Estimating Corporate Scope 1 Emissions Using Tree-Based Machine Learning Methods

Maida Hadziosmanovic
Concordia University
Montreal QC H3G 2J1, Canada
m_hadzio@live.concordia.ca

Elham Kheradmand
University of Montreal
Montreal QC H3T 1J4, Canada
elham.kheradmand.nezhad@umontreal.ca

Nazim Benguettat
Concordia University
nazim.benguettat@concordia.ca

H. Damon Matthews
Concordia University
damon.matthews@concordia.ca

Shannon M. Lloyd
Concordia University
shannon.lloyd@concordia.ca

Abstract

Companies worldwide contribute to climate change, emitting significant amounts of greenhouse gases (GHGs). Yet, most do not report their direct or Scope 1 emissions, resulting in a large data gap in corporate emissions. This study aims to fill this gap by training several decision-tree machine learning models to predict company-level Scope 1 emissions. Our results demonstrate that the Extreme Gradient Boosting and LightGBM models perform best, where the former shows a 19% improvement in prediction error over a benchmark model. Our model is also of reduced complexity and greater computational efficiency; it does not require meta-learners and is trained on a smaller number of features, for which data is more common and accessible compared to prior works. Our features are uniquely chosen based on concepts of environmental pollution in economic theory. Predicting corporate emissions with machine learning can be used as a gap-filling approach, which would allow for better GHG accounting and tracking, thus facilitating corporate decarbonization efforts in the long term. It can also impact representations of a company's carbon performance and carbon risks, thereby helping to funnel investments towards companies with lower emissions and those making true efforts to decarbonize.

1 Introduction

In light of the climate crisis and the significant amounts of greenhouse gases (GHGs) emitted by companies, stakeholders have been pressuring on them to disclose their GHG emissions [1,2]. While companies that conduct carbon footprints typically report their emissions voluntarily [3], some countries or regions have mandatory reporting requirements. However, such schemes are often limited to certain industries and/or companies emitting beyond some threshold [4,5]. Research has shown that less than 5% of public companies actually disclose their direct (Scope 1) GHG emissions [6], and it is likely that even less disclose indirect (Scope 2 and 3) emissions, as demonstrated by disclosure patterns in prior research [7][1]. This gap in corporate emissions data presents several

[1]A company's carbon footprint is typically calculated according to three emissions categories: Scope 1 (direct emissions from sources owned or controlled by the company), Scope 2 (indirect emissions from the generation

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2021.

problems. First, it makes it difficult to reconcile company-level emissions with industry, national, and global GHG estimates, which is essential for assessing decarbonization efforts accurately [9]. Second, since many companies have begun announcing ambitious emissions targets and claiming emissions reductions [10,11], there is a greater need to track and corroborate such claims in order to hold companies accountable for their polluting activities. Finally, the lack of company-level GHG data poses challenges for evaluating carbon performance and carbon risk accurately for investment portfolio construction [12], in turn impacting where investments are funneled. Ideally, investments would be directed to companies making true efforts to decarbonize, yet in reality, this is difficult to ensure due to missing GHG accounts.

1.1 Related Works

To address this gap in corporate emissions data, GHG estimation models have cropped up as potential solutions. Data providers, such as Morgan Stanley Capital International (MSCI), CDP (formerly Carbon Disclosure Project), and Thomson Reuters have developed models to estimate emissions based on various company features [12,13]. The CDP uses statistical regression techniques, while MSCI and Thomson Reuters use simple calculations that rely on data availability of a company's energy figures or historical emissions, or simply industry-averaged data [14,15,16]. The academic research is limited to two statistical models (17,18) and three machine learning (ML) models [19, 13, 20]. Statistical approaches are constrained by the motive of making inferences about populations from specific samples [21], relying on "in-sample goodness-of-fit rather than the out-of-sample prediction accuracy," [13]. In contrast, ML models can predict data by finding patterns in complex datasets, allowing for greater out-of-sample prediction accuracy [21, 13, 19] use the Light Gradient Boosting Machine (GBM) to estimate Scope 1 and 2 emissions, training the model on over 24,000 company-year GHG observations and over 1,000 predictor variables from the Bloomberg Terminal, with no clear justification for the choice of predictors. A large set of features renders the model complex and difficult to replicate, while including unnecessary features in a model may cause overfitting or deteriorate performance. Nguyen et al. [13] use a much smaller sample (>2,000 company-year GHG observations), and rely on a specific selection of predictors to estimate Scope 1, 2, 3, and total emissions. Their model is developed as a meta-learner (Elastic Net) which aggregates predictions from six base-learners. Of the base learners, the best performing is the tree-based learners, which include Extreme Gradient Boosting (XGBoost) and Random Forest. Serafeim and Caicedo [20] compare Random Forest and Adaptive Boosting (AdaBoost) to estimate Scope 3 emissions exclusively, finding that AdaBoost performs best. Overall, research on ML techniques for estimating company-level emissions is in its early stages and there is a need for further improvement and exploration in this area.

