Filling in the blanks*

*Estimating corporate emissions using machine learning.

2024 UPDATE

DR KIERAN BROPHY AND MIN LOW

esgbook EthiFinance

ESG Book is now part of EthiFinance. This document was originally published by ESG Book prior to the integration and may reference ESG Book branding, products, or services that are being transitioned to EthiFinance. Please visit ethifinance.com for more details.





Executive Summary

ESG Book's newly upgraded Emissions Estimation Model is a powerful tool that allows investors to enhance their understanding of financed emissions.

Using machine learning, the latest version of ESG Book's model estimates Scope 1, 2, and 3 emissions - including all 15 categories of Scope 3 - for over 45,000 companies worldwide, with coverage dating back to 2013.

ESG Book's model is comprised of 900 sub-models, with each developing relationships between readily available company-level data and emissions for a given industry, geography, and emissions Scope.

Our approach, which has been tested against conventional multi-variable regression models alongside other machine learning models, has been found to predict emissions more accurately across each Scope.

Introduction

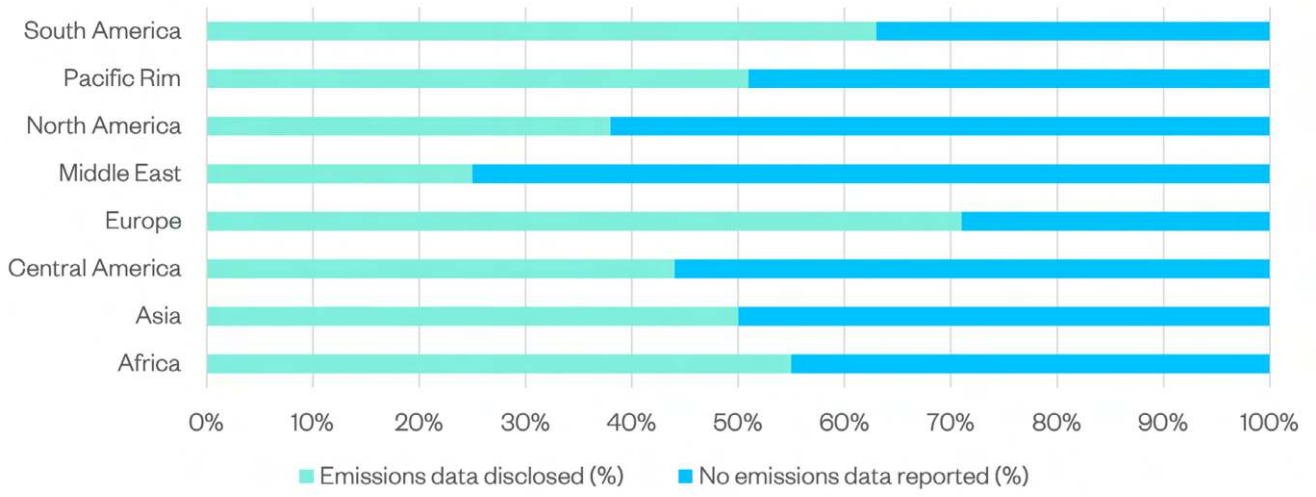
The number of climate-related regulations and frameworks is increasing every day. At the same time, corporate disclosure of greenhouse gas (GHG) emissions, particularly those in line with the Greenhouse Gas Protocol[a] (GHGP), remains poor. Of more than 6,200 companies tracked by ESG Book, less than 50% report Scope 1 emissions in line with the GHGP. As shown in Figure 1, this value varies considerably across sectors and countries.

Another challenge is the higher level of Scope 1 and Scope 2 disclosure versus Scope 3. Most emissions from carbon intensive sectors, such as energy, airlines, and automobiles fall under Scope 3 - i.e. the products and services it sells (downstream emissions) and its supply chain (upstream emissions). Therefore, without Scope 3 emissions data it is easy to draw unhelpful conclusions.

For example, a solar energy company could have a higher carbon intensity in its Scope 1 and 2 emissions than a large multinational energy company benefitting from economies of scale and efficiency in Scopes 1 and 2. However, the majority of emissions in the energy industry come from Scope 3. Therefore, not taking Scope 3 into consideration could lead to inconsistencies and inaccurate conclusions.

Increasingly, investors are realising the importance of measuring the emissions associated with their portfolio companies. However, actual disclosure of corporate emissions, particularly those in line with the GHGP remains low. Less than 4,000 of the world's leading companies currently report their emissions in line with the GHG Protocol for Scope 1 and 2 emissions.

Figure 1. Regional breakdown of the disclosed emissions of companies that ESG Book tracks the climate data of.



In order to fill this gap, ESG Book have released an Emissions Estimation model, developed from a best-in-class machine learning based model.

Prediction model(s)

Estimating the emissions of companies requires an understanding of the relationships between several readily available data points (sector, region, revenue, asset value, energy consumption, or emissions) and using these relationships to predict the emissions for companies where this data is missing. These data points are termed model predictors. Several ways of identifying the best relationship between model predictors and emissions exist. ESG Book tested several approaches to identify the best relationship between model predictors and emissions. The following section (overleaf) outlines potential model predictors and describes which would act as a good predictor of a company's emissions.

a.https://ghgprotocol.org

Model predictors

Emissions data from companies that have reported in line with the GHGP are used as model predictors. Just under 4,000 companies have reported Scopes 1 and 2 emissions in line with the GHG Protocol, whereas approximately 3,000 report Scope 3.

Predictors that indicate the size of a company within an industry, as well as indicators that improve the performance of the model have been chosen. Although the Pearson correlation between predictors and emissions per scope was analysed, this approach was not formally included as a basis for choosing a predictor since the relationship between predictor and emission might not be a linear one.

Non-linear effects could include; when a company is scaling-up, it might have more capital to allocate to insulate its buildings, requiring less energy per unit goods produced or services rendered, which would therefore result in a non-linear emission relationship. A relatively small number of model predictors were needed for the model to generate good results. Recent studies have had similar findings[1,2].

The chosen predictors are shown in Table 1. As reasoned in previous studies, a logarithmic transformation of each numeric predictor variable is taken before input into a model[1].

Table 1. An overview of the predictor variables used in the predictor models. Revenue (scaled) refers to revenue which is scaled to account for inflation. The scope of emissions being estimated is not used as a predictor variable, but the other scopes are.

*** Data not used as predictor variables, but in defining regressions.**

Category	Variable	Units
Industry classification	Industry Industry grouping* Economic sector*	Categorical
Geographic classification	Region Country	Categorical
Financial metrics	Total assets Total equity Gross income Market capitalisation Revenue Revenue (scaled) Operating expenditure	log (million USD$)
Other metrics	Number of employees Energy consumption	log (people) log (kilowatt-hour)
Emissions metrics	Scope 1 Scope 2 Scope 3 Scope 1+2	log (tonnes CO2e)

It is important to note that emissions can be highly dependent on both industry and geography. For example, aviation emissions have a unique emissions profile compared to other industries (emissions per dollar earned for an aviation company will look quite different to a rail company).

However, aviation emissions do not vary significantly by geography. For example, an aircraft will have approximately the same emissions in North America as it does in South-East Asia. The model must not therefore include aviation emissions in the broad 'Transportation' industry category. Whereas the model does not need to make meaningful adjustments for location. Conversely, Electric Utilities can be very geographic dependent. For instance, the emissions profile of utilities in China will vary considerably to the emissions profile of utilities in the UK.

These examples demonstrate how important it is that the model can accommodate both sectoral and geographic differences in emissions.

Company specific model

Several types of model were examined to see which would yield the most accurate results. Firstly, we look at a company-specific model. This is a relatively simplistic model for companies which have disclosed their emissions for a year, or multiple years, but not others.

To estimate the emissions for those years which are not disclosed, a linear scaling approach is used. This is a model that linearly scales emissions backwards or forwards in time based on revenue and the number of employees.

Predicting a given company's emissions using this approach can be summarised by the following:

$$Emissions_y = Emissionsx \frac{(\alpha + \beta)}{2}$$

Where ‘y’ is the year we want to estimate, and ‘x’ is the closest year to ‘y’ when a company has disclosed:

$$Where \quad \alpha = \frac{revenue_y}{revenue_x} \text{ and } \beta = \frac{employees_y}{employees_x}$$

Although the relationship between emissions and revenue and employees might not be exactly linear for the aforementioned reasons, it was found to be accurate in predicting a specific company's emissions. This approach can estimate the emissions of circa 4,800 companies from 2013 to present year.

