Working Paper 22-080

Machine Learning Models for Prediction of Scope 3 Carbon Emissions

George Serafeim
Gladys Vélez Caicedo

Harvard Business School

Machine Learning Models for Prediction of Scope 3 Carbon Emissions

George Serafeim
Harvard Business School

Gladys Vélez Caicedo
Harvard Business School

Working Paper 22-080

Funding for this research was provided in part by Harvard Business School.

Machine Learning Models for Prediction of Scope 3 Carbon Emissions

George Serafeim and Gladys Velez Caicedo[*]

Harvard Business School

Abstract

For most organizations, the vast amount of carbon emissions occur in their supply chain and in the post-sale processing, usage, and end of life treatment of a product, collectively labelled scope 3 emissions. In this paper, we train machine learning algorithms on 15 reported types of scope 3 emissions. The models utilize as inputs widely available financial statement variables, scope 1 and 2 emissions, and industrial classifications. We find that most reported scope 3 emission types can be predicted with higher accuracy using Adaptive Boosting machine learning algorithms relative to linear regression models and other supervised machine learning algorithms.

Keywords: *carbon emissions, climate change, environment, carbon accounting, machine learning, artificial intelligence.*

[*] George Serafeim is the Charles M. Williams Professor of Business Administration at Harvard Business School. Gladys Velez Caicedo is a research associate at Harvard Business School. Funding for this research was provided by the Division of Faculty and Research Development at Harvard Business School. We are solely responsible for any errors in this manuscript.

Introduction

An organization’s carbon emissions comprise three categories: scope 1, 2, and 3 (Ranganathan et al., 2004). Scope 1 emissions are direct emissions produced by a firm through the operation of its owned assets in its value chain. For example, airlines produce scope 1 emissions by burning jet fuel while flying their owned aircraft. Scope 2 emissions are indirect emissions generated from electricity, steam, heat, and cooling usage in firm operations and purchased from an external utility provider. Grocery stores produce scope 2 emissions by running appliances, light fixtures, and refrigeration equipment. Scope 3 emissions (often referred to as value chain emissions) are indirect emissions generated from all other activities conducted using assets not owned by the reporting company that are involved in the production and usage of the reporting company’s product or service. Scope 3 emissions are produced from a broad diversity of sources and processes including emissions associated with purchased goods and services, employees commuting, waste generated from operations, processing of sold products, and use of sold products (Bhatia et al., 2010).

To date, among the firms interested in measuring their carbon emissions, most have spent their time quantifying scope 1 and 2 emissions. Those emissions are often the most intuitive to comprehend because they are directly tied to the reporting firm’s actions and the simplest to quantify. Scope 1 emissions can be measured using internal activity metrics while to quantify scope 2 emissions, a firm must sum its utility consumption by energy source and multiply by a set of emission conversion factors. Moreover, reductions of scope 1 and 2 emissions are more within the span of control of a company, while reductions in scope 3 emissions require changes in supplier and customer behavior. Hence, when firms begin to manage their emissions, they tend to gravitate toward measuring, reporting, and targeting scope 1 and 2 emissions due to ease of calculation and higher degree of control. By comparison, scope 3 emissions are more difficult to understand and quantify (Cheema-Fox et al., 2021). The data needed to quantify these emissions frequently comes from third-parties or secondary sources. These business partners may not even collect such data, in which case, the reporting company may need to locate average secondary source data, contributing to estimation difficulty. Challenges such as these are common in quantifying the different types of scope 3 emissions.

Despite the challenges in measuring and mitigating scope 3 emissions, these emissions are significant. First, it is common for scope 3 emissions to account for a substantial share of a firm’s total emissions footprint (Klaaßen & Stoll, 2021). Automobile manufacturing is an example of an industry in which scope 3 emissions dominate corporate emissions profiles. Running factories and operating assembly lines to produce internal combustion engine vehicles (ICEVs) generates a significant amount of scope 1 and 2 emissions, however, it is consumers driving vehicles (the use of sold products) that comprises the vast majority of a vehicle manufacturer’s carbon footprint. The same can be said for the food products industry, which has significant upstream emissions from agricultural activities and transportation and distribution. Absent reporting scope 3 emissions, the narrative surrounding a company's climate-related financial risk may be incomplete or misleading. For example, a company with very large upstream scope 3 emissions operating in an industry where suppliers have power and can pass costs to their customers (i.e. the company), could find themselves bearing the costs of future carbon taxes not only on their scope 1 emissions but also on their upstream scope 3 emissions. This would be especially damaging for firms that have limited ability to pass their own costs to customers, in industries with low customer switching costs. Similarly, for companies with large downstream scope 3 emissions, technological risk might be heightened as cost effective alternative products might be preferred by customers (e.g., Electric Vehicles vs ICEVs).

Perhaps equally significant is the role of scope 3 emissions as a potential driver of change across supplier and consumer networks. If firms are held accountable for their scope 3 emissions, firms will likely push existing business partners to reduce their emissions or seek out new business partners with better emissions performance. We already observe evidence of this taking place. Walmart, to reduce its scope 3 emissions footprint, has launched an initiative to engage with its suppliers to accelerate their transition to renewable energy and sustainable practices.[1] Coordinated efforts of this kind have the potential for swift, widespread impact.[2] In sum, scope 3 emissions represent a substantial future climate transition risk due to their sheer volume, but also represent a significant opportunity to motivate future emissions reductions.

Most firms, especially smaller financially constrained firms, do not have the resources or influence to engage in costly scope 3 measurement exercises. For example, the Securities and Exchange Commission (SEC) rule on climate disclosure mentions that scope 3 measurement often triples the professional service fees that a firm would need to pay relative to measuring scope 1 and 2 emissions.[3] Those firms would benefit from a low cost and time efficient solution to the scope 3 measurement challenge that allows them to establish a rough approximation of their emissions and improve the quality of those measurements over time as they have more resources. Moreover, data providers already estimate scope 3 emissions, using linear models and heuristics, given the lack of disclosure by most firms.[4] The estimation of those emissions could benefit from the application of machine learning algorithms. Therefore, the goal of this paper is to report the prediction accuracy of machine learning algorithms using a set of variables that are widely available in financial statements or by firms reporting the easier to measure scope 1 and 2 emissions. The algorithms train on reported scope 3 emissions, thereby using the information set available.

[1] Walmart. Sustainability Hub. Project Gigaton.

[2] McKinsey & Company. Strategy & Corporate Finance Practice Research. How to succeed with carbon reduction initiatives.

[3] SEC. The Enhancement and Standardization of Climate-Related Disclosures for Investors. Proposed Rule.

[4] Refinitiv. ESG Carbon Data and Estimate Models.
MSCI ESG Research. Filling the Blanks: Comparing Carbon Estimates Against Disclosures. Comparing Carbon Estimates Against Disclosures.

Results

Features

We incorporate features that are widely available across publicly listed firms. We note that there could be more elaborate models that incorporate many more features, such as industry-specific energy production data, than those we use in this study that could improve further the prediction of scope 3 emission types. For our purposes, we prioritize the practicality of the models and for that reason we choose features that are widely available.

Those features can be classified into five groups. The first group includes nominal variables; the sub-industry classification and home country of a firm. We download Global Industry Classification Standard (GICS) sub-industry data from Compustat via Wharton Research Data Services (WRDS) and Bloomberg. Where GICS data is unavailable from Compustat, we use data from Bloomberg. We merge GICS data onto CDP data using ISIN. Since 2002 over 8,000 companies have disclosed climate-related data to the CDP (CDP, 2020). The organization's primary method of data collection is the dissemination of its annual Climate Change Questionnaire. [5]

The next three groups all include accounting and financial statement variables in addition to a market valuation multiple. The second group includes variables that represent a stock of resources, such as those found in the balance sheet (e.g., total assets). The third group includes variables that represent flows of resources, such as those found in the income statement (e.g., sales). The fourth group includes variables that represent ratios (e.g., profitability margin, operating efficiency). We source annual financial metrics from Worldscope via Wharton Research Data Services (WRDS). We merge financial data onto CDP data using International Securities Identification Number (ISIN) and year. We download year-end foreign exchange rates from Bloomberg to convert local currency values in the CDP data to USD. We merge exchange rates onto CDP data using ISIN and year.

The fifth group contains scope 1 and 2 emissions metrics. We source these data from CDP. We compile corporate data from Climate Change Questionnaires published from 2013 through 2020. All non-ratio variables are logged to mitigate skewness in their distribution. Summary statistics for variables in groups 2-5 are presented in **Table 1**.

Table 1

Feature	Observations	Mean	Median	Standard Deviation
Stock Variables				
*log*Total Assets	9,013	22.88	22.84	1.41
*log*Net Property Plant and Equipment	9,013	21.38	21.47	1.87

[5] CDP. Climate Change Questionnaire. Reporting Guidance 2020.

*log*Common Equity	9,013	21.88	21.88	1.41
*log*Market Capitalization	9,013	22.64	22.65	1.49
*log*Number of Employees	9,013	9.58	9.67	1.64
Flow Variables				
*log*Sales	9,013	22.48	22.49	1.48
log Cost of Goods Sold	9,013	21.73	21.90	2.19
*log*Selling, General & Administrative Expenses	9,013	20.68	20.71	1.64
*log*Operating Expense	9,013	22.33	22.36	1.51
*log*Operating Income	9,013	19.42	20.09	4.00
*log*Capital Expenditures	9,013	19.45	19.54	2.06
Ratios				
Return on Sales	9,013	0.13	0.10	0.12
Asset Turnover	9,013	0.85	0.75	0.55
Capital Intensity	9,013	0.10	0.05	0.16
Capital Renewal	9,013	0.19	0.16	0.15
Age of Capital Assets	9,013	11.10	6.66	16.91
Inventory Turnover Ratio	9,013	16.91	3.67	48.27
Market to Book	9,013	1.70	1.36	0.99
Emissions Metrics				
*log*Scope 1	9,013	11.64	11.54	2.96
*log*Scope 2	9,013	11.70	11.91	2.16

Table 1 | Summary statistics for the sample set of predictive input features. All observations have non-missing values for Scope 1 and Scope 2 emissions. Financial variables that are missing yet required input features are imputed using a k-nearest neighbor algorithm.

Reporting of Scope 3 Emission Types

Unlike the reporting of scope 1 and 2 emissions, scope 3 emissions are disclosed by type. There are 15 Scope 3 types and in this analysis these are known as targets. **Table 2** shows the percentage of total firm-years with emissions reported, emissions reported with mostly primary data (more than 80%) sourced from suppliers or customers, and emissions reported as zero for each of the 15 Scope 3 types. Several patterns emerge. First, emission types that are more idiosyncratic to the operating and strategic choices of individual firms tend to be less reported or reported as zero (i.e., franchises, investments, processing of sold products, or leased assets). Outside of this group of idiosyncratic scope 3 emissions types, we observe that more companies are reporting emissions, using primary data and as non-zero values, for upstream rather than downstream (i.e., purchased goods vs use of sold products, upstream vs downstream transportation and distribution), and for types that are easier to measure and where firms have more control (i.e., business travel, fuel and energy, waste generation from firm operations, employee commuting).