1.2 Objective and contribution

This study addresses this need by training a series of ensemble models based on decision trees for the estimation of company-level Scope 1 emissions. We train tree-based models exclusively because they have resulted in the greatest accuracies in prior relevant studies. Our results show that XGBoost performs best, showing an improvement in mean absolute error (MAE) of 19% compared to the Scope 1 base-learner by Nguyen et al. [13]. We contribute to the literature in two key ways. First, in contrast to other existing ML models, our model is of reduced complexity and computational cost, as it is based on a smaller number of features for which data is more commonly reported and accessible. Second, we make a theoretical contribution, showing that explanations of climate change as an externality in economic theory can be used to direct the choice of predictors of corporate GHG emissions in ML estimation models.

2 Data and methodology

Our overall approach is depicted in Figure 1, and is described throughout this section.

of purchased electricity, steam, heat, or cooling); and Scope 3 (all other indirect emissions resulting from sources not owned or controlled by the company) [8]

2.1 Data collection and feature selection

We first collected target variable data (Scope 1 emissions) for all public companies globally from the Bloomberg Terminal for the fiscal years 2018-2020, removing emissions reported as zero. We then collected feature data from the Bloomberg Terminal [22], United Nations Treaty depositary [23], and Climate Change Laws of the World database [24]. To select predictor variables for the models, we derived leading questions by drawing on economic theory to identify the potential reasons for varying levels of direct GHG emissions at the company-level and why companies might act to reduce these emissions. The answers to these questions guided our choice of feature variables. We also considered common features used in prior works and the level of data availability of each feature in our sample, excluding features that had greater than 50% missing data within our initial sample. Our features included variables that relate to industry, physical assets, company size, energy consumption, profitability, liquidity, corporate climate initiatives, climate change management in the company, characteristics of the board of directors, company location, multi-nationality, and presence of carbon regulations in the company's locations. A comprehensive list of the features, their sources, and the leading questions derived from economic theory are displayed in Tables 2 and 3 in Appendix A.1.

2.2 Data pre-processing

Next, we pre-processed data by removing outliers, and transforming numerical non-ratio and percentage variables to a logarithmic scale (see Appendix for further details). We then estimated missing values of numerical predictor variables using the k-nearest neighbors algorithm. We chose the optimal value of k based on the performance of a Random Forest model for the prediction of Scope 1 emissions (see Appendix A.3). Our final data set after preprocessing included $13,041$ company-year observations.

2.3 Tree-based models

In this study, we focus on decision tree models rather than other model types (e.g., neural networks, k-nearest neighbors) for several reasons. Previous studies that experimented with ML models to estimate company emissions have consistently shown that tree-based models performed best on a range of predictors, especially financial metrics [19,13,20]. In addition, in terms of functionality, tree-based models are known to handle categorical data, they are well-suited for non-linear relationships, and they are largely immune to multicollinearity between predictive feature data [25,26].

We ran the following tree-based algorithms in Python: CatBoost, XGBoost, LightGBM, AdaBoost, and Random Forest. We ran two iterations of the CatBoost algorithm; for the first, we used the original encoding solution provided for categorical data in CatBoost on the following variables: industry, subregion of domicile, and subregion of risk[2]. For the second, we applied one-hot encoding on these categorical variables. Since the other tree-based models do not have a predetermined encoding methodology for handling categorical data, we used one-hot encoding on these variables. Our training set was on 80% of the dataset, and the hold-out was 20% of the dataset. We used 10-fold cross validation to fine-tune the hyperparameters for each model.

3 Results and Discussion

Our main results in Table 1 present different performance evaluation metrics for each model, including root mean squared error (RMSE), mean squared error (MSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and adjusted R^2 values. We compare these results to those of the XGBoost Scope 1 emissions base-learner developed by Nguyen et al. [13] (henceforth, the benchmark model), since this is the only other academic study to our knowledge that has trained a model to estimate corporate Scope 1 emissions.