Industry specific model(s)

Since emissions data for the vast majority of companies does not exist, a relationship between predictor variables and emissions within an industry and geography must be developed.

The following 3 models were compared:

01 A conventional regression, specifically a Ridge Regression

02 An Adaptive Boosting decision tree model, and finally

03 An Extreme Gradient (XG) Boosted decision tree model

To determine if a model was performing better than the others, the data was split in the following sets with different proportions: 70% training, 15% validation, and finally 15% holdout.

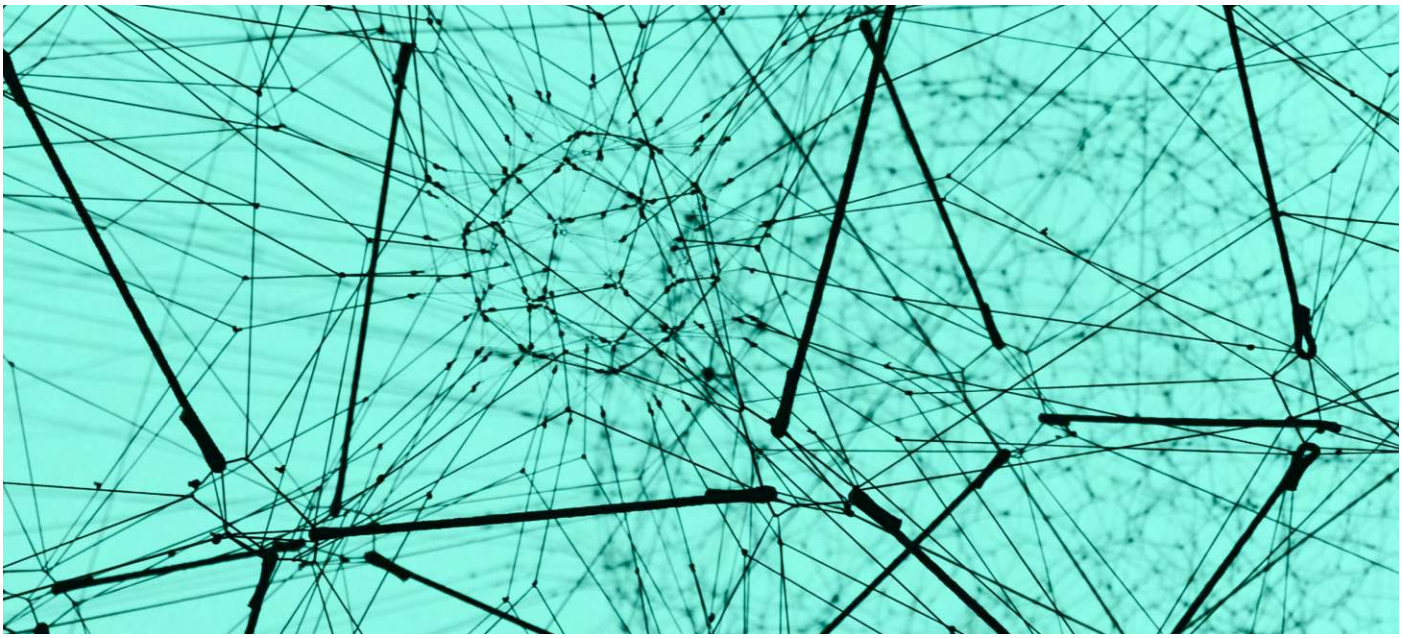
The performance of the trained models are assessed by predicting against the holdout set and comparing the estimated emissions from the models against the actual emissions of companies.

Ridge Regression

A Ridge regression is a multivariable regression where the loss function is the linear least squares function[3]. This means that the regression tries to minimise the difference between disclosed emissions and emissions predicted by a linear approximation.

Ridge regressions are used in data sets where there are more predictor variables than observations, as it is a better predictor than Ordinary Least Squares. This is because the Least Squares method cannot tell which predictor variable is more or less useful in predicting emissions, hence reducing the accuracy of the model. Importantly, it can also handle multicollinearity, whereby predictor variables are correlated, which can cause inaccuracy in a regression.

The Ridge regression is implemented using Python Sklearn library RidgeCV[b], an inbuilt Ridge regression with built in cross-validation. Different values for α, the constant that controls the variance of the estimates, were trialled and the optimal value chosen per regression.



Machine learning models

Two separate machine learning models were tested: Adaptive Boosting and eXtreme Gradient Boosting.

Importantly for our analysis, both are able to handle non-linear relationships between emissions and predictor variables.

Both models use decision trees (a branching method to demonstrate every possible output for a specific input, which evaluates to being either true or false) to determine the optimal relationship between predictor variables and emissions. Both predictive models learn from the mistakes of more simplistic models. An initial model is created, then a second is created that tries to reduce the errors of the previous model.

With machine learning models, parameters known as hyperparameters must be chosen carefully as they can control a model's learning process. Hyperparameter tuning was carried out using Optuna[c]. The model is run for a specified number of trials, stratified per industry using the training and validation sets, and hyperparameters that yield the lowest errors to the 15% validation set are used in the final prediction model.

b. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.RidgeCV.html
c. https://optuna.org

Adaptive Boosting

A recently published paper recommends Adaptive Boosting as a predictor model[1] compared to various other machine learning models to predict Scope 3 emissions[4].

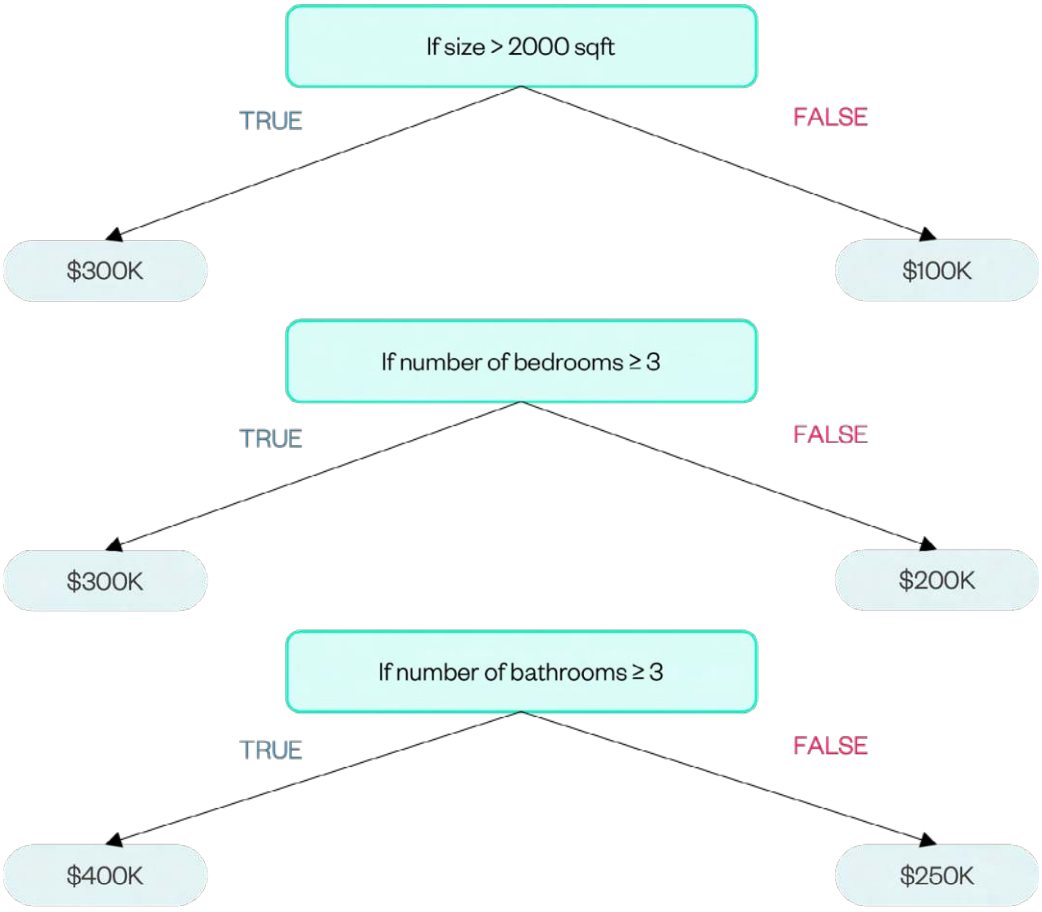
The Adaptive Boosting algorithm (AdaBoost) is a sequential decision tree that, as illustrated in Figure 2, uses shallow decision trees (known as decision stumps) of only one node and two leaves. AdaBoost uses a forest of such stumps rather than trees and 'adapts' by varying the importance of each stump to reduce errors and build the optimal relationship.

