IFC-ECCBSO-Bank of Spain Workshop on "New insights from financial statements"

17 October 2024

It's in the financials, stupid! But is it certain?[1]

Christian Haas, Ulf Moslener and Sebastian Rink,
Frankfurt School of Finance and Management

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the European Committee of Central Balance Sheet Data Offices (ECCBSO), the Bank of Spain, the BIS, the IFC or the other central banks and institutions represented at the event.

It's in the Financials, Stupid! But is it Certain?

Christian Haas[1], Ulf Moslener[1] and Sebastian Rink[1]

[1]Frankfurt School of Finance & Management

This version: September 2024

Abstract

Sustainability data is increasingly relevant for multinational enterprises (MNEs), financial institutions, and researchers. However, sustainability data remain incomplete, fragmented, or scarce. In our paper, we propose a novel approach to address this challenge using machine learning (ML) to predict sustainability metrics from readily available financial data. This method allows for a more detailed and accurate assessment of sustainability in MNEs and their global value chains. Our approach is tested using a comprehensive dataset of financial and sustainability information at the company level. The results indicate that ML is helpful in predicting key sustainability metrics, such as corporate carbon emissions and water discharge. However, users should reflect on the specific use case when applying ML since model performance can vary sectorally, spatially, and temporally. In addition, we develop a metric to assess the uncertainty of the predictions and find that it can substantially affect the model output. Regulators should build on our findings to encourage the use of ML-generated sustainability data while also requiring more transparency from data providers and model users.

Keywords: corporate carbon emissions, corporate sustainability information, financial data, machine learning

JEL Codes: C53, Q54, Q56

We would like to thank Christina E. Bannier and Maurice Dumrose for thorough feedback as well as participants in the Sustainable Finance Research Platform brown bag lunch for helpful comments. The authors acknowledge funding from the project safe Financial Big Data Cluster (safeFBDC) by the Federal Ministry for Economic Affairs and Climate Action.

I. Introduction

Sustainability commitments such as the Paris Climate Agreement (United Nations, 2015) and the Kunming Biodiversity Declaration (Kunming Declaration, 2021) show the ambition to transform economies worldwide. As a result, multinational enterprises (MNEs), financial institutions, and researchers increasingly require sustainability information. Availability and quality of sustainability data throughout the dimensions environment (E), social (S), and governance (G) are more important than ever. However, these data are incomplete, fragmented, or scarce. In this paper, we present a novel approach to using machine learning to fill these sustainability gaps.

MNEs have different reasons for improving the sustainability of their products and services such as innovation capacity (Acemoglu, 1997; Balasubramanian and Lee, 2008), regulatory and investor pressure (Slager et al., 2023), or financial performance (Kim and Starks, 2016; Adams and Ferreira, 2009). A key component in the evaluation and management of sustainability in MNEs is the availability and quality of sustainability data for MNEs, their subsidiaries, and their global value chains (GVCs) (Marano et al., 2024). As reported and audited data are only gradually and incompletely becoming available, MNEs are forced to rely on other data sources.

Financial institutions and regulators have received increasing attention to sustainability in recent years. In particular, internationally operating banks and investors have committed to steer their portfolios in line with sustainability goals (Bolton et al., 2022) while retail investors display some preference for green financial products (Bauer et al., 2021). An increasing awareness of the financial risk associated with sustainability has led many regulators to strengthen sustainability-related risk management regulation (ECB, 2022). Today, financial markets reflect biodiversity and climate-related risks to some extent (Ilhan et al., 2021, 2023; Giglio et al., 2023; Garel et al., 2023). These developments require increasing amounts of sustainability data from investees and lendees.

Research on sustainability and business practices has grown in recent years (e.g. Marano et al. (2024) and Starks (2023)). Many empirical studies require company-level sustainability data. Sufficient data quality and data availability, as well as understanding modeling assumptions for non-reported data, is required in the process.

Currently, sustainability at the company level is typically measured by ESG ratings. However, these ratings diverge in terms of measuring specific ESG aspects (Berg et al., 2022). Given the difficulty of measuring company sustainability (Edmans, 2023), granular physical indicators would allow different actors to assess corporate sustainability more inde-

pendently, allowing greater diversity of views and, as a result, potentially more efficient functioning of markets, for example, through capital allocation and GVC engagement. These granular physical indicators include corporate carbon emissions, the corporate biodiversity footprint, corporate resource usage (water, primary materials, etc.), and diversity measures such as gender ratios.

However, granular physical sustainability data remain incomplete, fragmented, or scarce. Corporate ESG disclosure around the world is currently evolving but far from being established (Krueger et al., 2024). It is unlikely that full sustainability information on GVCs will be available soon. Corporate carbon reporting seems to be the most evolved, but it remains incomplete (Busch et al., 2022). This can lead to frictions in the efficient use of these data for decision making. Therefore, methods are needed to fill the data gaps.

Data providers and researchers have realized this need and have provided different modeling approaches to estimate ESG data. To date, this work has typically been limited to corporate carbon emissions. Methods include simple estimations, which use a direct proportional relationship between a company's size (e.g., revenue, number of employees) and the corporate carbon footprint, and regression setups, where emissions are regressed against a variety of operational and financial predictors to encapsulate a company's business model, scale, and technological practices (e.g., Goldhammer et al. (2017) and Griffin et al. (2017)). However, these approaches suffer from simplicity and thus potential biases for out-of-sample predictions, as well as a lack of a generally accepted theory on how accounting data drive corporate emissions.

Machine learning can help overcome these shortcomings, even beyond corporate carbon footprints. Companies are complex systems (Loughran and McDonald, 2023). They have different ages, sizes, product lines, geographic presences, financing structures, and company cultures. These aspects may all play a role for companies' sustainability metrics, and it is very likely that relationships between and within company-level metrics are nonlinear. Machine learning with its ability to search large non-parametric algorithmic spaces (Jordan and Mitchell, 2015) seems well positioned for this environment.

Previous work employing machine learning to predict sustainability data already shows promising results, but is limited in metrical and methodological scope. Nguyen et al. (2021) and Nguyen et al. (2022) estimate corporate Scope 1-3 emissions using a set of regression-based supervised learning algorithms and achieve up to 30% improvement compared to parametric approaches. Other estimates of environmental data at the company level remain rare. Tian (2023) estimates the water efficiency of the companies. The literature is largely silent on the estimation of social aspects at the company level. Governance

aspects are discussed, especially diversity aspects, with Ranta and Ylinen (2023) using text-based machine learning to generate diversity indicators from social media posts and Khan et al. (2023) predicting board diversity from company characteristics. This diversity of approaches to applying machine learning requires the user to tailor new data sets, code environments, and machine learning algorithms for each sustainability metric. This is time-consuming, costly, and challenging in environments where a variety of sustainability data points are necessary, such as in disclosures by MNEs' global value chains, in risk management by banks, or research involving various company-level sustainability aspects.

In addition, even if such estimations are available, regulators normally do not accept their use by financial institutions or MNEs. The main reason is that only point estimates are provided and no additional information is available as to how reliable that estimation is at the firm or portfolio level. Point estimates as they are generated by machine learning models represent the conditional expectation of the predicted variable. Information about confidence intervals or coverage intervals is not provided. This is also a challenge when making causal inferences from these data in research.

We propose an approach that solves these challenges. Essentially, this is achieved through three characteristics or our approach: First, we restrict the independent variables (features) to a large, multidimensional but readily available set of company-level financial data supplemented by fundamental data such as industry, country, and company age. Second, the code architecture systematically integrates the search for the best-performing algorithms in combination with the appropriate data preprocessing before those combinations (pipelines) are optimized and then fed into a final meta-model training. Third, our approach incorporates considerations about prediction uncertainty. Based on recent developments in the machine learning literature (e.g. Tibshirani et al. (2019) and Barber et al. (2023)), we add two prediction intervals to our framework. The first refers to the probability that *on average* in all predictions the true value lies within the interval (*marginal coverage*). The second refers to the probability that the true value lies within the interval *for a specific set* of input values (*conditional coverage*).

We apply our setting to consider three questions:

Q1: To what extent can we derive corporate sustainability data from corporate financial data using machine learning?

Q2: How does the prediction performance change for different dimensions, such as time, region or sector?

Q3: What can uncertainty measure reveal about the quality of the point estimates within

our approach?

In order to demonstrate that a lot of information on the sustainability metrics is actually captured "in the financials" (Q1), we test our approach on the sustainability metrics Scope 1 and Scope 2 emissions, air pollution (NOx emissions), water discharge, and female board share. Despite the versatility of the approach (i.e. just one dataset for all indicators), the predictions seem reasonable for these different metrics. The performance of the framework varies between sustainability indicators and seems to work better for non-truncated data. In doing so, we contribute to the literature by (i) confirming that financial data contain sustainability relevant information and (ii) demonstrating that our approach outperforms existing machine learning-based approaches.

We show that the performance of the models also varies between the time, region, and sector dimensions (Q2). For the earlier years in the dataset (2005 until about 2015), the predictions tend to be better than in the following years. When looking at different world regions, we also observe variations with the predictions for Europe being better than those in other parts of the world. With respect to the different sectors, the approach performs differently and is dependent on the sustainability indicator predicted. In agriculture, for example, the prediction of water discharge is comparatively good, while that of air pollution is relatively bad. This is vice versa for transportation. These findings indicate that the literature in the field should not only rely on the global means of model performance metrics but should make it conditional on the respective use case.

The issue of prediction uncertainty (Q3) is to date not considered in the existing literature related to the imputation of missing sustainability data. We fill this gap by adapting recent developments in the technical literature on machine learning. We find that the *conditional coverage* provides relatively precise information about the confidence intervals of the predictions. As such, this metric should be reported. The intervals are dependent on the coverage needed, that is, the desired confidence a user would like to have her predictions. Using a 68% coverage versus a 95% coverage as examples, we find that a higher risk tolerance at 68% would lead to very small intervals, whereas a low risk tolerance would result in large intervals. This leads us again to deduce that the consideration of use case-specific parameters is important when using machine learning in general and specifying confidence intervals in particular.

We conclude by making the case for the use of machine learning to fill data gaps, as it is a cost-efficient way to generate company-specific sustainability data while maintaining high standards for data integrity. Our suggested approach should help regulators to accept the use of estimations as they are more qualified. Additionally, our approach will help MNEs

and financial institutions meet the disclosure requirements on their GVCs and portfolios, as well as enable them to run more granular analyses including the consideration of risk tolerances in sustainability data predictions.

The remainder of the paper is structured as follows. The next section outlines the methodology, including the introduction of the uncertainty measure. Section III presents the results along the three questions. Section IV concludes.

II. Methodology

Predicting cardinal sustainability data is a supervised ML regression task. Here, we suggest a versatile approach to using regression-based machine learning to predict a variety of company-level sustainability data only from financial data. This approach could help make machine learning a more widely used tool in sustainability and business research (Bosma and van Witteloostuijn, 2024) as well as regulation and industry. In addition, we propose a method to estimate the prediction uncertainty next to the traditional model performance criteria based on point estimates.

A. Model Space

To ensure the versatility of our approach, we define a large model space that we search using a Bayesian approach[1] to maximize our chance of capturing the "best" model for a given sustainability metric. We maximize the coverage of the model space by first training and optimizing a large set of model pipelines (running up to 9,600 trials per sustainability metric[2]), then optimizing the best-performing model configurations, and finally using them in meta-model training, see Figure 1. We evaluate the performance throughout the process using the Mean Squared Error (MSE) and double 10-fold cross-validation.

Base Model Training - Pipeline Selection: The first step of base model training consists of the search for suitable model configurations (pipeline elements per learner) for the sustainability metric at hand by exploring a variety of data preprocessing techniques along with different regression learners as input to hyperparameter optimization. In this step, we run up to 25 optimization trials per model configuration.

Data preprocessing is an essential aspect of the modeling process as it can significantly influence the performance of predictive models. In the literature on sustainability data prediction, this step is usually treated separately from the actual model training. Here, we use data preprocessing as a hyperparameter for optimization in itself. In doing so, we expand the considerations of the no free lunch (NFL) theorem (Wolpert and Macready, 1997) to preprocessing in our setup. The options for data preprocessing in our setup span missing indicator flags, imputation methods (mean, median, iterative), outlier removal strategies (none, winsorization), scaling techniques (standard, robust), transformation approaches (none, quantile) and feature selection methods (none, Lasso). We introduce

[1] Unlike grid or random search, Bayesian optimization utilizes past evaluation results to choose the next set of hyperparameters, efficiently narrowing down to the best possible model settings (Snoek et al., 2012).

[2] We have devised an early stopping mechanism to boost computational efficiency. For more information, see Appendix B.

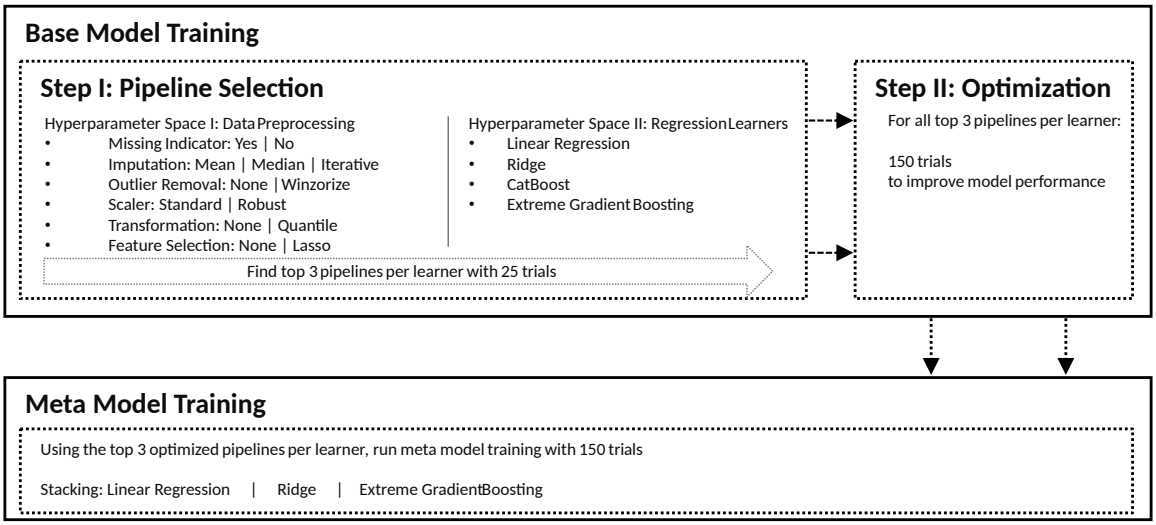


Figure 1. Machine Learning Approach

The figure illustrates the process of Base Model Training, including pipeline selection and optimization using Bayesian hyperparameter tuning, followed by Meta Model Training with a stacked regression approach. Evaluation uses Mean Squared Error (MSE) and double 10-fold cross-validation. Base model training involves selecting top pipelines per learner, followed by further optimization. Meta model training uses the top pipelines in different stacking configurations.

the missing indicator flag (a binary feature) to allow the model to learn from the reporting behavior of a company. The imputation methods are relevant for handling missing data, a common issue in financial datasets, and can affect the model's bias and variance. Imputation strategies such as mean and median are simple and widely used, while iterative methods can provide a more sophisticated approach that accounts for correlations between features (van Buuren, 2007). Outlier removal and scaling enhance the robustness and stability of the models, particularly in financial applications where outliers can represent noise (Aggarwal, 2013). Feature engineering and selection further refine the model by introducing new features that could capture non-linear relationships or selecting the most relevant features to avoid overfitting (Iguyon and Elisseeff, 2003).

In the space of regression learners, the approach contemplates linear regression, ridge regression, extreme gradient boosting (XGBoost) and CatBoost. Linear regression and ridge are fundamental techniques with ridge introducing regularization to manage multicollinearity and overfitting (Hoerl and Kennard, 1970). They are relevant in this study for comparing the training output with methods that are used regularly in econometrics. However, it is unlikely that these learners form the basis of the "best" performing models in the realm of machine learning. On the other hand, Gradient Boosting algorithms, including XGBoost (Chen and Guestrin, 2016) and CatBoost (Prokhorenkova et al., 2018), are powerful ensemble methods that have shown high performance on a wide range of prediction tasks, particularly in the presence of non-linear and complex relationships. One

of the two algorithms is expected to perform best in the context of sustainability data prediction.

Compared to Support Vector Machines (SVMs), K-Nearest Neighbors (KNN), neural networks, and simpler tree methods, XGBoost and CatBoost bring a blend of depth and breadth to the modeling process. SVMs, while effective for small to medium-sized datasets, can be outperformed by tree-ensemble methods in handling large and complex data sets (Fernández-Delgado et al., 2014). KNN suffers from the curse of dimensionality and is inherently slower in making predictions due to its instance-based nature (Beyer et al., 1998). Neural networks, although powerful for large-scale and complex non-linear relationships, require extensive tuning and larger datasets to generalize effectively without overfitting (LeCun et al., 2015). This could reduce the versatility of our setup. Simpler tree methods such as CART or C4.5 can provide interpretable models but usually lack the predictive power of boosted ensembles, which aggregate multiple trees to reduce variance and bias (Breiman, 1996). Therefore, for our regression tasks using financial data, where the data can be noisy and feature relationships complex, XGBoost and CatBoost are likely to be good fits.

Base Model Training - Optimization: The second step involves optimizing the top three model configurations per learner and target variable, selected according to their initial performance in Step I. Each model configuration undergoes 150 Bayesian hyperparameter optimization trials. With this step, we substantially expand the optimization efforts to ensure that the model performance is close to "best".

Meta Model Training: Finally, meta-model training is applied using the three best-performing models per learner. The approach employs stacking, which combines the predictions of multiple models by training a meta-learner, often leading to performance improvements (Wolpert, 1992).

B. Model Performance Metrics

We evaluate the performance of the top three models from both base and meta-training using performance assessment metrics as summarized in Table I. We evaluate the point estimates using these metrics.

Global assessments encompass a set of quantitative measures designed to evaluate the predictive performance of global models. The metrics in this category include the accuracy distribution, which examines the variability in prediction accuracy through the distribution of relative errors between predicted and actual values. The mean absolute error (MAE) and the mean squared error (MSE) quantify the average prediction error and

Metric	Measure/Value	Description
Global Assessment		
Accuracy distribution	Distribution	Refers to the distribution of relative errors between predictions and actual values.
Mean Absolute Error(MAE)	Scale	The average of the absolute errors between the predicted and actual values, representing average prediction error.
Mean Squared Error(MSE)	Numeric value	The average of the squared errors, emphasizing larger errors more than MAE. Also used in the model training as loss function.
R^2 (R-squared)	Numeric value	Measures the proportion of the variance in the dependent variable that is predictable from the independent variables.
Local Assessment		
Quintile allocation	Percentages	Percentage of correctly allocated predictions to actual quintiles (or allocation of n quintiles deviations).
Context	MSE	Evaluation of the model for sectorally, temporally, or spatially local performance.