Overall, the best performing model is XGBoost, followed by LightGBM and Random Forest. The RMSE values are 1.30, 1.31, and 1.32 respectively, and are lower than that of the benchmark model, which is 1.40. Our adjusted R^2 (A-R2) for the XGBoost, LightGBM, and Random Forest models are also higher than the benchmark model, indicating that our set of predictor variables may better

[2]See the section "Transforming categorical features to numerical features" in the online CatBoost supporting documents for further information (https://catboost.ai/en/docs/concepts/algorithm-main-stages_cat-to-numberic)

Table 1: Prediction performances of the trained models

Model	**RMSE**	**MSE**	**MAE**	**MAPE**	**A-R2**	**MAE** (Nguyen et al. 2021)	**MAE-I** (%)
CatBoost-1	1.43	2.03	0.96	0.32	0.81	-	6.80
CatBoost-2	1.41	1.99	0.96	0.29	0.82	-	6.80
XGBoost	1.30	1.69	0.83	0.29	0.84	1.03	19.42
AdaBoost	1.94	3.77	1.38	0.36	0.66	-	−33.98
LightGBM	1.31	1.73	0.86	0.30	0.84	-	16.50
Random Forest	1.32	1.74	0.87	0.30	0.84	1.03	15.53

explain the variation in the output (target) variable. We also see an overall improvement in MAE (MAE-I) by 19.42% in XGBoost and 16.50% in LightGBM, compared to the benchmark model.

Figure 3 shows the top 20 features by overall feature importance in the XGB model based on mean absolute SHapley Additive exPlanations (SHAP) values. SHAP can be used as a way to interpret ML models, explaining the impact or contribution of a feature on the model's prediction [27]. The most important feature in the XGBoost model is energy consumption (SHAP value of 1.58), followed by gross property, plant and equipment (GPPE) (SHAP value of 0.41). Prior studies have also demonstrated the importance of energy and assets in predicting emissions [13, 20].

To understand the effect of features on the entire dataset, we present a bee swarm plot in Figure 4, which shows the impact of the top 20 features by importance on the prediction when a SHAP value is negative (meaning a negative contribution to the prediction), zero (meaning no contribution to the prediction), and positive (meaning a positive contribution to the prediction). For example, higher values of energy consumption (logEnergyConsumption) will have a positive (increasing) contribution to the prediction of Scope 1 emissions, and lower values will have a negative contribution. For companies in the Financials, Technology and Real Estate industries, (binary variables where 1 indicates a company is in the industry and 0 indicates it is not), the prediction is negatively impacted. In contrast, the Utilities and Energy industry features have a positive contribution to the prediction. The overall importance of a company's industry for estimating carbon footprints has been established in prior studies as well [13,20]. Other features show less clarity with respect to the direction of the contribution, such as the age of assets (logAssetage), since both high and low values show both positive and negative contributions. Future directions of this research may look to use SHAP results to further streamline the choice of features in the model.

Altogether, our study establishes the usefulness of tree-based models for estimating corporate emissions. Our results point to a significant improvement in the accuracy of our XGBoost model compared to the benchmark model for predicting Scope 1 emissions. As our model does not employ meta-learners and uses less, more commonly available features, we have shown that Scope 1 corporate emissions can be estimated with models of lower complexity and greater computational efficiency. Moreover, we have shown that model results could be improved with a feature selection methodology that is founded on economic theory. In the absence of widely available corporate Scope 1 emissions data, this model can be used as a gap-filling approach. This is important in the context of climate change because it would allow for better GHG accounting and tracking, and, following the adage of "what gets measured, gets managed", would facilitate corporate decarbonization efforts. It could also contribute to more accurate representations of a company's carbon performance and risks, thereby supporting investments directed at companies making mitigation efforts and lowering their GHG emissions.

4 References

[1] Chithambo, L., Tingbani, I., Agyapong, G. A., Gyapong, E., & Damoah, I. S. (2020). Corporate voluntary greenhouse gas reporting: Stakeholder pressure and the mediating role of the chief executive officer. Business Strategy and the Environment, 29 (4), 1666-1683.

[2] Liesen, A., Hoepner, A. G., Patten, D. M., & Figge, F. (2015). Does stakeholder pressure influence corporate GHG emissions reporting? Empirical evidence from Europe. Accounting, Auditing & Accountability Journal, 28(7), 1047-1074. https://doi.org/10.1108/AAAJ-12-2013-1547

[3] Depoers, F., Jeanjean, T., & Jérôme, T. (2016). Voluntary Disclosure of Greenhouse Gas Emissions: Contrasting the Carbon Disclosure Project and Corporate Reports. J Bus Ethics 134, 445–461. doi:10.1007/s10551-014-2432-0

[4] Organisation for Economic Co-operation and Development and Climate Disclosure Standards Board (OECD and CDSB). (2015). Climate change disclosure in G20 countries: Stocktaking of corporate reporting schemes. https://www.oecd.org/environment/cc/g20-climate/collapsecontents/Climate-Disclosure-Standards-Board-climate-disclosure.pdf

[5] European Commission. (2018). Commission Implementing Regulation (EU) 2018/2066 of 19 December 2018 on the monitoring and reporting of greenhouse gas emissions pursuant to Directive 2003/87/EC of the European Parliament and of the Council and amending Commission Regulation (EU) No 601/2012. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02018R2066-20210101

[6] Hadziosmanovic, M., Lloyd, S. M., Bjørn, A., Paquin, R. L., Mengis, N., & Matthews, H. D. (2022). Using cumulative carbon budgets and corporate carbon disclosure to inform ambitious corporate emissions targets and long-term mitigation pathways. Journal of Industrial Ecology. 1-13. https://doi.org/10.1111/jiec.13322

[7] Ryan, J., & Tiller, D. (2022). A Recent Survey of GHG Emissions Reporting and Assurance. Australian Accounting Review, 101(32), 181–187.