A simple analogy for this is trying to determine the price of a house based on physical attributes about the house such as number of bedrooms and bathrooms. The hyperparameters that were tuned were the number of estimators (maximum number of tree stumps at which boosting is terminated) and the learning rate (the weight applied to each regressor at each boosting iteration). A higher learning rate impacts the contribution of each decision stump.

Previous studies have indicated that noisy data can lead to poor performance in AdaBoost models[5]. In ESG Book's emissions dataset, it can be assumed that there will be at least some noise due to differences in companies' emissions profiles within industries.

The Adaptive Boost regression is implemented using Python Sklearn library AdaBoostRegressor[d].

Figure 2. Overview of Adaptive Boosting (AdaBoost) model that uses a sequential set of decision stumps to predict house prices given the number of bedrooms and number of bathrooms. AdaBoost 'adapts' by varying the importance of each stump to reduce errors and build the optimal relationship .



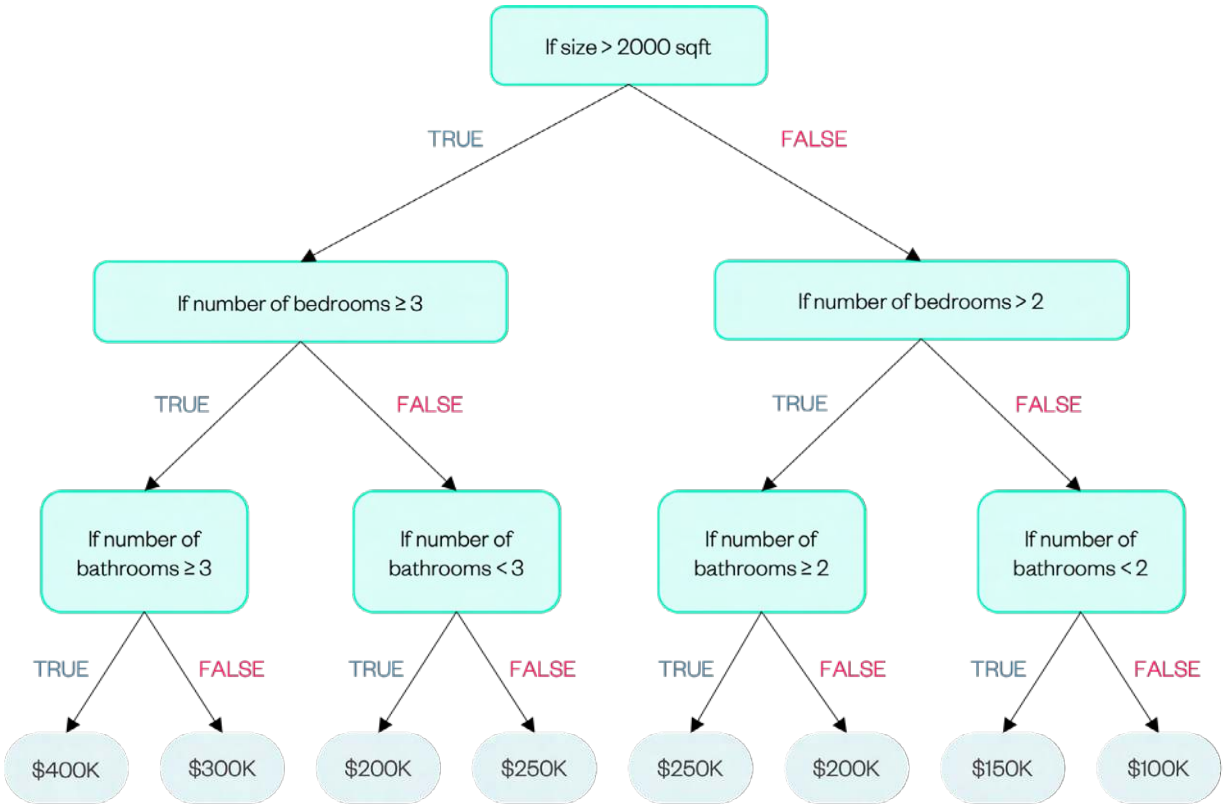
d. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostRegressor.html

Extreme gradient boosting

XGBoost is a relatively new approach that utilises the concept of gradient tree boosting[6]. Instead of the decision stumps in Adaptive Boosting, XG Boosting grows a decision tree sequentially and learns from previous iterations. The algorithm determines the optimal values for each leaf and minimises the overall error of the tree. This is called gradient boosting because it uses a gradient descent algorithm to minimize the loss (penalty for bad prediction) when adding new trees.

XG Boosting uses full decision trees to determine the relationship between input data and emissions. The house price analogy in XG Boost would take the form of Figure 3.

Figure 3. Overview of the Extreme Gradient Boosting (XGBoost) model that uses a set of decision trees to predict house prices given the number of bedrooms and number of bathrooms.



As well as its ability to build non-linear relationships, the XG Boost model can handle missing input data from a company. Therefore instead of interpolating (creating an estimate model for inputs into an emissions estimate model will lead to compounding model error), or filling with zeros (which can also lead to model error), some inputs can be left blank.

One of the added complexities of XG Boost is that it has more hyperparameters that need tuning. Here, the hyperparameters tuned were max depth (maximum depth of a decision tree), learning rate, number of estimators (number of trees in the forest), column sample by tree (subsample ratio of columns when constructing each decision tree), subsample (subsample ratio of the training instances), and gamma (minimum loss reduction required to make a further partition on a leaf node of the tree.).

The XG regression is implemented using the XGBoost library for Python. Further information on the model and hyperparameters can be found on their webpage[e].

e. https://xgboost.readthedocs.io/en/stable/index.html

Table 2. A summary of some of the key characteristics of the models trialled during research and development.

Model	Key characteristics	Pros	Cons
Ridge Regression	Linear regression model	Can handle multicollinearity Easier to explain predictions from model predictors	Assumes linear relationships Empty predictor values must be interpolated or replaced by zeros
Adaptive Boosting	Decision tree-based machine learning model	Can handle multicollinearity Robust to overfitting in low noise datasets Has only a few hyperparameters that need to be tuned	Empty predictor values must be interpolated or replaced by zeros Poor performance with noisy data
eXtreme Gradient Boosting	Decision tree-based machine learning model	Can handle multicollinearity Allows for complex dependencies between reported emissions and predictors Can handle empty predictor values	More difficult to understand and visualize More hyperparameters need to be tuned

Heirarchical approach

As previously highlighted, relationships between predictors and emissions can vary considerably between industries and geography. Although a company's industry, region and country are included as variables in each prediction model, a further optimisation approach was explored.

The model would start vague both in terms of industry classification and geography, create a predicter model, then iterate to a more granular industry classification and geography, and then create a predicter model again.