Table 2

log (Target)	Non-missing Observations	% of Observations Non-missing	N with Primary Data=>80%	% of reported with Primary Data=>80%	N Zeros	% of reported as Zero	Mean	Median	Standard Deviation
Scope 3 Type									
Busines Travel	7,728	86%	4,238	55%	78	1%	8.53	8.69	2.18
Purchased Goods	5,454	61%	1,805	33%	321	6%	11.68	12.92	4.35
Fuel and Energy	5,308	59%	2,152	41%	416	8%	9.81	10.31	3.93
Waste Generated in Operations	5,306	59%	2,066	39%	370	7%	7.66	8.19	3.24
Employee Commuting	5,005	56%	1,250	25%	332	7%	8.28	8.84	3.13
Upstream Transportation and Distribution	4,729	52%	1,768	37%	436	9%	9.63	10.48	3.99
Downstream Transportation and Distribution	3,857	43%	1,435	37%	524	14%	9.16	10.36	4.48
Capital Goods	3,474	39%	695	20%	647	19%	9.05	10.96	4.88
Use of Sold Products	3,374	37%	867	26%	665	20%	11.48	13.56	6.45
End of Life Treatment of Sold Product	2,793	31%	537	19%	716	26%	7.60	8.87	5.36
Upstream Leased Assets	2,133	24%	563	26%	1,011	47%	4.32	3.95	4.47
Downstream Leased Assets	2,014	22%	537	27%	943	47%	4.93	4.84	5.08
Investments	1,827	20%	485	27%	918	50%	5.65	0.00	6.02
Franchises	1,781	20%	212	12%	1,374	77%	0.46	0.00	2.24
Processing of Sold Products	1,691	19%	250	15%	953	56%	5.07	0.00	6.19

Table 2 | Summary statistics for the sample set of predicted targets. Business travel is the most reported Scope 3 emission type, due to its ease of calculation, however, calculations are highly dependent on designated emission factors corresponding the assumed mode of transportation. The mode of transportation assumption is also applicable in the calculations of employee commuting as well as downstream and upstream transportation and distribution, which are highly industry dependent. Equally, although there is a higher level of reporting within fuel and energy related activities and waste generated in operations, these responses have significant variation dependent on country-based regulations, access to fuel type and waste emission factors. Purchased goods and services is the second most reported Scope 3 type, however, firms of varying size and supply chain complexity choose to either calculate complete or only partial portions of their purchased goods and services value chain ranging from raw materials to intangible support services. Some firms choose to calculate these emissions based on a life cycle analysis (LCA) of the total value chain meanwhile other firms limit reported emissions to direct suppliers. Due to this, firms with similar financial and industry profiles may report emissions that are inconsistent with one another, and this proves to be challenging for a predictive model to accurately reproduce. Upstream and downstream leased assets also present challenges in terms of inconsistent reporting. Upstream and downstream leased assets are dependent on a firm's business model by way of operational or managerial decision-making. This leads to low reporting and high intra-industry variation. Therefore, firms with similar financial and industry profiles may report upstream and downstream leased assets emissions that are not comparable across identical sub-industries leading to a source of prediction error. Franchises and Investments Scope 3 types are highly industry dependent which is a primary driver of low response rates and relevancy. In addition, these two Scope 3 types depend on the firm's business model, as well as the operational or financial control boundary as defined by the GHG protocol. Use of sold products, processing of sold products and end of life treatment of sold products present challenges in calculation given that the boundary that separates these types is often not clear to firms and carry assumptions dependent on end user behavior. Within the sample set, this calculation challenge is reflected in the low response rate, large standard deviation relative to the mean, as well as the high variance exhibited by the log transformed distributions of these three Scope 3 types. In general, given the complex challenges in reporting denotes above, controlling for the sample set's bias and variance inherent to certain reporting methods upon a backdrop of irreducible error was a central concern. The total reducible error was mitigated by tuning the machine learning model to the optimal set of hyperparameters and controlling for certain sub-sets of total data such as cutting the sample training set by primary responses or excluding zero reported emissions.

Models

We present prediction metrics for primary models alongside benchmark models described in detail in the methodology section. We focus on Random Forest and Adaptive Boosting (AdaBoost) algorithms as primary models. While both algorithms are ensemble learning algorithms, random forest uses the concept of bagging while AdaBoost the concept of boosting (Dietterich, 2000). To benchmark these models, we report model statistics for linear regression estimators using ordinary least squares and gamma general linear models. Moreover, we report results for the k-nearest neighbors (k-NN) algorithm, a commonly used baseline for evaluating non-parametric, supervised machine learning model performance (Pedregosa et al., 2011).

Prediction Metrics

We report three distinct metrics that assess the predictive accuracy of the models. Given the characteristics and flaws of each metric, evaluating the models across multiple statistics increases the robustness of our conclusions. First, we report the Root Mean Squared Logarithmic Error (RMSLE). Lower values mean lower percentage errors in predicted emissions. RMSLE penalizes underestimated predictions more than overestimated predictions. Second, we report the R-squared between predicted and reported values (R^2). Higher values mean that more of the variation in reported target values is explained by the input features. However, it does not provide a measure of prediction accuracy. Third, for non-zero reported values we report the mean absolute percentage error (MAPE). Lower values mean lower percentage error. In contrast to RMSLE, MAPE penalizes overestimates more relative to underestimates given that errors are divided by reported values.[6] The models are trained on a training set comprised of 80% of the total samples. All reported prediction metrics are assessed on a holdout test set comprised of 20% of the total samples not previously seen by the model. The test train data set split is initialized through a pseudorandom number generator with seed 1.

In **Table 3**, across models, we observe a decline (increase) in RMSLE (R^2). Across all emission types, the average R^2 increases from 46% for OLS to 68% for k-NN to 75% for random forest and to 78% for AdaBoost. Restricting the sample to observations where companies are using mostly primary data from suppliers and customers to estimate their scope 3 emissions in each type increases further the across type average R^2 to 83%.

[6] Sci-kit learn. Metrics and scoring: quantifying the quality of predictions. Documentation.

Table 3

Scope 3 Emissions Category Type		All Data including Zeros				Primary Data>=80% with Zeros
		OLS	KNN	RF	AdaBoost	AdaBoost
Business Travel	RMSLE	1.57	1.31	1.24	1.24	0.87
	R^2 Score	69.2%	79.7%	82.2%	82.2%	89.3%
Capital Goods	RMSLE	4.62	3.86	3.47	3.26	1.53
	R^2 Score	41.6%	65.2%	73.0%	76.8%	89.9%
Downstream Leased Assets	RMSLE	4.75	3.87	3.30	2.97	2.02
	R^2 Score	41.1%	66.9%	77.3%	82.2%	86.4%
Downstream Transportation and Distribution	RMSLE	4.07	3.49	3.00	2.92	1.71
	R^2 Score	40.8%	62.0%	74.0%	75.5%	81.7%
Employee Commuting	RMSLE	2.75	2.38	2.27	2.19	1.41
	R^2 Score	50.6%	66.3%	70.1%	72.5%	79.5%
End of Life Treatment of Sold Products	RMSLE	4.66	3.64	3.40	3.10	1.77
	R^2 Score	49.9%	73.7%	77.5%	81.8%	90.1%
Franchises	RMSLE	1.98	1.34	1.30	1.28	2.56
	R^2 Score	16.7%	74.3%	76.4%	77.2%	84.1%
Fuel and Energy	RMSLE	3.07	2.70	2.37	2.25	1.79
	R^2 Score	59.4%	70.7%	78.2%	80.8%	81.5%
Investments	RMSLE	5.37	4.35	3.82	3.55	1.76
	R^2 Score	44.5%	68.9%	77.0%	80.6%	91.0%
Processing of Sold Products	RMSLE	5.63	4.71	4.17	4.09	4.26
	R^2 Score	46.6%	67.2%	75.6%	76.7%	70.3%
Purchased Goods	RMSLE	3.68	2.94	2.72	2.55	1.93
	R^2 Score	54.8%	74.2%	78.5%	81.4%	82.0%
Upstream Leased Assets	RMSLE	4.33	3.83	3.27	3.14	2.12
	R^2 Score	25.6%	51.7%	68.4%	71.2%	82.2%
Upstream Transportation and Distribution	RMSLE	3.32	2.99	2.69	2.56	1.93
	R^2 Score	50.4%	62.7%	71.6%	74.6%	73.3%
Use of Sold Products	RMSLE	5.59	4.57	4.26	3.74	2.65
	R^2 Score	47.6%	69.4%	74.2%	80.9%	83.5%
Waste Generated in Operations	RMSLE	2.86	2.30	2.13	2.06	1.62
	R^2 Score	49.4%	71.4%	76.1%	78.0%	79.3%

Table 3 | Summary statistics for the full sample set of predicted targets across the benchmark and primary models. All models are trained on 80% of full sample set and the output metrics denote the prediction performance on a holdout test set comprised of 20% of the total samples not previously seen by the model. The four first columns of output metrics are tested on a randomized 20% of the total sample not previously seen by the model. The fifth column of output metrics are tested on a randomized 20% of the sample set where greater than or equal to 80% of the reported data is sourced from direct suppliers.

Table 4 reports similar models but excludes reported values of zero. This allows to report results for Gamma-GLM and to report MAPE. We observe that AdaBoost reaches the lowest RMSLE and MAPE and highest R^2, across almost all types. Across type average MAPE declines from 43% and 46% for OLS and GLM to 37% for k-NN, 33% for random forest and 27% for AdaBoost. Using only reported emissions estimated mostly with primary data decreases further MAPE to 15% for AdaBoost. However, in some cases using the smaller sample of observations leads to higher RMSLE and lower R^2, suggesting a trade-off between higher quality data and sample size. We find that the model can significantly improve prediction accuracy metrics when report zero values are excluded from the data set.

Table 4

Scope 3 Emissions Category Type		All Data without Zeros					Primary Data>=80% without Zeros
		OLS	GLM	KNN	RF	AdaBoost	AdaBoost
Business Travel	RMSLE	1.33	1.36	1.08	0.98	0.90	0.77
	R^2 Score	74.3%	73.1%	84.1%	87.1%	89.2%	90.6%
	MAPE	14.1%	14.8%	10.3%	9.2%	7.5%	13.8%
Capital Goods	RMSLE	2.23	2.21	1.69	1.49	1.39	1.47
	R^2 Score	49.9%	51.5%	75.3%	81.5%	84.2%	85.4%
	MAPE	23.2%	23.2%	15.0%	13.7%	10.7%	12.6%
Downstream Leased Assets	RMSLE	2.59	2.69	2.16	1.91	1.76	1.62
	R^2 Score	45.7%	38.5%	67.0%	75.5%	79.7%	83.4%
	MAPE	29.7%	31.7%	22.0%	19.9%	15.3%	17.5%
Downstream Transportation and Distribution	RMSLE	2.36	2.31	1.83	1.66	1.55	1.17
	R^2 Score	48.7%	51.8%	73.7%	78.9%	81.8%	89.0%
	MAPE	30.0%	30.9%	21.9%	20.4%	16.8%	8.4%
Employee Commuting	RMSLE	1.73	1.75	1.37	1.24	1.19	1.32
	R^2 Score	66.4%	65.2%	80.6%	84.3%	85.6%	78.1%
	MAPE	25.9%	26.7%	18.6%	15.5%	13.6%	8.8%
End of Life Treatment of Sold Products	RMSLE	2.74	2.85	2.32	1.97	1.81	1.50
	R^2 Score	60.7%	56.4%	74.1%	82.2%	85.1%	89.3%
	MAPE	48.7%	57.3%	37.0%	32.0%	22.1%	11.8%
Franchises	RMSLE	2.57	2.82	2.10	1.55	1.18	1.41
	R^2 Score	65.2%	55.5%	78.5%	89.0%	93.8%	86.8%
	MAPE	33.3%	37.1%	21.4%	17.2%	8.8%	6.3%
Fuel and Energy	RMSLE	1.88	1.91	1.51	1.45	1.34	1.39
	R^2 Score	75.8%	74.9%	85.2%	86.4%	88.4%	87.2%
	MAPE	99.9%	105.8%	66.2%	63.8%	55.0%	11.9%
Investments	RMSLE	2.42	2.43	1.69	1.49	1.35	1.75

