Bank and International Energy Agency	Region	"United States and Canada", "Europe", "Asia / Pacific", "Africa / Middle East" or "Latin America and Caribbean"				Categorical
		Revenue group	"High income", "Upper middle income", "Lower middle income" or "Low income"				Categorical
		CO_2 tax regulations	"No CO_2 Law", "Subnational Implemented", "National Implemented" or "Regional Implemented"				Categorical
			Min	**Median**	**Max**	**Units**	
		CO_2 emissions	1	1,098	9,302	Million ton	Numerical
		Carbon intensity of energy mix	13	81	158	tons CO2 / TJ	Numerical
		CO_2 emissions per GDP	0.05	0.27	1.95	kg / USD	Numerical
		CO_2 emissions per GDP @ purchasing power parity	0.07	0.27	0.75	kg / USD	Numerical
Financial metrics at company level	FactSet, Refinitiv and Worldscope	Revenues	0.006	1,531	500,343	Million USD	Numerical
		Number of employees	1	5,450	2,300,000		Numerical
		Total assets (inc. Financials)	0.01	2,986	2,804,677	Million USD	Numerical
		Gross property plant and equipment	0.01	973	539,114	Million USD	Numerical
		Capex	0.0002	77	52,953	Million USD	Numerical
		Age of assets	0.001	19	99	Years	Numerical
Energy indicators at company level	Bloomberg	Energy production (1.8% data coverage)	0.1	27,559	653,900	GWh	Numerical
		Energy consumption (27.3% data coverage)	0.01	709	2,380,450	MWh	Numerical

Source: Thibaut Heurtebize, Frederic Chen, François Soupé, and Raul Leote de Carvalho. The Journal of Impact and ESG Investing, Winter 2022, 3 (2) 36-54; DOI: https://doi.org/10.3905/jesg.2022.1.059. For illustration purposes only.

Model base learners and meta methodology

Linear Regression Models

Examples: Ordinary Least Squares, Ridge Regression, Lasso Regression, Elastic Net

Choice: **Elastic Net -** linear regression model, avoids overfitting by reducing number of (as Lasso Regression) while minimizing size of coefficients (as Ridge Regression)

Pros: predictions are easy to explain from the model
Cons: not always as accurate as non-linear models

Non-linear Regression Models

Examples: Random Forest, Extremely Randomized Trees, Gradient Boosting

Choice: **Extremely Randomized Trees -** a version of **Random Forests with** additional layer of randomness when building the **Decision Trees,** generating many random split proposals and taking the best split available instead of building a **Decision Tree** that splits observations optimally

Pros: can model more complex interactions between predictors and predictors and is often more accurate
Cons: predictions can be hard to explain

Basic Combinations

Mean Combination, Median Combination, Maximum prediction combination

Choice: Maximum prediction combination for conservatism but also because of fit better with the training data set

Source: Thibaut Heurtebize, Frederic Chen, François Soupé, and Raul Leote de Carvalho. The Journal of Impact and ESG Investing, Winter 2022, 3 (2) 36-54;
DOI: https://doi.org/10.3905/jesg.2022.1.059. For illustration purposes only.

Model construction and evaluation

In sample & Out of sample

- *In sample*: universe of companies reporting emissions and with high data quality
- *Out of sample* universe of companies not reporting emissions or with poor quality data. It is used to compare model predictions with predictions from other models: S&P Global Trucost and Bloomberg, two data vendors

Model training

- Log-transformed reported emissions (normal distributed) are used. Many predictors are also log-transformed.
- Multiple rounds of cross-validation performed on different partitions
- In sample data was partitioning into 80% for training, 10% for validation and 10% for testing.
- Iterative approach to detect and correct errors in data, either with reported emissions or predictors, by investigating the predictions that fall further from reported data at each iteration

R^2 for model evaluation

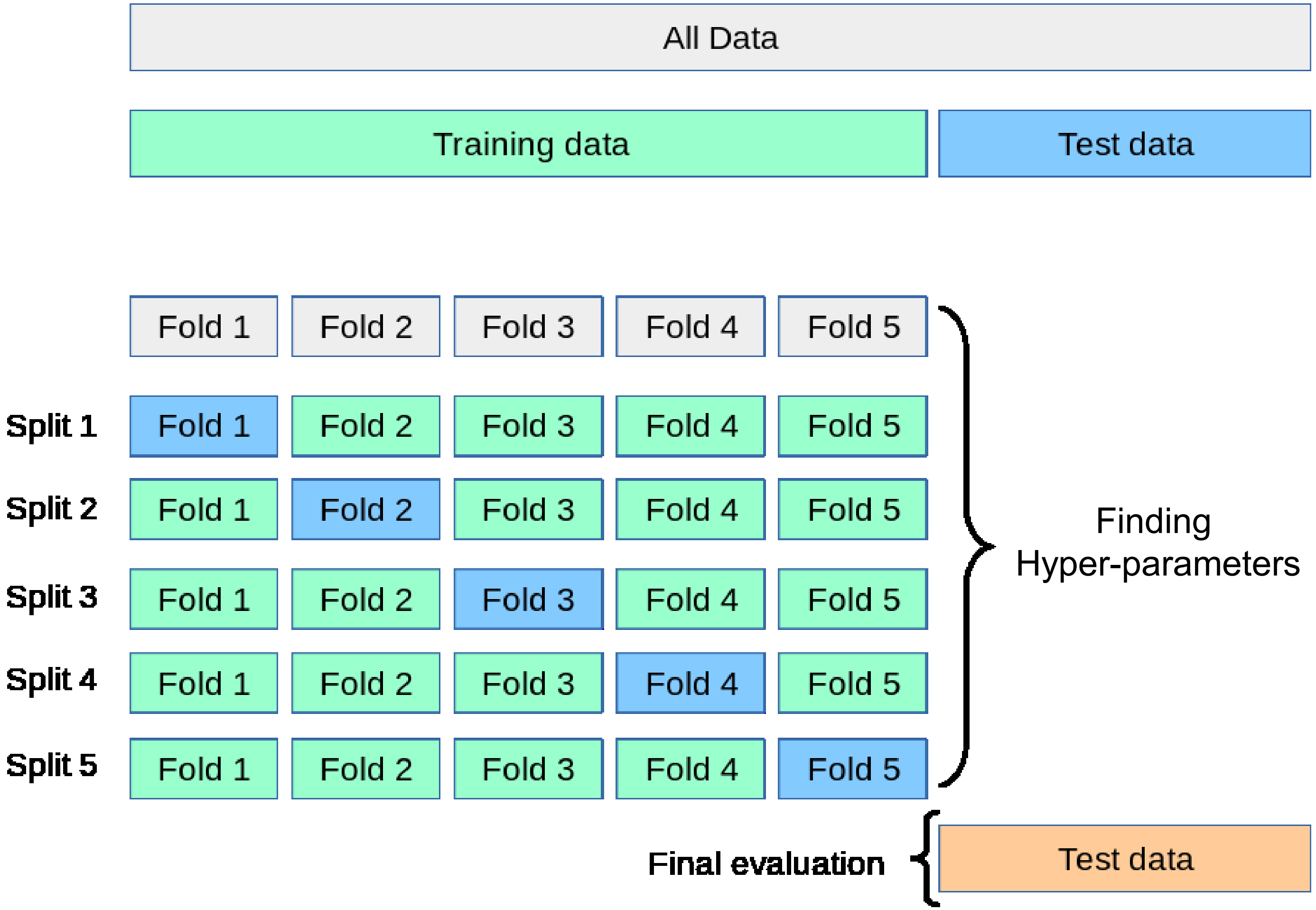
- Measures the accuracy of the model in sample