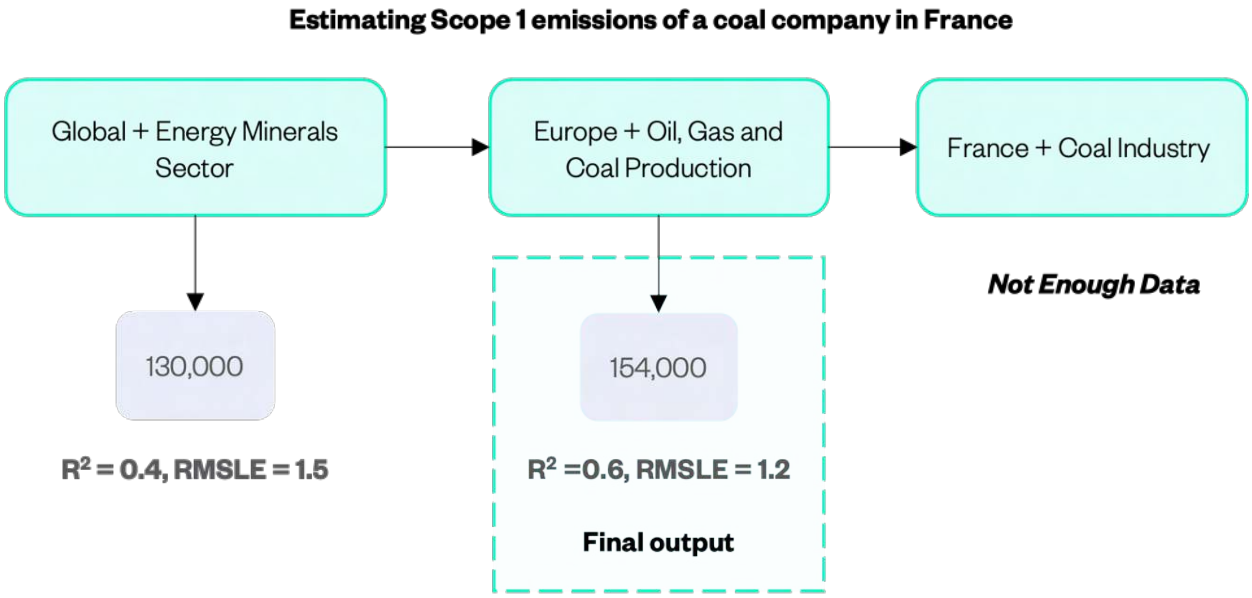
This approach is summarised in Figure 4. To estimate the emissions for a French coal production company, the model would start at energy minerals (which groups oil, gas and coal production together) at a global scale. A relationship between input data and emissions is developed and the predicted results compared to the holdout set, deducing how good that relationship is likely to be.

Next, the model model examines oil, gas and coal production in Europe, before finally looking at coal production companies in France. If there are enough datapoints, a relationship can be worked out. Based on the hold out test set, the winning relationship is deemed to be the one which has the best correlation and lowest errors.

This means that each predictor 'model' consists of over 800 sub-models which seek to develop relationships for different industry classifications, geography and emissions scope. To determine the winning relationship/feature, the predicted emissions are compared to the disclosed emissions in the hold out test set by generating two distinct metrics: the coefficient of determination (R2) and Root Mean Square Log Error (RMSLE). If R2 is higher and RMSLE is lower for a new predicter model iteration compared to the previous, then that relationship/feature is chosen before moving on to the next model iteration.

RMSLE was chosen as an error metric for an important reason: one of the biggest issues in creating an emissions estimation model would be systemically under-estimating emissions. RMSLE penalises under-estimation more than over-estimation. It also is a relative error, so that errors in predicting higher emitters are not unduly punished. R2 is taken from sklearn's r2_score[f].

Figure 4. Flow chart of the model's hierarchical approach.



New: Scope 3 category estimation

In our upgraded Emissions Estimation Model, we expand the model to also estimate the 15 categories of Scope 3 emissions. Due to the sparseness and inconsistency in Scope 3 category disclosures (only 35% of companies within our universe discloses their Scope 3 category emissions, and even less report the categories that are material to them[g]), we have chosen not to train the XGBoost model on each Scope 3 category.

Instead, we employ an approach that uses the most material Scope 3 category disclosures as a basis for estimation.

This is done by first developing a proprietary mapping, based on academic and industry research, which determines the Scope 3 categories that are material to an industry. Using this mapping, we filter for companies which are reporting what is material to them, and use this set of 'good disclosure' companies to obtain an indication of the mean relative contribution of each Scope 3 category to total Scope 3 emissions per industry. Finally, we apply the mean contribution ratio of each Scope 3 category to the estimated Scope 3 total of a company to estimate the individual Scope 3 category emissions.

As there are also gaps in the Scope 3 category reporting for companies that do disclose their Scope 3 category emissions, we also use a similar approach as above to estimate emissions to fill in these gaps, where the mean contribution ratio of each Scope 3 category is applied to the reported Scope 3 total emissions to obtain the estimated Scope 3 category emissions for unreported categories.

The reported Scope 3 total emissions is used instead of the estimated figure as this provides a better constrain on the category estimation given that there is already a company-specific Scope 3 emissions baseline.

f. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html
g. For further reading, you can read our Insight into current company Scope 3 reporting https://www.esgbook.com/scope-3/

Results

Comparison of models

Tables 3 and 4 show the results when comparing model predictions to the disclosed emissions in the 15% hold out set. Linear scaling proved to be the most accurate approach for prediction when historic disclosed emissions for a company are available.

Table 3. Company specific model results for each emissions scope. Median relative error is the median percentage error and median log error is the median absolute log_{10} error in emissions estimates. The righthand column refers to the number of companies that had predicted emissions that were within -50% and +100% of disclosed emissions.

Model	Scope	Median relative error (%)	Median log error log_{10}(tCO2e)	% within -50% and +100% of disclosed emissions
Linear Scaling	Scope 1	13	0.06	93
	Scope 2	15	0.06	94
	Scope 3	14	0.06	96

For the industry-specific estimation model, the machine learning models are comprehensively more accurate compared to the Ridge Regression. For the machine learning models, the effects of filling empty input data with zeros on the results are compared. Of the two machine learning models, the XG Boost model without zeros (meaning empty input data remains empty), consistently outperformed Ada Boost and XG Boost with Zeros, across scopes.

Table 4. Industry model results. The right-hand column refers to the number of companies that had predicted emissions that were within -50% and +100% of disclosed emissions.

Model	Scope	Median relative error (%)	Median log error log_{10}(tCO2e)	% within -50% and +100% of disclosed emissions
Ridge regression Incl. zeros	Scope 1	204	0.61	22
	Scope 2	119	0.47	25
	Scope 3	844	1.05	15
Adaptive Boost Incl. zeros	Scope 1	77	0.39	42
	Scope 2	75	0.36	43
	Scope 3	225	0.75	28
XG Boost Incl. zeros	Scope 1	71	0.38	43
	Scope 2	57	0.28	50
	Scope 3	161	0.61	26
XG Boost No zeros	Scope 1	66	0.34	45
	Scope 2	62	0.28	50
	Scope 3	99	0.61	28

Furthermore, the hierarchical approach was shown to consistently yield more accurate results than by choosing a global geographic definition and a static industry classification with country and industry only defined in predictor variables.

Full results on this can be found here.

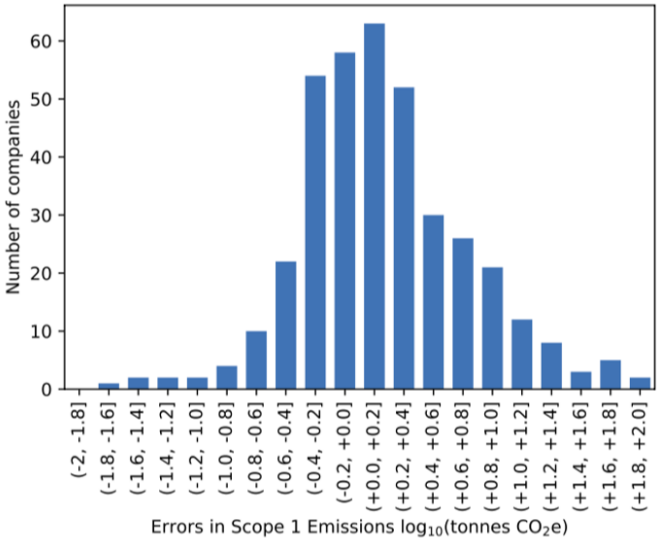
Therefore, it was clear that for ESG Book's emissions dataset a hierarchical XG Boost approach was optimal.