	R^2 Score	60.4%	60.0%	83.2%	87.1%	89.6%	82.4%
	MAPE	20.2%	21.1%	13.4%	11.8%	9.0%	20.8%
Processing of Sold Products	RMSLE	2.77	2.71	2.07	1.74	1.59	3.15
	R^2 Score	51.3%	54.2%	76.8%	84.2%	87.0%	42.2%
	MAPE	18.6%	18.7%	12.4%	11.0%	8.5%	15.6%
Purchased Goods	RMSLE	2.68	2.70	2.06	1.88	1.66	1.72
	R^2 Score	58.9%	58.2%	78.3%	82.4%	86.6%	87.0%
	MAPE	31.6%	33.2%	23.3%	21.0%	16.1%	20.2%
Upstream Leased Assets	RMSLE	2.25	2.32	2.18	1.84	1.87	1.34
	R^2 Score	47.8%	42.6%	52.7%	69.5%	68.3%	84.1%
	MAPE	156.2%	159.9%	183.2%	146.8%	139.1%	34.0%
Upstream Transportation and Distribution	RMSLE	2.11	2.13	1.77	1.48	1.37	1.54
	R^2 Score	59.3%	58.5%	74.0%	57.3%	87.2%	81.6%
	MAPE	63.0%	67.7%	67.0%	82.6%	55.9%	16.5%
Use of Sold Products	RMSLE	2.66	2.69	1.84	1.65	1.36	1.57
	R^2 Score	63.6%	62.3%	84.6%	87.8%	91.9%	90.3%
	MAPE	21.4%	22.1%	13.4%	12.9%	9.5%	6.9%
Waste Generated in Operations	RMSLE	1.98	2.05	1.57	1.49	1.42	1.41
	R^2 Score	65.1%	62.2%	79.9%	82.2%	83.8%	83.3%
	MAPE	31.2%	34.0%	23.1%	21.5%	18.5%	15.8%

Table 4 | Summary statistics for a partial subset excluding reported zero values from sample set of predicted targets across the benchmark and primary models. All models are trained on 80% of the partial subset excluding reported zero values from sample set and the output metrics denote the prediction performance on a holdout test set comprised of 20% of the subset not previously seen by the model. The four first columns of output metrics are tested on a randomized 20% of the partial subset not previously seen by the model. The fifth column of output metrics are tested on a randomized 20% of the partial subset where greater than or equal to 80% of the reported data is sourced from direct suppliers. For Scope 3 Types with a sufficiently large sample subset size, the model can more accurately predict emission values closer to the reported values than for the model which includes reported zero values for emissions, as in the distinction between Table 3 and Table 4. The reason for the improved accuracy in predicting non-zero values may be due to firms reporting zero value emissions within categories that they deem not relevant to their business model or not yet evaluated rather than the true value of those emissions being zero. This reporting distinction may be at times arbitrary and thus the model has difficulty in reproducing a zero prediction which results in a lower overall prediction accuracy when reported zero emissions are included in the training set. In addition, model prediction accuracy decreases and mean absolute percentage error increases for some Scope 3 types when the sample set becomes too small to be statistically significant due to the exclusion boundary of primary data exclusive of zero reported emissions values.

Figure 1 | AdaBoost model using all data excluding zeros test set distribution of Scope 3 emissions by type

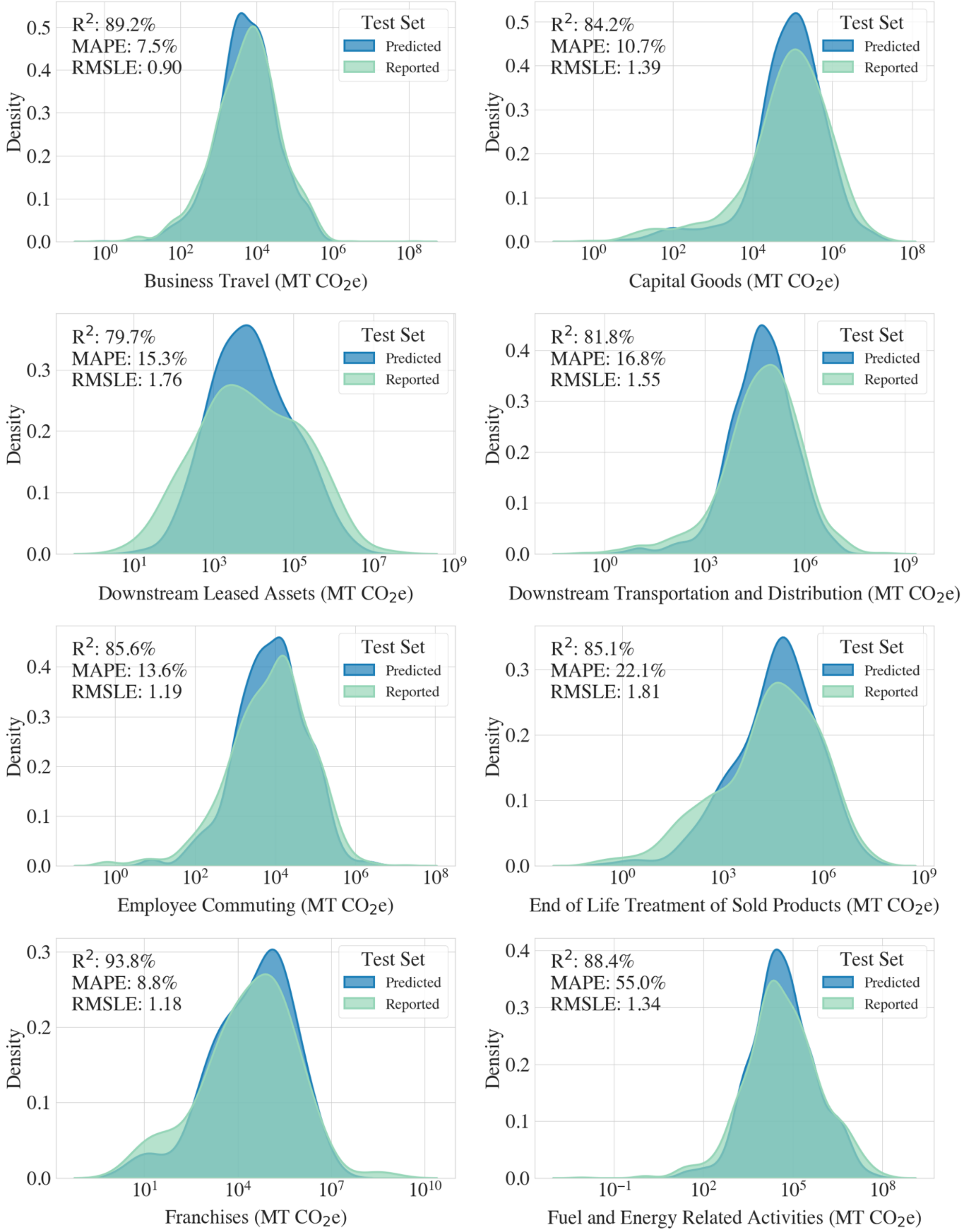

Figure 1 | The visual distributions in metric tons (MT CO$_2$e) of reported and predicted values from the AdaBoost model using a subset of data excluding reported zero values for each one of the 15 Scope 3 types. Distributions correspond to the output metrics in column 5 of Table 4. The data shows that the model produces more leptokurtic distributions compared to the distribution of reported values. In general, with the exclusion of reported zeros the model predicts a higher number of firm's emissions as being concentrated around the mean and avoids the upper/lower tails.

Figure 1 cont. | AdaBoost model using all data excluding zeros test set distribution of Scope 3 emissions by type

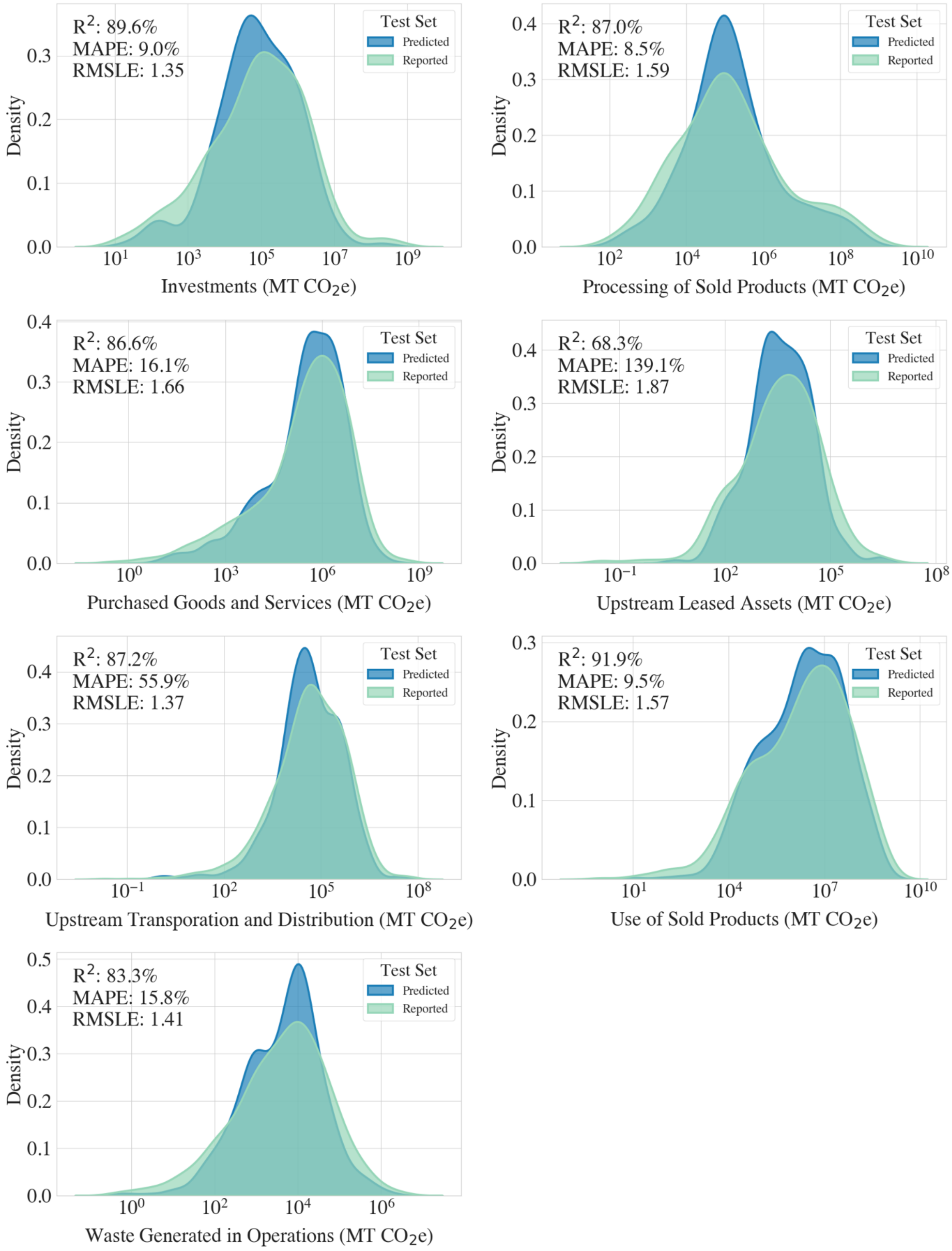

Figure 1 cont. | The visual distributions in metric tons (MT CO_2e) of reported and predicted values from the AdaBoost model using a subset of data excluding reported zero values for each one of the 15 Scope 3 types. Distributions correspond to the output metrics in column 5 of Table 4. The data shows that the model produces more leptokurtic distributions compared to the distribution of reported values. In general, with the exclusion of reported zeros the model predicts a higher number of firm's emissions as being concentrated around the mean and avoids the upper/lower tails.

