

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

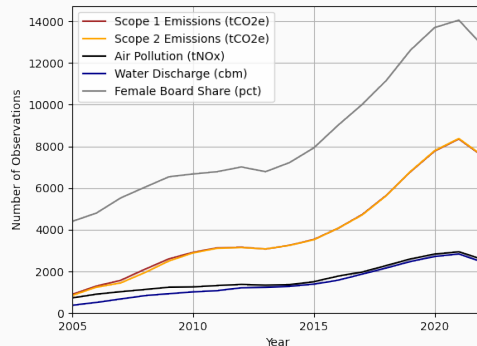


Figure 1: Reported Sustainability Data by Companies

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

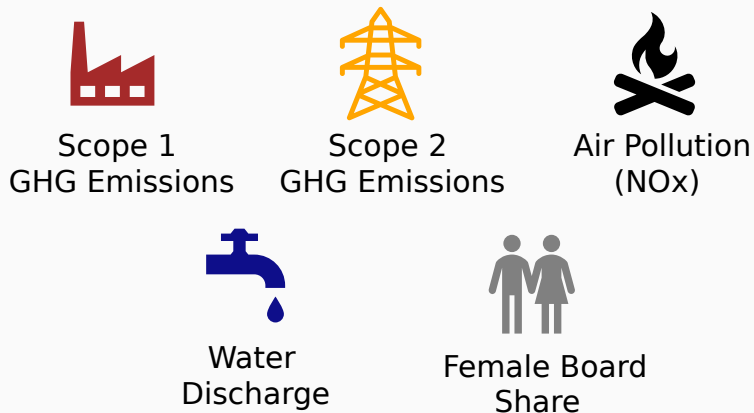


Figure 2: Target Variables

Training Approach for Point Estimates

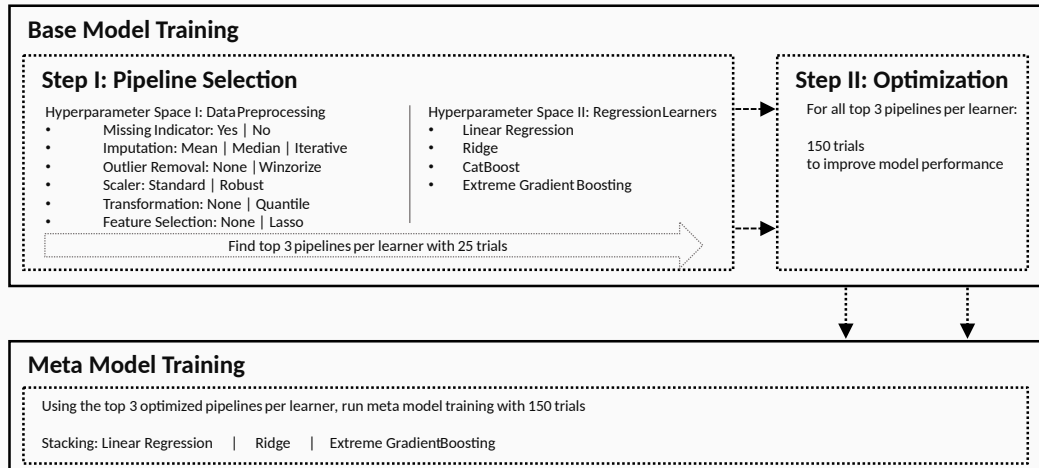


Figure 3: Point Estimate Training Approach using MSE

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:

$$I(x) = [\hat{y}_{\tau/2}(x) - \hat{r}, \hat{y}_{1-\tau/2}(x) + \hat{r}], \text{ where} \\ \hat{r} = \text{Quantile}\left(\frac{[(n_{cal}+1)(1-\alpha)]}{n_{cal}}, \{c_1, \dots, c_{n_{cal}}\}\right)$$

