Predicting CO2 Emissions: Analyzing the Impact of Environmental Commitments

Sponsor: Clarity AI

Georgia Institute of Technology

Forecasting Team 4

Table of Contents

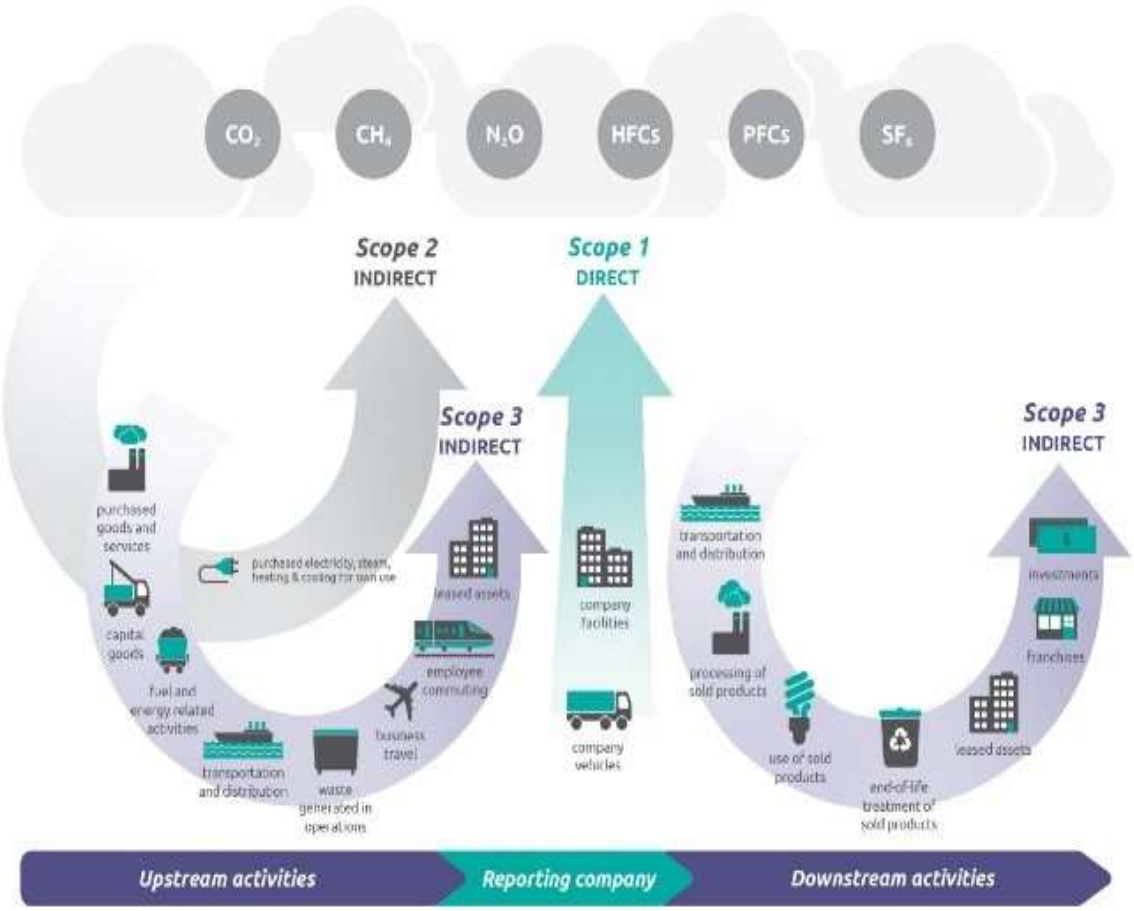
1. Background

Scope 1 Greenhouse Gas (GHG) emissions norms are critical in today's changing climate globally. These emissions are direct green house gases that a company makes directly from its operations like running factories and vehicles.

According to the Paris agreement, to keep the global warming no more than 1.5 deg Celsius, emissions need to be reduced by 45% by 2030, from 2010 levels and reach net-zero by 2050. There are around 195 countries that are part of Paris climate agreement targets for 2030.

Companies are keen to understand the trends of Scope 1 emissions in the upcoming years to optimize investing strategies to meet the Paris climate agreement. This project will investigate such data and find the methods to predict and forecast the emissions for various industries and individual companies.