Table 5 reports results for the random forest and AdaBoost algorithms in two ways. First, it reproduces the existing results where each scope 3 type model is trained and tested separately, which is referred to as the 'By Type' model below. We report those results for benchmarking purposes. Second, it produces results for a singular model where every scope 3 type is combined into one model and we introduce one more feature, a nominal variable, representing the scope 3 type and is referred to as the 'Singular' model below. This model allows us to understand if there is a benefit in prediction accuracy and model fit by allowing different scope 3 emission types to be estimated in the same model. We do not observe clear patterns emerging that would allow us to conclude that the singular model either underperforms or outperforms the type-specific models.

Table 5

		All Data							
		RF with Zeros		RF without Zeros		AdaBoost with Zeros		AdaBoost without Zeros	
		By Type	Singular	By Type	Singular	By Type	Singular	By Type	Singular
Business Travel	RMSLE	1.24	1.23	0.98	0.91	1.24	1.24	0.90	0.82
	R^2 Score	82.2%	83.1%	87.1%	88.6%	82.2%	82.9%	89.2%	90.8%
	MAPE	N/A	N/A	9.2%	9.5%	N/A	N/A	7.5%	8.0%
Capital Goods	RMSLE	3.47	3.04	1.49	1.61	3.26	2.62	1.39	1.58
	R^2 Score	73.0%	78.5%	81.5%	77.4%	76.8%	84.6%	84.2%	78.5%
	MAPE	N/A	N/A	13.7%	22.8%	N/A	N/A	10.7%	19.6%
Downstream Leased Assets	RMSLE	3.30	3.37	1.91	2.15	2.97	2.97	1.76	2.17
	R^2 Score	77.3%	75.1%	75.5%	64.5%	82.2%	81.4%	79.7%	63.4%
	MAPE	N/A	N/A	19.9%	22.1%	N/A	N/A	15.3%	17.6%
Downstream Transportation and Distribution	RMSLE	3.00	2.88	1.66	1.64	2.92	2.89	1.55	1.56
	R^2 Score	74.0%	77.2%	78.9%	79.0%	75.5%	77.0%	81.8%	81.2%
	MAPE	N/A	N/A	20.4%	14.1%	N/A	N/A	16.8%	12.0%
Employee Commuting	RMSLE	2.27	2.08	1.24	1.32	2.19	2.15	1.19	1.28
	R^2 Score	70.1%	75.9%	84.3%	81.9%	72.5%	73.9%	85.6%	83.2%
	MAPE	N/A	N/A	15.5%	15.8%	N/A	N/A	13.6%	12.6%
	RMSLE	3.40	3.09	1.97	1.99	3.10	2.67	1.81	1.90

End of Life Treatment of Sold Products	R^2 Score	77.5%	80.1%	82.2%	80.6%	81.8%	85.7%	85.1%	82.5%
	MAPE	N/A	N/A	32.0%	35.5%	N/A	N/A	22.1%	25.8%
Franchises	RMSLE	1.30	1.37	1.55	1.62	1.28	1.28	1.18	1.62
	R^2 Score	76.4%	75.3%	89.0%	81.9%	77.2%	78.8%	93.8%	81.9%
	MAPE	N/A	N/A	17.2%	17.5%	N/A	N/A	8.8%	12.4%
Fuel and Energy	RMSLE	2.37	2.35	1.45	1.46	2.25	2.18	1.34	1.41
	R^2 Score	78.2%	78.8%	86.4%	85.5%	80.8%	82.1%	88.4%	86.6%
	MAPE	N/A	N/A	63.8%	12.7%	N/A	N/A	55.0%	10.7%
Investments	RMSLE	3.82	3.63	1.49	1.76	3.55	3.48	1.35	1.71
	R^2 Score	77.0%	81.0%	87.1%	75.0%	80.6%	82.6%	89.6%	76.6%
	MAPE	N/A	N/A	11.8%	11.8%	N/A	N/A	9.0%	10.3%
Processing of Sold Products	RMSLE	4.17	3.89	1.74	2.14	4.09	3.79	1.59	1.85
	R^2 Score	75.6%	77.2%	84.2%	65.8%	76.7%	78.5%	87.0%	75.9%
	MAPE	N/A	N/A	11.0%	15.7%	N/A	N/A	8.5%	11.7%
Purchased Goods	RMSLE	2.72	2.75	1.88	1.76	2.55	2.49	1.66	1.64
	R^2 Score	78.5%	77.1%	82.4%	84.1%	81.4%	81.6%	86.6%	86.4%
	MAPE	N/A	N/A	21.0%	22.1%	N/A	N/A	16.1%	16.6%
Upstream Leased Assets	RMSLE	3.27	3.33	1.84	1.46	3.14	3.20	1.87	1.43
	R^2 Score	68.4%	66.3%	69.5%	78.2%	71.2%	69.4%	68.3%	79.2%
	MAPE	N/A	N/A	146.8%	13.0%	N/A	N/A	139.1%	10.2%
Upstream Transportation and Distribution	RMSLE	2.69	2.44	1.48	1.59	2.56	2.25	1.37	1.49
	R^2 Score	71.6%	79.4%	57.3%	81.5%	74.6%	82.7%	87.2%	84.1%
	MAPE	N/A	N/A	82.6%	26.8%	N/A	N/A	55.9%	24.2%
Use of Sold Products	RMSLE	4.26	3.81	1.65	1.64	3.74	3.34	1.36	1.37
	R^2 Score	74.2%	80.6%	87.8%	87.3%	80.9%	85.4%	91.9%	91.3%
	MAPE	N/A	N/A	12.9%	112.0%	N/A	N/A	9.5%	108.3%
Waste Generated in Operations	RMSLE	2.13	2.05	1.49	1.44	2.06	1.95	1.42	1.35
	R^2 Score	76.1%	76.9%	82.2%	83.8%	78.0%	79.4%	83.8%	86.0%
	MAPE	N/A	N/A	21.5%	52.1%	N/A	N/A	18.5%	50.0%

Table 5 | Summary statistics for model performance using the full sample set and partial sample subset including and excluding reported zero values from sample set of predicted targets across the two primary Random Forest and AdaBoost models using the targeted by Scope 3 type as compared to the singular model where all Scope 3 types are used as a singular input predictive feature. All models are trained on 80% of the partial subset excluding reported zero values from sample set and the output metrics denote the prediction performance on a holdout test set comprised of 20% of the subset not previously seen by the model. In general, the AdaBoost model prediction accuracy metrics show marginally improved performance in comparison to the Random Forest model. Both the AdaBoost and Random Forest models have comparable performance between the targeted by Scope 3 type as compared to the singular model. Although no general trend emerges, the singular model underperforms or overperforms the targeted by type model.

Figure 2 | Cumulative distribution function of the relative error for models run as a singular model

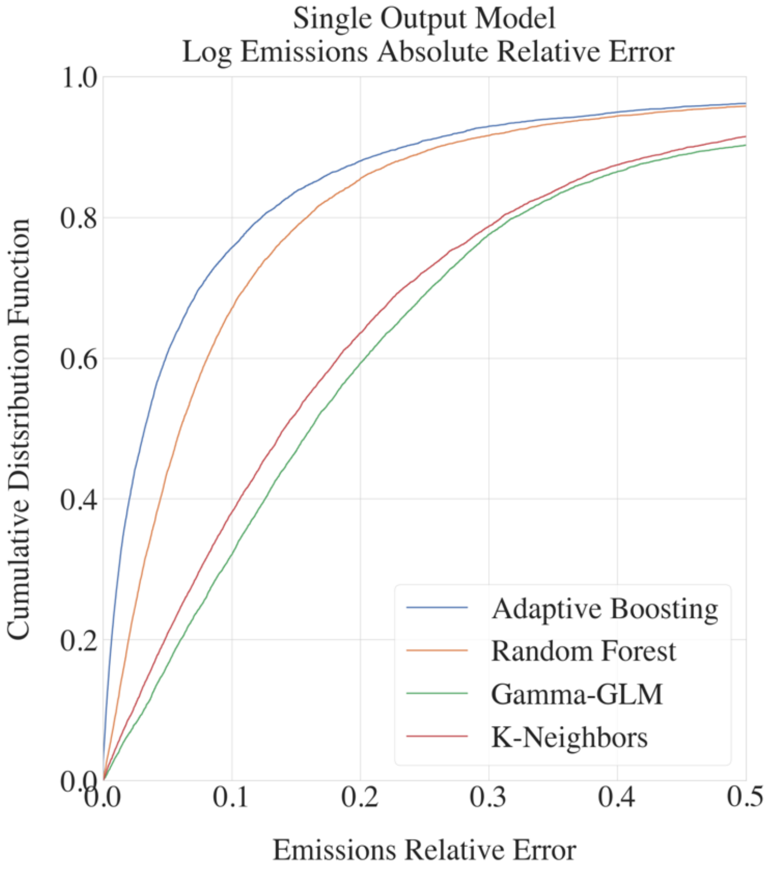

Figure 2 |Cumulative distribution function (CDF) of the relative error (y_predicted - y_reported)/y_reported) for four machine learning algorithms tested as a singular model. The CDF represents the probability that the relative error in emissions takes a value less than or equal the value along the horizontal axis. Both ensemble models outperform the Gamma-GLM and the k-NN models significantly, with the AdaBoost model performing better than the Random Forest model. Adaptive Boosting is the best performing model, where 80% of the predictions have a relative error below 12%. The relatively simple Gamma-GLM model performs the worst of the 4 models and has 80% of predictions below 32%.

Feature Importance

Having documented that the AdaBoost model provides improved predictive accuracy as measured against comparable decision tree models as well as linear models, we analyze which features are important for different scope 3 types. Scope 1 emissions, number of employees, scope 2 emissions, inventory turnover, and SG&A expenses stand out as the most important features. Scope 1 emissions are particularly important in predicting emissions from waste in operations, fuel and energy, processing of sold products and end of life treatment of sold products.

Several flow and stock variables achieve high levels of feature importance depending on the type. Some relationships are expected as they directly and intuitively influence emissions. For example, the number of employees achieves high feature importance for emissions from employee commuting given that these emissions will be a function of the number of employees, alongside commuting distance, and the carbon intensity of the method of commuting. Capital expenditures have high feature importance for capital goods as these emissions will increase the more capital goods a firm purchase, all else equal. SG&A expenses have high feature importance for business travel, as these emissions will be larger the more employees travel for business, an expense that is recorded within SG&A for most companies.