[8] World Resources Institute and World Business Council for Sustainable Development (WRI & WBCSD). (2004). A Corporate Accounting and Reporting Standard. https://ghgprotocol.org/sites/default/files/standards/ghg-protocol-revised.pdf

[9] Luers, A., Yona, L., Field, C. B., Jackson, R. B., Mach, K. J., Cashore, B. W., ... & Joppa, L. (2022). Make greenhouse-gas accounting reliable—build interoperable systems. Nature.

[10] NewClimate Institute. (2022). Corporate Climate Responsibility Monitor. https://newclimate.org/sites/default/files/2022-06/CorporateClimateResponsibilityMonitor2022.pdf

[11] Science Based Targets (SBT). (2020). Companies taking action. https://sciencebasedtargets.org/companies-taking-action

[12] Gurvich, A., & Creamer, G. G. (2021). Overallocation and Correction of Carbon Emissions in the Evaluation of Carbon Footprint. Sustainability, 13(24), 13613.

[13] Nguyen, Q., Diaz-Rainey, I., and Kuruppuarachchi, D. Predicting corporate carbon footprints for climate finance risk analyses: A machine learning approach. Energy Economics, 95:105129, 2021. ISSN 0140-9883. doi: https://doi:org/10:1016/j:eneco:2021:105129

[14] CDP. (2020). CDP Full GHG Emissions Dataset Technical Annex III: Statistical Framework. https://cdn.cdp.net/cdp-production/comfy/cms/files/files/000/003/028/original/2020_01_06_GHG_Dataset_Statistical_Framework.pdf

[15] MSCI. (2016). Filling the Blanks: Comparing Carbon Estimates Against Disclosures. https://www.msci.com/documents/10199/139b2ab7-c95f-4f09-9d33-fdc491c5316e

[16] Refinitiv. Refinitiv ESG, Carbon Data and Estimate Models. https://www.refinitiv.com/content/dam/marketing/en_us/documents/fact-sheets/esg-carbon-data-estimate-models-fact-sheet.pdf

[17] Goldhammer, B., Busse, C., & Busch, T. (2017). Estimating corporate carbon footprints with externally available data. Journal of Industrial Ecology, 21(5), 1165-1179.

[18] Griffin, P. A., Lont, D. H., & Sun, E. Y. (2017). The relevance to investors of greenhouse gas emission disclosures. Contemporary Accounting Research, 34(2), 1265-1297.

[19] Han, Y., Gopal, A., Ouyang, L., & Key, A. (2021). Estimation of Corporate Greenhouse Gas Emissions via Machine Learning. arXiv preprint arXiv:2109.04318.

[20] Serafeim, G., & Velez Caicedo, G. (2022). Machine Learning Models for Prediction of Scope 3 Carbon Emissions. Available at SSRN.

[21] Bzdok, D., Altman, N. & Krzywinski, M. (2018). Statistics versus machine learning. Nat Methods 15, 233–234. https://doi.org/10.1038/nmeth.4642

[22] Bloomberg, L. P. (2022). Bloomberg database. Bloomberg Terminal.

[23] United Nations Treaty Collection. (2016). 7. d Paris Agreement (Chapter XXVII). United Nations. https://treaties.un.org/Pages/ViewDetails.aspx?src=TREATY&mtdsg_no=XXVII-7-d&chapter=27&clang=_en

[24] Grantham Research Institute on Climate Change and the Environment and Sabin Center for Climate Change Law (GRICC & SCCCL). (2022). Climate Change Laws of the World database. Available at: https://climate-laws.org

[25] Malehi, A. S., & Jahangiri, M. (2019). Classic and bayesian tree-based methods. In P. Vizureanu. Enhanced Expert Systems (pp. 27-52). IntechOpen. doi: 10.5772/intecopen.79092

[26] Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 1310-1315). Ieee.

[27] Marcílio, W. E., & Eler, D. M. (2020). From explanations to feature selection: assessing shap values as feature selection mechanism. 2020 33rd SIBGRAPI conference on Graphics, Patterns and Images (SIBGRAPI) (pp. 340-347). IEEE. https://doi.org/10.1109/SIBGRAPI51738.2020.00053

[28] Ayres, R. U., & Kneese, A. V. (1969). Production, consumption, and externalities. The American Economic Review, 59(3), 282-297.