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

y_i = log-transformed of reported emissions of company *i* in the industry
$\hat{y}_i$ = log-transformed of predicted emissions of company *i* in the industry
$\bar{y}$ = average of log-transformed reported emissions of all reporting companies in the industry

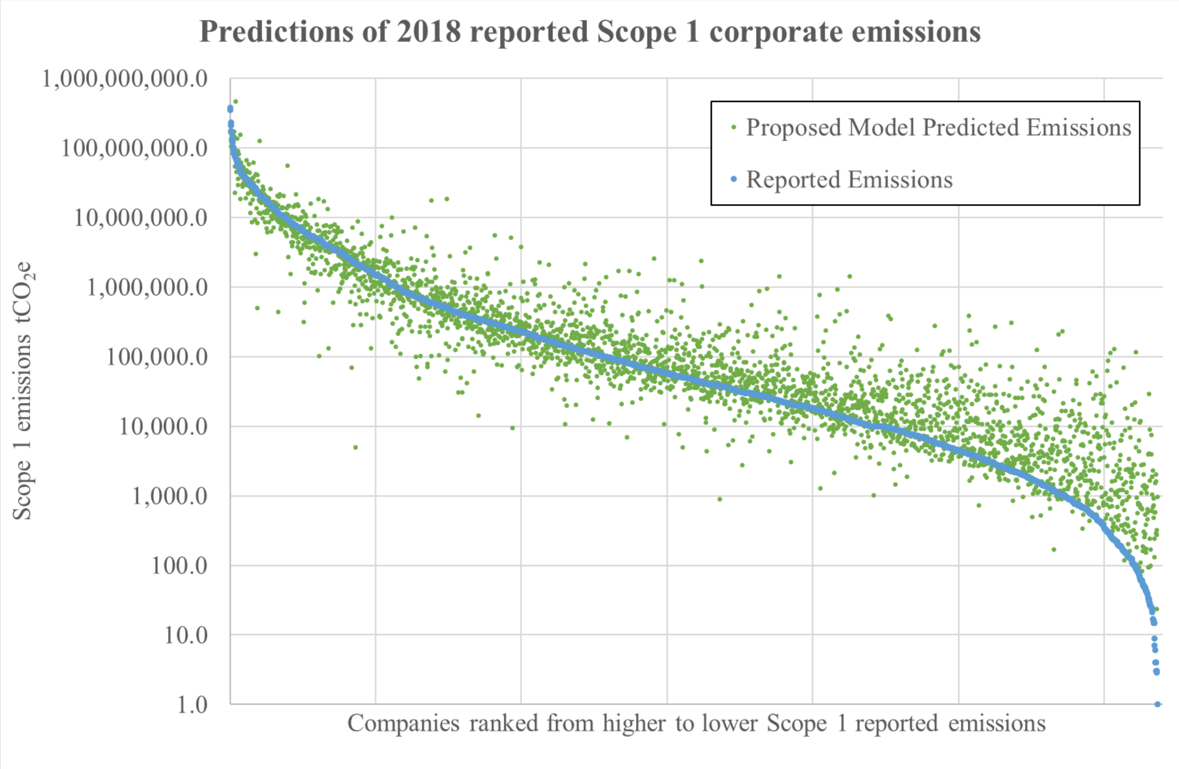
Source: Thibaut Heurtebize, Frederic Chen, François Soupé, and Raul Leote de Carvalho. The Journal of Impact and ESG Investing, Winter 2022, 3 (2) 36-54; DOI: https://doi.org/10.3905/jesg.2022.1.059. For illustration purposes only.