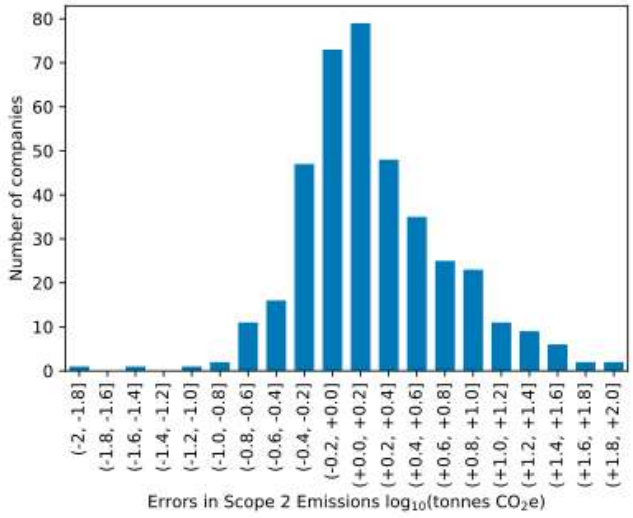
XG Boost model results

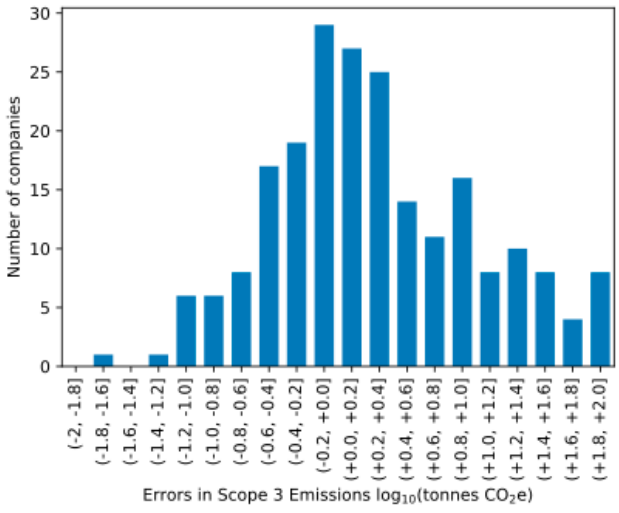
Approximately half of the companies in the hold out set were predicted within -50% and 100% of disclosed emissions (45% for Scope 1 emissions and 50% for Scope 2 emissions). This number drops to 28% for Scope 3 emissions, due to the smaller amount of disclosed Scope 3 emissions, and inconsistency in Scope 3 reporting.

In terms of model error, due to the inclusion of RMSLE as a basis for choice of prediction model, the median $\log_{10}$(predicted emissions) – $\log_{10}$(disclosed emissions) in the hold out test set is slightly positive across scopes. This means that more often than not, the model slightly overestimates emissions rather than underestimates. A positive feature since an overestimation of emissions incentivises companies to report their actual emissions. The full logarithmic error distribution per Scope is shown in Figure 5.

Figure 5. Logarithmic errors in predicted emissions across all scopes. Predicted emissions are compared to the hold out test set of disclosed emissions.



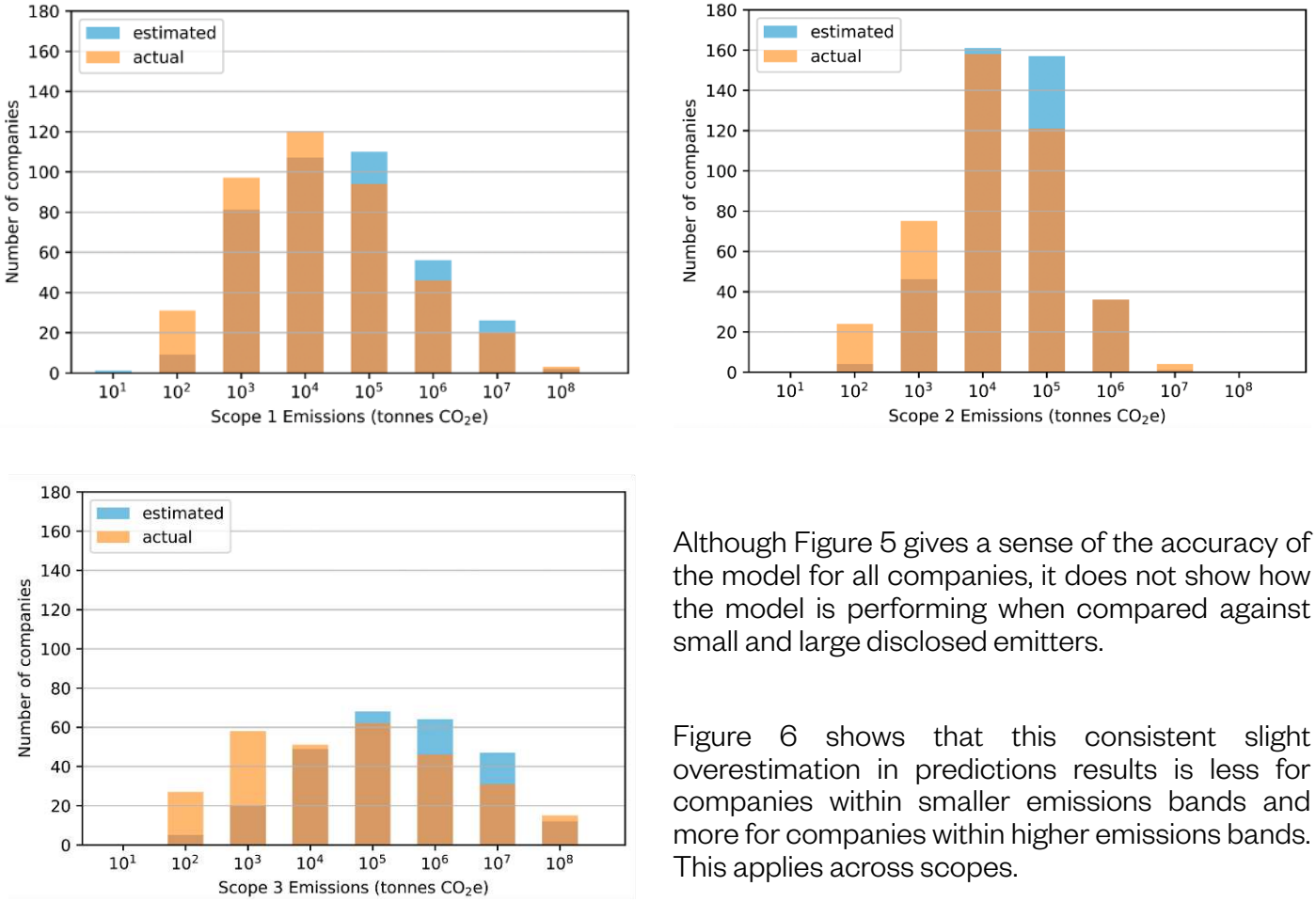




Although Figure 5 gives a sense of the accuracy of the model for all companies, it does not show how the model is performing when compared against small and large disclosed emitters.

Figure 6 shows that this consistent slight overestimation in predictions results is less for companies within smaller emissions bands and more for companies within higher emissions bands. This applies across Scopes.

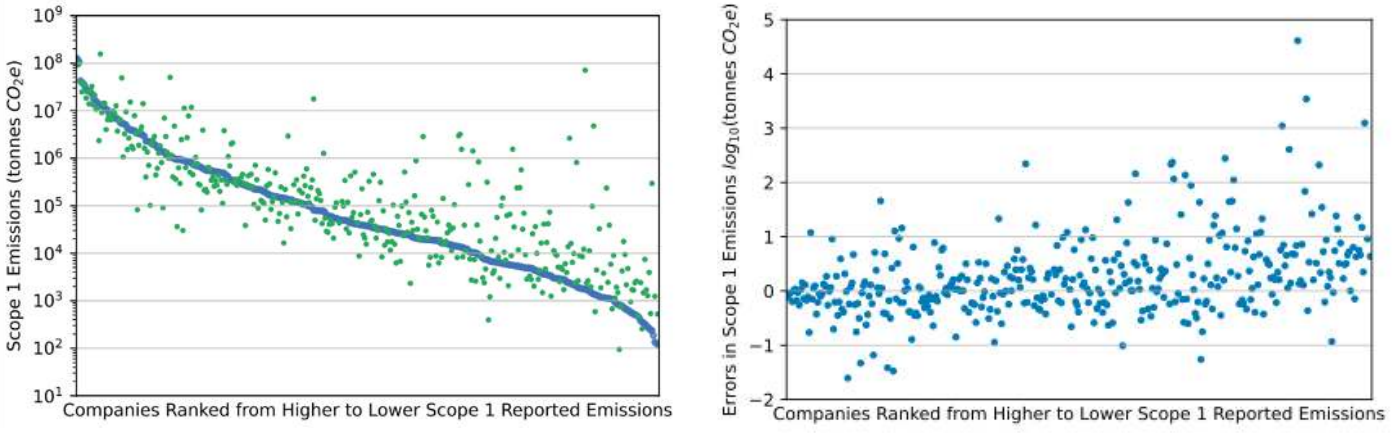
Figure 6. Distribution of estimated and disclosed emissions across emissions magnitudes in the hold out test set.