This table reports a summary of the performance metrics used to assess the point estimates of trained models. Global and local in this context do refer to full or partial results of the model, not to geographic coverage.

Table I. Performance Assessment Metrics for Machine Learning Models

the average of squared errors, respectively, with MSE placing greater emphasis on larger errors. We use MSE as the metric in our loss function and thus as our main measure of model performance. The R-squared (R^2) measures the proportion of variance in the dependent variable that can be predicted from the independent variables, offering insight into the explanatory power of the model.

Local assessments show how the models perform in their operational contexts. This category includes quintile allocation, which evaluates the model's ability to accurately rank predictions within specific quintile brackets. This could be useful when seeking best-in-class investment strategies (Edmans et al., 2022) or when reducing exposure to highly emitting companies in portfolios (Rink et al., 2024). The context examines the performance of the model in sector, space, and time. This should ensure transparency about performance differences prevalent in other sustainability data sets (Dobrick et al., 2023).

C. *Prediction Uncertainty*

Estimating sustainability data from corporate financial data is a task that is inherently characterized by significant uncertainty. Point estimates generated by classical statistical or machine learning models represent the conditional expectation of the target variable - in our case the conditional mean - but do not provide information about the uncertainty or variability in the possible range of prediction values. However, quantifying this uncertainty is crucial to make reliable statements about the accuracy of the estimation and associated risks, to possibly exclude certain ranges from use (*measure not estimate*), or to opt for more conservative or optimistic values instead of the point estimates of the conditional expectation.

A key challenge in accurately and consistently assessing the uncertainty of and between different models is the construction of valid prediction intervals. Our selection of approaches and methods for generating prediction intervals is, therefore, guided by the goal of achieving (sufficiently) good marginal coverage and conditional coverage.

Marginal coverage refers to the probability that, on average across all predictions, the actual target value lies within the prediction interval. This property ensures that the prediction intervals are correct on average across the entire distribution of input data. It is particularly useful for making consistent statements about uncertainty across different models and is relevant in scenarios where models are applied for multiple predictions. Marginal coverage does not guarantee accurate coverage for every individual prediction, but ensures that the average coverage meets the desired level.

Conditional coverage refers to the probability that the true target value lies within the predicted interval for a specific set of input values. This property is stronger than marginal coverage and is crucial when individual-level uncertainty quantification is needed, such as in company-specific predictions. Achieving conditional coverage requires that the prediction intervals are accurately calibrated for each possible input, capturing the uncertainty for that particular instance. However, this is generally more challenging to achieve, especially in the presence of complex data structures or heteroskedasticity.

In our data-driven approach, where financial data are used as predictors of sustainability-related information, additional challenges arise due to the (completely) unknown relationship between these variables (absence of theory). This uncertainty with respect to the underlying distribution suggests a preference for methods that do not rely on strong assumptions about the distribution.

In the literature on uncertainty quantification, there are various approaches to quantify-

ing the uncertainties of model prediction (Soize, 2017; Abdar et al., 2021). A common method involves using scalar uncertainty measures, such as the standard deviation (error), which provides a general measure of the spread of predictions. However, these methods offer only a global perspective on the underlying uncertainty and typically fail to adequately account for the variability of uncertainty across different areas of the distribution. Moreover, theoretical guarantees regarding marginal coverage and conditional coverage are only available under strong assumptions about the distribution and are typically not empirically validated (Palmer et al., 2022).

Quantile regression (Koenker and Bassett, 1978) offers an alternative approach that allows the calculation of prediction intervals using the estimation of quantiles of the target variable. This method is particularly useful when asymmetric uncertainties in predictions cannot be ruled out. A central advantage of quantile regression is that it asymptotically ensures both marginal coverage and conditional coverage for sufficiently large (approaching infinite) sample sizes (Chernozhukov et al., 2009; Romano et al., 2019).

Recently, conformalized prediction has been developed (Vovk et al., 2005; Lei and Wasserman, 2014; Lei et al., 2018). The underlying methods provide theoretical guarantees for marginal coverage in finite samples without making strong assumptions about the distribution.[3]

We employ a conformalized version of the quantile regression (Romano et al., 2019). This approach retains the favorable properties of quantile regression concerning the adaptivity of the prediction intervals and (asymptotically) conditional coverage, while also providing theoretical guarantees for marginal coverage in finite samples under the assumption of i.i.d. data. As with our systematic approach to identifying the optimal model for predicting the conditional mean, this approach to uncertainty quantification aims to be sufficiently close to optimal to allow meaningful conclusions.

Specifically, our approach involves generating symmetric prediction intervals with a target coverage rate of $\tau \in \{68\%, 95\%\}$ to assess prediction uncertainty. Given the size of the datasets, we employ a split method in which the test data set is divided into a calibration set and a (uncertainty) test set. The training of the models using an adapted loss function is performed on the training set. The calibration set is then used to compute conformal scores, which are subsequently employed to construct prediction intervals that achieve the desired coverage levels. The test set is used to evaluate the coverage and analyze prediction uncertainty. This approach is intended to balance the trade-off between

[3]Our approach builds on work that assumes exchangeability, which is satisfied under the assumption of i.i.d. data. For conformalized prediction approaches beyond exchangeability, see Tibshirani et al. (2019) and Barber et al. (2023).

statistical efficiency and computational efficiency. We implement this approach in four steps:

1. *Quantile Regression:* We train the best base models for each learner as well as the meta models on the training data set using the loss function $L_\alpha(\hat{y}_\alpha, y) = (y - \hat{y}_\alpha)\,\alpha \mathbb{1}\{y > \hat{y}_\alpha\} + (\hat{y}_\alpha - y)\,(1-\alpha)\,\mathbb{1}\{y \leq \hat{y}_\alpha\}$, with $\alpha \in \{(1-\tau)/2, (1+\tau)/2\}$. y is the target variable and $\hat{y}$ the (quantile) prediction of the target variable.

2. *Quantile Prediction:* We then use the trained models to predict the conditional quantiles $\hat{y}_\alpha(x)$ for each observation in the calibration data set. x is a vector of the predictor variables for each observation.

3. *Conformal Scores:* Based on these quantile predictions, we determine conformal scores, $c(x, y) = \max\{\hat{y}_{(1-\tau)/2}(x) - y, y - \hat{y}_{(1+\tau)/2}(x)\}$ for each observation in the calibration data set.

4. *Rectifying Quantiles:* Defining $\hat{r} = \text{Quantile}\left(\frac{\lceil (n_{cal}+1)\tau \rceil}{n_{cal}}, \{c_1, ..., c_{n_{cal}}\}\right)$ we can derive conditional prediction intervals

$$I(x) = \left[\hat{y}_{\tau/2}(x) - \hat{r}, \hat{y}_{1-\tau/2}(x) + \hat{r}\right]$$

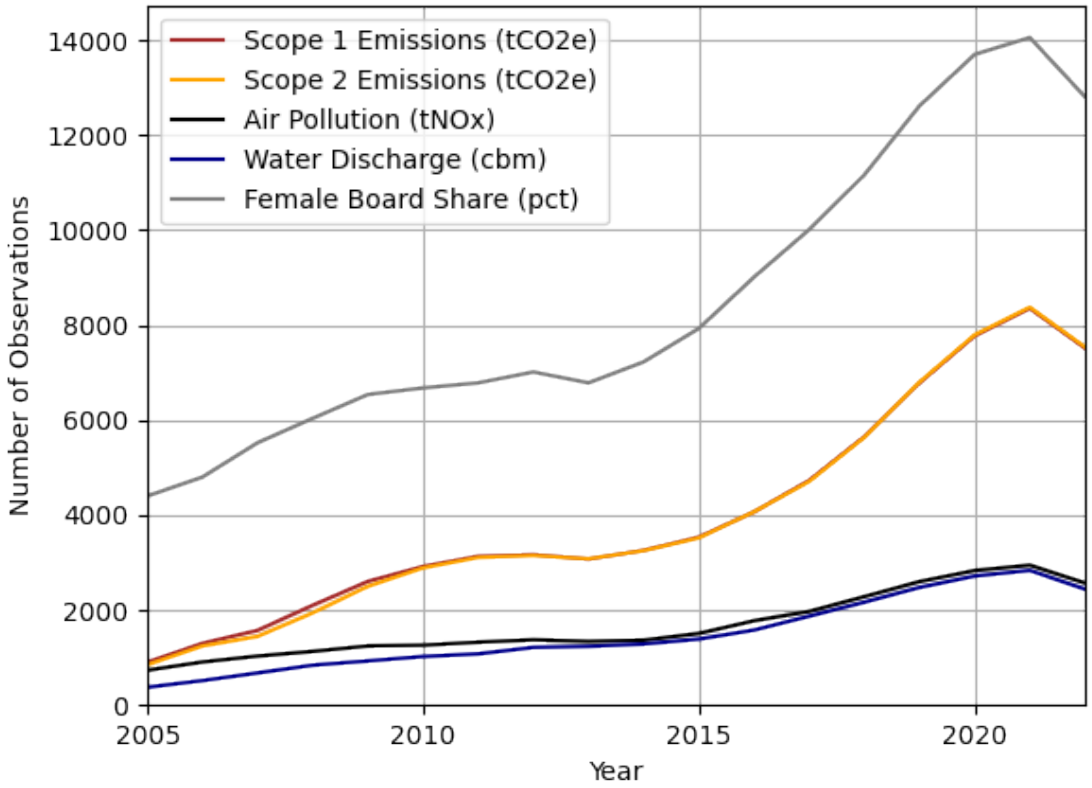
for each observation.

D. Data

Our data set comprises a comprehensive collection of company-year observations for listed equities obtained from the London Stock Exchange Group (LSEG) Data and Analytics database. This data set is global in scope, encompassing data from 95 countries, which provides a diverse and extensive foundation for modeling and analysis. To maintain the integrity and reliability of the results, only reported sustainability data are included in this study, thereby avoiding the potential biases introduced by the LSEG Data and Analytics' estimation models or unaudited data. This conservative approach ensures that the analysis is based solely on real and verifiable metrics.

As shown in Figure 2, the data set includes key sustainability indicators such as Scope 1 and Scope 2 emissions, air pollution, water discharge, and female board share (our "target variables"). Data availability has recently increased substantially, enabling ML applications in the field.

Table II provides a detailed summary of the data, highlighting the breadth and depth of



The figure illustrates the availability of reported sustainability data by listed companies over time.

Figure 2. Reported Sustainability Data over Time

the data set. The data set time period runs from 2005 to 2022.

- Scope 1 and Scope 2 Emissions: The dataset contains nearly 50,000 observations each for these emissions categories, covering 83 sectors across 83 countries, with data available for more than 8,000 companies. We select these indicators to benchmark against other studies and due to the relevance of climate change to business.
- Air Pollution and Water Discharge: These variables have fewer observations, reflecting the more limited availability of environmental data. However, they still provide significant coverage, with data on more than 3,000 companies from more than 60 countries. These indicators are included to reflect emerging topics in business and finance research such as biodiversity, blue economy, and livable cities.
- Female Board Share: This social governance indicator is well-represented, with over 100,000 observations across 95 countries and 86 sectors, providing the richest data set in our analysis. It should demonstrate that our approach is applicable beyond

environmental sustainability data.

The data set includes 212-256 predictor variables, depending on the specific sustainability metric, with data completeness ranging from 57% to 65%. This breadth of variables offers a comprehensive view of company characteristics and operational contexts.

Dataset	Scope 1 Emissions	Scope 2 Emissions	Air Pollution	Water Discharge	Female Board Share
General Information					
Number of observations	47685	47320	20980	18426	108834
Number of sectors	83	83	76	74	86
Number of countries	83	83	63	63	95
Number of companies	8391	8335	3389	3098	14406
Start year	2005	2005	2005	2005	2005
End year	2022	2022	2022	2022	2022
Number of predictor variables	240	240	213	211	255
Data completeness (in %)	63.86	63.77	65.72	65.62	57.92
Target Variable Information					
Mean	3786965	1051686	19914	187082423	16.00
Standard deviation	26440465	48614503	179304	1318742226	14.00
Minimum	0.00	0.00	0.00	0.00	0.00
Maximum	4421000000	7386660000	14042000	26877900000	100.00
Target Variable Information					
Log (1+value) Mean	10.75	11.07	6.31	14.98	2.17
Log (1+value) Std	3.55	2.72	3.28	3.60	1.40
Log (1+value) Min	0.00	0.00	0.00	0.00	0.00
Log (1+value) Max	22.21	22.72	16.46	24.01	4.62

This table presents summary statistics for the dataset, including Scope 1 and Scope 2 emissions, air pollution, water discharge, and female board share. The table details general information such as the number of observations, sectors, countries, companies, the time period (2005-2022), and data completeness rates. It also includes target variable information, such as means, standard deviations, and the range (minimum to maximum) of the absolute and log-transformed values.

Table II. Summary Statistics

The features selected for this study focus on fundamental and financial data, excluding direct sustainability-related metrics (e.g., energy use) to challenge the model's ability to infer sustainability data solely from financial and fundamental data. The feature set includes the following. A complete list of the variables including summary statistics is presented in Appendix E.

- Financial data (e.g., income statement, balance sheet, cash flow metrics), capturing idiosyncratic company characteristics such as size, age, innovative capacity, and asset intensity.
- Fundamental data
  - Industry indicators to account for general emission intensity trends within specific sectors.
  - Spatial variables that reflect policy and socio-economic operating conditions.

  - Company age that reflects the general development stage of a company.

- Temporal effects captured by the inclusion of year variables.

Data preprocessing steps included retrieving data via the LSEG Data and Analytics API, filtering out non-listed equity observations, and excluding entries without date variables or those with quarterly reporting. The target variables were logarithmically transformed to improve predictive performance and for highly correlated features (correlation >= 99%) only one feature is retained to mitigate multicollinearity. Furthermore, features with very low data availability (missingness >= 99%) were excluded, as imputation was not feasible in these cases.

This rigorous data selection and processing framework prior to the pipeline preprocessing steps ensures that the resulting model is robust and reliable, providing information on how financial metrics can be used to predict sustainability data.

III. Results

A. It's in the Financials, Stupid!

Our analysis reveals that sustainability data can be predicted from corporate financial data with a high degree of accuracy. Complex and non-parametric machine learning models significantly outperform linear models in predicting sustainability data from financial data, see Table III. This indicates that the underlying structure of sustainability data is quite complex and cannot be adequately captured by simple linear relationships.

Specifically, our best-performing models achieve mean squared errors (MSE) in the range of 0.6-1.9 and mean absolute errors (MAE) between 0.3 and 0.7. These results are substantially better compared to a benchmark study on corporate carbon emissions by Nguyen et al. (2022), which reported MAE values around 1.1 and 0.8 for Scope 1 and Scope 2 emissions where we achieve 0.6 and 0.5 respectively. Moreover, our models demonstrate high explanatory power. The best performing models consistently explain more than 90% of the variance in the sustainability data sample. This high level of performance underscores the potential of advanced machine learning techniques in enhancing the predictive accuracy of sustainability metrics based on financial data. Hence, it's in the financials, stupid!

The performance of the prediction varies between the different sustainability metrics. The performance of the female board share is weaker in relative terms compared to environmental variables. This indicates that performance cannot always be improved in

Target Variable	Metric	Base Learner				Meta Learner		
		Linear Regression	Ridge	CatBoost	XGBoost	Linear Regression	Ridge	XGBoost
Scope 1 Emissions (tCO2e)	MAE	1.261	1.240	0.648	0.662	0.610	0.609	0.589
	MSE	3.401	3.252	1.579	1.617	1.622	1.622	1.589
	R2	0.713	0.726	0.867	0.864	0.863	0.863	0.866
Scope 2 Emissions (tCO2e)	MAE	1.252	1.118	0.513	0.627	0.556	0.556	0.527
	MSE	3.400	2.959	1.163	1.269	1.275	1.275	1.204
	R2	0.540	0.599	0.842	0.828	0.827	0.827	0.837
Air Pollution (tNOx)	MAE	1.501	1.444	0.695	0.905	0.672	0.672	0.650
	MSE	4.598	4.320	1.923	2.262	1.907	1.907	1.909
	R2	0.564	0.590	0.818	0.786	0.819	0.819	0.819
Water Discharge (cbm)	MAE	1.782	1.782	0.278	0.623	0.275	0.275	0.273
	MSE	7.528	7.522	1.055	1.526	1.063	1.063	1.053
	R2	0.443	0.443	0.922	0.887	0.921	0.921	0.922
Female Board Share (pct)	MAE	0.876	0.876	0.473	0.501	0.448	0.448	0.428
	MSE	1.253	1.249	0.598	0.638	0.595	0.595	0.592
	R2	0.347	0.349	0.688	0.667	0.690	0.690	0.691

This table reports the performance metrics of the machine learning models Linear Regression, Ridge, CatBoost, and XGBoost, used to predict sustainability data. The models are evaluated based on their Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R2) values in five target variables Scope 1 Emissions, Scope 2 Emissions, Air Pollution, Water Discharge, and Female Board Share. Target variable is in log + 1 format.

Table III. Global Model Performance

predicting sustainability data from financial data by increasing the sample size. It also highlights the difficulty that some models seem to have in dealing with truncated data (by definition, the female board share is scaled between 0% and 100%).

The relative errors in the predictions remain non-negligible. To show the applicability of the predicted data to different use cases, we dissect the results at the local level in the next step.

B. Model Performance Varies

In subsequent analyses, we show the results for the best-performing model for each of the sustainability metrics. These are base XGBoost models for Scope 1 and 2 emissions, a meta Ridge model for air pollution, and meta XGBoost models for water discharge and female board share.

Quintile Analysis

First, we evaluate the performance of our models using a quintile-based approach. Specifically, we divide the data set into quintiles with roughly the same number of observations based on the size of the target variable[4]. This allows us to assess how well our models perform across different ranges of sustainability metrics.

The quintile-based approach shows that the models generally perform the worst in the

[4]Note that female board share has nearly 30% of observations equaling zero. Therefore, in this case, we manually adjusted the binning. As a result, more observations are in Quintile 1 in this case.

smallest quintile, which includes the lowest values of the target variables; see Table IV. This pattern is consistent in the five target variables analyzed. A key factor contributing to this performance disparity is the inherent difficulty in predicting values close to zero, which is only necessary in the smallest quintile.