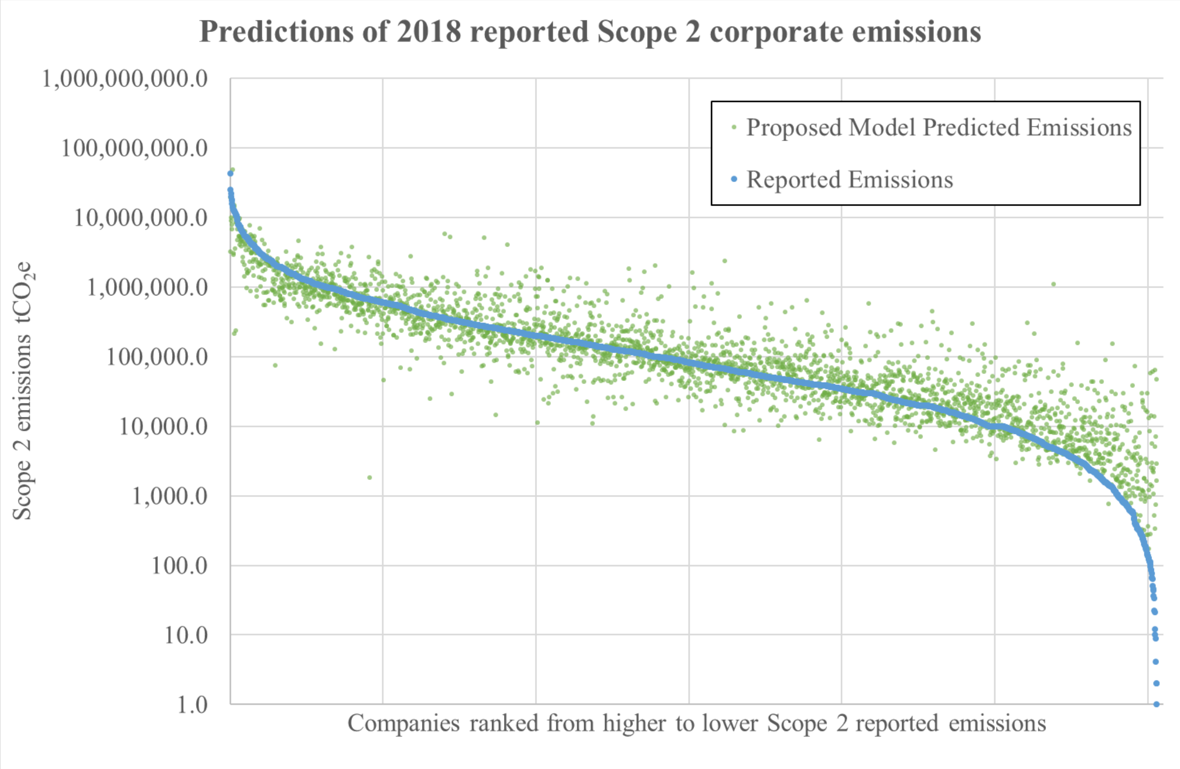
Cross validation



Source: Thibaut Heurtebize, Frederic Chen, François Soupé, and Raul Leote de Carvalho. The Journal of Impact and ESG Investing, Winter 2022, 3 (2) 36-54; DOI: https://doi.org/10.3905/jesg.2022.1.059. For illustration purposes only.

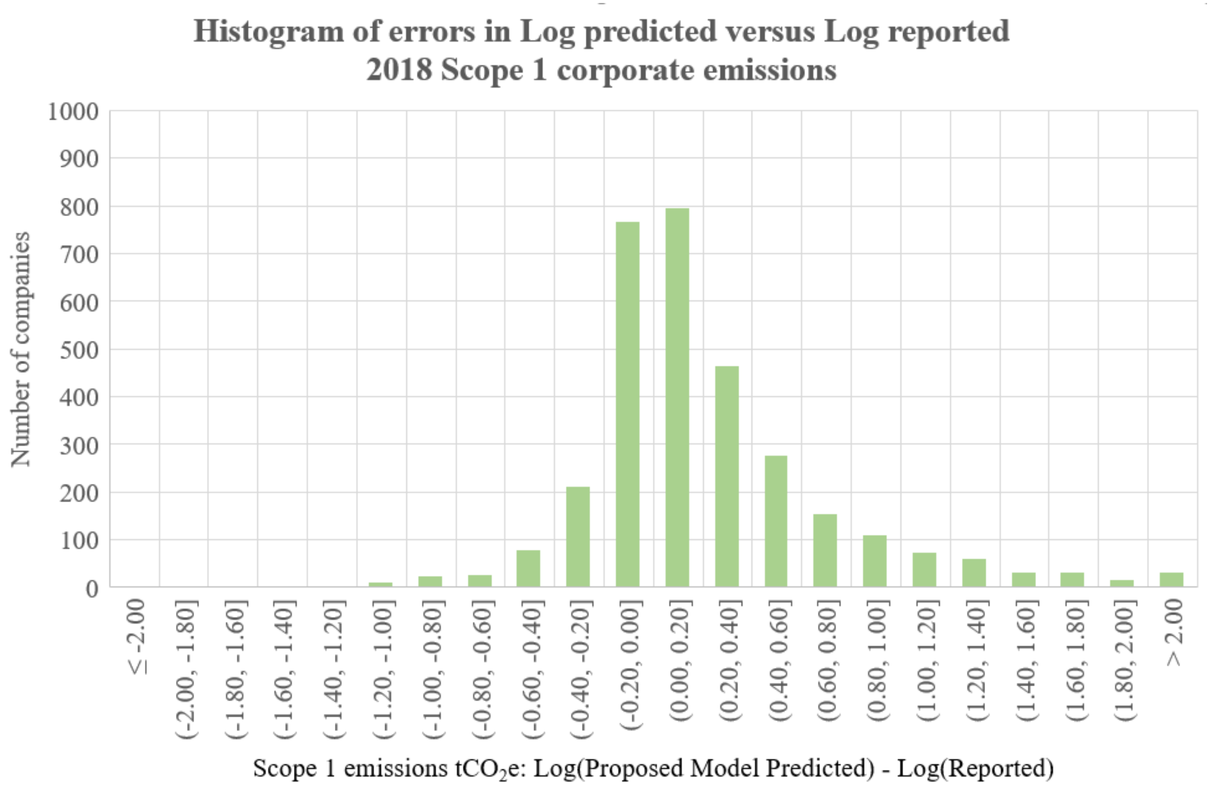
Corporate GHG emission models: in-sample