[29] United Nations Statistics Division. (2022). Country and Area Codes (M49). https://unstats.un.org/unsd/methodology/m49/overview/

A Appendix

A.1 Feature Selection: An economic perspective on emissions predictors

We used economic theory to guide our choice of features for the models. Economic theory recognizes environmental issues, including GHG pollution, as an externality. Externalities can be understood as market failures resulting from the inefficiency of the production and consumption of goods and services [28]. In the corporate sector, the quantities and varying production mechanisms of different products and services impact these GHG emissions by-products, as well as a company's capacity to manage these by-products. This is represented by the predictors associated with questions 1, 2, and 3 (Table 2). Economic theory resolves the issue of externalities by finding ways to internalize the costs of pollution. Pressures and incentives to internalize these costs vary, and can be internal (e.g., the board) or external (e.g., government imposed emissions allowances) to the company. The predictors associated with questions 4 and 5 (Table 3) represent such internal and external pressures.

A.2 Feature Details

Highest level at which climate change is managed This predictor includes 5 categories:

- Board/subset of Board/committee appointed by Board
- Subset of Board/Committee appointed by Board
- Manager/Officer
- No individual or committee
- Unknown

Table 2: Feature selection supported by economic theory

Question number	Leading questions (economic perspective)	Explanatory feature	Predictor variable	Source of data
1	The production of which products or services impact emissions (externalities)?	Industry	International Classification, Benchmark (ICB) Industry	Bloomberg Terminal
		Physical assets	Gross Property, Plant, and Equipment; Asset age; Capital expenditure (CapEx)	Bloomberg Terminal
2	How does the quantity of products or services produced impact emissions (externalities)?	Company size	Revenue	Bloomberg Terminal
		Company size	Number of employees	Bloomberg Terminal
		Energy consumption	Energy consumption (MWh)	Bloomberg Terminal
3	What is a company's capacity to internalize (mitigate) emissions?	Profitability and liquidity	Return on assets (ROA); Return on equity (ROE); EBITDA margin; Free cash flow (FCF); Cash flow per share (CFPS); Cash flow from operations (CFO)	Bloomberg Terminal

Some fields did not stipulate one of these categories explicitly or exactly, but included more detailed information. To simplify such information in these fields, we assessed each answer that was given in this field manually, categorizing them in one of the five categories listed above. For example, the answer:

> "*Whilst climate change is not considered at Board level, the Legal Director is responsible for monitoring emerging regulations including climate change and the Finance Director is responsible for monitoring energy and business travel costs. Both are IG Group Board members.*"

was assessed as belonging to the category "Subset of Board/Committee appointed by Board".

Country of domicile, country of risk, and associated subregions Bloomberg defines Country of Domicile as the location of management. Country of Risk is determined differently for every company, but may largely depend on where the company generates its highest amount of revenue, or its primary currency [22]. Country of domicile and country of risk were provided as ISO codes. We thus converted all ISO codes to country names. Where no country was indicated for either country domicile or country of risk, the entire row was removed. The subregion of the country was then assigned according to the United Nations subregion categories [29]. Subregions were used instead of countries to reduce the cardinality of this categorical variable.

Average asset age in years This predictor was calculated as accumulated depreciation/depreciation expenses. Data on accumulated depreciation and depreciation expenses were taken from the Bloomberg Terminal. Where one of these variables was not available, average asset age was not calculated.

Table 3: Feature selection supported by economic theory (Continued)

Question number	**Leading questions (economic perspective)**	**Explanatory feature**	**Predictor variable**	**Source of data**
4	What are the internal pressures to internalize (mitigate) emissions?	Voluntary climate initiatives	Whether an emissions target has been set; Whether an internal price of carbon is set	Bloomberg Terminal
		Climate change management in the company	Climate change policy; Highest level at which climate change is managed	Bloomberg Terminal
		Influence or characteristics of the board of directors	Percent of women on the board	Bloomberg Terminal
5	What are the external pressures to internalize (mitigate) emissions?	Company location	Global subregion according to country of domicile; Global subregion according country of risk; Paris signatory according to country of domicile; Paris signatory according to country of risk	Bloomberg Terminal; United Nations Treaty Collection Depositary
		Multi-nationality	Percent of revenue from foreign sources	Bloomberg Terminal
		Presence of carbon regulations	Presence of carbon tax in country of domicile; Presence of carbon tax in country of risk; Presence of emissions trading scheme in country of domicile; Presence of emissions trading scheme in country of risk	Climate Change Laws of the World Database

Percent of revenue from foreign sources In cases where the percent was reported as >100, this was adjusted to 100%.