Target Variable	Metric	Q1	Q2	Q3	Q4	Q5
Scope 1 Emissions (tCO2e)	MAE	1.112	0.571	0.476	0.466	0.618
	MSE	3.922	0.903	0.718	0.757	1.593
Scope 2 Emissions (tCO2e)	MAE	1.004	0.377	0.404	0.335	0.446
	MSE	3.517	0.413	0.500	0.405	0.982
Air Pollution (tNOx)	MAE	1.137	0.464	0.451	0.698	0.612
	MSE	4.074	0.780	0.803	2.015	1.861
Water Discharge (cbm)	MAE	0.532	0.222	0.205	0.185	0.221
	MSE	3.146	0.467	0.482	0.362	0.804
Female Board Share (pct)	MAE	0.692	0.301	0.315	0.344	0.401
	MSE	1.376	0.254	0.293	0.331	0.448

This table presents the performance metrics of the best-performing machine learning model across different quintiles of the target variables. The dataset is divided into quintiles based on the size of each sustainability metric Scope 1 Emissions, Scope 2 Emissions, Air Pollution, Water Discharge, and Female Board Share. Performance is measured using Mean Absolute Error (MAE) and Mean Squared Error (MSE) for each quintile, labeled Q1 through Q5, where Quintile 1 (Q1) represents the smallest values of the target variable, while Quintile 5 (Q5) represents the largest values. All input values are in log + 1 format.

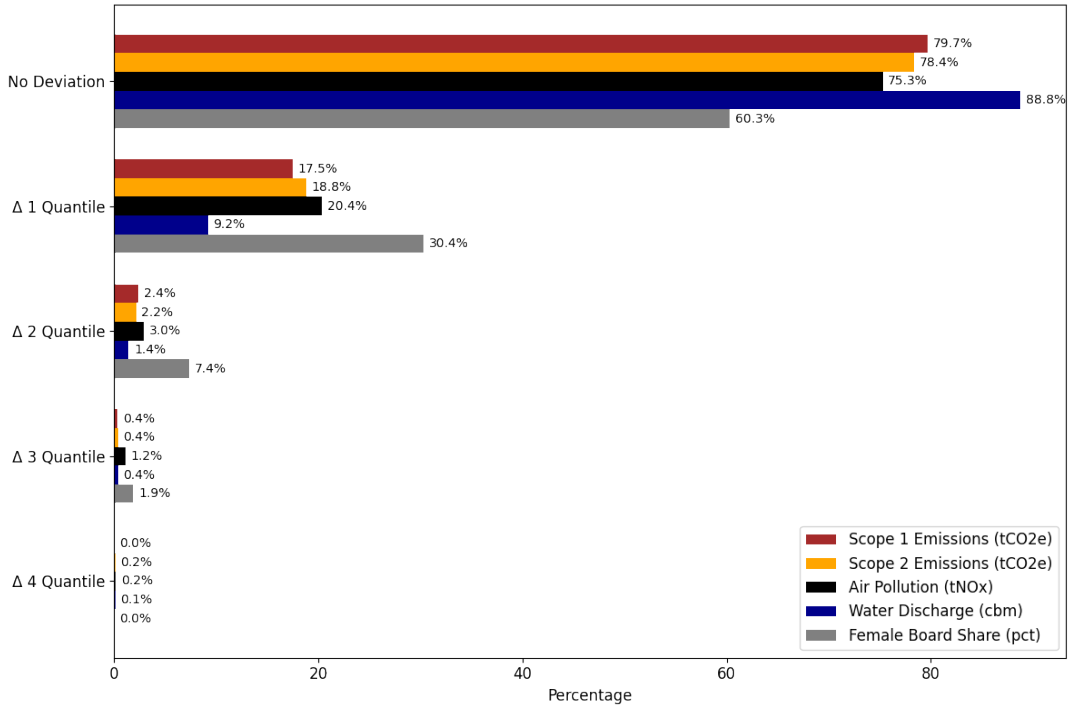
Table IV. Performance of Best Models by Quintiles

The performance variation between the four remaining quintiles is relatively modest, with some target variables showing a slight decrease in performance in the fifth (largest) quintile. We do not observe any particular quintile consistently performing the best or second worst across all metrics, and the deviations in performance are generally small. This suggests that the models maintain a stable performance level in most of the predicted data distributions.

Finally, we categorically examine the degree of deviation between the predicted and actual quintiles. We find that our predictions fall into the right quintile 60.3% (female board share) to 88.8% (Scope 1 emissions) of the time, see Figure 3. In most cases of deviation, the predicted quintile deviates by only one quintile from the actual value. This underscores the overall robustness and performance of our models for categorizing companies based on our continuous predictions.

In practical terms, for an MNE, these findings indicate that the models are reliable for identifying high-impact areas within GVCs. This can be particularly useful for companies seeking to prioritize their strategic focus and allocate resources effectively to areas with the greatest impact on sustainability.

Next, we analyze the performance of the model in the temporal, spatial, and sectoral



The pie charts display the proportion of observations of absolute deviating quintiles for the key variables: Scope 1 and Scope 2 emissions, air pollution, water discharge, and female board share.

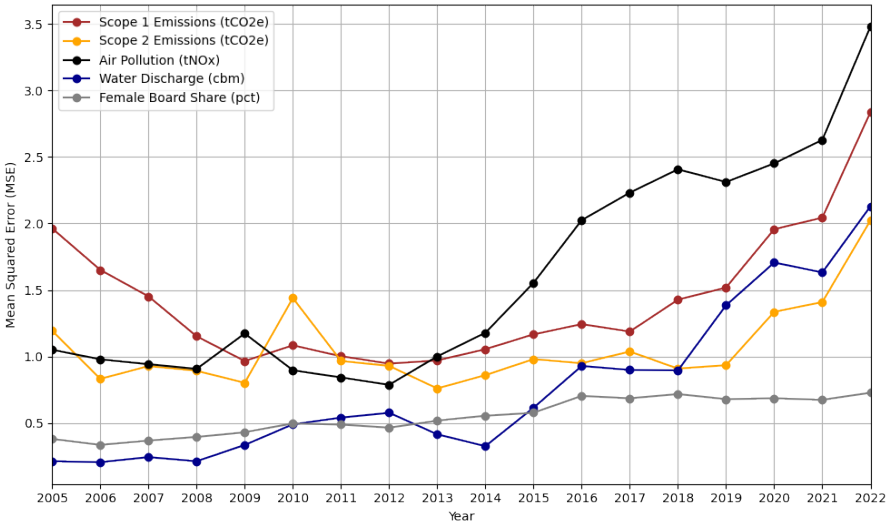
Figure 3. Number of Deviating Quintiles

dimensions to assess how our models perform under different conditions.

Temporal Dimension

Our analysis indicates that the models perform best during the period up to 2015, with a slight decrease in performance in subsequent years; see Figure 4. This decline is relatively minor for all target variables except air pollution. We attribute it to the increased variability in the data set over time, that is, more companies reporting their sustainability data. Although more data should generally help in training machine learning models, higher heterogeneity among reporting companies (business models, sizes, technology mix, etc.) make it more difficult for the algorithm to predict with the same performance level, especially if the heterogeneity grows faster than the data availability.

When applying machine learning models to real-world scenarios, it is essential to consider these temporal variations. Users should ensure that the specific use case is well understood so that the models can be appropriately adapted. Particularly when focusing on current data, it might be necessary to apply variations or modifications to the models trained on older data. One potential approach is to incorporate a discount factor in the loss function that adjusts for temporal discrepancies.



This figure illustrates the model performance over time (2005–2022) for five different variables: Scope 1 emissions, Scope 2 emissions, air pollution, water discharge, and female board share. The ordinate represents the MSE, showing how prediction performance has evolved for each variable across the years.

Figure 4. Temporal Model Performance

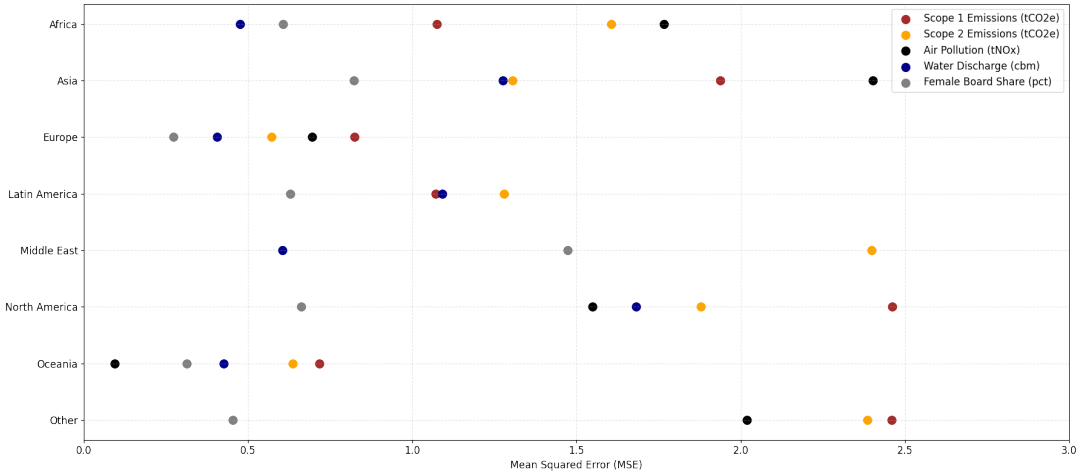
Spatial Dimension

In the spatial analysis, we observe some variation in model performance between different geographic regions; see Figure 5. In particular, our models tend to perform better in Europe (and Oceania) than in other parts of the world. This discrepancy could be partially attributed to stricter and more comprehensive reporting regimes in Europe, resulting in a larger and more reliable dataset (Krueger et al., 2024), that is, reporting is more homogeneous due to mandatory guidelines that improve model performance in these regions (or reduce heterogeneity). However, despite these differences, the models still perform well in all regions. For MNEs looking to optimize global supply chains, our approaches are applicable, although some caution is advised.

Sectoral Dimension

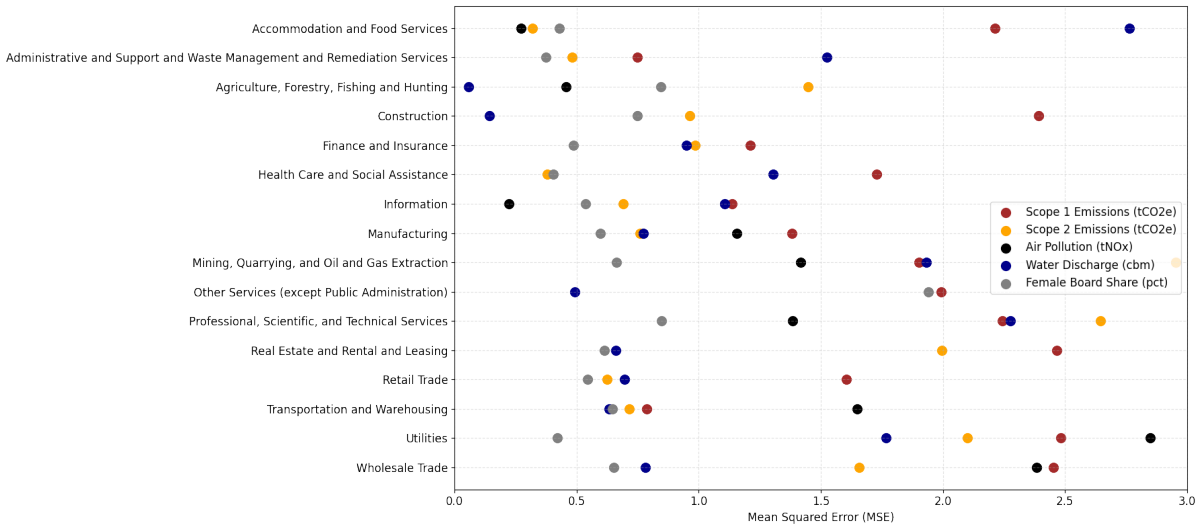
Regarding the sectoral dimension, we find minor variations in model performance across different industries, with some outliers; see Figure 6. For most industries, the models perform reliably, suggesting that they are well suited for applications that involve analyzing portfolios spanning multiple sectors and industries, such as those of banks or asset managers. However, if a particular sector is of special importance to an economic actor, it may be beneficial to use a weighted loss function tailored to sectoral importance.

In summary, while there is some variation in the model performance across the temporal, spatial, and sectoral dimensions, the models generally perform well in all three areas.



This figure shows the model performance across different global regions, as measured by the MSE for the five target variables: Scope 1 emissions, Scope 2 emissions, air pollution, water discharge, and female board share. Each region is represented along the ordinate, while the MSE values are plotted as dots.

Figure 5. Spatial Model Performance



This figure displays the MSE of model predictions across industries for the five target variables: Scope 1 emissions, Scope 2 emissions, air pollution, water discharge, and female board share. Each dot represents the MSE for a particular variable within a specific industry.

Figure 6. Sectoral Model Performance

However, the variation underscores the importance of understanding the specific use case and tailoring the models accordingly to ensure optimal performance for the task at hand.

C. Beware The Prediction Uncertainty

The analysis of prediction uncertainty across the target variables reveals heterogeneous patterns. For smaller coverage requirements, the uncertainty is generally negligible,

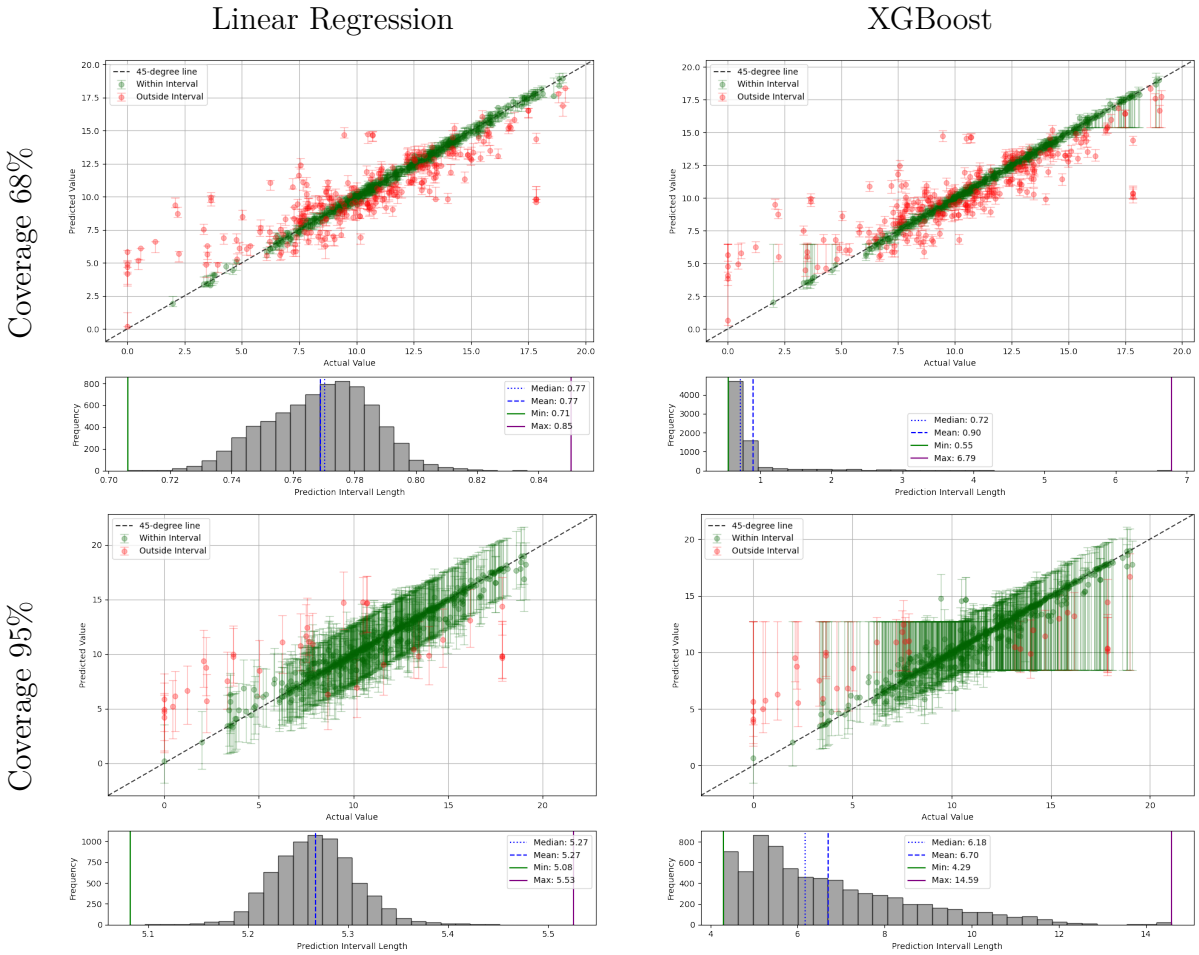
whereas for larger coverage requirements, it becomes significant. Our conformalized quantile regression approach demonstrates empirically valid coverage (with an absolute difference between empirical and target coverage $\leq 0.68\%$ for all variables and coverage rates). The prediction uncertainty remains largely symmetrical, balancing under- or overestimation of prediction uncertainty compared to other standard uncertainty measures. In this section, we show the results for Scope 1 Emissions; see Appendix D for the results for the other target variables.

Figure 7 illustrates the prediction uncertainty in four settings, including conformalized quantile regression using linear regression and XGBoost and with coverage rates of 68% and 95%. The figure demonstrates that, while prediction uncertainty always exists, it is relatively low for the 68% coverage rate. This suggests that users who are not highly risk-averse can use the point estimates without encountering (on average) significant uncertainty. However, at the 95% coverage rate, uncertainty increases substantially, indicating a wider range of possible outcomes, which may be of concern to users who require greater confidence in their predictions.

In addition, the choice of the learner to estimate the uncertainty affects the results. Linear regression models tend to produce on average narrower prediction intervals with relatively constant interval lengths across the distribution of the target variable. In contrast, XGBoost exhibits a more adaptive behavior, with wider prediction intervals towards the extremes of the target variable values. For lower actual values, the uncertainty intervals become larger, particularly skewing toward values above the actuals. Similarly, for the largest values, the intervals widen, though skewed downward, toward lower-than-actual values.

Figure 8 further examines the distribution of the prediction intervals, confirming their general symmetries. In this figure, we use predictions of the conditional mean from the (best performing) XGBoost meta-leaner. Linear regression models show tighter distributions of interval lengths, while XGBoost produces more widely spread intervals.

This figure also presents alternative uncertainty measures for comparison. In addition to the conformalized quantile-based measure, a more naive approach is shown using simple one- and two-standard deviations (σ) from the conditional mean, which are typically used for coverage levels 68% and 95%. This naive approach results in significantly larger intervals compared to the conformalized quantile-based method in most settings. In addition, an alternative method has been employed that uses (standard) quantile regression (ν). This method underestimates the uncertainty compared to our conformalized approach most of the time. This leads us to conclude that the conformalized quantile-based



This figure displays the prediction uncertainty for Scope 1 Emissions. The settings are for the learners linear regression and XGBoost and the target coverage rate of 68% and 95%. For better readability, the actual vs. predicted plots display only 2% of the total observations, randomly selected.