Although Figure 5 gives a sense of the accuracy of the model for all companies, it does not show how the model is performing when compared against small and large disclosed emitters.

Figure 6 shows that this consistent slight overestimation in predictions results is less for companies within smaller emissions bands and more for companies within higher emissions bands. This applies across scopes.

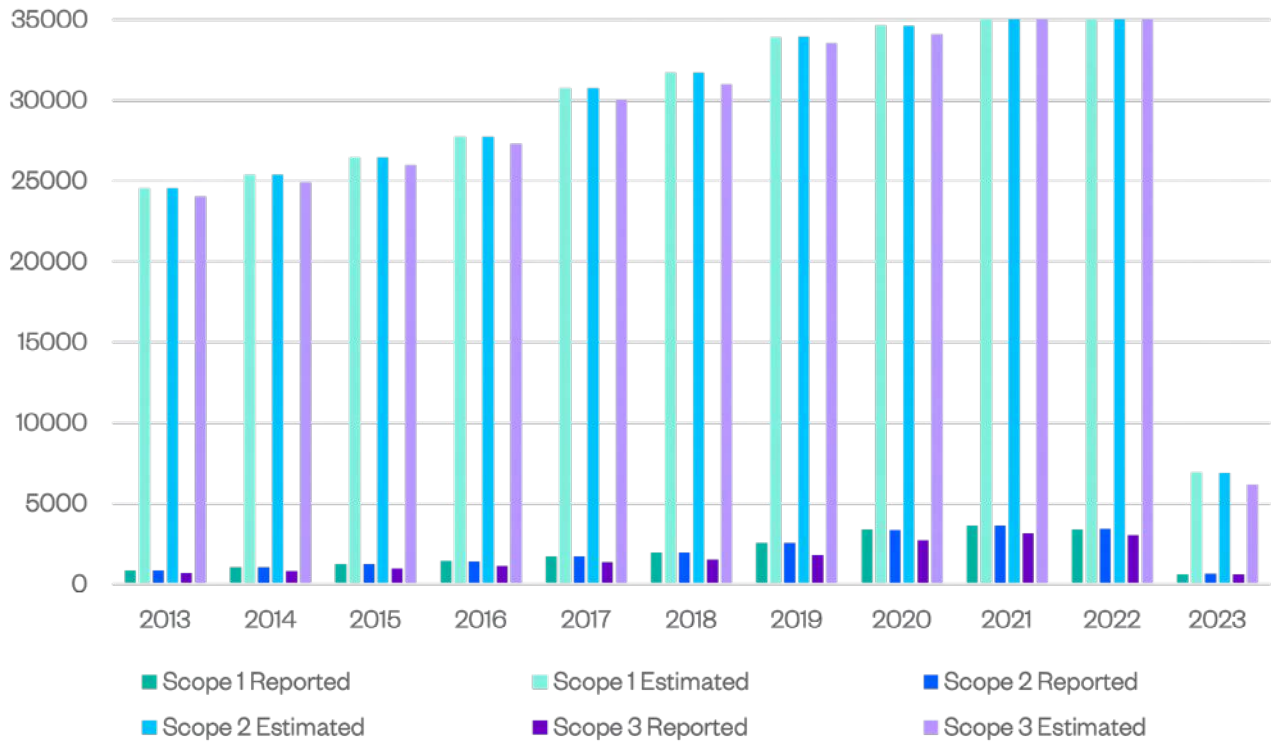
Figure 7. Scatter plot of (a) predicted versus disclosed emissions and (b) model error in predictions: $\log_{10}$(predicted emissions) – $\log_{10}$(disclosed emissions).



Coverage

The final output is the estimated emissions for approximately 45,000 unique companies globally for Scope 1, 2 and 3 emissions, giving a combined disclosed and estimated coverage of >96% across major indexes globally. It should be noted that 2023 data will increase in coverage as more reported data is collected and input data is available for estimation. Coverage for private companies is also possible where a minimum of 6 predictors for a company (as well as industry and country) are known. For more information, please contact info@esgbook.com.

Figure 8. ESG Book's Emissions Estimation Model coverage across emission scopes over 2013 to 2023



New: Historical Estimates

For the latest version of the Emissions Estimation Model, we estimate Scope 1, 2, and 3 (total and category) emissions going back to 2013. This allows users to have a view on the estimated trend in company emissions with time over the past 10 years.

Two approaches are used to generate historical estimates:

01 For companies that have reported emissions data for one or more years, the company-specific approach is used and the reported emissions are linearly scaled to fill in any missing gaps. This approach is chosen preferentially as this constrains the emission estimates to a company-specific baseline and is more accurate than a general industry-specific estimate.

02 For companies that do not have any emissions disclosed, the industry specific approach is used where the XG Boost model uses predictor variables corresponding to the year in question to generate an estimate of the company's emission for that year and scope of emissions individually.

Confidence

Five levels of confidence ratings provide an overview of how accurate an estimation is. Confidence ratings are based on how well the model was able to predict the hold out test set for a given industry and geography. The criteria for each confidence category is given in Table 5.

Table 5. Definition of each confidence category.

Confidence category	Coefficient of determination (R2)	Root Mean Square Log Error (RMSLE)
Very High	$R^2 \geq 0.5$	RMSLE ≤ 1.1
High	$R^2 \geq 0.5$	1.1 < RMSLE ≤ 1.6
	$0.5 > R^2 \geq 0$	RMSLE ≤ 1.1
Medium	$R^2 \geq 0.5$	RMSLE > 1.6
	$0.5 > R^2 \geq 0$	1.1 < RMSLE ≤ 1.6
	$R^2 < 0$	RMSLE ≤ 1.1
Low	$0.5 > R^2 \geq 0$	RMSLE > 1.6
	$R^2 < 0$	1.1 < RMSLE ≤ 1.6
Very Low	$R^2 < 0$	> 1.6

Generally, Scope 1 and Scope 2 emissions can be estimated to a high degree of confidence, where an average of 40% of Scope 1 estimations and 56% of Scope 2 estimations over 2013 to 2022 fall in the High or Very High confidence categories.

Crucially, an average of 62% of Scope 1 estimations and 73% of Scope 2 estimations can be estimated with a Medium, High or Very High confidence over the same historical timeperiod.