Overview of GHG Protocol scopes and emissions across the value chain



Source: WRI/WBCSD Corporate Value Chain (Scope 3) Accounting and Reporting Standard (PDF), page 5.

2. Objective

In this project, the goal is to study the yearly data of various companies around the world and forecast their performance to predict if they would meet the 2030 emission targets. The goal is also to help enhance the precision of Clarity AI predictions, enabling us to forecast sustainability indicators by analyzing historical trends and patterns.

3. Primary Research Questions

- ✓ How do different forecasting models perform in capturing the dynamic nature of sustainability metrics over time?
- ✓ What are the key drivers and factors influencing the variability of sustainability indicators over time, and how can they be effectively captured by forecasting models?

- ✓ Do organizations with explicit commitments towards achieving global net zero emissions, or targets to reduce CO2 emissions, differ in their forecasted sustainability trajectories with respect to those without such commitments and targets?

4. Data

Data consisted of two sets – training and validation data. Training data consisted of two target variables as well as 13 independent variables.

Feature Name	Data Type	Description	Unique Values
Clarity id	Categorical	Identifier for different companies within the dataset. Each 'clarity_id' represents a unique category, which can affect CO2 emissions based on individual characteristics or operations.	1372
Metric	Numerical	The feature being measured, which is CO2 scope 1 emissions.	-
Metric year	Numerical	The year for which the data is recorded.	-
Provider code	Categorical	Identifier for the data provider or source.	2
Industry code	Categorical	Identifier for each industry.	155
Industry name	Categorical	Description for each industry.	155
Country code	Categorical	Identifier for the country where each company's headquarters is located.	57
CO2 emissions	Numerical	The CO2 emissions of the company for the given year.	-
Revenue	Numerical	The revenue earned by the company in the given year. This feature helps capture the scale of operations, as larger revenues might be associated with higher production activities and potentially higher emissions.	-
CO2 emissions intensity	Numerical	The ratio of emissions to revenue (Emission/Revenue).	-
Verification CO2directscope1	Binary	Indicates whether the CO2 emissions have been verified by a third party.	2

Target_emissions	Binary	Indicates whether the company has set emission targets. These targets could influence actual emissions as companies strive to meet their goals.	2
Policy_emissions	Binary	Represents whether the company has policies related to emissions in place.	2
Sbti_alignment	Binary	Indicates whether the company's targets are aligned with the Science-Based Targets initiative (SBTi).	2
Nz_statement	Binary	Indicates whether the company has made a Net Zero Statement.	2

[Table 1: Feature Summary with Unique Values]

The dataset used in this analysis is structured as panel data, comprising observations from multiple companies over a period of 21 years, from 2002 to 2022. Each company in the dataset is represented across varying years, allowing for the analysis of temporal trends and patterns within individual companies as well as across the entire dataset. This panel format facilitates the study of how CO2 emissions and related factors evolve over time, both at the company level and in aggregate.

The initial data size for the training set was 18284 data points covering 20 companies. We found that there were 1372 companies, 155 industries having headquarters in 57 countries.

5. Data Preprocessing

To address the skewness observed in the co2_emission_raw and revenue features, we applied a log transformation. This method effectively normalized the distributions of these features, successfully mitigating their skewness. However, after the log transformation, the revenue column showed NaN values for 5 observations due to negative values in the original data. These were backfilled using data from the subsequent year to maintain data integrity.

In addition to this, missing values were found in five variables: verification_co2directscope1, targets_emissions, policy_emissions, sbti_alignment, and nz_statement. Specifically, 15,034 data points were missing for verification_co2directscope1, 153 for targets_emissions, 22 for policy_emissions, 7,977 for sbti_alignment, and 2,643 for nz_statement. When we initially removed all the missing data points, we were left with only 9,792 data points—almost half of the original dataset. As this was not a viable solution, we decided that imputing the missing data was the best approach. Given that approximately 82% of the data was missing for verification_co2directscope1, we could not impute this value and thus opted to remove the column. For the other variables, we explored two imputation methods, which will be