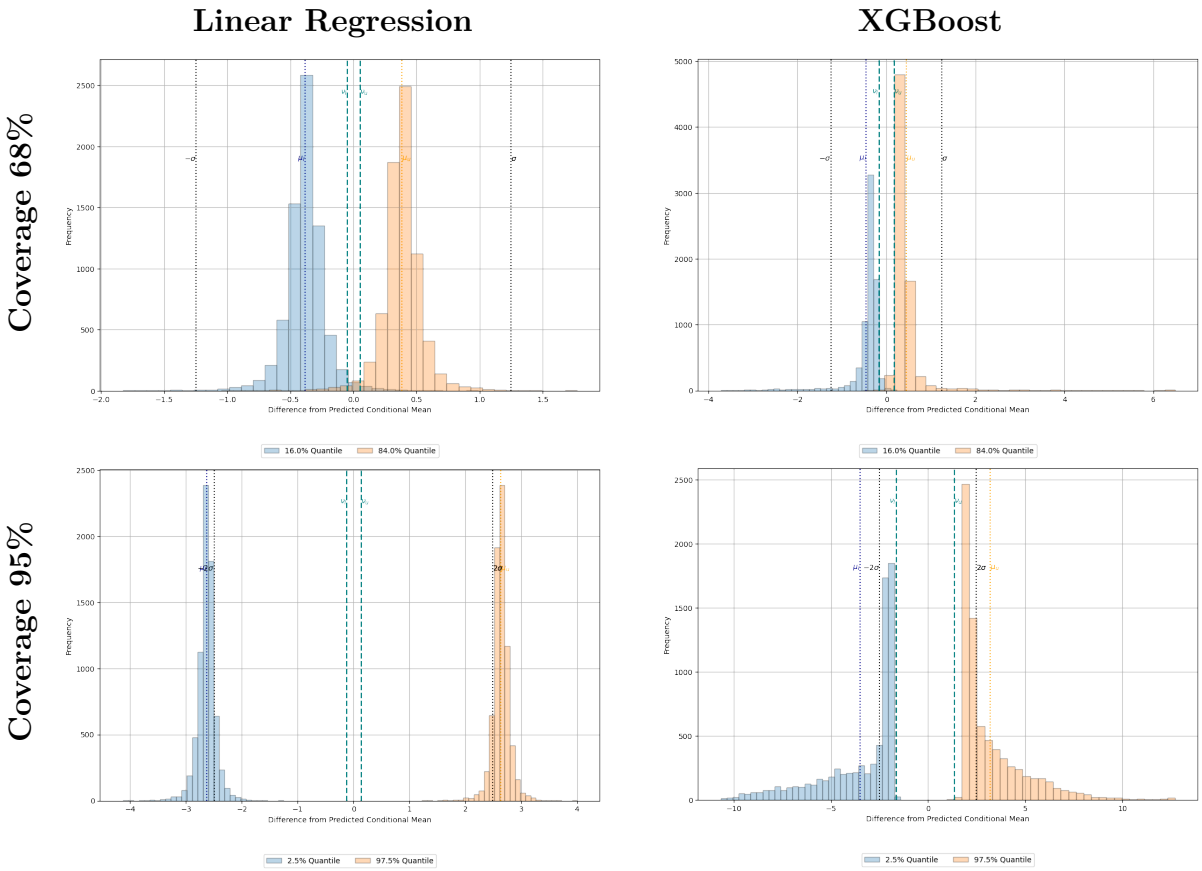
Figure 7. Prediction Uncertainty in Different Settings for Scope 1 Emissions

method is well suited to test prediction uncertainty.

The results of the uncertainty measures for other target variables than Scope 1 emissions show consistent results for environmental variables; however, for the variable representing the female board share (see Appendix D), which is truncated between zero and one, the uncertainty distributions differ, reflecting the specific characteristics of the data.

IV. Conclusion

This study demonstrates that ML can serve as an effective tool for generating company-level sustainability data using only financial data. The findings of this research have several important implications for policymakers, MNEs, and the financial industry.



This figure displays the deviation from the predicted conditional mean for Scope 1 Emissions in different settings. The settings are for uncertainty estimates based on linear regression and XGBoost and for targeted coverage rates of 68% and 95%. In all settings, the conditional mean is predicted by the best-performing XGB meta-model. The black dotted lines represent the prediction intervals using one and two standard deviations (σ), respectively. The green dashed lines represent the mean lower (upper) quantile prediction from standard quantile regression, ν_l (ν_u). Finally, the blue and orange bars show the distribution of quantile predictions for the lower and upper quantile from conformalized quantile regression. The mean of lower (upper) predictions is represented by the blue and orange dotted lines, μ_l (μ_u).

Figure 8. Deviation from Conditional Mean for Scope 1 Emissions

First, policymakers should reconsider the prevailing trend of relying solely on sustainability-related raw data within industry and the financial systems. Although the use of raw data has its merits, particularly in ensuring accuracy, there are instances where the transaction costs associated with obtaining such data are prohibitively high. In these cases, the use of machine learning-generated data offers a viable alternative to using generalized metrics, such as industry averages. This approach can improve the granularity and relevance of sustainability data for decision making.

Second, the study highlights the potential variations and uncertainties inherent in any prediction model, particularly those related to sustainability data. To address this, there is a need for increased transparency requirements for both data providers and companies

when using machine learning or other modeling techniques to generate sustainability data. This transparency is crucial to ensure that the uncertainties associated with these models are well understood and that any biases in the underlying data are clearly communicated. Policymakers could enhance transparency standards, similar to the current efforts for ESG ratings in the European Union (General Secretariat of the Council, 2024), and extend these standards throughout the industry to ensure consistency and reliability.

Third, our research emphasizes the importance of considering the specific use case when applying machine learning models. The performance of these models can vary significantly depending on temporal, spatial, or sectoral factors, which must be taken into account to ensure the accuracy and relevance of the generated data. Adjustments to the loss function can improve the suitability of the model for specific applications, but these considerations should be made transparent to users, illustrating the methodology behind the data generation process.

It is important to acknowledge the limitations of our research. In particular, questions remain regarding the generalizability of our models. The models developed in this study are trained primarily on data from MNEs, which may not be directly applicable to government organizations, non-profit entities, or small and medium-sized enterprises (SMEs). These other parts of the economy may require different data sets, and it remains uncertain whether sufficient data is available to support the development of ML models in these contexts. Future research should explore these gaps, potentially identifying alternative approaches to address the limitations in data availability for these use cases.

In conclusion, this study illustrates that machine learning has a significant role to play in making sustainability data more accessible and cost-effective. By improving the availability and quality of sustainability data through ML, companies and policymakers can make better informed decisions, ultimately advancing sustainability initiatives throughout the global economy.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., Makarenkov, V., Nahavandi, S., 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. doi:10.1016/j.inffus.2021.05.008.

Acemoglu, D., 1997. Training and Innovation in an Imperfect Labour Market. The Review of Economic Studies 64, 445. doi:10.2307/2971723.

Adams, R.B., Ferreira, D., 2009. Women in the boardroom and their impact on governance and performance. Journal of Financial Economics 94. doi:10.1016/j.jfineco.2008.10.007.

Aggarwal, C.C., 2013. Outlier analysis. volume 9781461463962. doi:10.1007/978-1-4614-6396-2.

Balasubramanian, N., Lee, J., 2008. Firm age and innovation. Industrial and Corporate Change 17. doi:10.1093/icc/dtn028.

Barber, R.F., Candès, E.J., Ramdas, A., Tibshirani, R.J., 2023. Conformal prediction beyond exchangeability. The Annals of Statistics 51. doi:10.1214/23-AOS2276.

Bauer, R., Ruof, T., Smeets, P., 2021. Get Real! Individuals Prefer More Sustainable Investments. Review of Financial Studies 34. doi:10.1093/rfs/hhab037.

Berg, F., Kölbel, J.F., Rigobon, R., 2022. Aggregate Confusion: The Divergence of ESG Ratings. Review of Finance 26. doi:10.1093/rof/rfac033.

Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U., 1998. When is "nearest neighbor" meaningful?, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). doi:10.1007/3-540-49257-7{_}15.

Bolton, P., Kacperczyk, M., Samama, F., 2022. Net-Zero Carbon Portfolio Alignment. Financial Analysts Journal 78. doi:10.1080/0015198X.2022.2033105.

Bosma, B., van Witteloostuijn, A., 2024. Machine learning in international business. Journal of International Business Studies 55, 676–702. doi:10.1057/s41267-024-00687-6.

Breiman, L., 1996. Bagging predictors. Machine Learning 24. doi:10.1007/bf00058655.

Busch, T., Johnson, M., Pioch, T., 2022. Corporate carbon performance data: Quo vadis? Journal of Industrial Ecology 26. doi:10.1111/jiec.13008.

van Buuren, S., 2007. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research 16. doi:10.1177/0962280206074463.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. doi:10.1145/2939672.2939785.

Chernozhukov, V., Hansen, C., Jansson, M., 2009. Finite sample inference for quantile regression models. Journal of Econometrics 152. doi:10.1016/j.jeconom.2009.01.004.

Dobrick, J., Klein, C., Zwergel, B., 2023. Size bias in refinitiv ESG data. Finance Research Letters 55, 104014. doi:10.1016/j.frl.2023.104014.

ECB, 2022. 2022 climate risk stress test. Technical Report. European Central Bank. Frankfurt am Main. URL: https://www.bankingsupervision.europa.eu/ecb/pub/pdf/ssm.climate_stress_test_report.20220708~2e3cc0999f.en.pdf, doi:10.2866/97350.

Edmans, A., 2023. Applying Economics—Not Gut Feel—to ESG. Financial Analysts Journal doi:10.1080/0015198X.2023.2242758.

Edmans, A., Levit, D., Schneemeier, J., 2022. Socially Responsible Divestment. SSRN Electronic Journal doi:10.2139/ssrn.4093518.

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems? Journal of Machine Learning Research 15.

Garel, A., Romec, A., Sautner, Z., Wagner, A.F., 2023. Do Investors Care About Biodiversity? SSRN Electronic Journal doi:10.2139/ssrn.4398110.

General Secretariat of the Council, 2024. Proposal for a Regulation of the European Regulation and the Council on the transparency and integrity of Environmental, Social and Governance (ESG) rating activities, and amending Regulation (EU) 2019 2088.

Giglio, S., Kuchler, T., Stroebel, J., Zeng, X., 2023. Biodiversity Risk. SSRN Electronic Journal doi:10.2139/ssrn.4420552.

Goldhammer, B., Busse, C., Busch, T., 2017. Estimating Corporate Carbon Footprints with Externally Available Data. Journal of Industrial Ecology 21. doi:10.1111/jiec.12522.

Griffin, P.A., Lont, D.H., Sun, E.Y., 2017. The Relevance to Investors of Greenhouse

Gas Emission Disclosures. Contemporary Accounting Research 34. doi:10.1111/1911-3846.12298.

Hoerl, A.E., Kennard, R.W., 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 12. doi:10.1080/00401706.1970.10488634.

Iguyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection.

Ilhan, E., Krueger, P., Sautner, Z., Starks, L.T., 2023. Climate Risk Disclosure and Institutional Investors. Review of Financial Studies 36. doi:10.1093/rfs/hhad002.

Ilhan, E., Sautner, Z., Vilkov, G., 2021. Carbon Tail Risk. Review of Financial Studies 34. doi:10.1093/rfs/hhaa071.

Jordan, M.I., Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. doi:10.1126/science.aaa8415.

Khan, H.u.R., Bin Khidmat, W., Hammouda, A., Muhammad, T., 2023. Machine learning in the boardroom: Gender diversity prediction using boosting and undersampling methods. Research in International Business and Finance 66, 102053. doi:10.1016/j.ribaf.2023.102053.

Kim, D., Starks, L.T., 2016. Gender Diversity on Corporate Boards: Do Women Contribute Unique Skills? American Economic Review 106, 267–271. doi:10.1257/aer.p20161032.

Koenker, R., Bassett, G., 1978. Regression Quantiles. Econometrica 46, 33. doi:10.2307/1913643.

Krueger, P., Sautner, Z., Tang, D.Y., Zhong, R., 2024. The Effects of Mandatory ESG Disclosure Around the World. Journal of Accounting Research doi:10.1111/1475-679X.12548.

Kunming Declaration, 2021. Declaration from the High-Level Segment of the UN Biodiversity Conference 2020 under the Theme: Ecological Civilization: Building a Shared Future for All Life on Earth. Technical Report. URL: https://www.cbd.int/doc/c/c2db/972a/fb32e0a277bf1ccfff742be5/cop-15-05-add1-en.pdf.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. doi:10.1038/nature14539.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., Wasserman, L., 2018. Distribution-Free Predictive Inference for Regression. Journal of the American Statistical Association 113. doi:10.1080/01621459.2017.1307116.

Lei, J., Wasserman, L., 2014. Distribution-free prediction bands for non-parametric regression. Journal of the Royal Statistical Society. Series B: Statistical Methodology 76. doi:10.1111/rssb.12021.

Loughran, T., McDonald, B., 2023. Measuring Firm Complexity. Journal of Financial and Quantitative Analysis , 1–28doi:10.1017/S0022109023000716.

Marano, V., Wilhelm, M., Kostova, T., Doh, J., Beugelsdijk, S., 2024. Multinational firms and sustainability in global supply chains: scope and boundaries of responsibility. Journal of International Business Studies 55, 413–428. doi:10.1057/s41267-024-00706-6.

Nguyen, Q., Diaz-Rainey, I., Kitto, A., Mcneil, B., Pittman, N.A., Zhang, R., 2022. Scope 3 Emissions: Data Quality and Machine Learning Prediction Accuracy. URL: https://deliverypdf.ssrn.com/delivery.php?ID=735105085006000008086072113005079106120009055009062036075099117085093102127079083094C pdf&INDEX=TRUE.

Nguyen, Q., Diaz-Rainey, I., Kuruppuarachchi, D., 2021. Predicting corporate carbon footprints for climate finance risk analyses: A machine learning approach. Energy Economics 95. doi:10.1016/j.eneco.2021.105129.

Palmer, G., Du, S., Politowicz, A., Emory, J.P., Yang, X., Gautam, A., Gupta, G., Li, Z., Jacobs, R., Morgan, D., 2022. Calibration after bootstrap for accurate uncertainty quantification in regression models. npj Computational Materials 8. doi:10.1038/s41524-022-00794-8.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A., 2018. Catboost: Unbiased boosting with categorical features, in: Advances in Neural Information Processing Systems.

Ranta, M., Ylinen, M., 2023. Board gender diversity and workplace diversity: a machine learning approach. Corporate Governance (Bingley) 23. doi:10.1108/CG-01-2022-0048.

Rink, S., Matheis, Y., Dumrose, M., 2024. Private Capital under the Paris Agreement: Institutional Investors and Company Decarbonization. Unpublished Work .

Romano, Y., Patterson, E., Candès, E.J., 2019. Conformalized quantile regression, in: Advances in Neural Information Processing Systems.

Slager, R., Chuah, K., Gond, J.P., Furnari, S., Homanen, M., 2023. Tailor-to-Target: Configuring Collaborative Shareholder Engagements on Climate Change. Management Science doi:10.1287/mnsc.2023.4806.

Snoek, J., Larochelle, H., Adams, R.P., 2012. Practical Bayesian optimization of machine learning algorithms, in: Advances in Neural Information Processing Systems.

Soize, C., 2017. Uncertainty Quantification. volume 47. Springer International Publishing, Cham. doi:10.1007/978-3-319-54339-0.

Starks, L.T., 2023. Presidential Address: Sustainable Finance and ESG Issues—Value versus Values. Journal of Finance 78, 1837–1872. doi:10.1111/jofi.13255.

Tian, M., 2023. Impact of Climate Water Risk on Corporate Operational and Capital Markets Performance: A Machine Learning Approach. Ph.D. thesis. University of Michigan. URL: https://deepblue.lib.umich.edu/bitstream/handle/2027.42/192421/mytian_1.pdf?sequence=1&isAllowed=y.

Tibshirani, R.J., Barber, R.F., Candès, E.J., Ramdas, A., 2019. Conformal prediction under covariate shift, in: Advances in Neural Information Processing Systems.

United Nations, 2015. Paris Agreement - UNFCCC. URL: https://unfccc.int/process-and-meetings/the-paris-agreement.

Vovk, V., Gammerman, A., Shafer, G., 2005. Algorithmic learning in a random world. doi:10.1007/b106715.

Wolpert, D., Macready, W., 1997. No free lunch theorems for optimization. IEEE Transactions on Evolutionary Computation 1, 67–82. doi:10.1109/4235.585893.

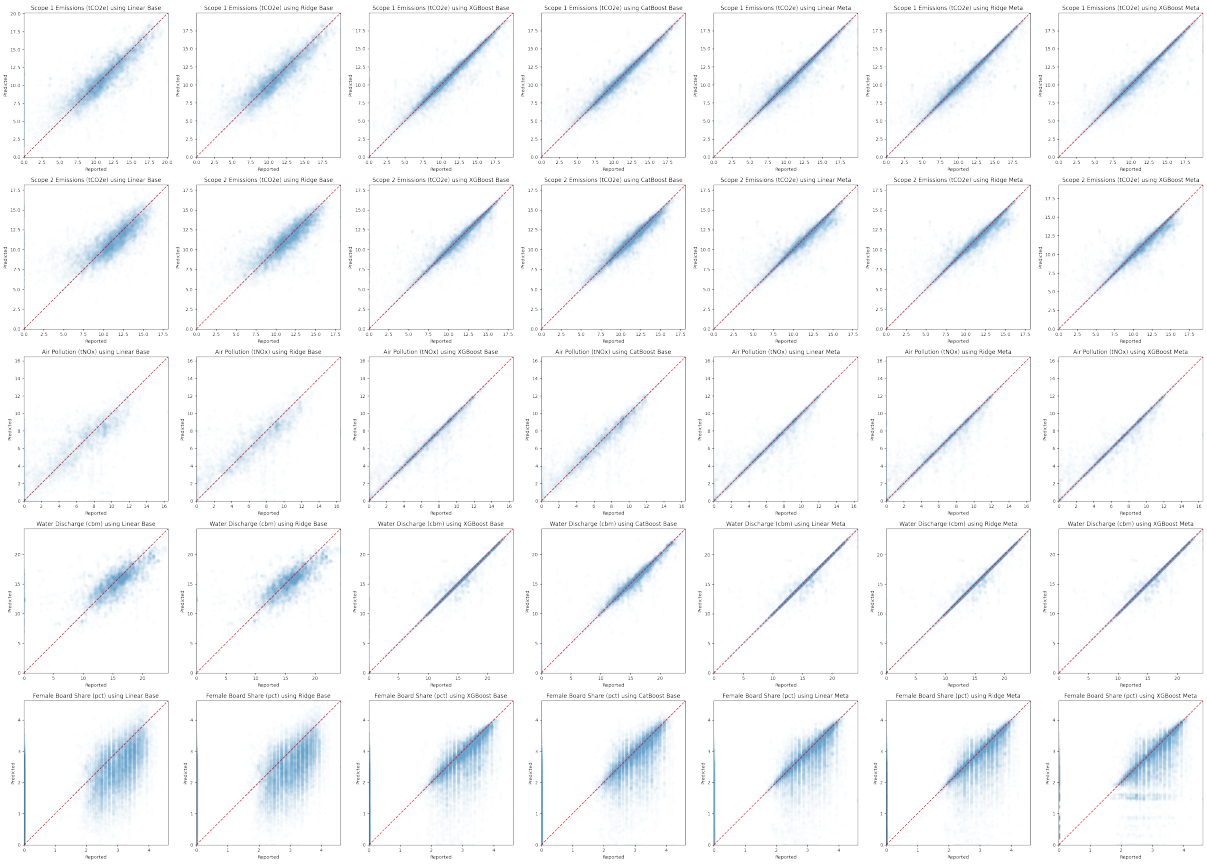
Wolpert, D.H., 1992. Stacked generalization. Neural Networks 5. doi:10.1016/S0893-6080(05)80023-1.