Whether an emissions target has been set This field was denoted as 'yes' or 'no' based on whether there was an emissions target reported in any of the fiscal years. Two fields in the Bloomberg terminal indicate whether an emissions target is set: one is based on CDP reporting, and the other is derived by Bloomberg.

Internal price of carbon This field denotes 'yes', or 'no', or in cases where nothing is reported in Bloomberg, we denoted the field as 'unknown'. This data was converted to an ordinal scale, ordered as: no (1), unknown (2), yes (3).

A.3 Data Pre-processing

Outliers Following initial data collection, we removed the 1st and 99th percentiles of the target variable data (Scope 1 emissions). For outliers in the feature data, we removed outliers manually using boxplot visualizations. We identified and removed outliers that appeared several orders of magnitude outside of the 25-75 percentile box and appeared isolated. However, if many data points appeared outside of the 25-75 percentile range, but were not isolated, we did not remove these. We endeavored to retain as much real data as possible.

Logarithmic transformations and scaling Following the approach of Serafeim and Caicedo (2022) and Nguyen et al. (2021), we applied a natural logarithmic transformation such that $z' = \log(z)$ on the target variable (Scope 1 emissions) and certain numerical predictor variables (GPPE, energy consumption, number of employees, asset age). We scaled these predictive variables up 1 so that the zeros in the dataset are handled prior to the logarithmic transformation. For free cash flow (FCF), cash from operations (CFO), and capital expenditure (CAPEX), we applied a logarithmic transformation such that $z' = \log(z + |\min(z)|)$ to handle the negative values in these variables. This scales the dataset upwards by the absolute of the minimum value. Also following the approach of Serafeim and Caicedo (2022), we did not apply logarithmic transformations on variables representing ratios or percentages. These include cash flow per share (CFPS), return on assets (ROA), return on equity (ROE), percent revenue from foreign sources, percent women on the board, and EBITDA margin.

Missing data We then imputed missing values for all numerical predictor variables (EBITDA margin, ROA, ROE, logRevenue, logGPPE, logCAPEX, logEnergyConsumption, logEmployees, logFCF, logCFO, CFPS, logAssetAge, Percent of women on the board, Percent of revenue from foreign sources) by using the k-nearest neighbors algorithm. We excluded the target variable in the imputation methodology, Scope 1 emissions, as we assume that this data would not be typically available for all companies. Figure 2 displays the root mean squared errors (RMSE) of this model with respect to different values of k (in the range of 1 to 39). We chose $k = 26$ to impute the missing data, as this resulted in the lowest error (equal to 1.35).