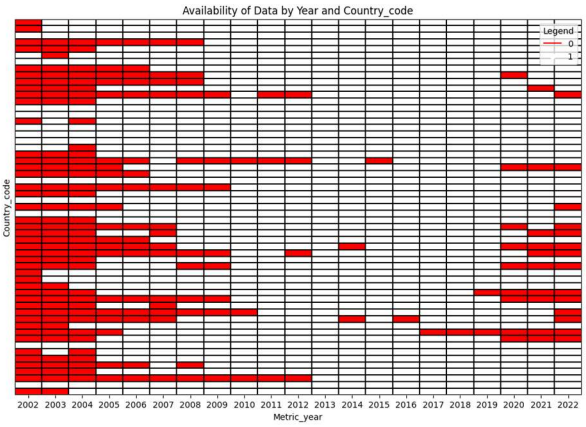
discussed further in the modeling section, including imputing missing values using the mode for each variable, grouped by company.

To streamline the dataset, we removed columns that were deemed irrelevant or redundant. The metric column was excluded as it provided no meaningful information. The clarity_industry_code was removed due to its redundancy with clarity_industry_name, and the co2directscope1_intensity column was removed to prevent data leakage. Finally, the revenue feature underwent standard scaling to ensure consistency in the modeling process.

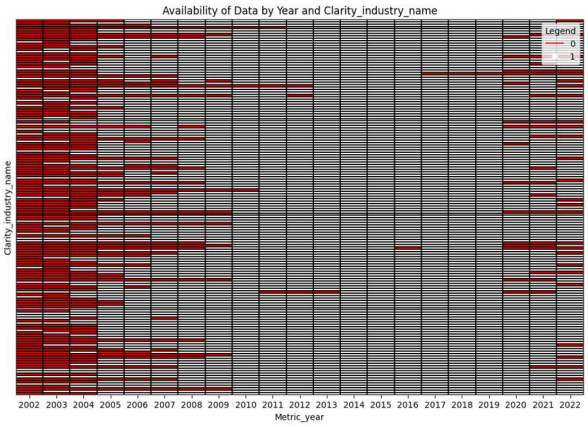
6. Exploratory Data Analysis

This section presents an EDA focused on CO2 emissions and revenue, categorized by country, industry, and company. The analysis includes visualizations of frequency distributions, average CO2 emissions, and revenue for each category. Additionally, heatmaps are provided to illustrate data availability over the years, offering a comprehensive overview of the dataset's structure and coverage. This EDA lays the groundwork for understanding the key factors influencing CO2 emissions and their relationship to revenue across different segments.

6.1 Availability Across Country, Industry, and Company

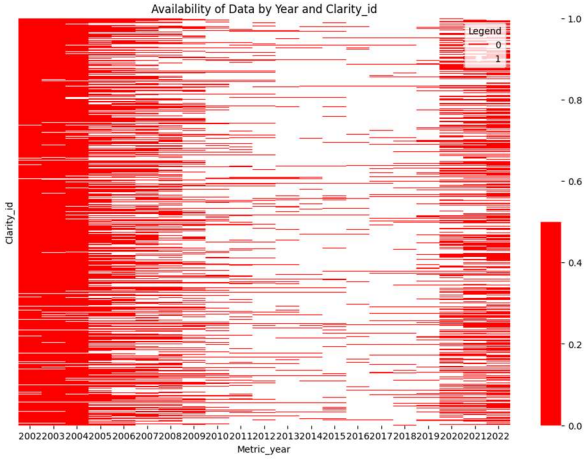


[Figure 1-1: Country]



[Figure 1-2: Industry]

The heatmaps illustrate the availability of data across years for countries, industries, and companies, revealing substantial gaps in data collection. The country-level heatmap shows that data availability is inconsistent, with several countries having many missing years, particularly from the mid-2000s onward (Fig. 1-1). Similarly, the industry-level heatmap indicates that certain industries have large periods without data, especially after 2008 (Fig. 1-2).