A. Additional Results

In this annex, we provide supplementary graphical representations of our model performance results.

A. Model Performance: Reported vs. Predicted Values

Figure 9 presents a comparison between reported and predicted values within our test data set for both baseline and meta-models in the five sustainability metrics. This comparison reaffirms that more complex models, such as XGBoost and CatBoost, as well as the meta-models, consistently outperform linear models like OLS and Ridge by a significant margin.

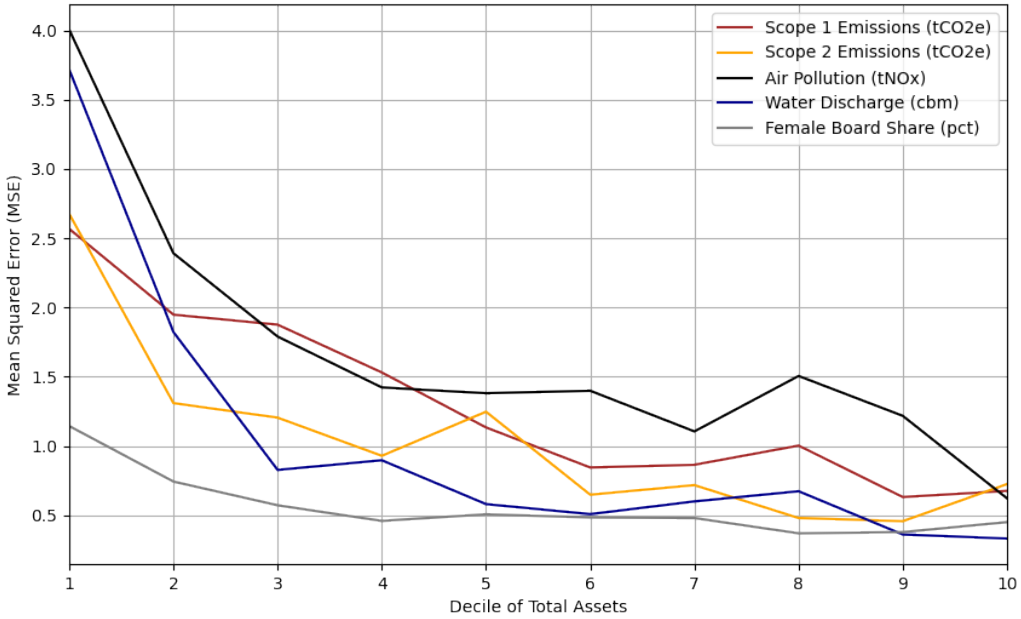


This figure displays the plots of the reported vs predicted variables for the five target variables vertically (Scope 1 emissions, Scope 2 emissions, air pollution, water discharge, and female board share) and the best performing model per learner horizontally (linear regression, ridge, CATBoost, XGBoost, meta-linear, meta-ridge, meta-XGBoost).

Figure 9. Plots reported vs predicted for best model per target variable and learner

B. Model Performance by Company Size

Figure 10 illustrates the variation in model performance relative to company size, represented by deciles of company size and the corresponding mean squared error (MSE) per decile. The results indicate that model performance is notably poorer in the lowest decile, which includes the smallest companies. Despite this, the performance remains superior compared to previous studies even in this decile. Following the initial decile, there is a sharp improvement in model performance, which then levels off with only minor variations around deciles 7 to 9.

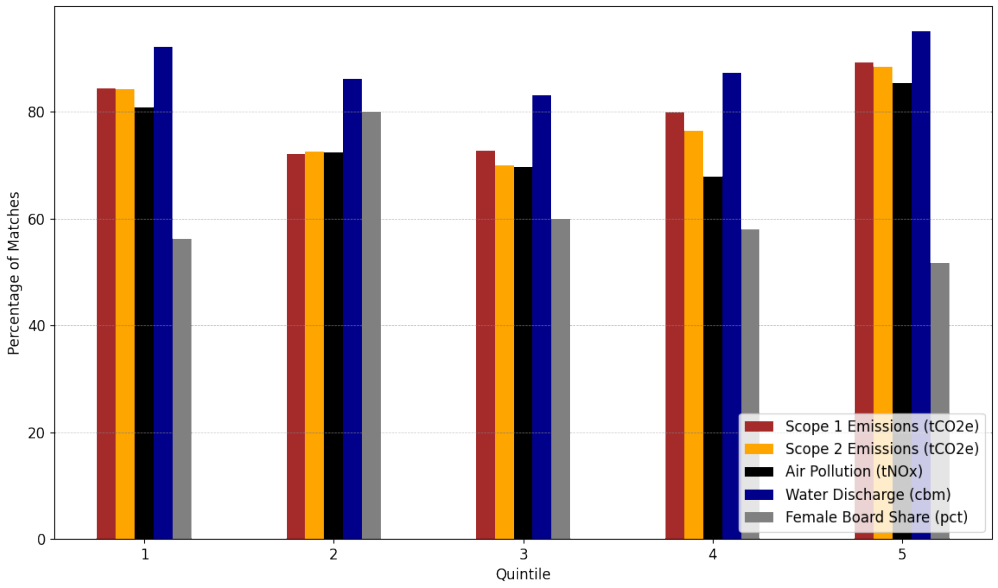


This figure displays the MSE of model predictions for company-size deciles for the the five target variables: Scope 1 emissions, Scope 2 emissions, air pollution, water discharge, and female board share.

Figure 10. Model Performance by Company Size

C. Quintile Analysis of Sustainability Metrics

Figure 11 presents the quintile results, also discussed in the main body of the paper, showing the percentage of matches for each sustainability metric within their respective quintile brackets. The primary finding is that the variation in model fit across quintiles is generally minimal for most sustainability metrics, with the exception of the female board share metric, where substantial variation is observed among the quintiles. This suggests that the predictive accuracy for the female board share is more sensitive to the distribution of data between different quintiles compared to other sustainability metrics.

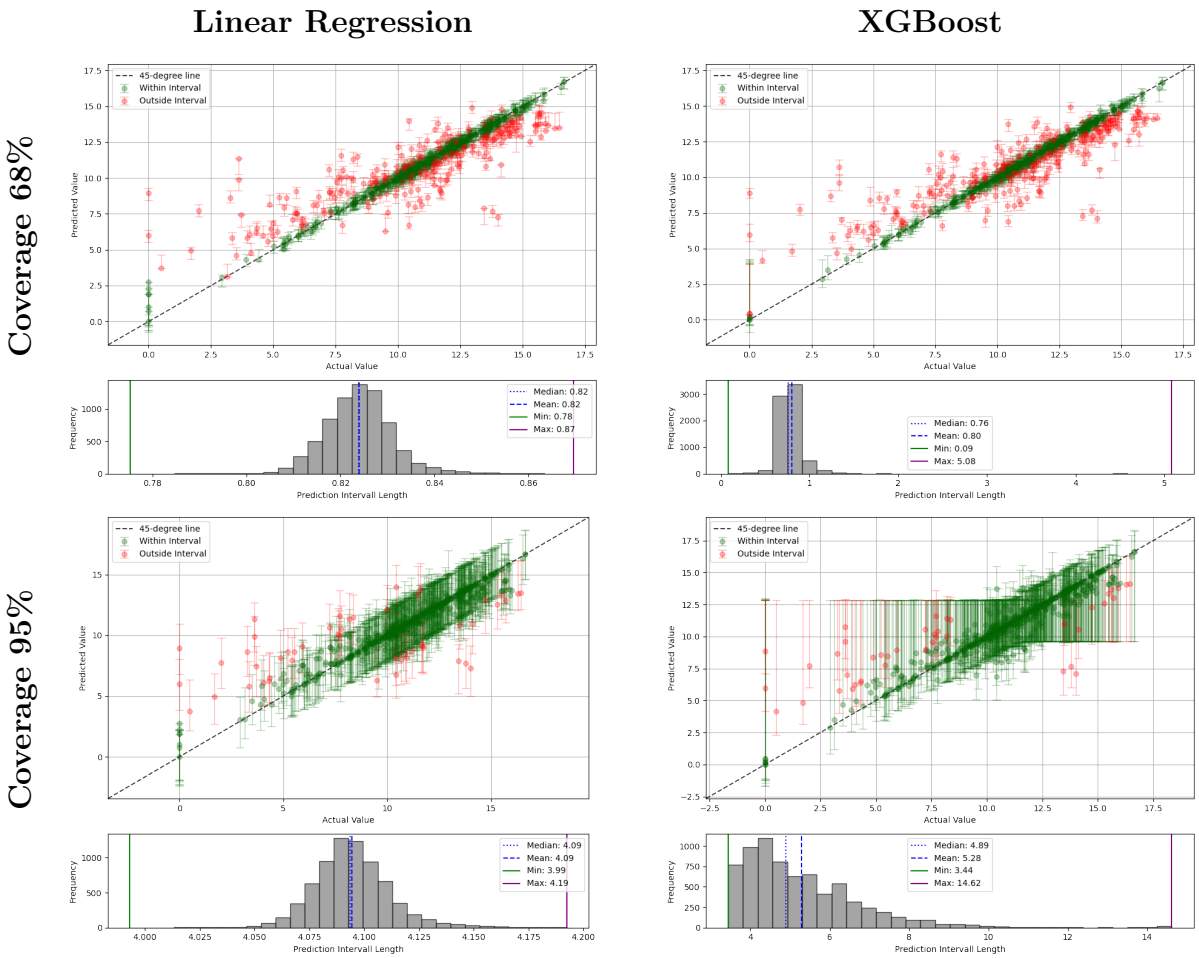


This figure displays the percentage of correctly allocated observations (predicted vs actual) per quintile for the five target variables: Scope 1 emissions, Scope 2 emissions, air pollution, water discharge, and female board share.

Figure 11. Percentage of Quintile Matches

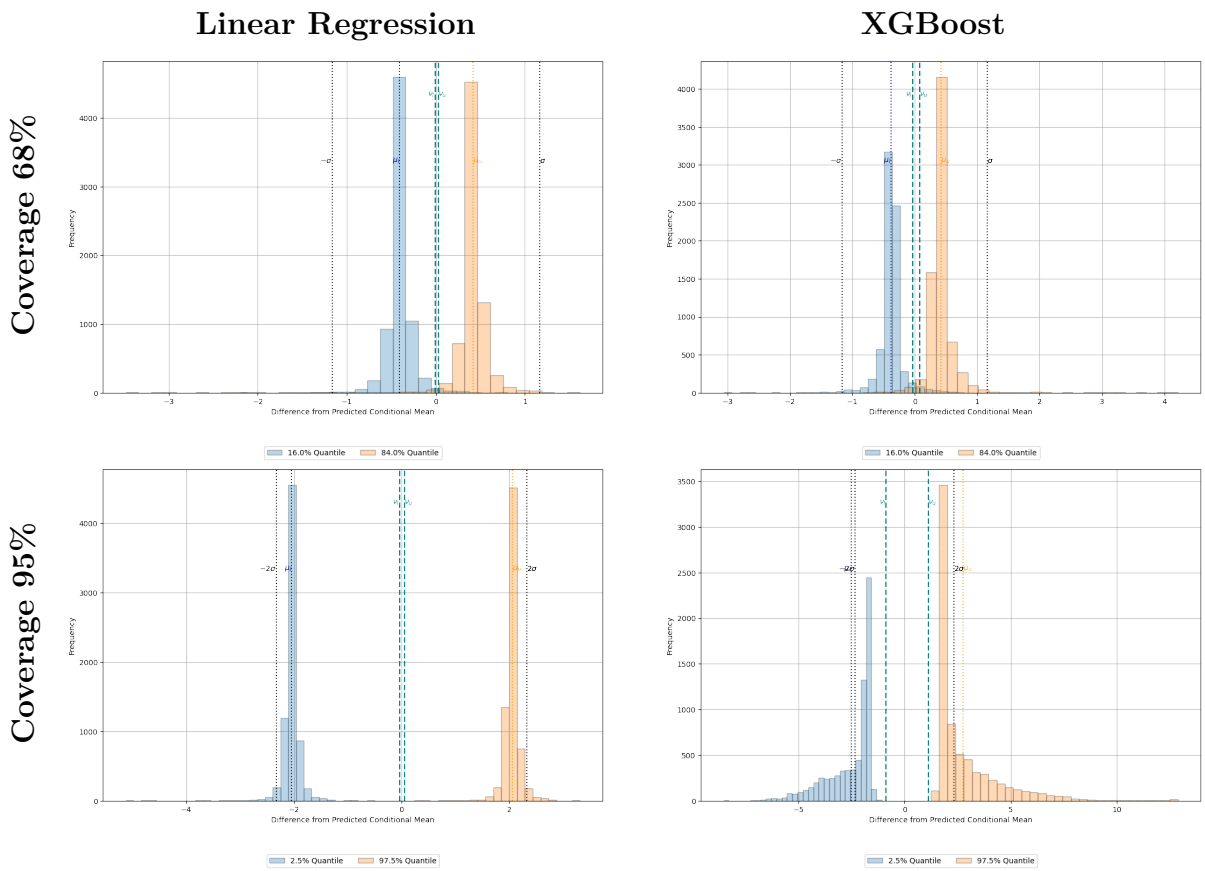
D. Prediction Uncertainty for Remaining Target Variables

Scope 2 Emissions



This figure displays conformalized prediction uncertainty for Scope 2 Emissions. The settings are for the learners linear regression and XGBoost and the coverage of 68% and 95%. For better readability, the figure shows 2% of total observations which are randomly selected.

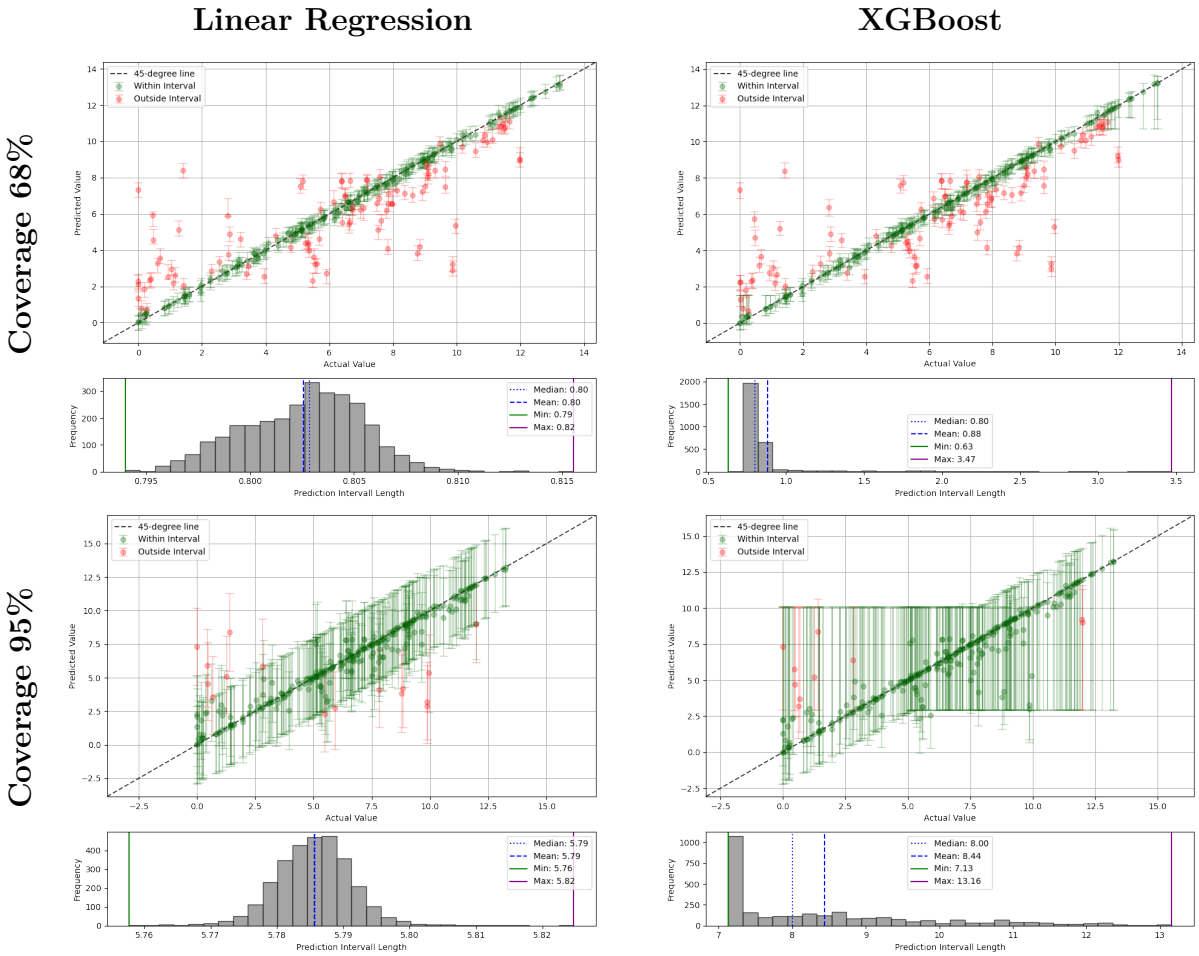
Figure 12.
Conformalized Prediction Uncertainty in Different Settings for Scope 2 Emissions



This figure displays the deviation for the conditional mean for Scope 2 Emissions in different settings. The settings are for uncertainty estimates based on linear regression and XGBoost and for targeted coverage rates of 68% and 95%. In all settings, the conditional mean is predicted by the best-performing XGB meta-model. The black dotted lines represent the prediction intervals using one and two standard deviations (σ), respectively. The green dashed lines represent the mean lower (upper) quantile prediction from standard quantile regression, ν_l (ν_u). Finally, the blue and orange bars show the distribution of quantile predictions for the lower and upper quantile from conformalized quantile regression. The mean of lower (upper) predictions is represented by the blue and orange dotted lines, μ_l (μ_u).