Ratio variables are less important except for inventory turnover, which is particularly important in predicting capital goods but also several other types. Among the nominal variables sub-industry membership is a more important feature than home country for most scope 3 types. Interestingly, sales, a variable often used to create carbon intensity metrics (i.e., carbon emissions per unit of sales) exhibits low feature importance apart from emissions from use of sold products.

Figure 3 | Feature importance heatmap of the AdaBoost targeted by Scope 3 type model using all data excluding zeros

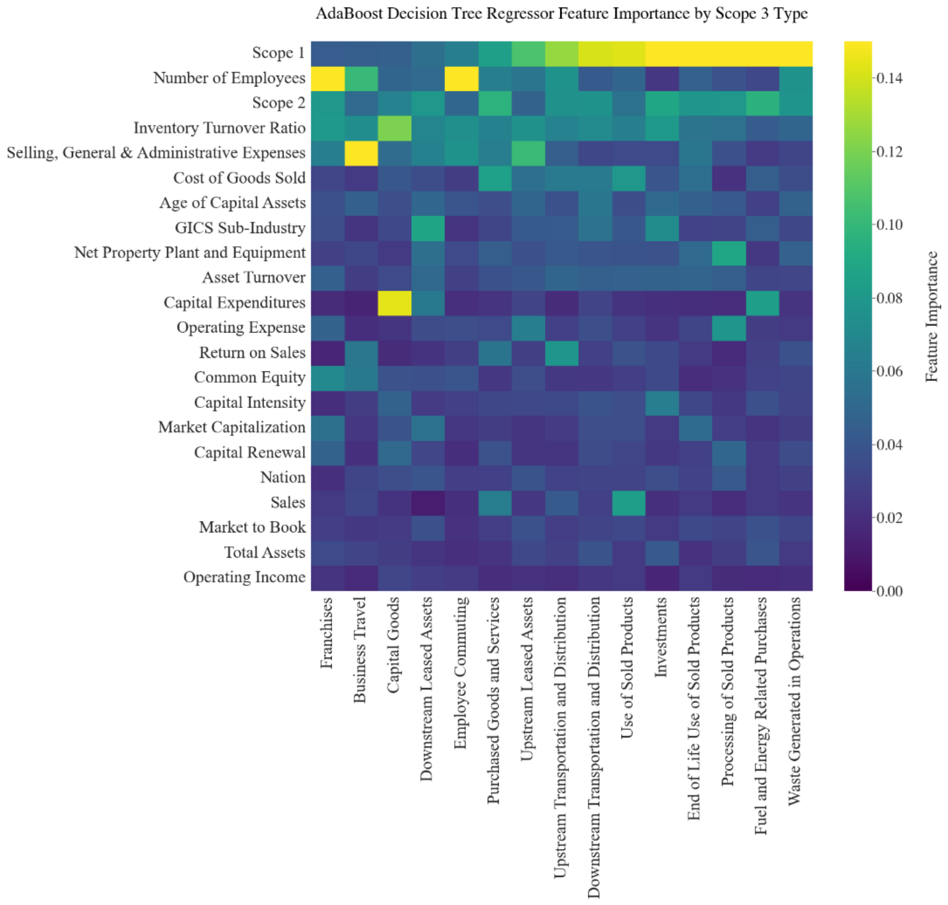

Figure 3 | Each predictive input feature is assessed between 0 and 1 that measures the average importance of each features in creating a decision tree results in more accurate predictive outcomes. The data shows that the most useful features in creating accurate predictions

are Scope 1 emissions, number of employees, scope 2 emissions, inventory turnover, and SG&A expenses. In general, the least useful features in creating accurate predictions are the remaining ratio variables, total assets, operational income, nation of domicile and sales.

Discussion

Current limitations within reported Scope 3 data include inconsistent and partial reporting across Scope 3 types (Klaaßen & Stoll, 2021). Moreover, most firms lack the resources and ability to measure their scope 3 emissions, given lack of data, control over decisions made by their suppliers or customers, and difficulty in calculation. This results in scope 3 measurement being completed mostly by large firms with plenty of financial resources.

A central aim of the machine learning approach presented in this paper is to leverage existing reporting of scope 3 emissions by firms who have invested the resources to calculate their emissions to train models and document their predictive ability. By leveraging the data of first movers within the emissions reporting landscape, we take a first steps towards estimating Scope 3 emissions using different models. We document that machine learning models trained on reported data can be a promising avenue to provide widespread access to estimates of scope 3 emissions for all 15 types. This in turn would allow companies that lack the resources to conduct detailed measurement of their scope 3 emissions to derive a first estimate of their emissions, upon which they can improve measurement practices, set targets for emissions improvement and design decarbonization strategies. In addition, a machine learning approach can enable investors that need data on a very large number of companies for portfolio construction and benchmarking purposes, to use the data as they evaluate the scope 3 emissions of companies.

Several caveats apply to our methodology and inferences. First, the models estimate scope 3 emission types using broadly available accounting data and the more easily calculated scope 1 and 2 emissions. While a strength of this approach is that a firm could retrieve all these data items with relative ease and as a result calculate its scope 3 emissions using machine learning models, the models miss other important features, such as idiosyncratic supply chain and product choices, that could influence scope 3 emissions. Moreover, the external validity of the predictions derived from the machine learning models might be limited if some firms have supply chains and product features that materially differ from those in the dataset on which the models are trained.

Conclusion

Scope 3 value chain emissions have rapidly become a central concern for companies and investors seeking to assess their climate risk exposure. Companies seek to use this data to understand and manage their climate risk. Investors seek to use this data as a proxy for assessing the climate transition risks facing their investments. Current limitations to accurately assessing and accessing this data include inconsistent and partial reporting across Scope 3 types as well as lack of resources to perform total value chain carbon emissions accounting. In this analysis, machine learning is applied as a tool for use by both companies and investors to create a complete, publicly accessible dataset for quantifying Scope 3 emissions across thousands of companies. In comparison to traditional linear regression models and naïve mean models currently applied within the industry, machine learning algorithms prove to be a cost-effective solution to improving prediction accuracy by leveraging the non-linear interactions between the input features and predicted targets. By leveraging the non-linear interactions and multi-collinearity of financial and emission features, the machine learning models applied in this analysis can successfully improve prediction accuracy where traditional regression models fail. The data finds that the average R^2 increases from 46% for OLS to 78% for AdaBoost when applied to the full sample set of reported Scope 3 emissions. In addition, restricting the sample to observations where companies are using mostly primary data from suppliers and customers to estimate their scope 3 emissions in each type increases average R^2 to 83% across the Scope 3 types. Furthermore, the data finds that AdaBoost reaches the lowest RMSLE and MAPE and highest R^2 across almost all types when the sample subset excludes reported zero emissions from the training set. Across all Scope 3 types the average MAPE declines from 43% to 27% for AdaBoost, a 16% average improvement from traditional linear models. Using only reported emissions estimated mostly with primary data decreases the average MAPE across Scope 3 types further to 15% for AdaBoost. Prediction accuracy has a marginal, but not significant, improvement using distinct machine learning algorithms. Most of the improvement is captured in the hyperparameter tuned decision tree model applied in this analysis.

Notwithstanding the improved accuracy achieved by machine learning algorithms to complete the data set, significant challenges persist in the Scope 3 emissions reporting landscape. Machine learning algorithms are only able to predict emissions that reflect the extent of the accuracy in the reported, or training, data. Given the challenges companies face in reporting their Scope 3 carbon emissions, the machine learning models trained on this sample set provide a first approximation for estimating total value chain emissions. In summary, applying machine learning models to build out a complete Scope 3 emissions data set is a cost-effective method to help drive first approximations of a corporate Scope 3 emissions.

Methods

Data and methodology

The purpose of this machine learning model is to create an open-access method for total value chain carbon emissions predictions at the firm level based on nominal, financial metrics in addition to scope 1 and 2 emissions data available to investors. The total value chain carbon emissions target variables are the 15 individual Scope 3 types as defined by the GHG protocol for publicly listed firms.

Data Cleaning and Pre-Processing To ensure a high-quality dataset, pre-processing steps are taken as follows: Scope 1 and Scope 2 reported emissions data are compiled from the CDP Climate Change Questionnaire between 2013-2020. Firms whose total Scope 1 or Scope 2 emissions are zero or missing are eliminated from the data frame. Scope 3 reported emissions data are disaggregated into 15 types and are compiled from the CDP Climate Change Questionnaire between 2013-2020. 23 firms that report values greater than 1 billion metric tonnes of emissions across any individual Scope 3 type are eliminated from the data frame, as these may be erroneous calculations. After eliminating these outliers, firms with total Scope 3 equal to or greater than 1 billion when summed across all types remain. 49 firms with missing GICS Industry classification are eliminated from the data frame. In addition, GICS 'Financials' Sector classified firms (29,581 entries ~ 1,972 unique firms) are eliminated from the data frame. This decision was made because most financial firms do not report scope 3 emissions associated with financed emissions and therefore most of the entries are close to zero. The Partnership for Carbon Accounting Financials is currently working to address this measurement challenge. 10 GICS Sector classifications remain (Energy, Industrials, Materials, Real Estate, Communication Services, Utilities, Information Technology, Consumer Staples, Health Care, Consumer Discretionary). Observations containing Scope 3 emission responses yet having missing entries within the requisite predictive financial features are imputed through a k-nearest neighbors imputer using the observation's 5 nearest neighbors. Each observation's missing financial features are imputed using the mean value from the sample's 5 nearest neighbors found in the training set neighbors by applying the Euclidean distance matrix (Rubinsteyn & Feldman 2016).

$$Euclidean\ Distance\ =\ |X\ -\ Y|\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

In addition, 2,059 firm-year observations with a total Scope 3 emissions reported as zero are held out to be predicted from the final trained model. Finally, to limit the effect of outliers, predictive features that are financial ratios, namely inventory turnover ratio, age of capital assets, return on sales, capital intensity, capital renewal, market to book, asset turnover are winsorized to the 1st and 99th percentile. Tree-based models are rule-based models that partition or bin the input space of predicted targets based upon the data boundaries (Maclin & Optiz 1999) . As such, rule-based models, such as gradient boosted trees and random forests, are largely incapable of extrapolating the predicted target values below or beyond the lower and upper limit range of the training data (Loh et al. 2007). This limitation involved in predicting a continuous output such as emissions is addressed by including reported emissions values that range from 0.5 to 1 x 10^9 metric tons output.

Feature and Target Transformation and Scaling

Both predictive features and predicted targets are distributed across multiple orders of magnitude. To determine the data distribution in order to apply the appropriate scaling, continuous predictive features are examined under a logarithmic transformation, $z' = log(z+1)$. The continuous financial features treated with a log-transform resulted in a better behaving normal-like distribution. This is to be expected as the studied financial variables are likely a result of multiplication of different factors, where a log normal distribution is expected, rather than an addition operation, where a normal distribution is expected (Aitchison & Brown 1957). In addition, the model prediction accuracy improved with the log-transformed financial features. The ratio financial features, such as inventory turnover ratio, are not transformed, as their distributions do not benefit from a log transformation.