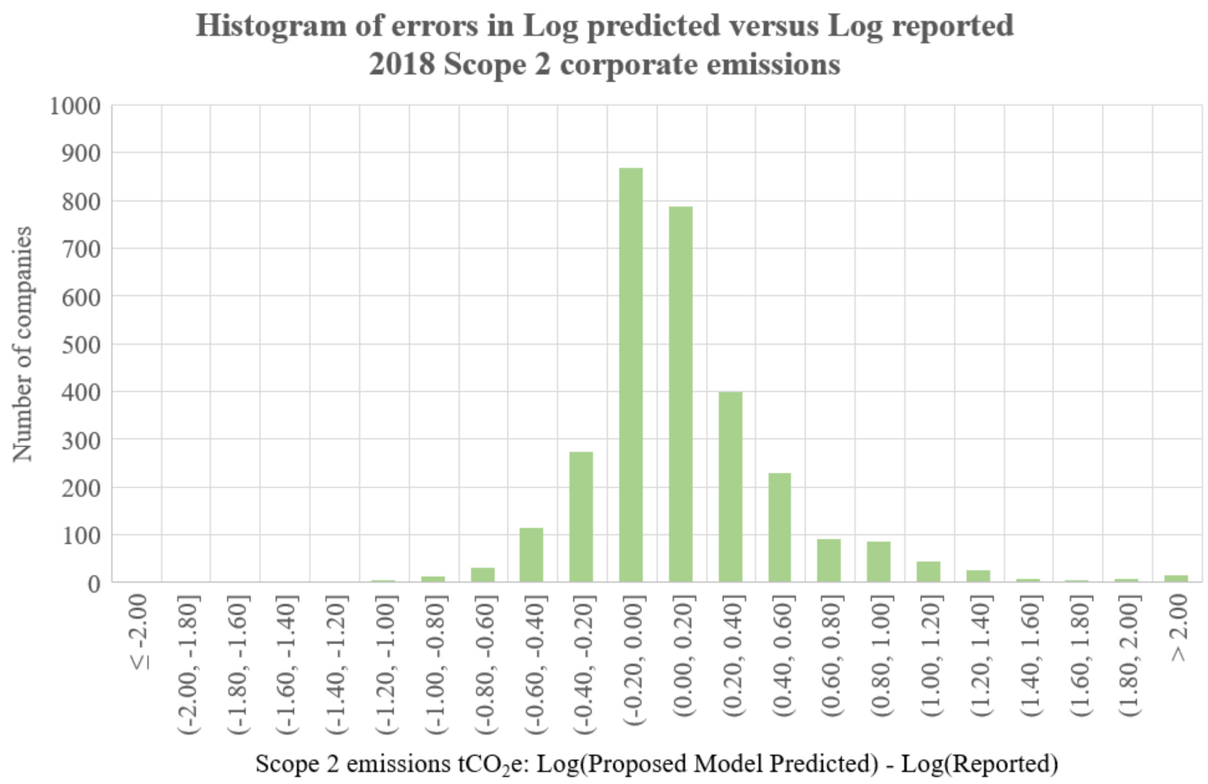




Source: Thibaut Heurtebize, Frederic Chen, François Soupé, and Raul Leote de Carvalho. The Journal of Impact and ESG Investing, Winter 2022, 3 (2) 36-54;
DOI: https://doi.org/10.3905/jesg.2022.1.059. Data for Dec-2018. Model data calculated by end of 2019. For illustration purposes only.

Corporate GHG emission models: in-sample





Source: Thibaut Heurtebize, Frederic Chen, François Soupé, and Raul Leote de Carvalho. The Journal of Impact and ESG Investing, Winter 2022, 3 (2) 36-54;
DOI: https://doi.org/10.3905/jesg.2022.1.059. Data for Dec-2018. Model data calculated by end of 2019. For illustration purposes only.

Corporate GHG emission models: in-sample

R^2 of the machine learning model for reported scope 1 emissions

GICS 2	Number of unique issuers	R^2 Maximum Prediction Combination	R^2 Extremely Randomized Trees	R^2 Elastic Net	Average scope 1 emissions (tCO_2e)	Share of total scope 1 emissions	Cumulative sum of share
Utilities	250	80%	89%	70%	21,122,718	40%	40%
Materials	547	84%	92%	78%	6,771,192	28%	67%
Energy	281	81%	88%	77%	9,046,819	19%	86%
Transportation	168	86%	94%	81%	5,230,486	7%	93%
Capital Goods	498	74%	85%	66%	689,198	3%	95%
Food Beverage & Tobacco	200	79%	89%	71%	727,153	1%	96%
Commercial & Professional Services	99	82%	85%	74%	991,566	1%	97%
Automobiles & Components	134	77%	85%	72%	451,435	0%	98%
Consumer Services	99	86%	93%	83%	523,669	0%	98%
Real Estate	256	65%	82%	55%	178,692	0%	98%
Food & Staples Retailing	57	76%	86%	64%	594,411	0%	99%
Diversified Financials	146	69%	79%	63%	202,358	0%	99%
Pharmaceuticals Biotechnology & Life Sciences	114	83%	90%	80%	215,333	0%	99%
Technology Hardware & Equipment	197	67%	78%	55%	114,656	0%	99%
Semiconductors & Semiconductor Equipment	95	68%	84%	60%	210,589	0%	99%
Household & Personal Products	42	74%	81%	66%	414,386	0%	100%
Consumer Durables & Apparel	143	59%	75%	49%	111,742	0%	100%
Retailing	120	72%	82%	68%	102,160	0%	100%
Telecommunication Services	93	77%	90%	66%	124,939	0%	100%
Banks	207	69%	85%	58%	51,071	0%	100%
Health Care Equipment & Services	89	69%	80%	68%	78,834	0%	100%
Media & Entertainment	74	75%	83%	72%	40,836	0%	100%
Software & Services	93	69%	80%	60%	25,935	0%	100%
Insurance	106	68%	85%	55%	18,320	0%	100%