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

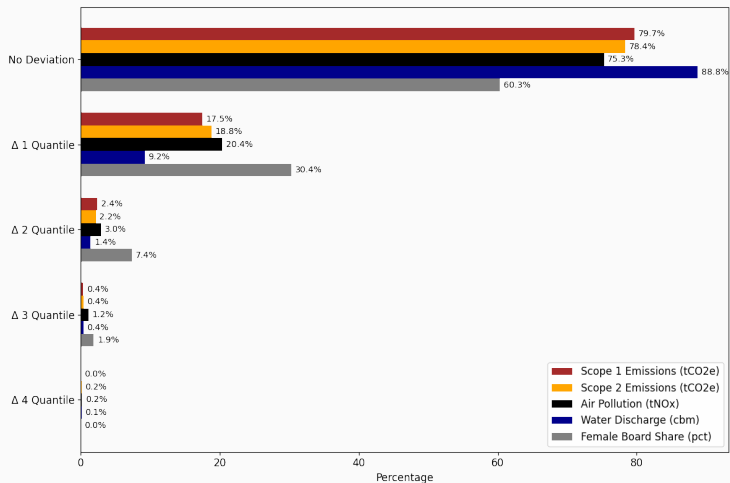


Figure 4: Number of Deviating Quintiles

Temporal Variation

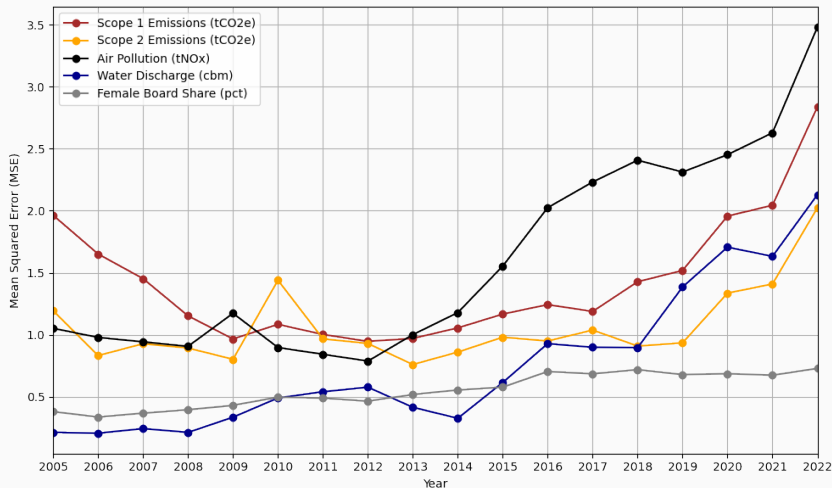


Figure 5: Temporal Model Performance

Spatial Variation

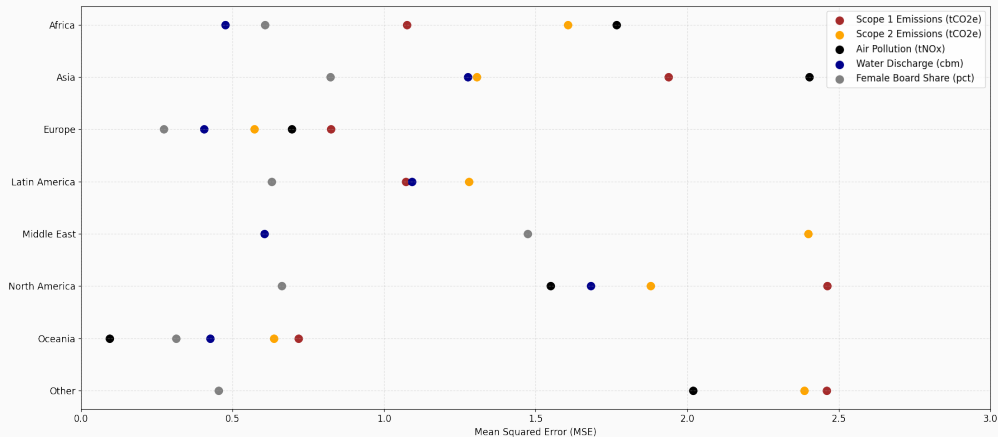


Figure 6: Spatial Model Performance

Sectoral Variation

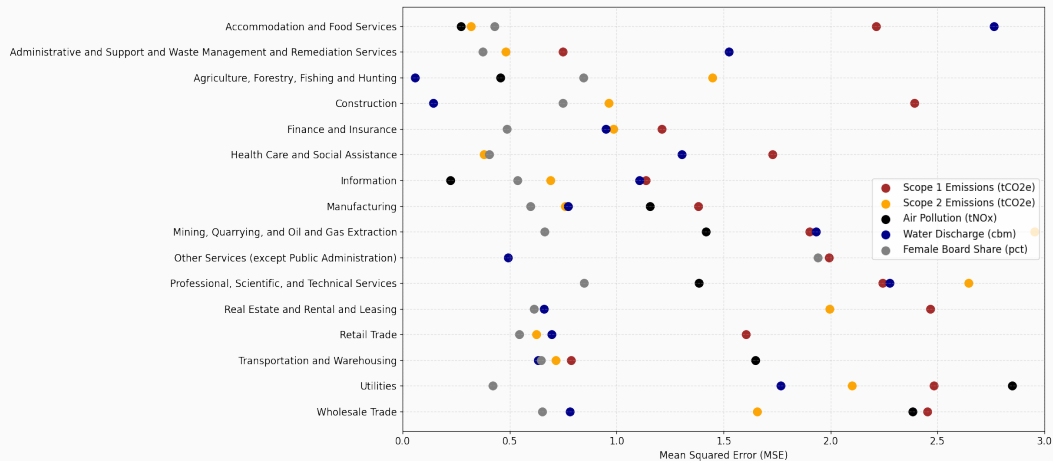


Figure 7: Sectoral Model Performance

Beware of Prediction Uncertainty!