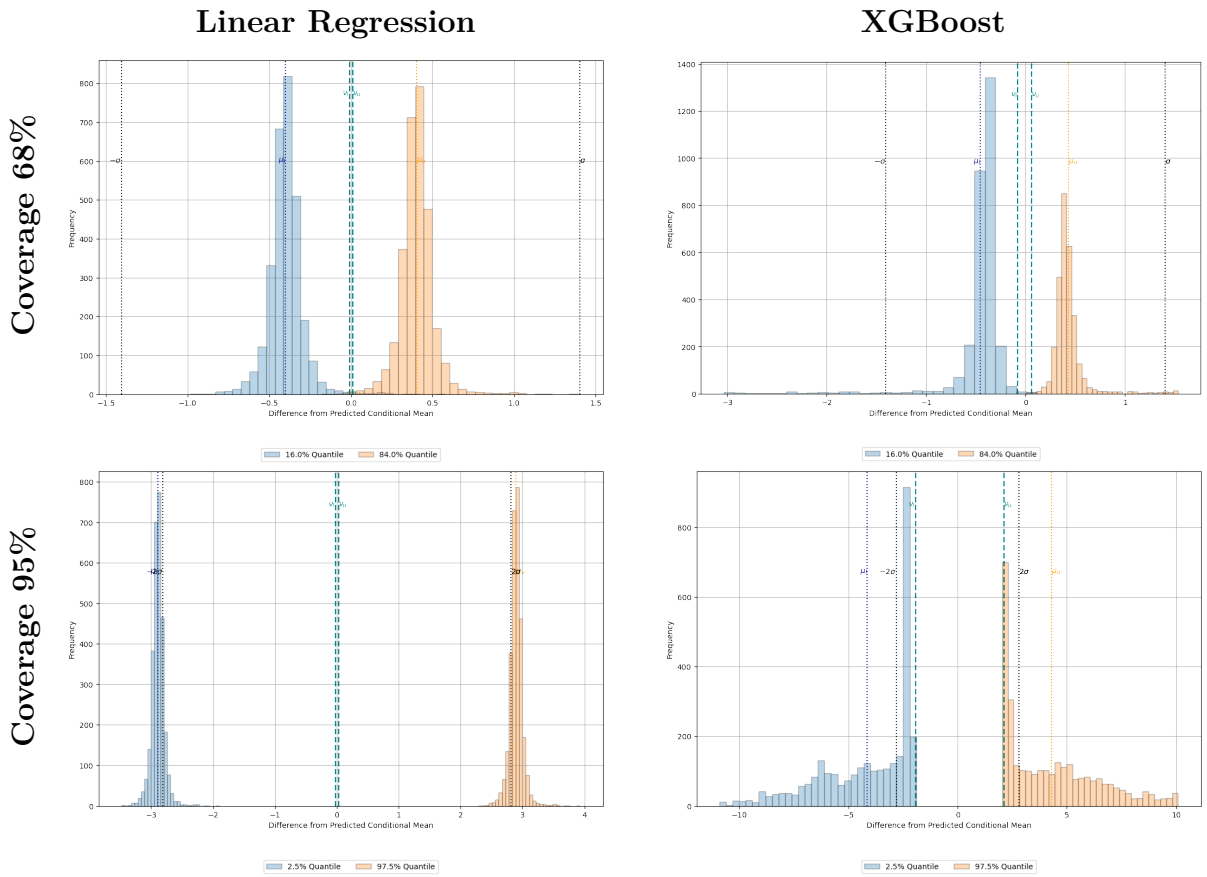
Figure 13. Deviation from Conditional Mean for Scope 2 Emissions

Air Pollution



This figure displays conformalized prediction uncertainty for NOx Emissions. The settings are for the learners linear regression and XGBoost and the coverage of 68% and 95%. For better readability, the figure shows 2% of total observations which are randomly selected.

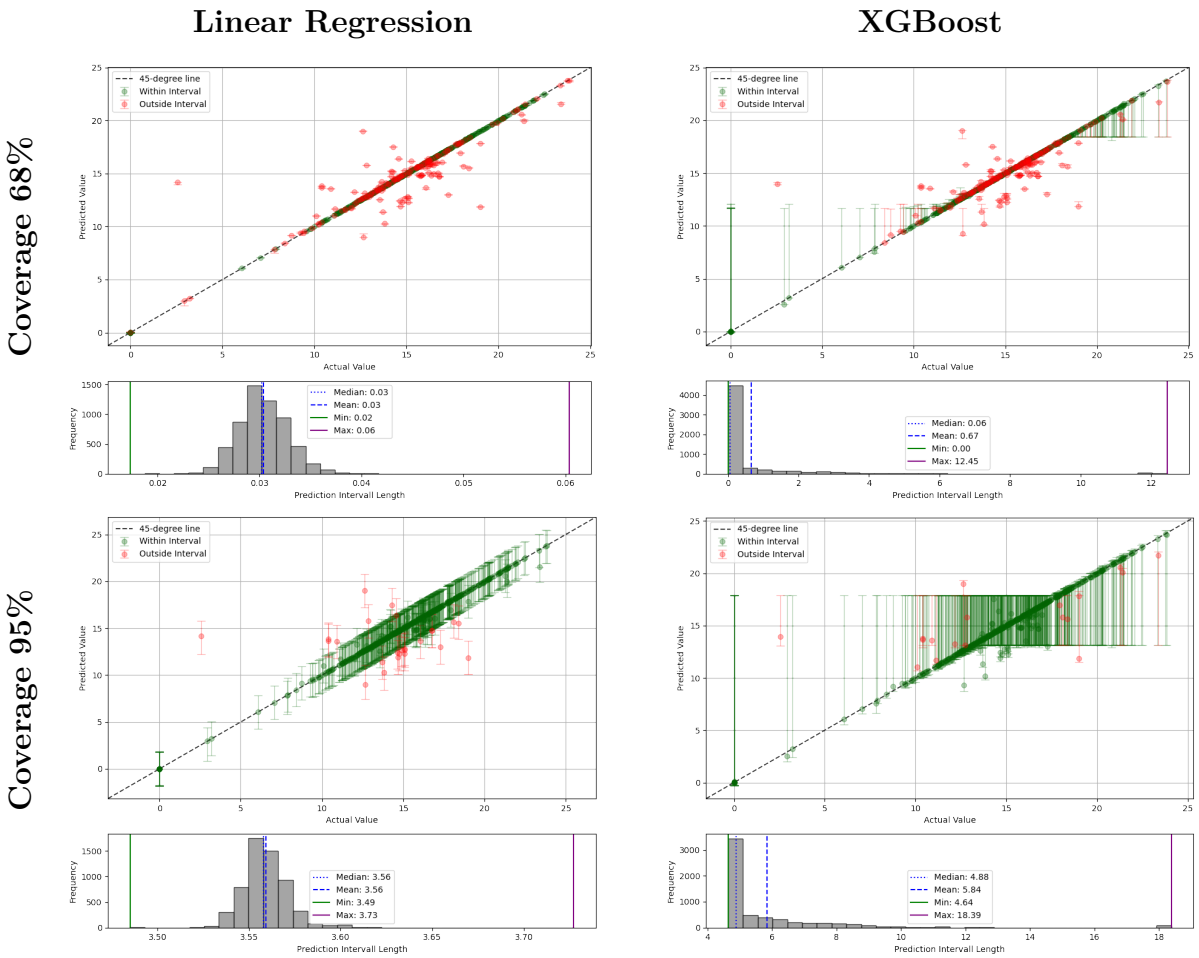
Figure 14.
Conformalized Prediction Uncertainty in Different Settings for NOx Emissions



This figure displays the deviation for the conditional mean for NOx Emissions in different settings. The settings are for uncertainty estimates based on linear regression and XGBoost and for targeted coverage rates of 68% and 95%. In all settings, the conditional mean is predicted by the best-performing XGB meta-model. The black dotted lines represent the prediction intervals using one and two standard deviations (σ), respectively. The green dashed lines represent the mean lower (upper) quantile prediction from standard quantile regression, ν_l (ν_u). Finally, the blue and orange bars show the distribution of quantile predictions for the lower and upper quantile from conformalized quantile regression. The mean of lower (upper) predictions is represented by the blue and orange dotted lines, μ_l (μ_u).

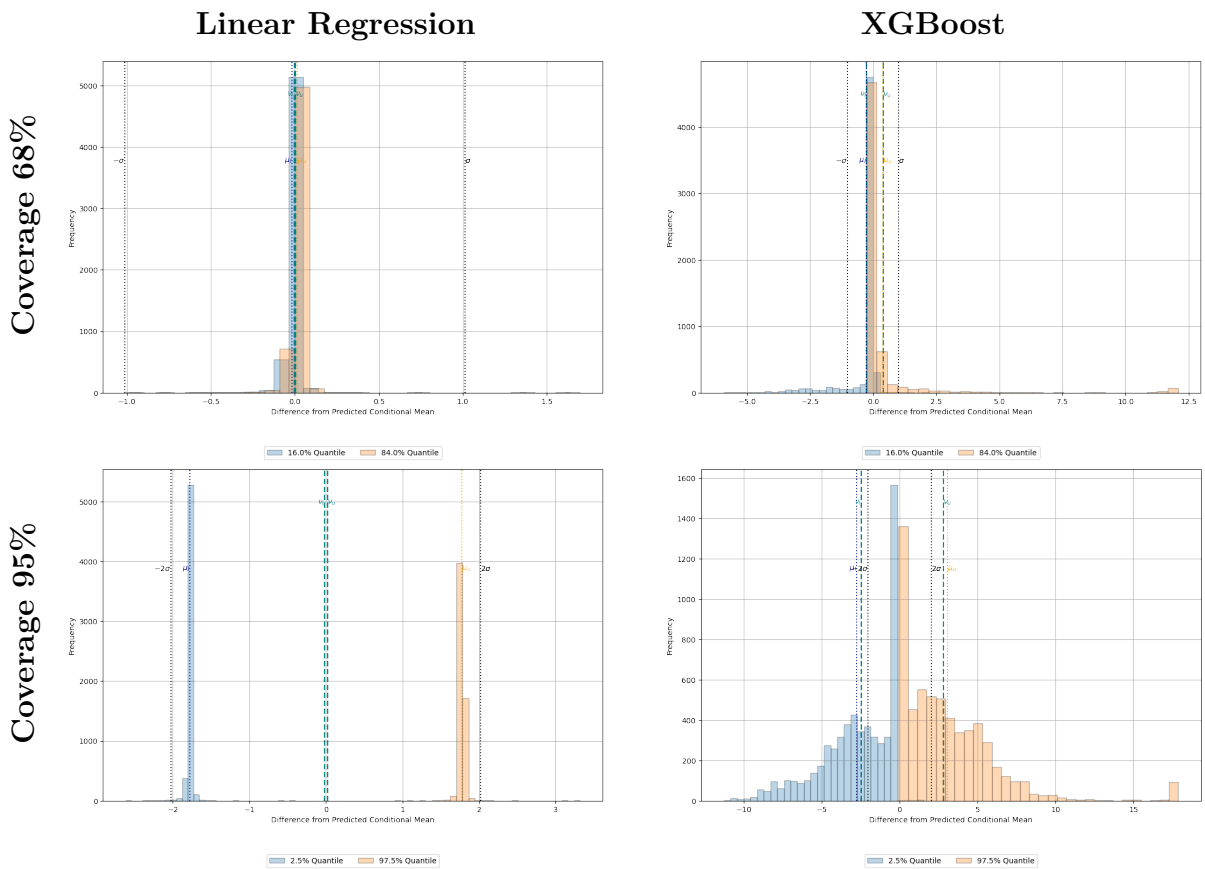
Figure 15. Deviation from Conditional Mean for NOx Emissions

Water Discharge



This figure displays conformalized prediction uncertainty for Water Discharge. The settings are for the learners linear regression and XGBoost and the coverage of 68% and 95%. For better readability, the figure shows 2% of total observations which are randomly selected.

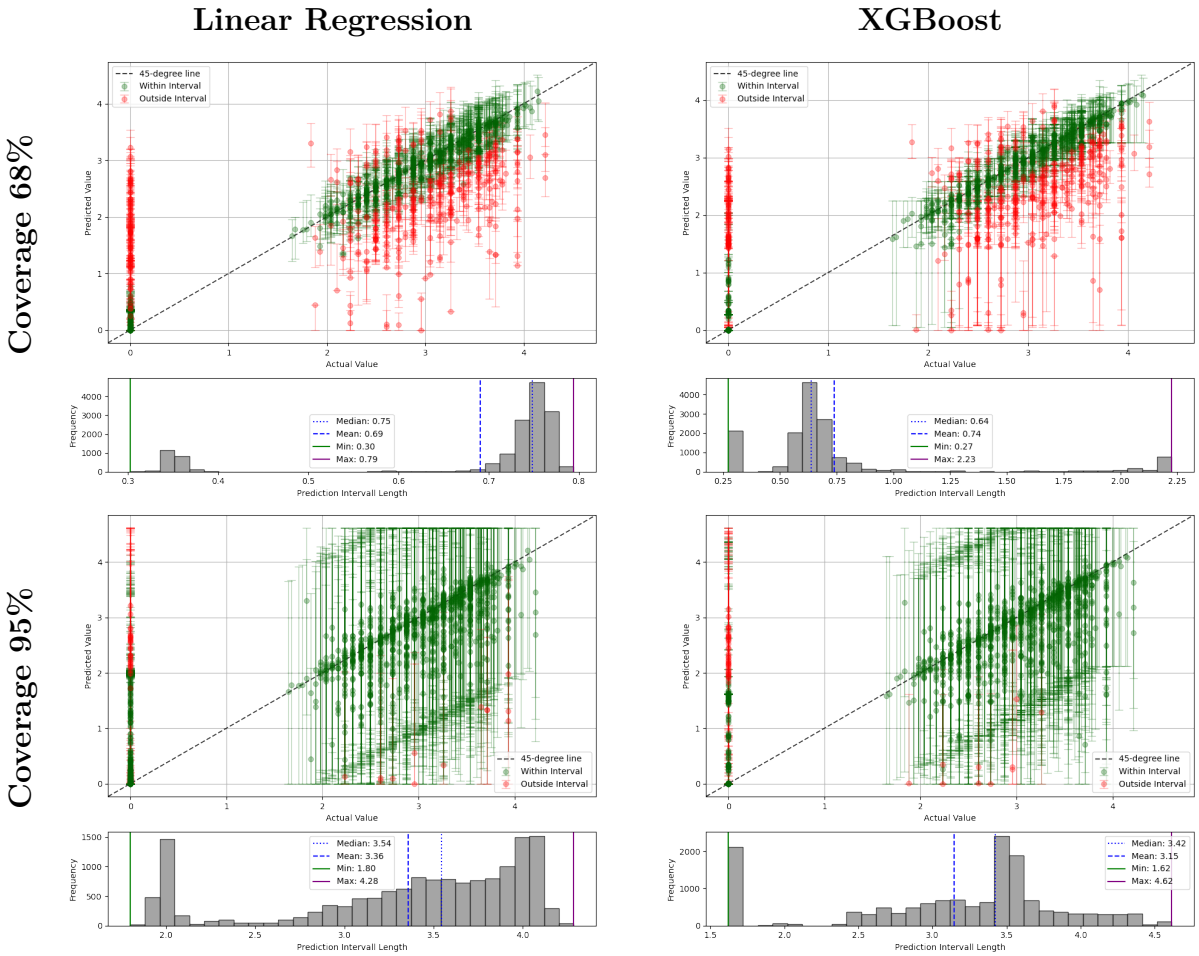
Figure 16.
Conformalized Prediction Uncertainty in Different Settings for Water Discharge



This figure displays the deviation for the conditional mean for Water Discharge in different settings. The settings are for uncertainty estimates based on linear regression and XGBoost and for targeted coverage rates of 68% and 95%. In all settings, the conditional mean is predicted by the best-performing XGB meta-model. The black dotted lines represent the prediction intervals using one and two standard deviations (σ), respectively. The green dashed lines represent the mean lower (upper) quantile prediction from standard quantile regression, ν_l (ν_u). Finally, the blue and orange bars show the distribution of quantile predictions for the lower and upper quantile from conformalized quantile regression. The mean of lower (upper) predictions is represented by the blue and orange dotted lines, μ_l (μ_u).

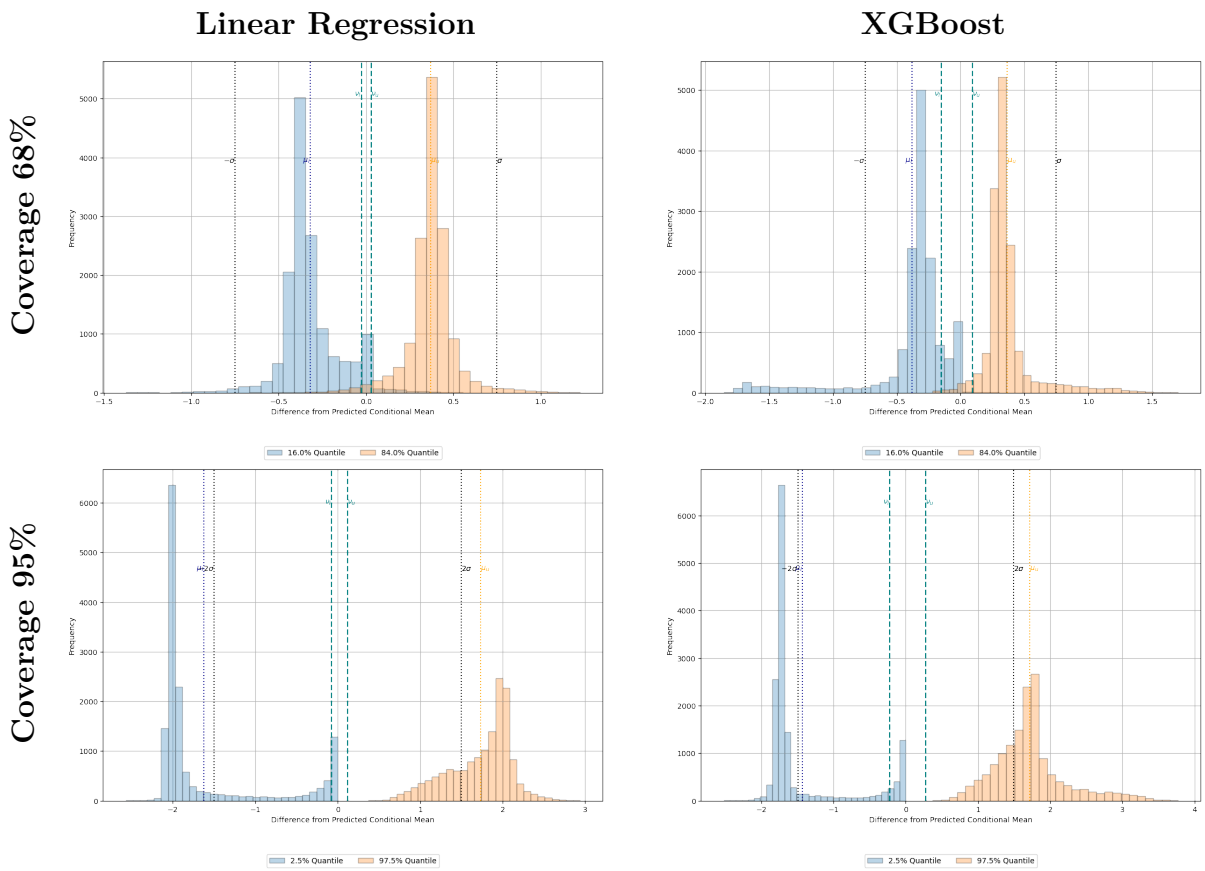
Figure 17. Deviation from Conditional Mean for Water Discharge

Female Board Share



This figure displays conformalized prediction uncertainty for Female Board Share. The settings are for the learners linear regression and XGBoost and the coverage of 68% and 95%. For better readability, the figure shows 2% of total observations which are randomly selected.