A.4 Figures

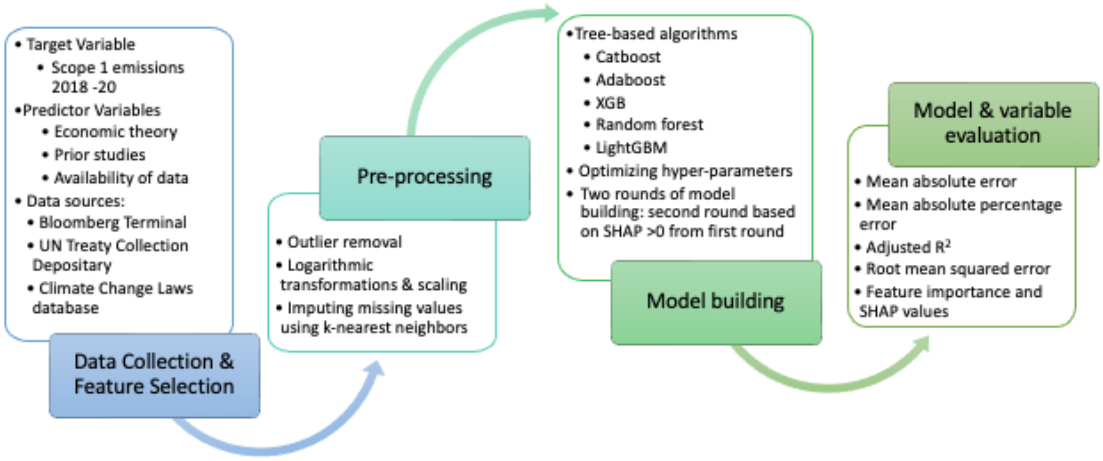

Figure 1: Overview of the methodological approach

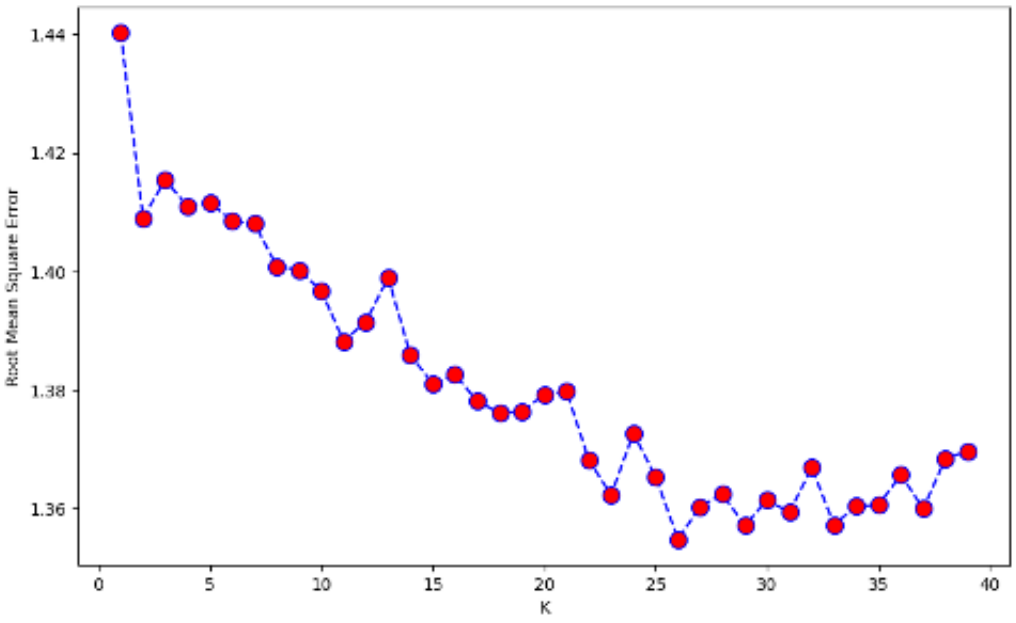

Figure 2: RMSE for different k values

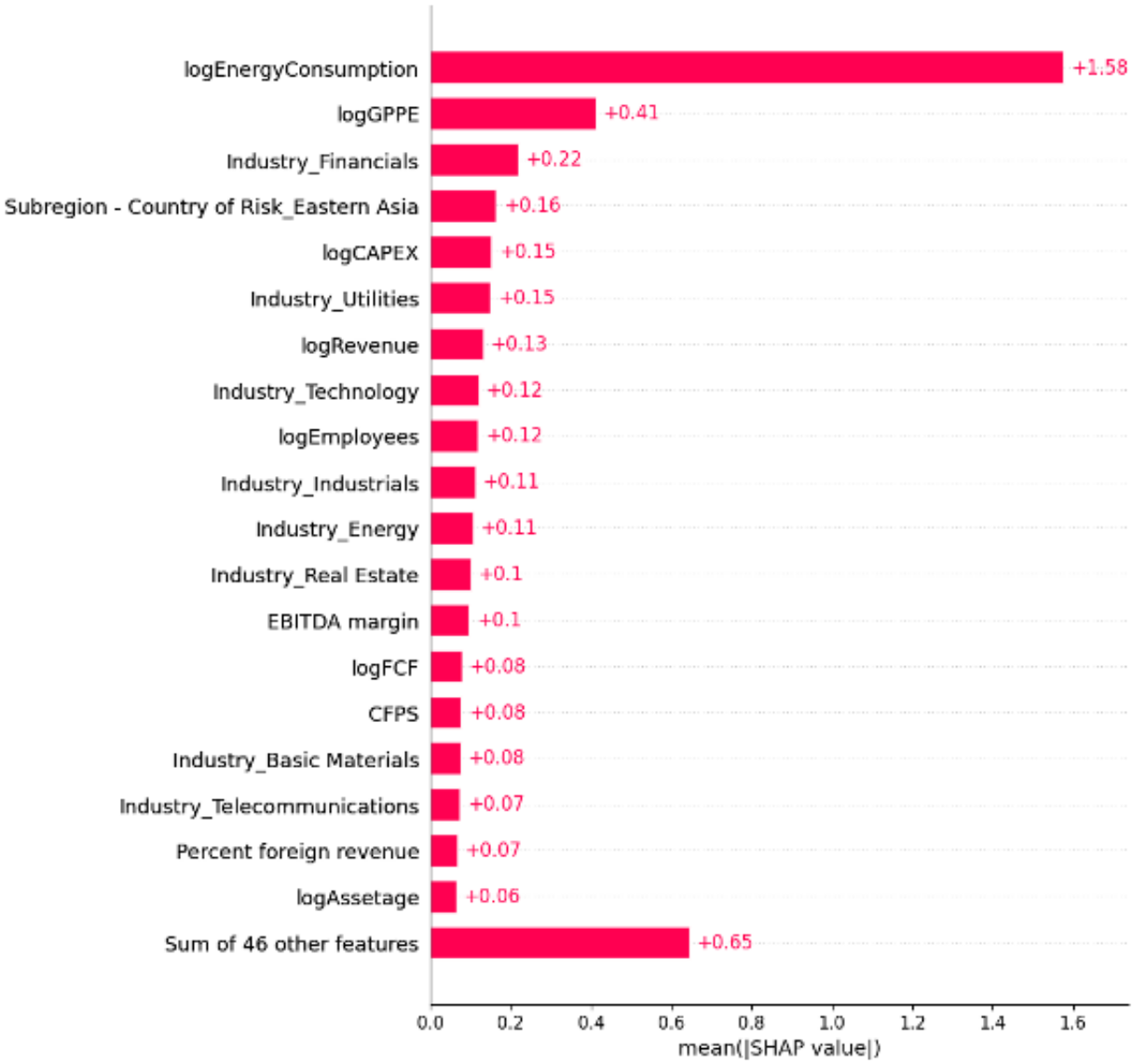