Once log-transformed, the 15 types of Scope 3 emission targets display a compound Poisson-gamma Tweedie distribution with a cluster of observations as a point mass at zero followed by continuous distribution. This distribution is representative of the landscape of reporting. This also results in an additional degree of difficulty in using regression models to train for a minimal error.

Distance-based models require feature scaling and normalization, such as OLS regression models and k-nearest neighbors (k-NN) (Pedregosa et al. 2011). Two primary methods are used for feature scaling dependent on the data distribution. Standard scaling, or Z-score normalization, is appropriate for normally distributed predictive features and it scales features to unit variance with a zero mean.

$$X'_{standard\ scaling} = \frac{X_i - \bar{X}}{\sqrt{Var(X)}}$$

Rescaling, or minimum-maximum normalization, is applied for non-normally distributed predictive features and rescales the features to fall between the range 0 and 1.

$$X'_{norm} = \frac{X_i - X_{min}}{X_{max} - X_{min}}$$

As a test for appropriate scaling, a normal quantile-quantile (Q–Q) plot is generated to compare the predictive feature with a randomly generated, independent standard normal data. The linearity of the graph is used to determine if features are normally distributed in order to apply the appropriate scaling.

The following features display a normal distribution as defined by the Q-Q plots:

- Total Assets
- Market Capitalization
- Operating Expense
- Selling, General & Administrative Expenses

All other continuous financial features are treated with a minimum-maximum normalization with the distance-based regression models. Standard scaling and normalization are not required for tree-based ensemble models, such as the Random Forest Regressor and AdaBoost Regressor, therefore, standard scaling and normalization of the features did not result in an improved model test score. In these instances, the models were trained with log-transformed features without scaling. Finally, nominal variables were encoded and tested through two different methods: ordinal encoding and one-hot encoding. Ordinal encoding produces a single column of features by encoding categorical features as an integer array to produce one feature of integers to values 0 to n - 1 categories. One-hot encoding encodes categorical features as binary vectors by first mapping the categorical features into integer values and subsequently encoding each integer as a binary vector. Model performance did not vary significantly between the two methods, therefore, ordinal encoding was selected. Ordinal encoding is implemented using Python Sklearn's pre-processing *OrdinalEncoder.*

Feature Selection

Predictive feature selection was performed through a combination of the following: selecting the most widely used variables within current regression techniques, the most operations-relevant financial features reported within the balance sheet and income statement, as well as critically applying recursive feature elimination through regularization techniques that penalize model complexity and prevent overfitting.

Recursive Feature Elimination

To understand the relationships between the predictive features and the target variables, we first examine the feature correlation matrix to develop an initial understanding of the Pearson correlation coefficients between features and targets.

Introducing multicollinearity exposes the model to excess noise that may negatively affect certain model predictions. Tree-based models, which are rule-based, are relatively resistant to the noise introduced through predictive feature multicollinearity (Friedman & Popescu 2008). As such, we keep the majority of features throughout each model to understand which features are most successful at predicting the individual 15 targets. However, this may become a stronger issue for linearized models that are used as benchmarks, where introducing excess collinearity possibly leads to overfitting or increased error. The impact of this can be mitigated through recursive feature elimination.

Feature selection was performed through a combination of selecting the most commonly reported financial and emission features as well as through recursive feature elimination. Features selected by hand represent both financially relevant as well as commonly used features in emissions predictions regression techniques as applied by emissions data providers. As a separate test to measure the quality of predictions of the selected features, recursive feature elimination (RFE) with cross validation was applied using a Lasso CV, Random Forest and Gradient Boost mask by voting. Recursive feature elimination does not make prior assumptions about the most relevant features but rather fits the model to all the features recursively and upon each loop eliminates the least predictive feature. RFE is useful to eliminate interdependencies and collinearity that may exist between the model features in order to reduce noise within the data (Pedregosa et al., 2011). In this instance, RFE was applied to the training data initially with all features fitting to three models: Lasso CV, Random Forest Regression and Gradient Boosted decision trees and each model voted on which features to keep or eliminate. The features with the most votes are retained and the features with the least votes are eliminated. Each Scope 3 type target is fitted independently with distinct features selected for each target. We conclude that the targets selected by hand on average agree with the targets selected through recursive feature elimination. By applying recursive feature elimination, some non-commonly used financial features were identified, such as inventory turnover ratio. Recursive feature elimination is implemented using Python Sklearn's feature_selection *RFE.*

After the data selection and pre-processing, 9,013 firm-year observations remain for 1,938 unique firms. Two general models are constructed: one model is structured to predict 15 Scope 3 emission types as individual targets while the other model is structured to predict one emissions output using the reported emission's attributed scope 3 type as an additional predictive feature. Each predicted target has access to the following predictive features: 2 nominal variables, 11 financial variables, 7 financial ratio variables and 2 emissions variables.

Data Subsets and General Models

Given the complexity of the reporting landscape, the selected models were run on different subsets of data that varied the quality of the data responses both in acceptable responses (zero or non-zero emissions) as well as in method of calculated emissions (primary supplier reported data or estimated reported data). The first method accepted zero emissions responses as ground truth entries for observations. Allowing zero entries in responses creates inconsistent patterns to base model predictions on. The model has difficulty in predicting zero responses while simultaneously taking into account non-zero responses. Although there are some discernable patterns within zeros responses, such as less relevant Scope 3 types within specific industries, these patterns extend at random to firms that have not yet calculated these emissions or firms with varied responses for Scope 3 type relevancy. As such, the model prediction produced by allowing zero emissions responses systematically underestimate Scope 3 emissions across all types. As a response to this reporting, a second data method was constructed to drop zero emissions responses as ground truth entries for samples. When zero emissions responses are dropped from the data set, model performance improves significantly, while in some cases, overestimating emissions.

In addition to constraining responses for zero and non-zero emissions, a primary data measure was constructed to control for the quality of responses. Primary data is reported by the firm as being calculated using supplier (or customer) sourced emissions data. The data set is further constrained by limiting the sample set to be 80% or greater percentage of primary data. The model trained using primary data performs well within Scope 3 types that have sufficient samples. Model performance can only be measured and validated on the test set, which comprises 20% of the sample set. The constraint on primary data yields a training set that is as low as 250 training samples and a test set that is as low as 50 samples, in the case of primary data entries for "Processing of Sold Products". This reduction in sample size makes it difficult to validate the model performance using primary data for low reported Scope 3 types.

Models: By Type vs Singular

Two primary approaches are taken towards constructing a general model. Each machine learning model is trained on a general targeted by type or singular model. Model 1 is a targeted by type model which creates an individual model for each of the 15 Scope 3 types, resulting in 15 fits. Model 2 is a singular model which uses the 15 Scope 3 types as an additional singular feature to the data set. Model 2 has consistent performance across Scope 3 types by being able to leverage a more extended data set, instead of having a reduced sample set to train on, as is the case with specific targets in the targeted by type Model 1. Model 1 displays improved performance on some individual Scope 3 types but performs poorly when the sample set is comprised of significantly reduced observations. Model 2 avoids both extremes by having the ability to generalize well using data across all Scope 3 types in one model.

Model Selection and Evaluation

Benchmarking

To assess the performance of current industry estimates from data providers, alternate prediction models are constructed. The performance of selected machine learning models, k-NN, Random Forest and AdaBoost, in comparison to existing prediction models used by data providers, such as OLS and Gamma-GLM models.

Ordinary Least Squares Regression

As a first measure of estimation between features and targets, an ordinary least squares linear regression is fitted to the data. A linear regression fits a linear model between features and targets to minimize the residual sum of squares between the observed targets, or Scope 3 types, in the dataset, and the targets predicted by the linear approximation. The *GammaRegressor* is implemented using Python Sklearn library with a log link function.

Generalized Linear Models, Gamma-GLM

The data provider, CDP, applies a multi-variable Gamma-Generalized Linear Model (Gamma-GLM) using revenue and activity information to estimate Scope 3 emissions.[7] Each of the 15 Scope 3 types has an independent multi-variable regression model where activity-revenue is the independent variable. The CDP model assumes that revenue is directly proportional to production and therefore proportional to emissions. The emissions associated with ‘Employee Commuting’ are estimated using the number employees and the emissions associated with ‘Capital Goods’ are estimated using capital expenditure, both as reported by the firm. In addition to these assumptions, CDP applies the CDP Activity Classification System (CDP-ACS) hierarchy to their regression model. This categorial classification system, developed by CDP, provides a framework that focuses on quantifying a company’s environmental impacts connected to its activities to ensure that the environmental impacts are as consistent as possible. Rather than general broad-based industry classifications, CDP firms are group by primary industry, sector, and "activity". The environmental granularity of the CDP-ACS allows firm "activity" to be used as a predictor variable. To reproduce a benchmark model according to CDP's Scope 3 emission model, a Generalized Linear Model with a Gamma distribution is implemented on the data set. A Gamma-GLM regressor is fit to each Scope 3 Type using all available financial and categorical features. The primary differences between the two approaches include industry classification, selection of "Applicable/Not Applicable" criteria and the use of collinear features. The aim of the machine learning model is to be able to produce accurate predictions using widely reported financial features as well as industry-based classifications. For this reason, we structure our Gamma-GLM benchmark model to fit to the CDP regression model while preserving the features used within the subsequent machine learning models. In the Gamma-GLM benchmark model, the GICS sub-industry

[7] CDP. Full GHG Emissions Dataset. Technical Annex IV: Scope 3 Overview and Modelling.

classification is used as a categorial feature, rather than the CDP-ACS, and all available firm financial features are applied. The *GammaRegressor* is implemented using Python Sklearn library with a log link function.

K-nearest neighbors

The k-nearest neighbors (k-NN) regressor is a non-parametric algorithm that calculates predictions based on a measure of similarity defined by the minimal distance between samples (Pedregosa et al., 2011). The k-NN regressor calculates the continuous target by taking the average of the k nearest neighbors where distance is evaluated using the default Minkowski metric. The *KNeighborsRegressor* is implemented using Python Sklearn library with a default of *n_neighbors = 5* and calculated as follows:

$$d_{a,b} = \left(\sum_{i=1}^{k} (|x_{ai} - y_{bi}|)^{q} \right)^{1/q}$$

where *q* = 1 for the Manhattan distance and q = 2 for the Euclidean distance.

Ensemble Methods

Ensemble methods are a category of non-parametric machine learning algorithms constructed to make prediction based on the combined collective action of different estimators. Tree-based ensemble methods are relatively robust against overfitting, are not heavily impacted by the multi-collinearity of input features and perform well within noisy data including outliers. In contrast to linear models, tree-based ensemble regressors need minimal data pre-processing and are not sensitive to scaling and normalization. The following models apply decision tree (CART) base estimators to create flexible models with reduced bias and variance (Maclin & Optiz 1999).

Random Forest Regressor

The Random Forest algorithm is a parallel ensemble learning meta-estimator that primary aims to decrease mode bias using bagging (Breiman 2001). In ensemble algorithms, Random Forest bagging methods build a randomized forest of a decision tree estimators on random subsets of the original training set drawn with replacement and aggregate individual predictions to form a final prediction. These methods reduce the variance and overfitting of the decision tree by introducing randomization into the construction and then build an ensemble using averaging of predictions. The *RandomForestRegressor* is implemented using Python Sklearn library with default parameters.