$R^2 > 80\%$ for emissive industries (>2 M tCO_2e)

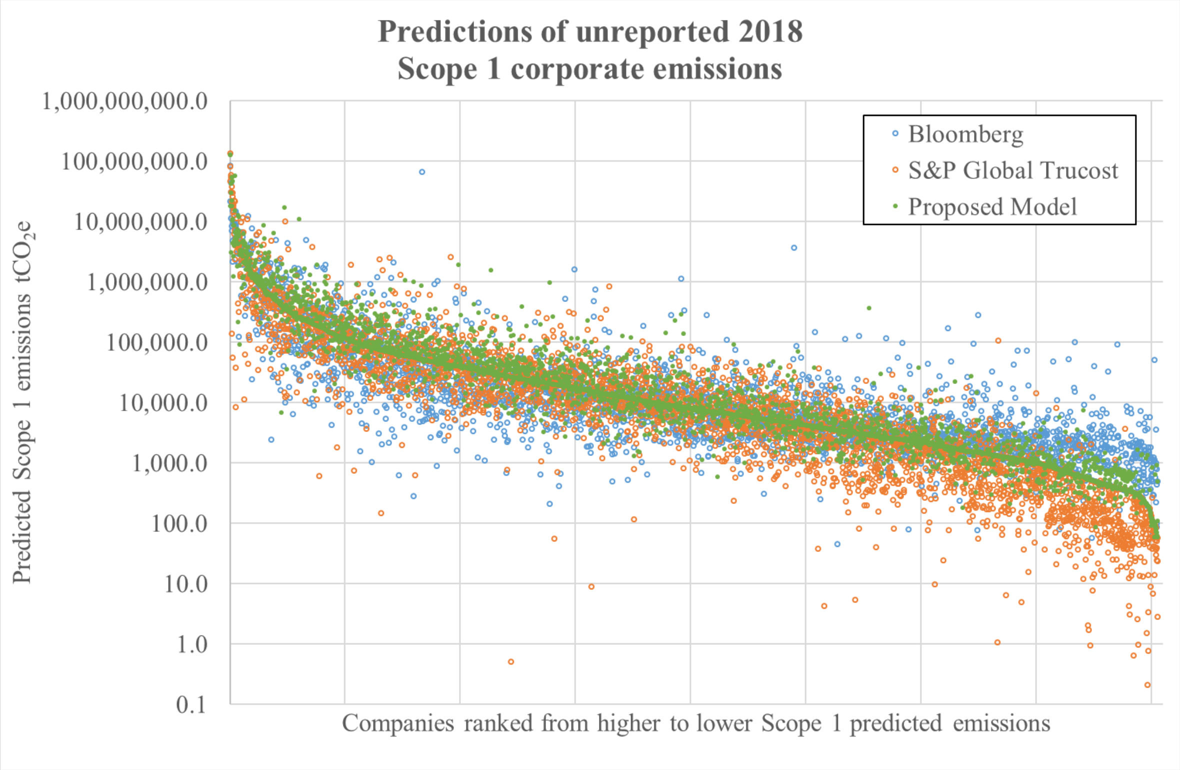
R^2 of the machine learning model for reported scope 2 emissions

GICS 2	Number of unique issuers	R^2 Maximum Prediction Combination	R^2 Extremely Randomized Trees	R^2 Elastic Net	Average scope 2 emissions (tCO_2e)	Share of total scope 2 emissions	Cumulative sum of share
Materials	531	76%	86%	61%	1,555,227	35%	35%
Energy	273	78%	86%	62%	1,081,211	13%	48%
Utilities	223	69%	69%	34%	997,460	9%	57%
Capital Goods	497	83%	87%	70%	308,777	7%	64%
Automobiles & Components	131	86%	88%	79%	883,613	5%	69%
Telecommunication Services	92	78%	89%	70%	1,046,732	4%	73%
Technology Hardware & Equipment	201	84%	93%	79%	443,214	4%	77%
Transportation	164	76%	94%	84%	483,537	3%	80%
Food Beverage & Tobacco	194	82%	88%	67%	390,322	3%	83%
Food & Staples Retailing	55	76%	90%	76%	1,227,244	3%	86%
Semiconductors & Semiconductor Equipment	96	77%	86%	67%	481,999	2%	88%
Consumer Services	99	86%	91%	67%	393,041	2%	90%
Retailing	122	88%	93%	78%	303,500	2%	91%
Real Estate	276	65%	92%	85%	120,302	1%	93%
Consumer Durables & Apparel	146	86%	87%	72%	220,952	1%	94%
Pharmaceuticals Biotechnology & Life Sciences	117	89%	84%	55%	265,355	1%	95%
Banks	222	78%	91%	81%	120,058	1%	96%
Household & Personal Products	44	86%	88%	81%	389,151	1%	97%
Software & Services	92	88%	95%	85%	162,513	1%	98%
Health Care Equipment & Services	89	85%	90%	85%	158,188	1%	98%
Media & Entertainment	79	83%	90%	80%	159,570	1%	99%
Diversified Financials	152	80%	83%	70%	66,254	0%	99%
Commercial & Professional Services	98	75%	83%	67%	87,704	0%	100%
Insurance	106	79%	90%	74%	54,924	0%	100%