Figure 18.
Conformalized Prediction Uncertainty in Different Settings for Female Board Share



This figure displays the deviation for the conditional mean for Female Board Share in different settings. The settings are for uncertainty estimates based on linear regression and XGBoost and for targeted coverage rates of 68% and 95%. In all settings, the conditional mean is predicted by the best-performing XGB meta-model. The black dotted lines represent the prediction intervals using one and two standard deviations (σ), respectively. The green dashed lines represent the mean lower (upper) quantile prediction from standard quantile regression, ν_l (ν_u). Finally, the blue and orange bars show the distribution of quantile predictions for the lower and upper quantile from conformalized quantile regression. The mean of lower (upper) predictions is represented by the blue and orange dotted lines, μ_l (μ_u).

Figure 19. Deviation from Conditional Mean for Female Board Share

B. Early Stopping

In our efforts to improve computational efficiency during model training, we implemented an early stopping mechanism. This mechanism is designed to reduce unnecessary computations during the hyperparameter tuning process, specifically within the ten-fold cross-validation phase of each trial.Figure 20 visually represents the early stopping process.

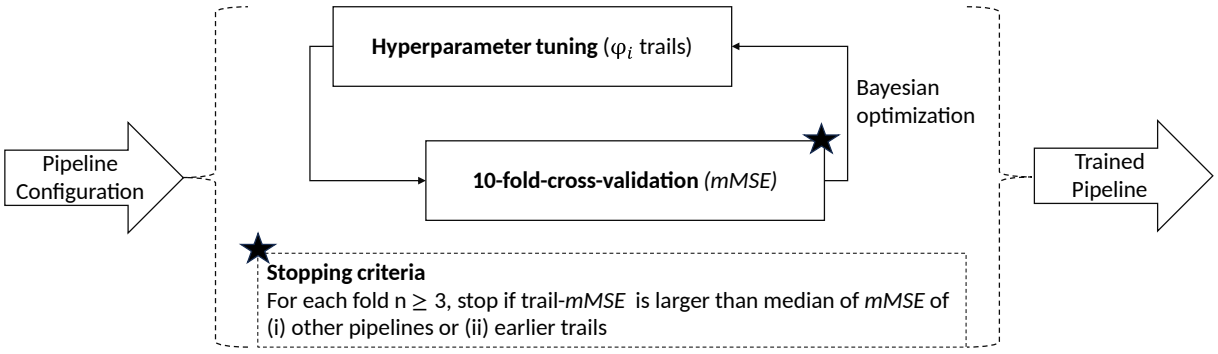
The early stopping approach is structured as follows:

1. Initialization of Trials: During the hyperparameter tuning process, each trial undergoes a tenfold cross-validation to evaluate the model's performance under different parameter settings.

2. Minimum Trial Requirement: To ensure that the model has sufficient opportunity to learn and stabilize, the early stopping mechanism is only considered after the first three folds of the cross-validation process have been completed.

3. Performance Comparison: After the third fold, the mechanism compares the mean squared error (MSE) obtained in the current fold with the median MSEs of previously evaluated models or earlier trials of the same model configuration.

4. Activation of Early Stopping: If the MSE of the current trial exceeds the average MSEs of the previous trials, the early stopping mechanism is triggered. When activated, this mechanism halts the remaining cross-validation folds for the current trial and proceeds directly to the next model configuration. This avoids further computation on a model configuration that is unlikely to outperform existing configurations.



This figure displays the early stopping mechanism that is applied throughout the code.

Figure 20. Early Stopping Approach

C. Computational Efficiency vs. Comprehensiveness Trade-off in Model Development

In this section, we elaborate on the trade-offs we encountered between computational efficiency and the comprehensiveness of our approach in model development.

Our approach to model configuration deliberately incorporates a trade-off between computational demands and the comprehensiveness of the analysis. By treating preprocessing steps as additional hyperparameters and employing a wide range of models and configurations, we place a significant burden on computational resources. This strategy, while resource-intensive, is aimed at enhancing the replicability of our results and ensuring that our methods are broadly applicable, independent of specific institutional contexts in which the data are generated.

The versatility offered by this approach can be valuable to the research community, as it allows for more generalized and adaptable modeling techniques.

However, this versatility comes at the cost of computational efficiency. Despite implementing early stopping mechanisms to reduce unnecessary calculations, the overall process remains resource-demanding. In the following, we provide an overview of the computational resources required for our approach:

- CPU: **2 x AMD EPYC Milan 7713 - 64-Core**
- GPU: **NVIDIA RTX Ada A6000**
- Time Consumption: **19.01 days** *(assuming serial execution)*
- Electricity Consumption: **87.94 kWh** *(estimation)*
- Early Stopping Effect: **active in 37.86% of 59,250 trials**

The above specifications highlight the scale of computational resources needed to execute our models effectively. The use of powerful CPUs and GPUs is essential to manage the large datasets and numerous trials involved in our comprehensive approach to discovering best performing models. Even with these resources, the time required to complete the model runs is considerable, emphasizing the trade-off between the depth of analysis and computational efficiency.

The introduction of early stopping mechanisms plays a crucial role in mitigating some of the computational burdens. As detailed in Annex B, early stopping is activated in approximately 30.02% of the 59,250 trials, significantly reducing the time and energy required to

train the models. This not only improves efficiency but also ensures that computational resources are allocated more effectively towards promising model configurations.

D. Insights and Recommendations for Future Researchers

In this annex, we provide guidance for future researchers who aim to replicate or build on our findings with reduced computational effort. By analyzing the performance of various model configurations and the factors contributing to the best results, we draw several conclusions that can inform more efficient model development in similar studies.

Table V summarizes the configurations of the best-performing models for each target variable based on the mean squared error (MSE) in the test data set. These best performing models reflect the combination of preprocessing steps and learner choices that yielded the most accurate predictions.

Target Variable	Missing Indicator		Imputation			Outlier Removal		Scaler		Transformation		Feature Selection	
	No	Yes	Mean	Median	Iterative	None	Winzorize	Standard	Robust	None	Quantile	None	Lasso
Scope 1 Emissions (tCO2e)	6	6	4	6	2	11	1	6	6	5	7	12	0
Scope 2 Emissions (tCO2e)	2	10	5	7	0	5	7	4	8	5	7	12	0
Air Pollution (tNOx)	4	7	5	6	0	8	3	1	10	8	3	11	0
Water Discharge (cbm)	3	9	8	2	2	10	2	4	8	7	5	12	0
Female Board Share (pct)	4	8	5	6	1	4	8	3	9	10	2	12	0

This table summarizes the best performing model setups along the dimensions missing indicator, imputation, outlier removal, scaler, transformation and feature selection for the five target variables: Scope 1 emissions, Scope 2 emissions, air pollution, water discharge, and female board share.

Table V. Summary of Winning Model Configurations

The analysis of the best performing models reveals several insights that can guide future research.

- Meta Learners and Complex Models: As stated in the main body of the paper, more complex models, such as meta learners, tend to outperform simpler linear models. Researchers should consider including these advanced techniques in their pipelines to achieve better performance.
- Pipeline Options: The study shows that there is no universal best configuration across all preprocessing steps. This suggests that researchers working with different data sets or under different use cases may need to experiment with the full range of pipeline options, rather than relying on a predetermined set of configurations. For example, there is no clear trend favoring a particular missing indicator or scaling method across all target variables.
- Imputation and Transformation Methods: While there are slight trends in the performance of certain imputation and transformation methods, these trends are not strong enough to recommend a specific approach universally. Researchers should evaluate the effectiveness of these methods on a case-by-case basis, as different variables may respond differently to the same preprocessing techniques.

- Feature Selection: The analysis indicates that, in our case, model configurations without feature selection generally performed better. This finding suggests that including more features, rather than reducing them, can enhance model performance. Future researchers should be cautious about overly aggressive feature selection, particularly when working with rich datasets, as it might lead to the loss of valuable information.

In addition to these findings on pipeline buliding, researchers might want to implement early stopping mechanisms, as discussed in Annex B, to reduce computational burden while ensuring that only the most promising configurations are fully explored.

E. Summary Statistics Predictors

Table VI. Summary statistics of predictor variables

The table presents summary statistics for the predictor variables. These statistics include the count of observations, the mean, standard deviation, minimum, maximum, and percentiles (25%, 50%, 75%) for each variable. The units for each variable are indicated in the "Unit" column, with common abbreviations such as B for billions, M for millions, and K for thousands.

	count	mean	std	min	25%	50%	75%	max	Unit
Accounts Payable - Long Term	18316	38.90	467.62	-7.88	0.01	0.11	0.93	18479.00	B
Avg. Payables Payment Days	80527	303.08	23949.38	-32528.25	35.47	56.43	93.25	5769517.74	Unit
Brands, Patents, Trademarks, Marketing (Gross)	27924	8.91	112.41	-0.02	0.01	0.08	0.49	5118.16	B
Capital Expenditures (Total)	101362	79.14	1009.41	-4.17	0.05	0.34	2.94	75161.53	B
Cash & Cash Equivalents	94014	155.92	1722.84	-7.43	0.10	0.66	6.18	110763.21	B
Cash & Cash Equivalents (Total)	103303	223.49	2980.01	-7.43	0.11	0.79	8.87	202104.93	B
Cash & Short-Term Investments	98451	222.83	2579.73	-0.01	0.15	0.95	9.98	124652.84	B
Computer Software (Net)	32490	11.22	113.29	-527.61	0.01	0.07	0.59	5744.00	B
Long Term Debt Issued (Cash Flow)	59479	193.77	2415.98	-111.28	0.08	0.85	8.18	145183.14	B
Long & Short Term Debt Issued (Cash Flow)	19552	1839.72	48455.29	-24.76	0.10	0.95	5.65	1787877.77	B
Short Term Debt Issued (Cash Flow)	11008	471.55	4511.91	-12.11	0.02	0.51	17.66	192778.56	B
Long Term Debt (Total)	96939	342.28	3355.23	-1302.34	0.23	1.95	15.46	183775.07	B
Total Debt	99726	556.41	4952.84	-0.12	0.35	2.78	24.86	205362.30	B
Depreciation & Amortization	13845	30.52	442.09	-6.14	0.03	0.18	1.05	14863.00	B
EBIT	104120	108.47	1348.85	-30329.63	0.08	0.66	5.65	60569.45	B
Part-Time Employees	3661	7.23	23.04	0.00	0.03	0.35	3.59	444.55	K
Equity Earnings/Loss (Pre-Tax, Nonrecurring)	41519	12.19	157.91	-2483.34	-0.00	0.01	0.24	7087.00	B
Short-Term Financial Assets	18291	203.63	2887.98	-738.25	0.01	0.17	1.93	92441.70	B
Net Financing Income/Expense	92886	-6.24	128.64	-10670.09	-0.24	-0.02	0.00	4740.00	B
Goodwill (Gross)	22899	46.16	313.86	-0.40	0.14	0.75	4.26	6958.30	B
Impairment - Financial Investments	14263	6.74	82.03	-459.28	0.00	0.01	0.15	2242.53	B
Impairment - Fixed Assets	50314	4.92	83.93	-1709.81	0.00	0.03	0.39	11072.70	B

Continued on next page

	count	mean	std	min	25%	50%	75%	max	Unit
Income Taxes	101788	28.05	340.50	-1855.99	0.01	0.12	1.23	16990.40	B
Intangible Assets (Accum. Amort. & Impair.)	34239	49.82	424.81	-52.59	0.04	0.26	1.55	18254.57	B
Intangible Assets (Gross)	34632	122.90	907.03	-60.63	0.17	1.08	6.80	37612.29	B
Intangible Assets (Net Cash Flow)	4464	18.25	193.18	-136.28	0.00	0.01	0.11	5615.31	B
Long-Term Investments	66338	105.56	1678.63	-1246.24	0.03	0.52	8.03	126930.42	B
Total Investments	85989	618.23	6893.10	-230.94	0.09	1.27	25.76	442925.68	B
Lending & Long-Term Deposits	7493	714.66	6413.50	-42.61	0.73	10.50	53.20	242431.95	B
Short-Term Loans & Receivables (Net)	40189	267.76	2113.23	-5.18	0.33	1.68	11.65	97072.45	B
Total Loans & Receivables	102232	1126.53	21694.84	-38.63	0.20	1.44	18.53	2531993.14	B
Net Cash Flow - Financing	103774	11.49	961.89	-27753.00	-0.97	-0.05	0.16	121530.63	B
Net Cash Flow - Investing	103526	-115.48	1569.11	-132477.05	-4.01	-0.40	-0.04	25880.94	B
Net Cash Flow - Operating	103874	128.39	1942.81	-80142.33	0.07	0.60	5.42	125791.99	B
Net Financial Income/Expense (Other)	17440	-1.20	29.88	-1517.46	-0.06	-0.01	-0.00	299.66	B
Net Income (After Tax)	51281	101.89	1246.50	-40408.49	0.11	0.86	6.74	44344.86	B
Nonrecurring Income/Expense	86092	-1.47	174.77	-10939.90	-0.24	-0.01	0.00	29391.89	B
Total Operating Expenses	99953	760.55	6541.08	-1143.05	0.68	4.23	31.37	227970.94	B
Total Other Assets	100956	122.21	1323.71	-41897.01	0.04	0.42	5.69	69247.18	B
Total Other Current Assets	30798	29.04	260.52	-14527.45	0.03	0.30	4.92	13300.06	B
Total Other Current Liabilities	89942	102.06	1122.25	-12625.25	0.04	0.40	4.97	150865.89	B
Total Other Noncurrent Liabilities	42108	115.75	2088.40	-1290.92	0.09	0.68	7.05	145414.99	B
Payables & Accrued Expenses	100432	131.24	1145.48	-0.69	0.14	0.94	8.45	77457.92	B
Plant, Machinery & Equipment (Gross)	73124	525.67	6823.00	-105.61	0.15	1.20	11.25	305445.00	B
PPE (Accum. Depreciation & Impairment)	88596	399.03	4681.38	-10168.69	0.16	1.27	11.25	227543.45	B
PPE (Gross)	92792	855.02	8895.91	-4920.35	0.46	3.59	33.30	377471.99	B
PPE (Net)	100907	437.36	4442.67	-991.57	0.20	1.84	17.45	207052.67	B
Other PPE (Gross)	65438	120.92	1761.60	-3128.77	0.03	0.31	3.70	169436.18	B
Total Provisions	89472	79.50	778.60	-166.19	0.06	0.49	4.88	44491.12	B
Long-Term Receivables & Loans	40536	100.80	1385.81	-31.92	0.01	0.14	1.48	46311.56	B
R&D Expense	32492	40.82	630.29	-136.86	0.04	0.20	1.99	22401.73	B
R&D Costs (Gross)	10642	75.39	592.23	-0.00	0.02	0.09	0.73	10374.45	B

Continued on next page

	count	mean	std	min	25%	50%	75%	max	Unit
Short-Term Debt & Notes Payable	54238	267.85	2309.18	-168.57	0.06	1.15	20.76	104779.30	B
Total Short-Term Investments	52532	131.21	1936.84	-3.40	0.03	0.33	3.48	95270.26	B
Total Assets	103976	2732.62	30953.65	0.00	2.05	12.61	137.79	1808429.76	B
Total Book Capital	103954	1154.24	10028.24	-66.81	1.29	7.70	73.41	442390.86	B
Total Current Assets	89457	505.03	4941.26	-57.13	0.50	3.03	26.51	283116.65	B
Total Current Liabilities	44789	516.29	3918.81	0.00	0.80	4.49	40.21	298411.51	B
Total Fixed Assets (Net)	82720	737.86	6722.46	-17420.67	0.59	3.86	31.22	283593.15	B
Total Liabilities	103952	2045.66	26076.03	-20.44	0.99	6.76	75.56	1700262.11	B
Total Shareholders' Equity	99822	689.10	7064.58	-11055.62	0.75	4.17	34.21	304899.93	B
Working Capital Change (Cash Flow)	98107	-26.13	623.51	-31484.37	-0.38	-0.01	0.08	35264.41	B
Avg. Receivables Collection Period (Days)	43635	95.47	919.42	-17463.43	41.87	63.48	90.74	152776.77	Unit
Full-Time Employees	79275	323.43	52542.63	0.00	1.68	6.93	23.05	10785002.89	K
R&D as % of Revenues	43026	287.33	8230.36	-34.17	0.68	2.58	7.06	598611.42	Unit
Turnover	62282	2.12	108.80	-17573.65	0.00	0.02	0.23	19941.57	B
Date of Incorporation	105929	1.98	0.03	1.81	1.97	1.99	2.00	2.02	K
Year	109085	2.01	0.01	2.00	2.01	2.02	2.02	2.02	K
Parent Shareholders' Equity (Total)	4123	482.85	1871.17	-63.52	6.26	55.35	326.36	37441.42	B
Business Financing Revenue (Other)	4278	45.33	542.50	-8261.00	0.00	0.04	1.33	11075.05	B
Inventories (Finished Goods)	22867	48.84	497.25	-4.83	0.03	0.22	1.57	17537.84	B
Long-Term Loans	9204	62.72	938.79	-0.00	0.01	0.19	1.91	28704.25	B
Short-Term Loans	10735	33.42	370.27	-0.00	0.01	0.22	1.96	21022.04	B
Total Noncurrent Liabilities	41426	202.81	2583.53	-99.82	0.10	0.85	6.30	131277.39	B

It’s in the Financials, Stupid! But is it certain?