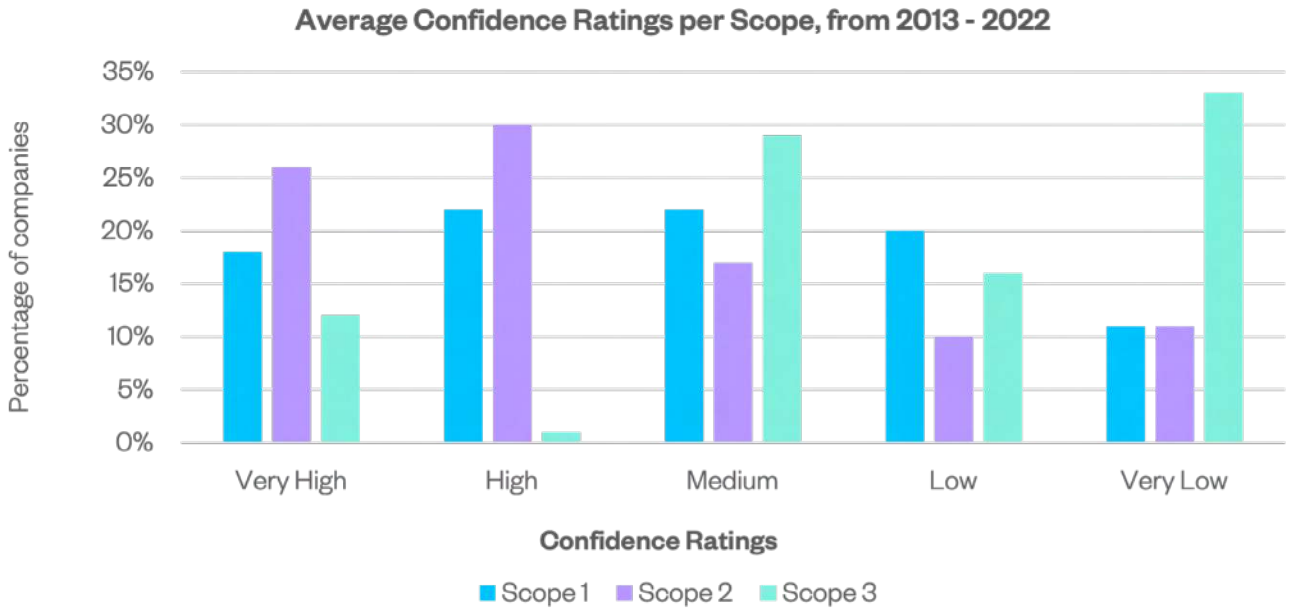
Unsurprisingly, we see varied results for the Scope 3 estimations, with an average of 13% of Scope 3 estimations in the High or Very High confidence categories. Nonetheless, an average of 42% of Scope 3 estimations can be estimated with a Medium, High or Very High confidence from 2013 to 2022.

This is due to the current reporting landscape, where Scope 3 emissions is often only disclosed for categories that are easier to measure (such as Scope 3 Category 6: Business Travel) or disclosed for a part of the company's operations (such as disclosing emissions for domestic operations only) rather than at a global scale.

Table 6. Percentage of companies in the top confidence categories by Scope.

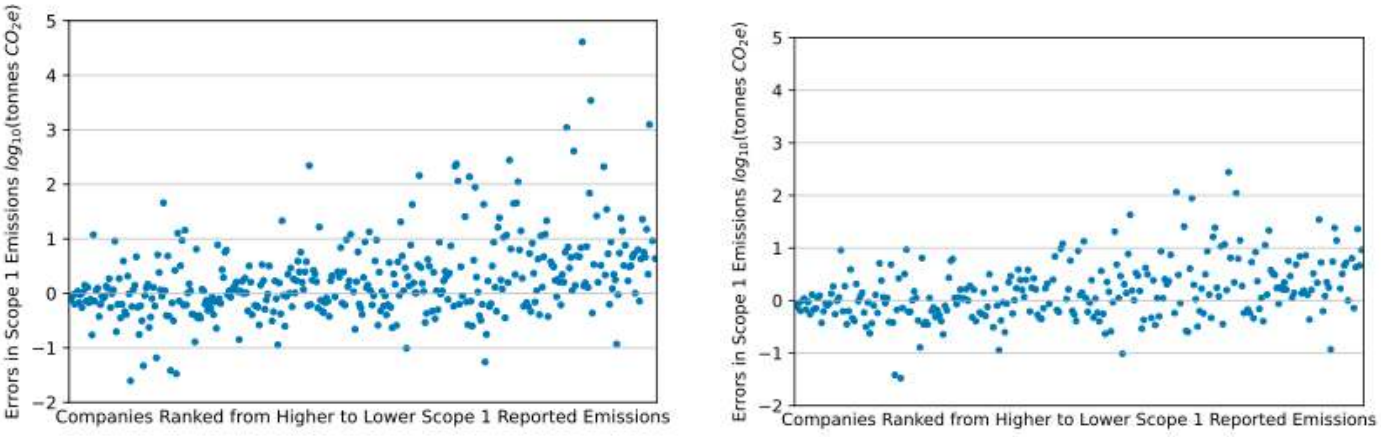
	Medium to Very High Confidence	High or Very High Confidence
Scope 1	62%	40%
Scope 2	73%	56%
Scope 3	42%	13%

Figure 9. Chart showing the percentage of companies falling into each confidence rating for Scope 1, Scope 2 and Scope 3 estimated emissions.



As shown in Figure 10, by removing 'Very Low' and 'Low' category results, more outliers are removed across all magnitudes of emissions. Median percentage error is reduced from 66% to 61%, log error is reduced from 0.34 to 0.31 $\log_{10}(tCO_2e)$, and estimates within -50% and +100% of disclosed emissions rose to 48% from 45%.

Figure 10. Scope 1 model error in prediction for full hold out test set (left), compared to when 'Very Low' and 'Low' results are removed (right).



However, the biggest improvement in results were seen for Scope 3; with median relative error decreasing from 99% to 78%, log error from 0.61 to 0.4 and 36% of results within -50% to +100% of disclosed emissions, up from 28%.

New: Predictor Type Indicator

Many frameworks ask to report on which predictor variables were used. Therefore, in our latest model, we have added an input data quality indicator, which analyses both the predictor variables and approach used to generate emission estimates to provide more transparency and context behind the estimation process. The indicator is inspired by the PCAF Data Quality Score[h] but adapted for machine learning based models. The indicator works as follows:

Emissions Type	Estimate approach	Emissions Type Indicator
Disclosed – company reported emissions	-	Reported
Estimated – company specific approach	Scaled based on revenue and number of employees	Estimated (scaled based on previous disclosure)
Estimated – industry specific approach	XGBoost model with both physical and financial input data	Estimated (physical and financial activity data)
Estimated – industry specific approach	XGBoost model with just financial input data	Estimated (financial activity data)

Importance of predictor variables

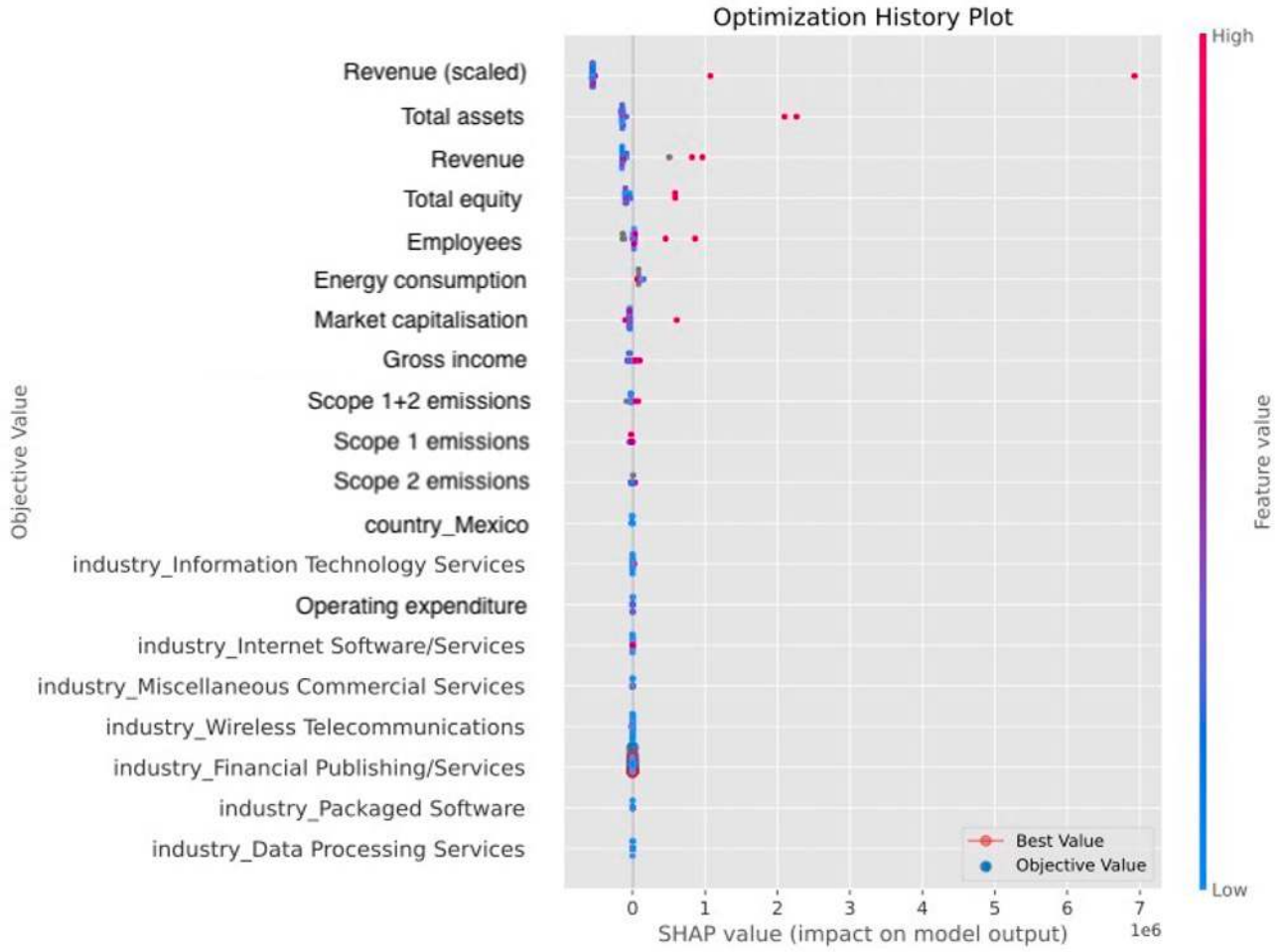
We also have compiled a dataset of which predictor variables were important to each prediction. This is done by employing SHAP (SHapley Additive exPlanations), a game theoretic approach that explains the output of machine learning models with Shapley values[i]. It is based on previous studies that investigated a unified approach to interpreting machine learning model predictions[i].