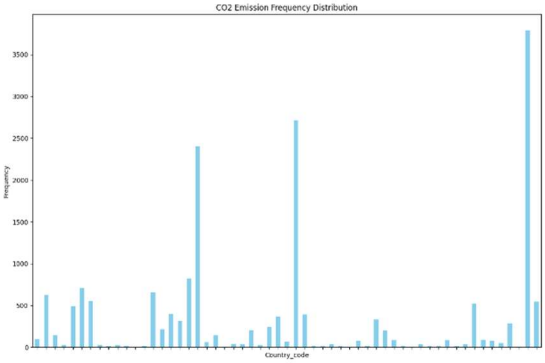
[Figure 1-3: Company]

The company-level heatmap is even more fragmented, showing significant gaps across many companies and years, with only a few companies consistently providing data over the entire period (Fig. 1-3).

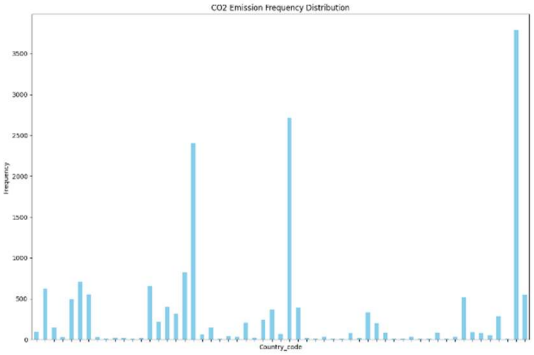
The significant data gaps across time and entities pose challenges for predictive modeling, particularly for time series analysis, which requires continuous data. The irregular and sparse nature of this dataset makes it difficult to apply time series models effectively, as any trends identified may be unreliable due to missing data. Additionally, the panel data structure, with its inconsistent data availability across entities, further complicates time series modeling.

Given these limitations, tree-based models like Random Forest or Gradient Boosting are more suitable. These models handle missing data more effectively and do not require continuous time series data. The data gaps also emphasize the importance of feature engineering, such as using imputation to fill missing values or incorporating temporal indicators. Overall, this dataset requires modeling approaches that can manage irregular and incomplete data while maintaining reliable predictions.

6.2 Frequency Plots Across Country, Industry, and Company



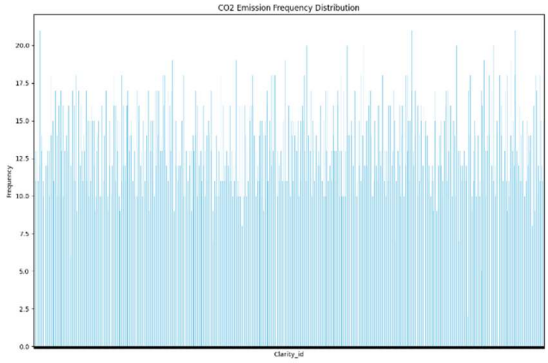
[Figure 2-1: Country]



[Figure 2-2: Industry]

The CO2 emission frequency plots reveal significant imbalances across countries, industries, and companies. At the country level, a few countries dominate the dataset, contributing the

majority of data points, while many others are underrepresented (Fig. 4-1). Similarly, the industry-level plot shows that certain industries are heavily represented, likely due to higher environmental scrutiny or greater industrial activity, while others contribute far fewer data points (Fig. 4-2). This imbalance across both countries and industries suggests a potential bias toward these dominant entities, which could skew analyses if not properly addressed.

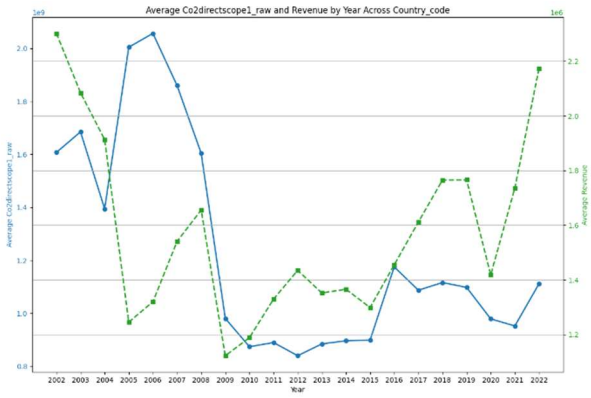


[Figure 2-3: Company]

The company-level plot, while more evenly distributed, still indicates that some companies have significantly more data points, pointing to potential outliers (Fig. 4-3).