Adaptive Boosting

The Adaptive Boosting algorithm (AdaBoost) is a sequential ensemble learning meta-estimator that primarily aims to decrease model bias using boosting (Freund & Schapire 1996). Boosting achieves minimal training errors by combining a series of weak base learners to create a collectively stronger predictor that minimizes the sum of squared error residuals of predictions. In this application, the model initializes the boosting algorithm on a sequential forest of decision trees paired to a linear loss function.

Similar to Random Forest, AdaBoost is an ensemble method that trains a forest of decision tree regressors. In the case of AdaBoost, only shallow trees, called decision stumps, are formed. Decision stumps are typically one node and two leaves and are known as weak learners. Weak learners focus on using only one variable input feature to make a prediction. Unlike Random Forest, in AdaBoost, the order in which weak learners are constructed is directed by the error of previous learners. In this way, AdaBoost guides and informs the construction of sequential decision stumps dependent upon the performance of the previous stumps, rather than independent, in determining predictions. Unlike other tree methods, in AdaBoost, not all stumps carry equal weight, or significance, used to form a prediction. Varying the decision stump significance allows AdaBoost to focus on correcting the residuals of difficult predictions in order of greatest to least sum of squared error residuals (Drucker 1997).

The samples within a data frame are initialized with equal importance called sample weights. All sample weights sum up to 1. The sample weight is increased or decreased depending on the prediction quality, which is the ability of the weak learner to correctly predict the target as measured by increasing to decreasing sum of squared residuals error for each decision stump predictor. The order in which decision stumps are selected to train on the data set is measured by the sum of squared residual errors. The best decision stump predictor which results in the smallest sum of square residuals total prediction error is used to begin the training and to subsequently create the sequential decision stumps. The total error of each predictor is the sum of square residuals total prediction error of the data set as measured by minimizing the AdaBoost linear loss function.

1. AdaBoost initializes decision stump regressors known as weak learners with equal sample weights.

   $$ {w_i}^{(1)} = 1 \quad i = 1, \ldots, N_1 $$

2. AdaBoost trains the decision stump weak learners based on the equal sample weights and obtains predictions

   $$ y_i^{(p)}(\mathbf{x}_i) \quad i = \mathbf{1}, \ldots, N_1 $$

3. AdaBoost measures the error in predictions of each weak learner by calculating the linear loss function for each training sample and averaging the loss over the training data set

$$ L_i = \frac{\left| y_i^{(p)}(\mathbf{x}_i) - y_i \right|}{D} \qquad \text{where} \quad D = sup \left| y_i^{(p)}(\mathbf{x}_i) - y_i \right| \quad i = \mathbf{1}, \ldots, N_1 $$

Calculate average loss: $\bar{L} = \sum_{i=1}^{N_1} L_i p_i$

4. AdaBoost determines the significance, β, to assign to each stump based on the magnitude of compensation of the previous error. The greater the compensation or reduced sum of squared residual error, the higher the measure of confidence in the predictor which results in a lower β or significance assigned to that decision stump.

$$\beta = \frac{\bar{L}}{1 - \bar{L}}$$

5. Sample weights are updated based on the loss function. The updated sample weight will increase for larger average loss and decrease for smaller average loss. The decrease in sample weight for smaller average loss reduces the probability that those samples will be chosen within the next training set for the next decision stump in the ensemble.
6. Each subsequent weak learner is informed by the errors of the previous weak learner and uses this information to determine the construction of subsequent weak learner.
7. Varying the sample weights to focus on predictions that are more difficult to estimate means that each sequential machine has a disproportionately more difficult subset of training samples to learn from. The average loss increases over iterations until the bound on the loss function $\bar{L} = 0.5$ is not satisfied and the algorithm terminates.

The AdaBoost algorithm approximates the expectation of the modeling and prediction error using the training set observation average over multiple experiments.

$\mathrm{y}^{(p)}(\mathbf{x})$ sample modeling error (ME) and prediction error (PE) are defined as follows:

$$PE = \frac{1}{N_2} \sum_{i=1}^{N_2} \left[y_i - y_i^{(p)}(\mathbf{x}_i) \right]^2 \qquad and \qquad ME = \frac{1}{N_2} \sum_{i=1}^{N_2} \left[y_i^{(t)} - y_i^{(p)}(\mathbf{x}_i) \right]^2$$

where $y_i^{(p)}(\mathbf{x}_i)$ is the prediction for the *i* th test sample and y_i the *i* th test sample where $y_i^{(t)}$ is the *i* th test sample ground truth (company reported value). N_1 is training set, $N_2$is the test set (Drucker 1997).

Model Evaluation

Feature Importance

Feature importance is a measure used to calculate the relative predictive performance score of each input feature in a model for each target. The score is indicative of the predictive strength of an input feature. Features with a higher feature importance have a larger impact on the model predictions relative to the other input features. Within the base estimator used with AdaBoost, a decision tree regressor, the feature importance is calculated using the mean and standard deviation of accumulation of the Gini index, or node impurity, decrease within each tree.

Model Evaluation Metrics

Model are evaluated along on three metrics:

- *R^2 Regression Score*
- *MAPE: Mean Absolute Percentage Error*
- *RMSLE: Root Mean Squared Log Error*

R^2 Regression Score

The R^2 regression score function evaluates the of goodness of fit of the model where the best possible score is 1.0 (Pedregosa et al., 2011). The R^2 regression score function represents the proportion of variance in the target, or dependent variable, that can be attributed to the independent variables, or input features, in the model. The R^2 regression score function provides measure of the goodness of fit between the data and the model and serves as a measure of how well unseen test samples are likely to be predicted by the model, through the proportion of explained variance. R^2 is not comparable between different datasets or across different targets given that variance is dataset dependent. The R^2 regression score function is implemented using Python Sklearn library and calculated as follows:

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

where the *i*-th sample model prediction is $\hat{y}_i$, y_i is the corresponding ground truth reported value and

$$\bar{y} \; = \frac{1}{n}\sum\nolimits_{i=1}^{n} y_i$$

MAPE: Mean Absolute Percentage Error

The mean absolute percentage error calculates a measure of prediction accuracy as a ratio of relative error between ground truth value and the predicted value (Pedregosa et al., 2011). The difference is divided by the corresponding ground truth value and this ratio is summed for every predicted sample. The mean absolute percentage error is implemented using Python Sklearn library and calculated as follows:

$$MAPE(y,\hat{y}) \; = \frac{1}{n}\sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{max(\epsilon, |y_i|)}$$

where ϵ is a small positive number to apply when $y_i = 0$

RMSLE: Root Mean Squared Log Error

The root mean squared log error score calculates a measure of difference between the expected value and the predicted value of samples to produce a risk metric corresponding to the expected value of the squared logarithmic error or loss (Pedregosa et al., 2011). RMSLE is most applicable to measure which have an exponential distribution whose values are products of multiplicative operations. RMSLE is robust in handling data outliers without drastically increasing the relative error, as happens in RMSE. RMSLE is primary calculating the relative error between predictive values and corresponding ground truth values and incurs a larger penalty for the underestimation of the ground truth sample values than overestimating. The root mean squared log error score is implemented using Python Sklearn library and calculated as follows:

$$RMSLE(y, \hat{y}) \; = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(1+y_i) - log(1+\hat{y}_i))^2}$$

K-fold Cross Validation

K-fold cross validation is applied to evaluate the model performance as an estimate of generalizability or expected model performance on unseen data (Pedregosa et al., 2011). The k-fold produces reduces both bias and variance towards any particular train-test split. The k-fold cross-validation procedure divides the data set into k non-overlapping folds. K-1 folds are used as a training set while the remaining data is used as a test set. A total of k models are fit and evaluated on the individual k test sets and the mean performance score with standard deviation is calculated. Two scoring metrics, mean absolute percentage error and root mean squared log error, are evaluated on multiple test set through k-fold cross-validation where k =10. The performance measure reported by k-fold cross-validation is average of the values computed in the 10-fold loop. The resulting CV score demonstrates the generalizability of the model or expected performance metric base on the averaged result across the k-fold iterations.

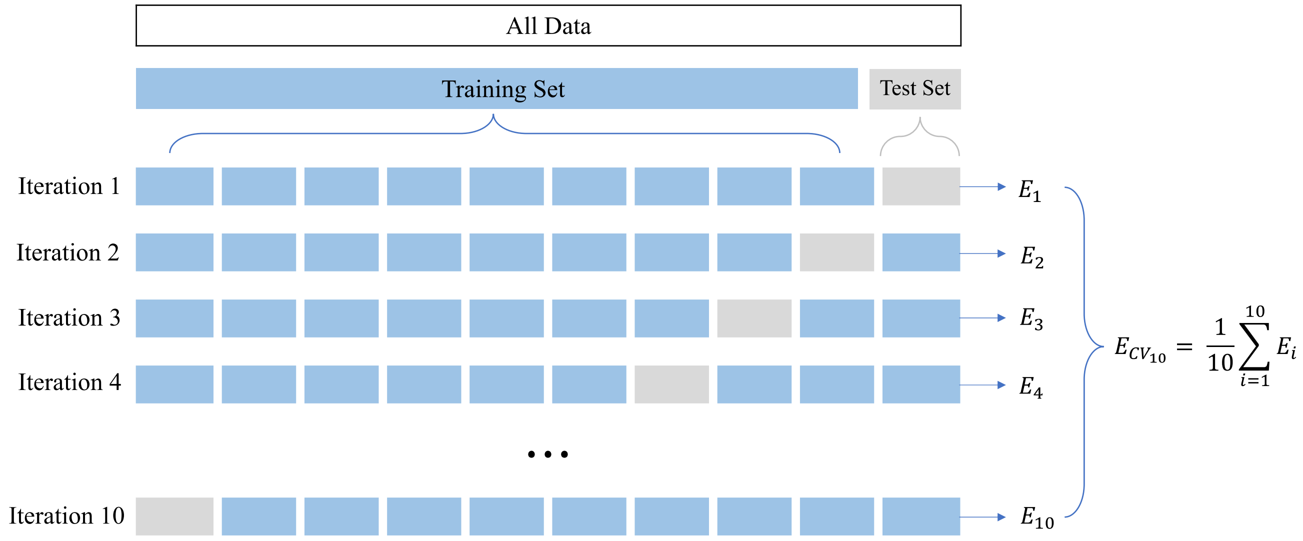

Figure 4 | The k-fold cross validation process reduces both bias and variance towards any particular train-test split. The k-fold cross-validation procedure divides the data set into k non-overlapping folds. K-1 folds are used as a training set (80%) while the remaining data is used as a test set (20%).