New Insights from Financial Statements

Christian Haas, Ulf Moslener & Sebastian Rink

October 2024

Frankfurt School of Finance & Management

Motivation

- Company-level sustainability data are **complex and non-linear**
- Machine Learning is only applied to GHG emissions so far
- **Increasing amounts of sustainability data** reported by companies

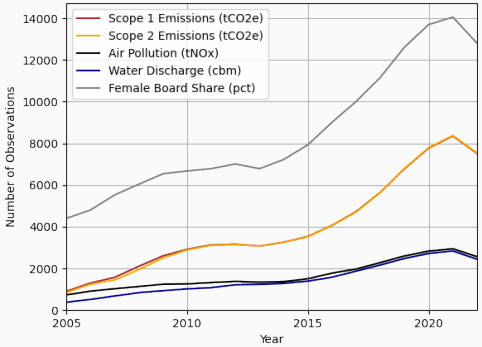


Figure 1: Reported Sustainability Data by Companies

Research Questions

1. To what extent can we **derive corporate sustainability data from corporate financial data only** using ML?
2. How does the prediction performance **change for different dimensions**?
3. **How certain are the point estimates** of the prediction models?

Methodology & Data

Target Variables

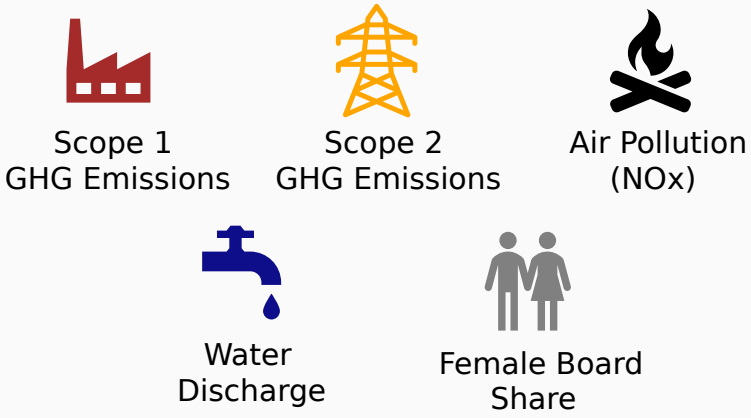


Figure 2: Target Variables

Training Approach for Point Estimates

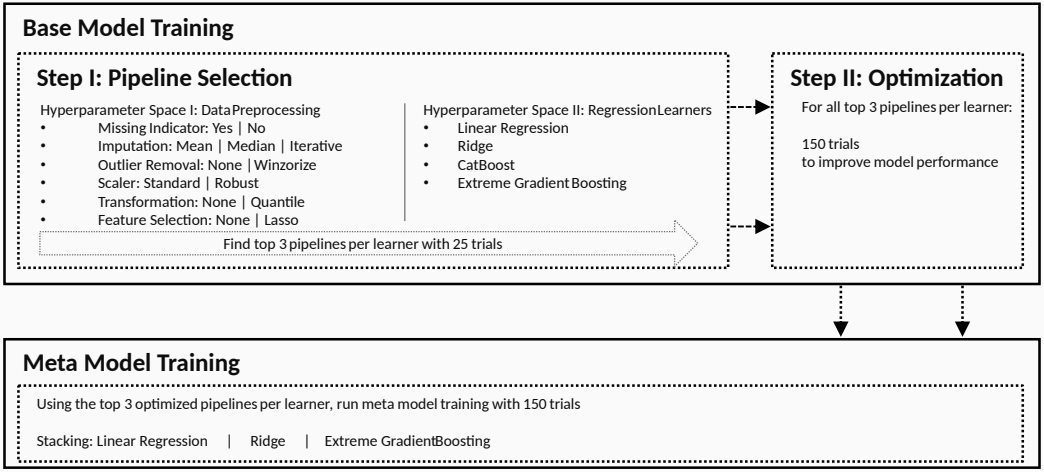


Figure 3: Point Estimate Training Approach using MSE

Stacking Stopping

Uncertainty Quantification

1. Quantile Regression:
   $L_\alpha(\hat{y}_\alpha, y) = (y - \hat{y}_\alpha)\,\alpha \mathbb{1}\{y > \hat{y}_\alpha\} + (\hat{y}_\alpha - y)\,(1-\alpha)\,\mathbb{1}\{y \leq \hat{y}_\alpha\}$
2. Quantile Prediction:
   $\hat{y}_\alpha(x)$
3. Conformal Scores:
   $c(x, y) = \max\{\hat{y}_{\tau/2}(x) - y, y - \hat{y}_{1-\tau/2}(x)\}$
4. Rectifying Quantiles for Intervals:
   $I(x) = \left[\hat{y}_{\tau/2}(x) - \hat{r}, \hat{y}_{1-\tau/2}(x) + \hat{r}\right]$, where
   $\hat{r} = \text{Quantile}\left(\frac{\lceil (n_{cal}+1)(1-\alpha) \rceil}{n_{cal}}, \{c_1, ..., c_{n_{cal}}\}\right)$

Dataset

Dataset	Scope 1 Emissions	Scope 2 Emissions	Air Pollution	Water Discharge	Female Board Share
General Information					
Number of observations	47685	47320	20980	18426	108834
Number of sectors	83	83	76	74	86
Number of countries	83	83	63	63	95
Number of companies	8391	8335	3389	3098	14406
Start year	2005	2005	2005	2005	2005
End year	2022	2022	2022	2022	2022
Number of predictor variables	240	240	213	211	255
Data completeness (in %)	63.86	63.77	65.72	65.62	57.92
Target Variable Information					
Log (1+value) Mean	10.75	11.07	6.31	14.98	2.17
Log (1+value) Std	3.55	2.72	3.28	3.60	1.40
Log (1+value) Min	0.00	0.00	0.00	0.00	0.00
Log (1+value) Max	22.21	22.72	16.46	24.01	4.62

Table 1: Summary Statistics

Results & Discussion

It's in the financials, stupid!

Target Variable	Metric	Base Learner				Meta Learner		
		Linear Regression	Ridge	CatBoost	XGBoost	Linear Regression	Ridge	XGBoost
Scope 1 Emissions (tCO2e)	MAE	1.261	1.240	0.648	0.662	0.610	0.609	0.589
	MSE	3.401	3.252	1.579	1.617	1.622	1.622	1.589
	R2	0.713	0.726	0.867	0.864	0.863	0.863	0.866
Scope 2 Emissions (tCO2e)	MAE	1.252	1.118	0.513	0.627	0.556	0.556	0.527
	MSE	3.400	2.959	1.163	1.269	1.275	1.275	1.204
	R2	0.540	0.599	0.842	0.828	0.827	0.827	0.837
Air Pollution (tNOx)	MAE	1.501	1.444	0.695	0.905	0.672	0.672	0.650
	MSE	4.598	4.320	1.923	2.262	1.907	1.907	1.909
	R2	0.564	0.590	0.818	0.786	0.819	0.819	0.819
Water Discharge (cbm)	MAE	1.782	1.782	0.278	0.623	0.275	0.275	0.273
	MSE	7.528	7.522	1.055	1.526	1.063	1.063	1.053
	R2	0.443	0.443	0.922	0.887	0.921	0.921	0.922
Female Board Share (pct)	MAE	0.876	0.876	0.473	0.501	0.448	0.448	0.428
	MSE	1.253	1.249	0.598	0.638	0.595	0.595	0.592
	R2	0.347	0.349	0.688	0.667	0.690	0.690	0.691

Table 2: Global Model Performance

Variation by Quintile

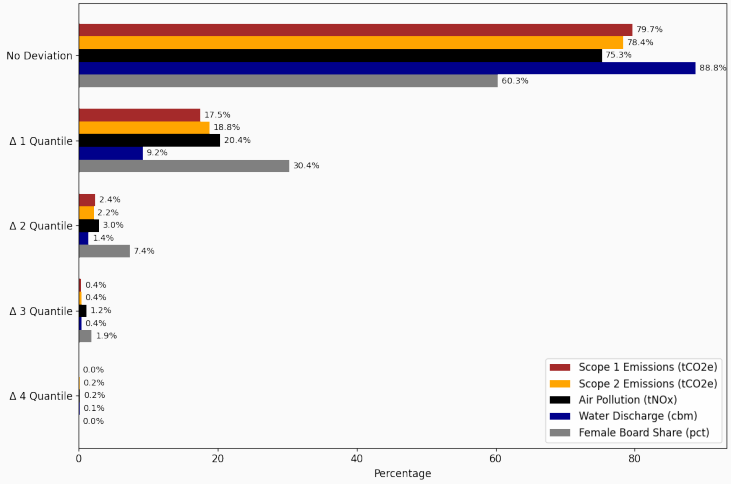


Figure 4: Number of Deviating Quintiles

Temporal Variation

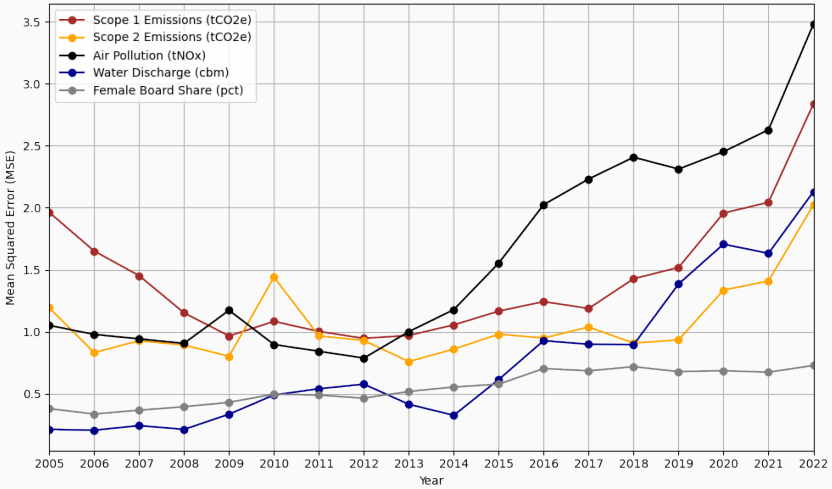


Figure 5: Temporal Model Performance

Spatial Variation

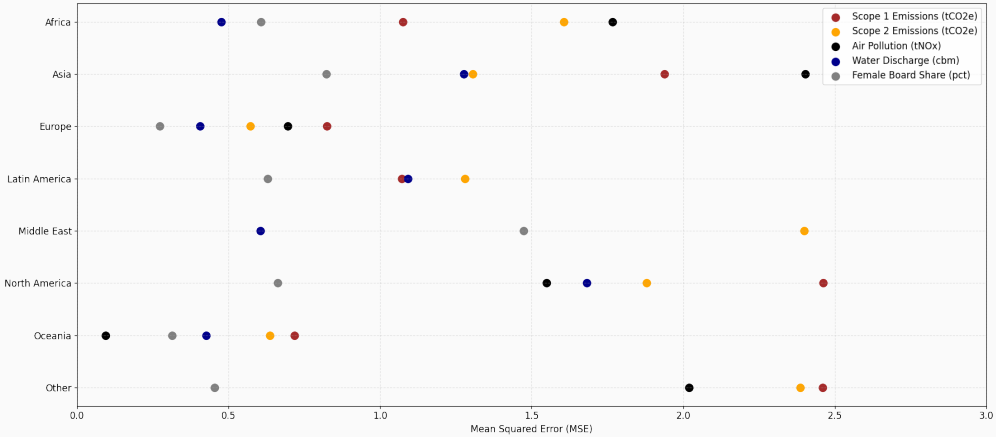


Figure 6: Spatial Model Performance

Sectoral Variation

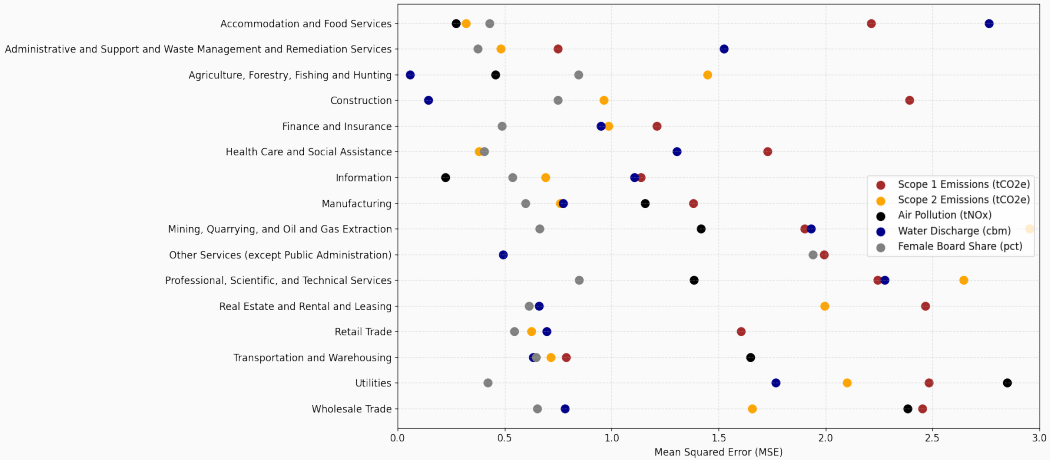


Figure 7: Sectoral Model Performance

Beware of Prediction Uncertainty!

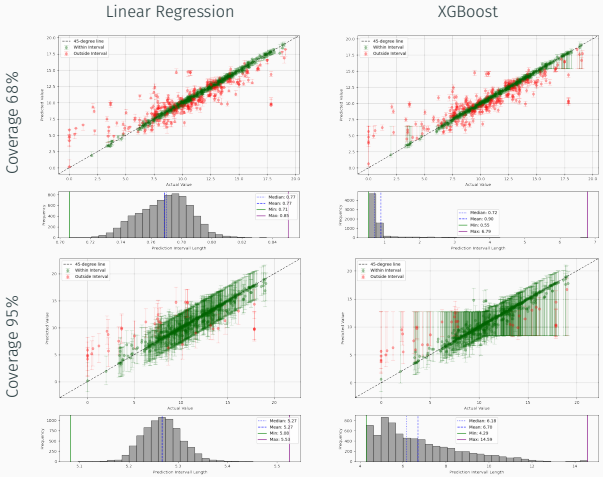


Figure 8: Prediction Uncertainty in Different Settings for Scope 1 Emissions

Reflect On Uncertainty Measure

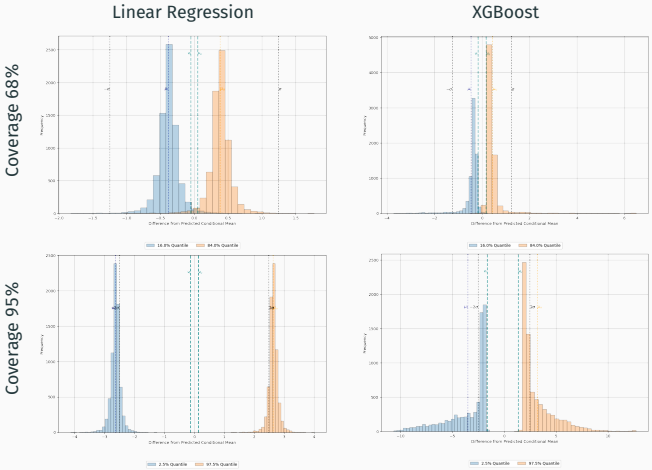


Figure 9: Deviation from Conditional Mean for Scope 1 Emissions

Conclusion

- **It's actually in the financials**, stupid!
- **Beware of prediciton variation and uncertainty!**
- Policy makers may take a more open stance in ML in sustainable finance, but the exact area of application matters
- Future research: local models & little reported sustainability metrics

Takeaways For Policy Makers

1. Machine learning can support financial institutions and companies in assessing sustainability risks and impacts with a high degree of predictive performance. As such, these **institutions should be allowed to use ML to predict sustainability data** as a supplement to the available reported data, especially if the costs of accessing the raw data are high.
2. Users of ML-predicted sustainability data should be required to **increase transparency on the quality of the predictions**, not only at the global level, but also in the dimensions time, space, and sector.
3. **Prediction uncertainty should be considered** by users of ML-predicted sustainability data and the respective assumptions / risk appetite should be made transparent.

References

Ismail, Shereen, Zakaria El Mrabet, and Hassan Reza (Dec. 2022). **“An Ensemble-Based Machine Learning Approach for Cyber-Attacks Detection in Wireless Sensor Networks”.** In: *Applied Sciences* 13.1, p. 30. ISSN: 2076-3417. DOI: *10.3390/app13010030*.

Appendix

Early Stopping

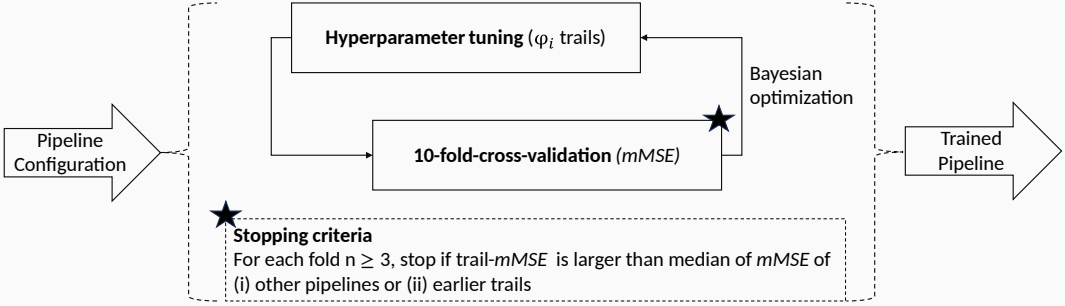


Figure 10: Early Stopping Approach

Efficiency

- CPU: **2 x AMD EPYC Milan 7713 - 64-Core**
- GPU: **NVIDIA RTX Ada A6000**
- Time Consumption: **19.01 days** *(assuming serial execution)*
- Electricity Consumption: **87.94 kWh** *(estimation)*
- Early Stopping Effect: **active in 37.86% of 59,250 trials**

Boosting, Bagging, Stacking

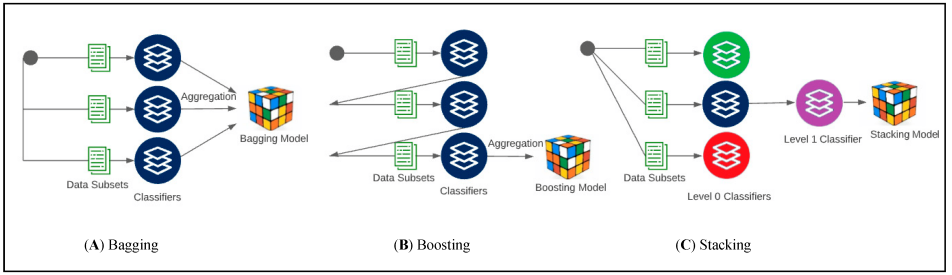


Figure 11: Approach to Boosting, Bagging, and Stacking by Ismail, El Mrabet, and Reza 2022