Figure 3: Top 20 important features in prediction of Scope 1 corporate emissions from XGBoost

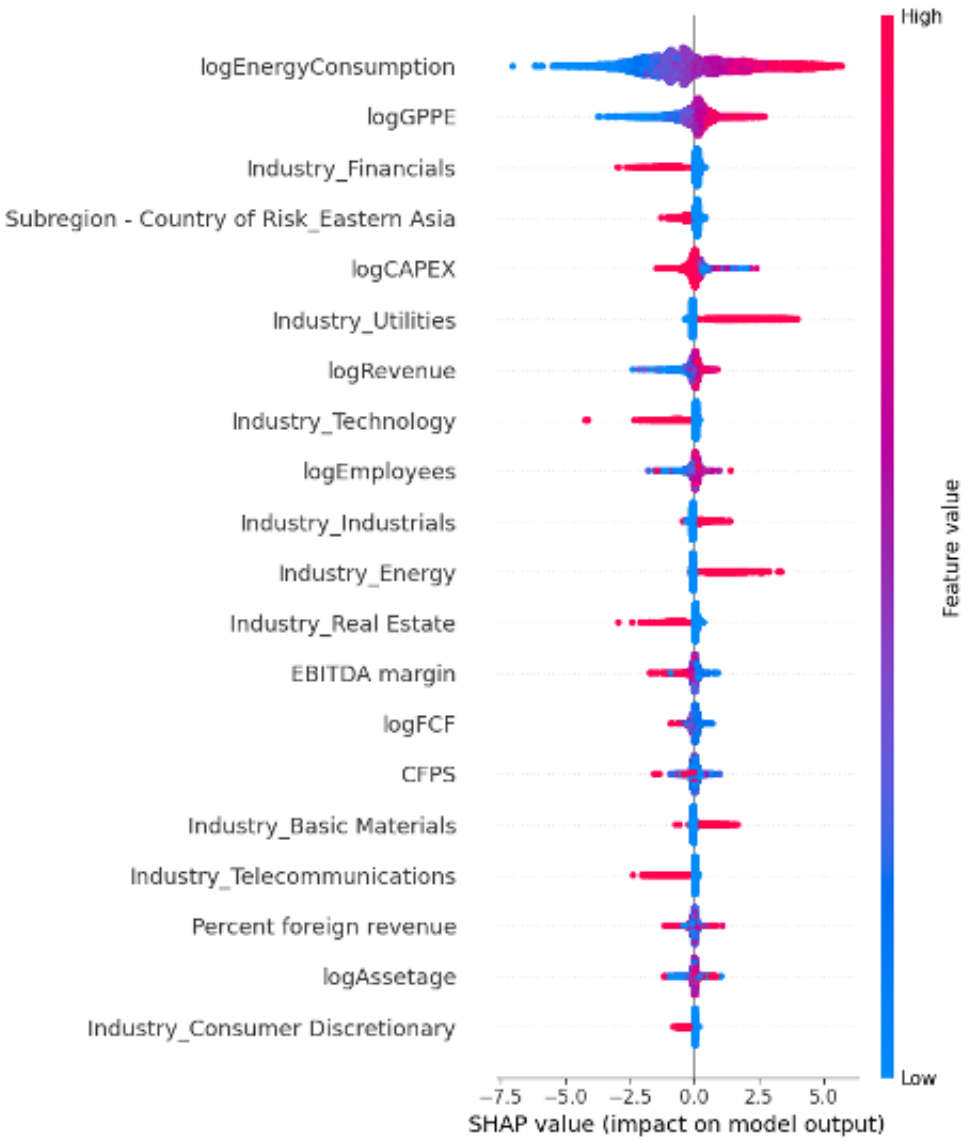

Figure 4: The bee swarm plot for the top 20 important features in prediction of Scope 1 corporate emissions from XGBoost