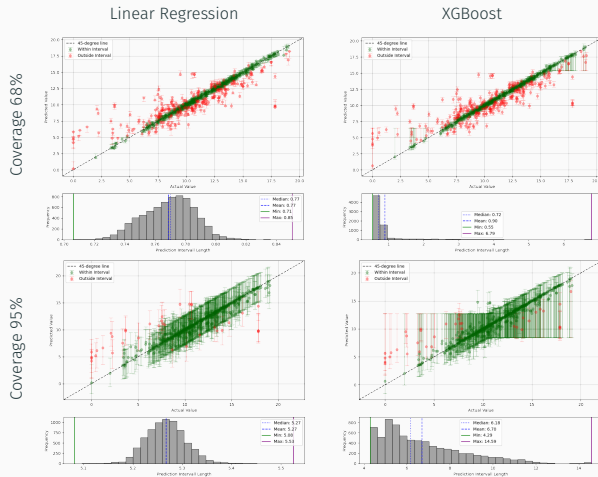


Figure 8: Prediction Uncertainty in Different Settings for Scope 1 Emissions

Reflect On Uncertainty Measure

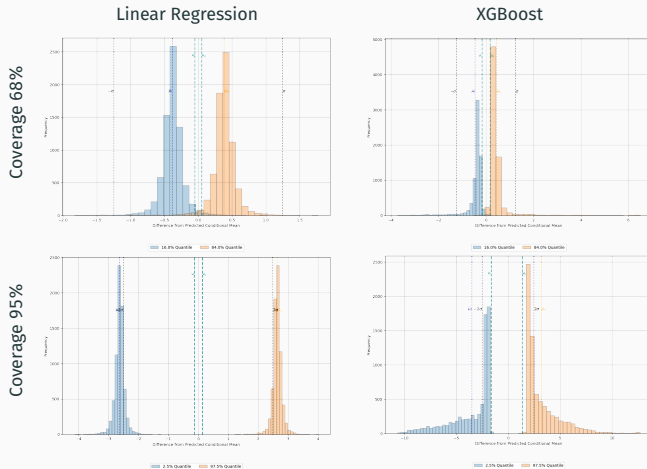


Figure 9: Deviation from Conditional Mean for Scope 1 Emissions

- It's actually in the financials, stupid!
- **Beware of prediction 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

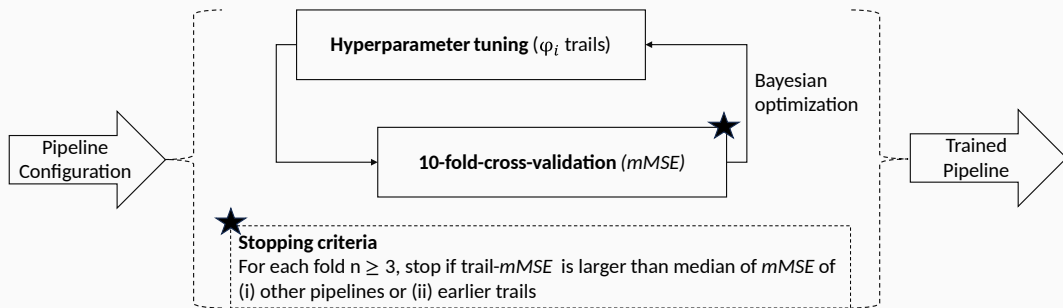


Figure 10: Early Stopping Approach

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

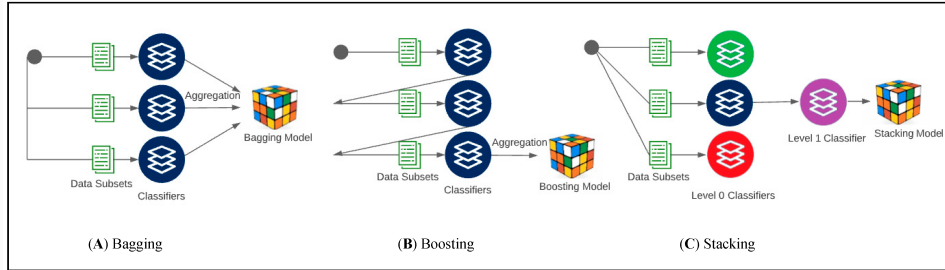


Figure 11: Approach to Boosting, Bagging, and Stacking by Ismail, El Mrabet, and Reza 2022