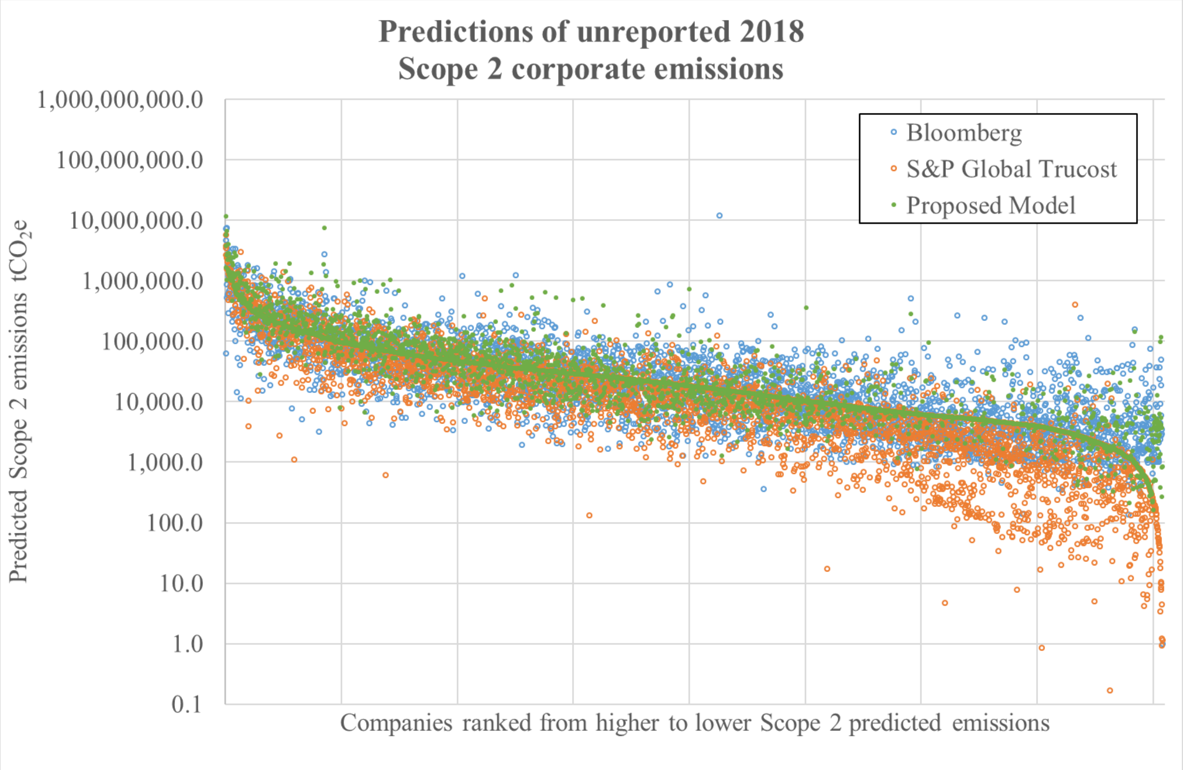
$R^2 > 75\%$ for 22 of the 24 industries

Source: Thibaut Heurtebize, Frederic Chen, François Soupé, and Raul Leote de Carvalho. The Journal of Impact and ESG Investing, Winter 2022, 3 (2) 36-54; DOI: https://doi.org/10.3905/jesg.2022.1.059. Data for Dec-2018. Model data calculated by end of 2019. For illustration purposes only.