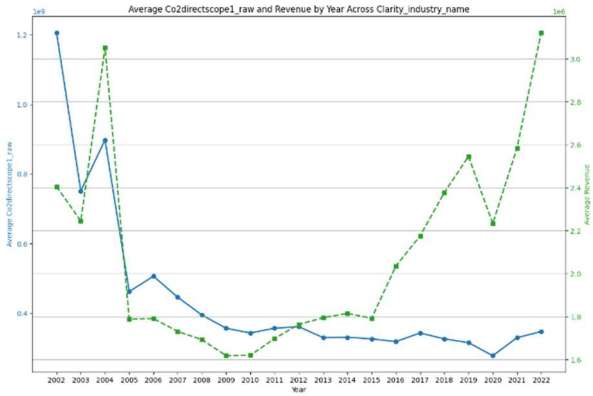
These patterns have important implications for developing a predictive model for CO2 emissions. The observed data imbalance and skewness, particularly across countries and industries, suggest that tree-based models (e.g., Random Forest or Gradient Boosting) are more appropriate. These models effectively manage skewed data distributions and are robust to outliers, making them well-suited for this dataset's variability. Additionally, the risk of overfitting, given the high variability and potential overrepresentation of certain entities, can be mitigated by the inherent structure of tree-based models, which are better equipped to generalize across diverse data subsets.

6.3 Yearly Average CO2 Emission and Revenue Plots Across Country, Industry, and Company

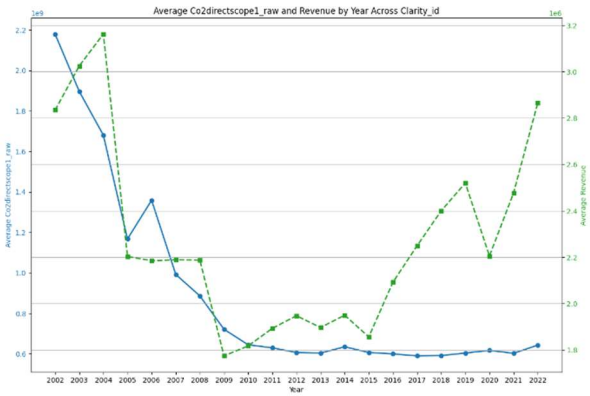


[Figure 3-1: Country]

The CO2 emission and revenue plots across countries, industries, and companies reveal significant trends in how these metrics have evolved over time. At the country level, CO2 emissions have generally decreased, while revenue patterns show more variability, suggesting that economic growth doesn't always correlate directly with emission levels.



[Figure 3-2: Industry]



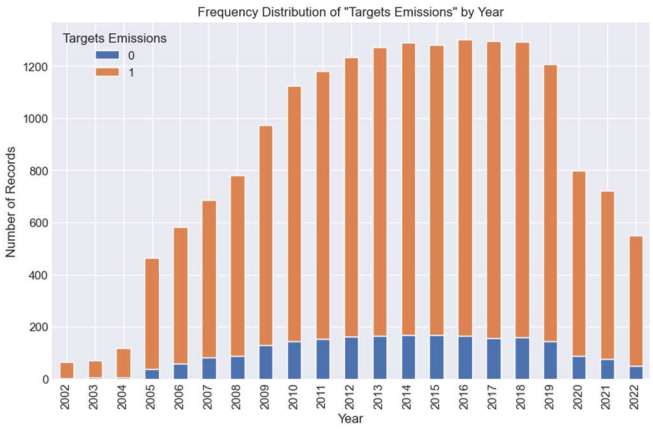
[Figure 3-3: Company]

Similarly, at the industry and company levels, CO2 emissions have stabilized or declined as revenues have increased, indicating a potential decoupling of economic performance from environmental impact. However, the fluctuations in emissions around economic downturns or regulatory changes highlight the continued influence of external factors on emission levels.