A sample SHAP output for a model is shown in Figure 10 (where objective value is the predictor variable). The most important variables are ranked on the y-axis in descending order, whilst the net effect each variable is having on model output, relative to an average, is conveyed on the x-axis.



h. Page 56 of PCAF's Global GHG Accounting & Reporting Standard https://carbonaccountingfinancials.com/files/downloads/PCAF-Global-GHG-Standard.pdf
i. https://shap.readthedocs.io/en/latest/index.html

Figure 10. Sample SHAP output for a model where objective value is thepredictor variable. The most important variables are ranked on the y-axis in descending order, whilst the net effect each variable is having on model output, relative to an average, is conveyed on the x-axis. Predictor variables beginning with 'industry' are categorical industry predictor variables.



Concluding remarks

ESG Book's Emissions Estimation Model solves a combination of critical methodological challenges in order to maximise both coverage and accuracy of the newly created dataset, such that it can be used to inform decision making by financial institutions and corporates in the context of decarbonisation objectives.

In the process of designing the model, multiple commonly used methods for emissions estimation have been tested, based on academic literature and elsewhere. It was found that a combination of company specific and industry specific models produced the most accurate results based on hold out test sets of disclosed emissions.

In summary:

- ESG Book's Emissions Estimation model estimates the Scope 1, 2 and, crucially, the 15 categories of Scope 3 GHG emissions for approximately 45,000 companies
- Estimates have 10+ years of history, from 2013 to the present year.
- Input variables have been screened so that they are reliable predictors of emissions.
- The model can process non-linear relationships between other data inputs and emissions.
- A methodology has been applied that allows for reliable estimates that can handle empty values in some of the data inputs.
- Inspired by the PCAF Data Quality Score, we have added an input data quality indicator, which analyses both the predictor variables used to generate emission estimates.
- Confidence levels are provided on the emissions estimates to allow this to be considered as part of the actionable output.
- For more information on how we can help you navigate corporate emissions with ESG Book's latest Estimated Emissions Model, please contact: info@esgbook.com.

References

1 Serafeim, G. & Vélez CaicedoGladys. Machine Learning Models for Prediction of Scope 3 Carbon Emissions. (2022).

2 Heurtebize, T., Chen, F., Soupé, F. & Carvalho, R. L. de. Corporate Carbon Footprint: A Machine Learning Predictive Model for Unreported Data. SSRN Electronic Journal (2022) doi:10.2139/ssrn.4038436.

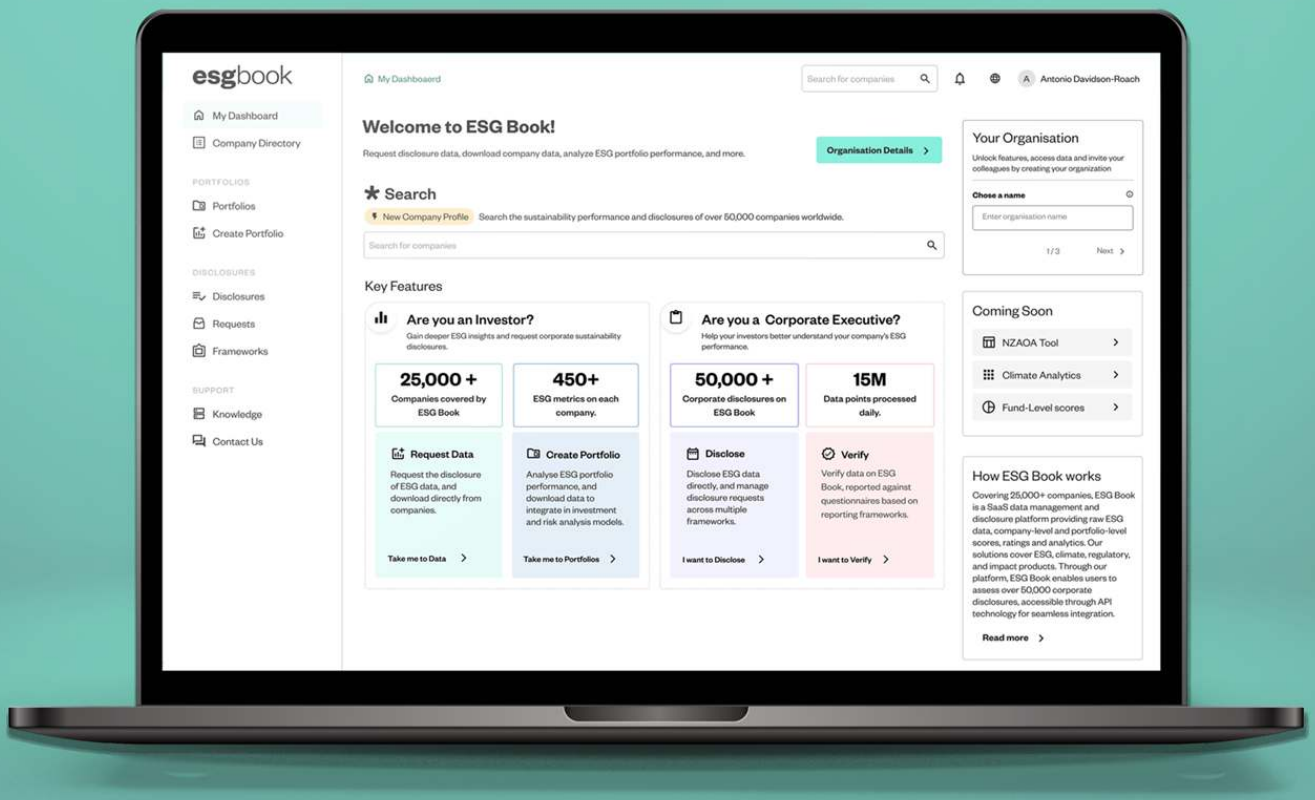
3 Hoerl, A. E. & Kennard, R. W. Ridge Regression: Applications to Nonorthogonal Problems. Technometrics 12, 69–82 (1970).

4 Freund, Y. & Schapire, R. E. Experiments with a new boosting algorithm. Machine Learning: Proceedings of the Thirteenth International Conference 148–156 (1996).

5 Oza, N. C. AveBoost2: Boosting for Noisy Data. in 31–40 (2004). doi:10.1007/978-3-540-25966-4_3.

6 Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785–794 (ACM, 2016). doi:10.1145/2939672.2939785.

Unlock the power of sustainabaility.



For more information, visit esgbook.com
or call us on +44 20 7113 3503

esgbook

ESG Book is now part of EthiFinance. This document was originally published by ESG Book prior to the integration and may reference ESG Book branding, products, or services that are being transitioned to EthiFinance. Please visit ethifinance.com for more details.