Corporate GHG emission models: out-of-sample



7,762 companies, 2018
compared to Bloomberg and S&P Trucost

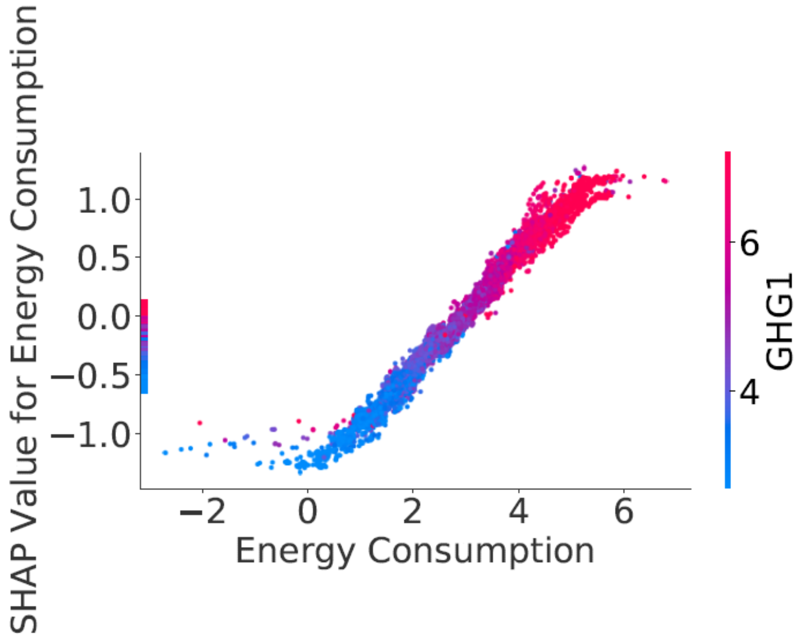


7,806 companies, 2018
compared to Bloomberg and S&P Trucost

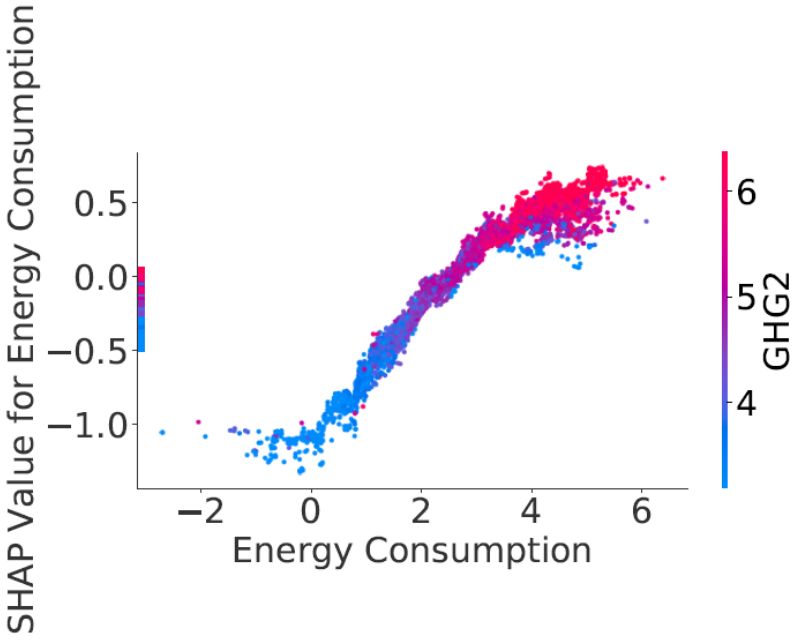
Source: Thibaut Heurtebize, Frederic Chen, François Soupé, and Raul Leote de Carvalho. The Journal of Impact and ESG Investing, Winter 2022, 3 (2) 36-54;
DOI: https://doi.org/10.3905/jesg.2022.1.059. Data for Dec-2018. Model data calculated by end of 2019. Bloomberg and S&P Trucost as of end 2020. For illustration purposes only.

Estimating contribution of predictors using SHAPley values

SHAP values are marginal contributions of predictors and calculated for each predictor
SHAP values for a specific predictor can be plotted against the predictor's values in the dataset



(**a**) Scope 1

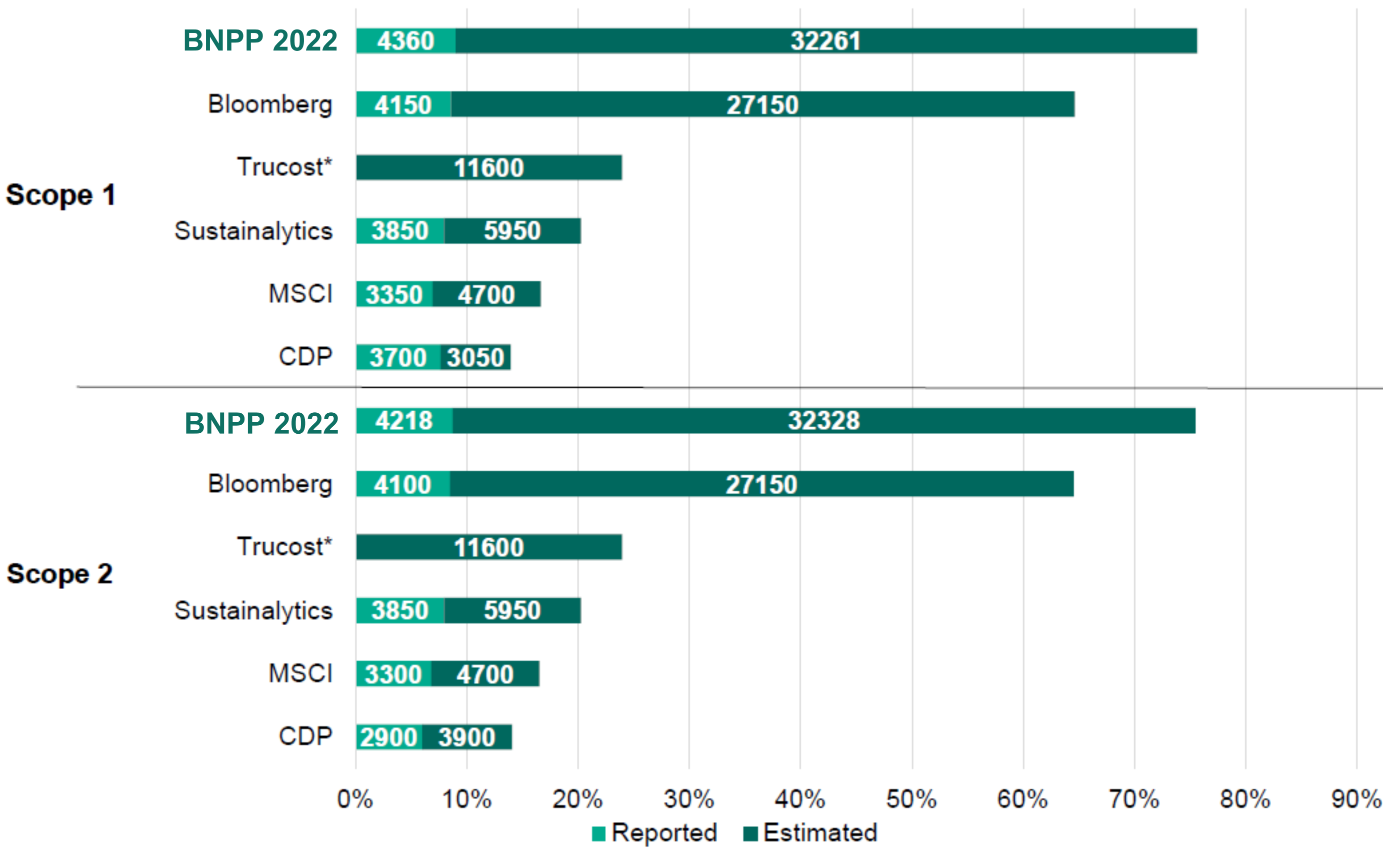


(**b**) Scope 2

Source: BNPP AM. Dec-2022. For illustration purposes only.

Model coverage vs data vendors

Coverage (48 429 larger issuers sample, as of August 2022)



Source : BNPP AM, Bloomberg, Trucost, Sustainalytics, MSCI, CDP. August 2022

Improvements over data providers:

1. **Larger coverage:**
   32k for BNPP ~4k for MSCI ~3k for CDP
2. **Higher accuracy**
   Good for all sectors and industries
   S1: Mean $R^2 = 83\%$ S2: Mean $R^2 > 75\%$
   test set: 30% companies with reported value in the last year
3. **Transparency**
   Based on SHAP values
4. **Full automation**
   Data cleaning and correction
5. **Flexibility**
   For future improvements or data correction

Targeting high quality outputs (~4k issuers)

- Highest quality data from **CDP level 7**:
  marked reported and reviewed by CDP
- Reported GHG emissions in the **CSR report** of the company:
  from Bloomberg

Using a small range of 21 input fields
available for 33k more issuers

Conclusions

Most companies do not report their GHG emissions

- How to predict carbon emissions for companies that have yet to report?