Figure 5 | Mean Absolute Percentage Error (MAPE) performance across 10- fold cross validation

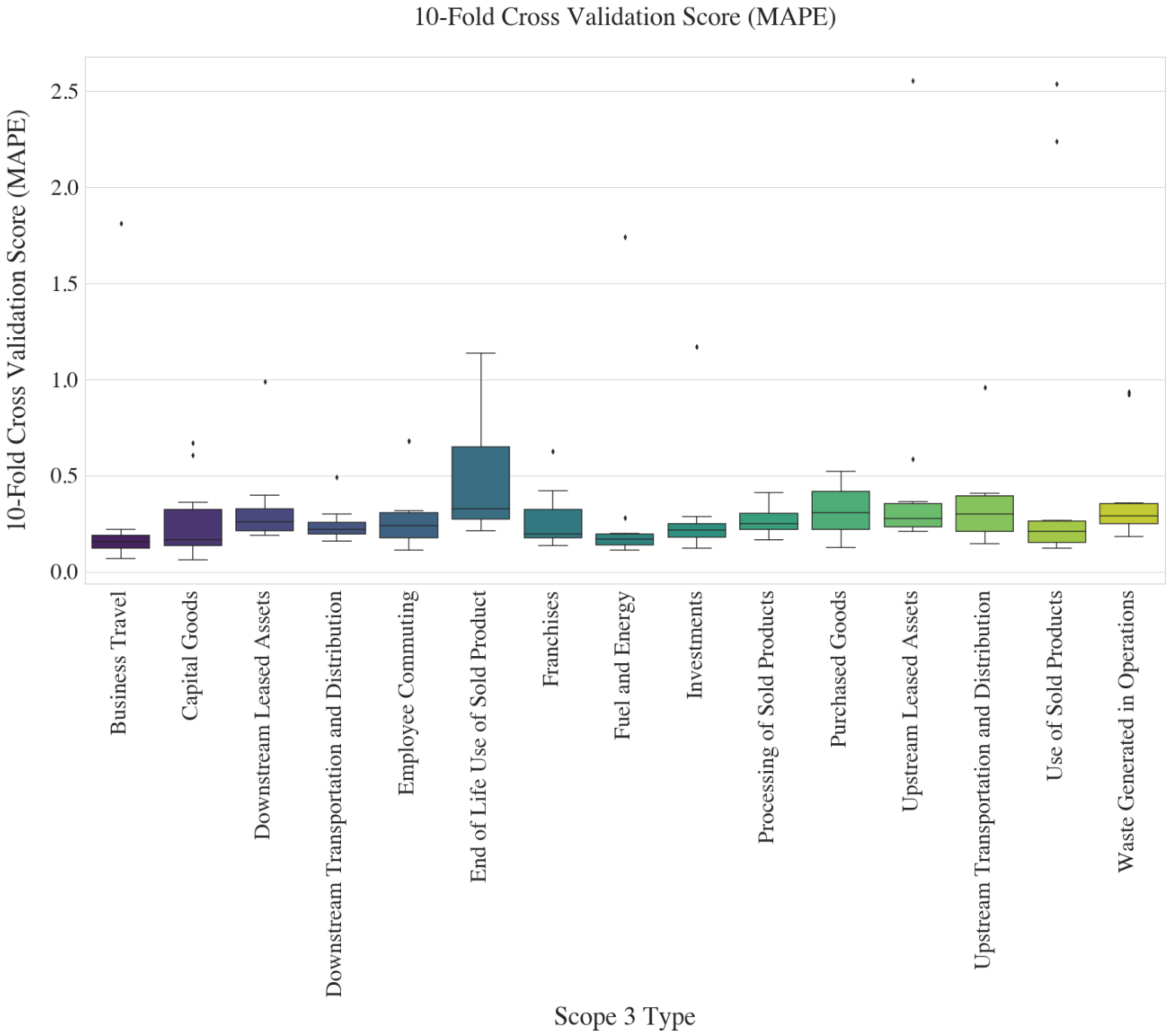

Figure 5 | MAPE boxplot of cross validation scores for each Scope 3 types. The model is run 10 times with randomly shuffled subsampling of the training and test data for each model fit. Given that the data set sample includes multiple firm year observations for any single firm, a constraint is applied such that firms only appear in either the training data or the test data subsets. We design a test train split to keep all firm year observations of any specific firm within either the test or the train split, but not both. The data shows that business travel, downstream transportation and distribution, employee commuting, processing of sold products, purchased goods and services repeatedly obtained similar mean absolute percentage errors across distinct subsets of train and test data splits. Use of sold products and upstream transportation and distribution perform very poorly in terms of predictive accuracy across shuffles train and test splits.

Figure 6 | Root Mean Squared Log Error (RMSLE) performance across 10- fold cross validation

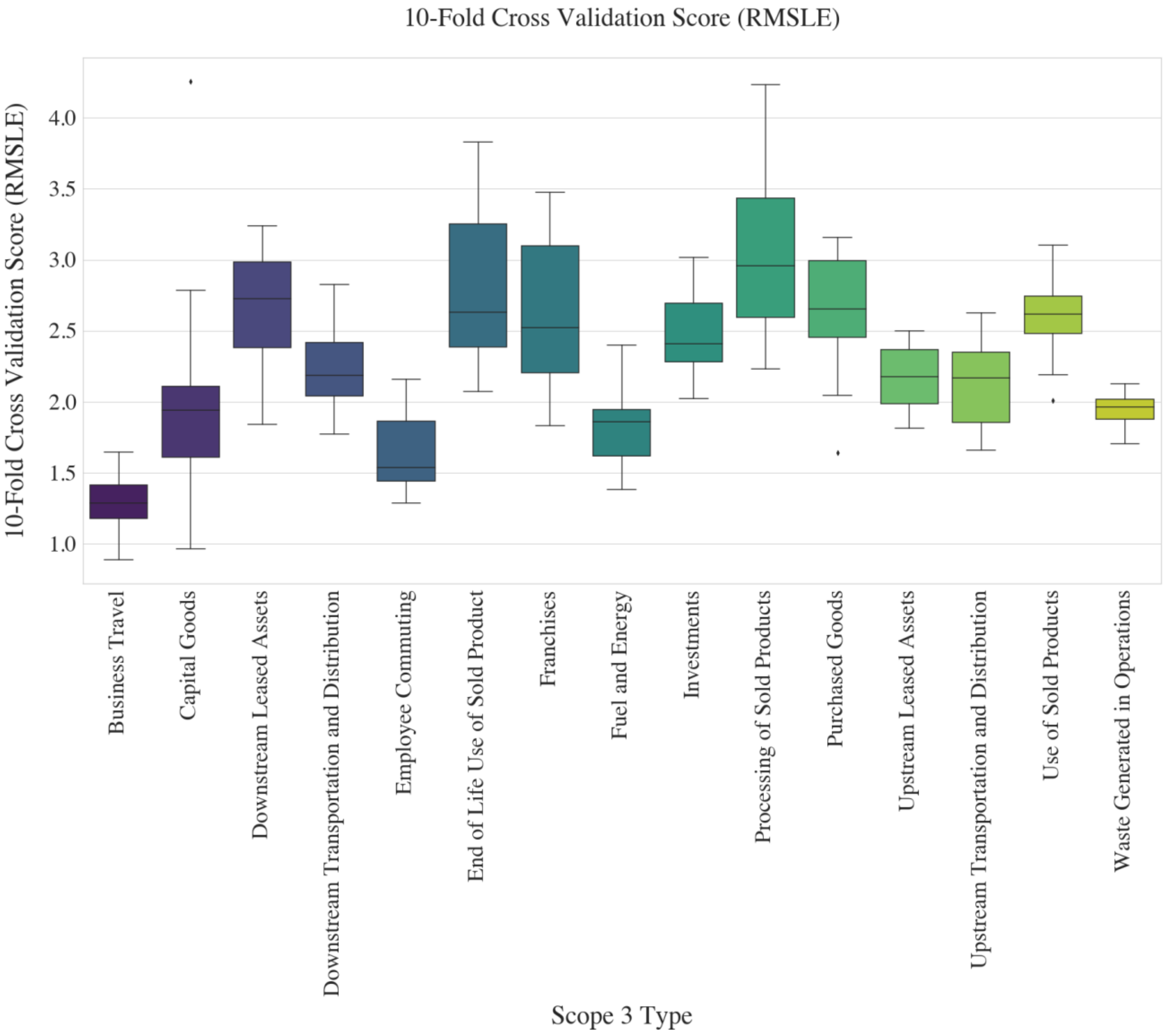

Figure 6 | RMSLE boxplot of cross validation scores for each Scope 3 types. The model is run 10 times with randomly shuffled subsampling of the training and test data for each model fit. Given that the data set sample includes multiple firm year observations for any single firm, a constraint is applied such that firms only appear in either the training data or the test data subsets. We design a test train split to keep all firm year observations of any specific firm within either the test or the train split, but not both. The data shows business travel, downstream transporation and distribution, employee commuting, fuel and energy and waste generated in operations have the smallest spread in log error across the 10 model fits. Wide variation in prediction error are found in end of life treatment of sold products, franchises and processing of sold products.

References

Aitchison, J., & Brown, J. A. C. (1957). The Lognormal Distribution, with Special Reference to Its Uses in Economics . In *https://doi.org/10.1086/258070* (Vol. 66, Issue 4). Cambridge University Press, University of Cambridge Department of Applied Economics. https://doi.org/10.1086/258070

Breiman, L. (2001). Random Forests. *Machine Learning 2001 45:1*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

CDP. (2020). *CDP North America Annual Report 2019-2020*. https://cdn.cdp.net/cdp-production/cms/reports/documents/000/005/234/original/CDP_NA_2019-20_Annual_Report.pdf?1591886351

Cheema-Fox, A., Realmuto LaPerla, B., Serafeim, G., Turkington, D., & Wang, H. (2021). (2021) Decarbonizing Everything. *Financial Analysts Journal*, *77*(3), 93–108. https://doi.org/10.1080/0015198X.2021.1909943

Dietterich, T. G. (2000). *An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization*. *40*, 139–157.

Drucker, H. (1997). Improving regressors using boosting techniques. *14th International Conference on Machine Learning*, 107–115.

Freund, Y., & Schapire, R. E. (1996). *Experiments with a New Boosting Algorithm*. http://www.research.att.com/

Friedman, J. H., & Popescu, B. E. (2008). Predictive Learning via rule emsembles. *The Annals of Applied Statistics*, *2*(3), 916–954. https://doi.org/10.1214/07-AOAS148

Janet Ranganathan, Laurent Corbier, Pankaj Bhatia, Simon Schmitz, Peter Gage, & Kjell Oren. (2004). *The Greenhouse Gas Protocol: A Corporate Accounting and Reporting Standard*. https://ghgprotocol.org/sites/default/files/standards/ghg-protocol-revised.pdf

Klaaßen, L., & Stoll, C. (2021). Harmonizing corporate carbon footprints. *Nature Communications 2021 12:1*, *12*(1), 1–13. https://doi.org/10.1038/s41467-021-26349-x

Loh, W.-Y., Chen, C.-W., & Zheng, W. (2007). Extrapolation errors in linear model trees. *ACM Trans. Knowl. Discov. Data*, *1*(6). https://doi.org/10.1145/1267066.1267067

Opitz, D., & Maclin, R. (1999). Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligence Research*, *11*, 169–198.

Pankaj Bhatia, Cynthia Cummis, Andrea Brown, David Rich, Laura Draucker, & Holly Lahd. (2010). Corporate Value Chain (Scope 3) Accounting and Reporting Standard . In *GHG Protocol*. https://ghgprotocol.org/sites/default/files/standards/Corporate-Value-Chain-Accounting-Reporing-Standard_041613_2.pdf

Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://scikit-learn.org/stable/about.html

Rubinsteyn, A., & Feldman, S. (2016). *fancyimpute: An Imputation Library for Python*. https://github.com/iskandr/fancyimpute

Yang, Y., Ingwersen, W. W., Hawkins, T. R., Srocka, M., & Meyer, D. E. (2017). USEEIO: A new and transparent United States environmentally-extended input-output model. *Journal of Cleaner Production*, *158*, 308–318. https://doi.org/10.1038/S41597-022-01293-7