We propose a framework based on statistical learning techniques

- Predicts scope 1 and 2 GHG emissions more accurately than existing models
- Shows high in-sample accuracy for all industries at GICS 2 level
- Uses iterative approach to detect and correct errors

Can be used to predict unreported GHG emissions with greater accuracy

The Journal of Impact and ESG Investing, Winter 2022, 3 (2) 36-54
DOI: https://doi.org/10.3905/jesg.2022.1.059

Corporate Carbon Footprint:
A Machine Learning Predictive Model for Unreported Data

Thibaut Heurtebize, Frederic Chen, François Soupé, and Raul Leote de Carvalho

Thibaut Heurtebize is a senior quantitative analyst in the Quant Research Group at BNP Paribas Asset Management in Paris, France. thibaut.heurtebize@bnpparibas.com

Frederic Chen is a data scientist at BNP Paribas Corporate and Institutional Banking in Paris, France. frederic.chen@bnpparibas.com

François Soupé is co-head of the Quant Research Group at BNP Paribas Asset Management in Paris, France. francois.soupe@bnpparibas.com

Raul Leote de Carvalho is deputy head of the Quant Research Group at BNP Paribas Asset Management in Paris, France. raul.leotedecarvalho@bnpparibas.com

KEY FINDINGS

- We address the question of how to predict carbon emissions for companies that have yet to report theirs, by proposing a framework based on statistical learning techniques that predicts Scope 1 and 2 corporate carbon emissions more accurately than others in the literature.
- The model shows high in-sample accuracy for all industries at the Global Industry Classification Standard (GICS) 2 level, which is explained by our choosing to model corporate emissions across the entire global universe while using industry as a variable.
- The use of an iterative approach to detect and correct errors in the data consumed by the model also plays an important role in explaining its greater accuracy.

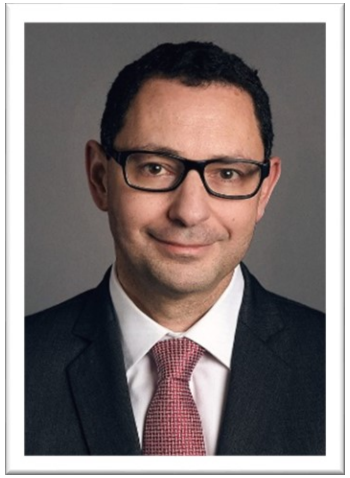
ABSTRACT

The authors propose a model based on statistical learning techniques to predict unreported corporate greenhouse gas emissions that generates considerably better results than existing approaches. The model uses one linear learner and one nonlinear learner only, which reduces its complexity to the minimum required. An iterative approach to detecting and correcting data significantly improves the model predictions. Unlike mainstream approaches, which tend to construct one model for each industry, we construct one single global model that uses industry as a factor. This addresses the problem of lack of breadth or lack of reported data in some sectors and generates practical results even for industries where other approaches have failed. We show results for Scope 1 and Scope 2 corporate carbon emissions. Adapting the framework to Scope 3 will be the focus of a future article.

The term "footprint" was introduced in the field of ecology by Wackernagel and Rees (1996) to measure how fast humans consume resources and generate waste compared with how fast nature can absorb our waste and generate resources. In turn, Wiedmann and Minx (2007) defined "carbon footprint" as a measure of the exclusive total amount of carbon dioxide emissions (CO_2) that is directly and indirectly caused by an activity or is accumulated over the life stages of a product. Today, the carbon footprint tends to account for all greenhouse gas (GHG) emissions caused by an individual, event, organization, service, place or product and is expressed in units of carbon dioxide equivalent (CO_2e).

The global warming potential (GWP) is the heat absorbed by any GHG in the atmosphere and is defined as a multiple of the heat that would be absorbed by the

Biography



Raul Leote de Carvalho, PhD

Deputy Head - Quant Research Group

Raul Leote de Carvalho is Deputy Head of the Quant Research Group at BNP Paribas Asset Management since October 2017.

This team with 30 quantitative researchers is responsible for supporting research and development of quantitative investment strategies, for providing quantitative inputs across all investment teams, for delivering quantitative services and innovation and for leveraging the use of alternative data and machine learning for investments. The team is based in Paris, Amsterdam, Hong Kong and London.

Raul joined BNPP AM in 1999 and held several positions related to quantitative research and investments in equities, fixed income and asset allocation covering the development of strategies, portfolio optimization, risk modelling and portfolio management.

Raul started his career in 1996 as a researcher in Computational and Theoretical Physics first at the University of Wuppertal, then at the Ecole Normale Supérieure de Lyon and finally at the University College of London. He earned a PhD in Theoretical Physics from the University of Bristol and an MSc in Condensed Matter Physics and a BSc in Chemistry both from the University of Lisbon.

Raul is Board Member and Chair of the Programme Committee of the Institute for Quantitative Research in Europe (Inquire Europe), Fellow of the Louis Bachelier Institute (France) and Member of the Institute of Physics (UK). He is also member of the editorial board of the Journal of Investing and referee for a number of journals. He passed the Investment Management Certificate in London in 2001.

Raul is based in Paris.

Disclaimer

BNP PARIBAS ASSET MANAGEMENT UK
5 ALDERMANBURY SQUARE
LONDON
EC2V 7HR

INVESTORS' corner

BNP PARIBAS
ASSET MANAGEMENT

The sustainable investor for a changing world