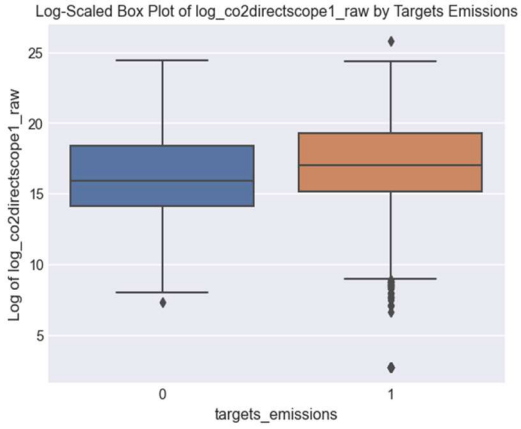
To better understand these patterns, it is important to explore how specific company commitments, such as emission targets, policies, and alignments with sustainability initiatives, are driving these trends. By visualizing and analyzing these commitment variables, we can assess their impact on CO2 emissions and evaluate whether companies' stated goals are effectively contributing to emission reductions.

6.4 Analysis of Commitment Variables and Their Impact on CO2 Emissions

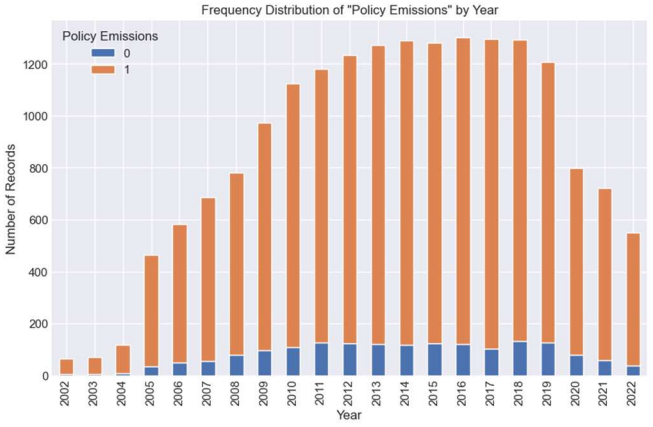
To further investigate the influence of company commitments on CO2 emissions, we have visualized and analyzed key commitment variables, including emission targets, policies, and alignment with sustainability initiatives. These visualizations, coupled with statistical tests like the Kruskal-Wallis test, provide insights into the effectiveness of these commitments in reducing emissions. The following visualizations illustrate the distribution of these commitment variables over time, as well as their impact on CO2 emissions, with statistical results such as the Kruskal-Wallis test highlighting the significance of these relationships.



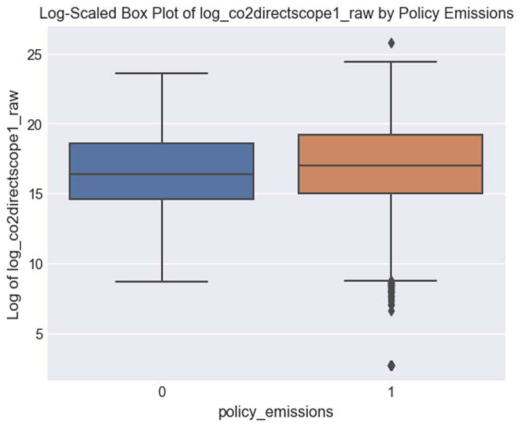
[Figure 4-1: Target Emissions]



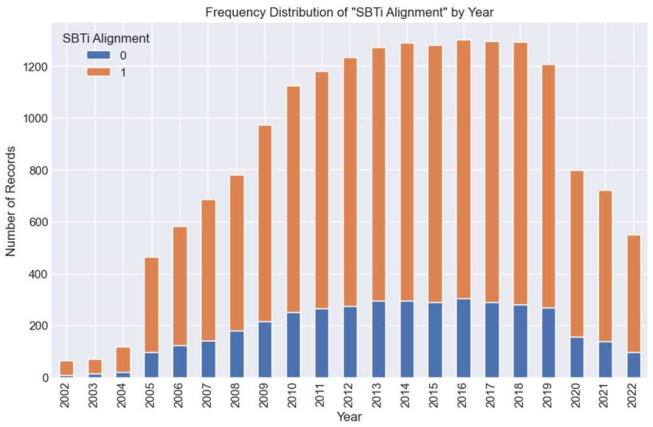
[Figure 4-2: Target Emissions]



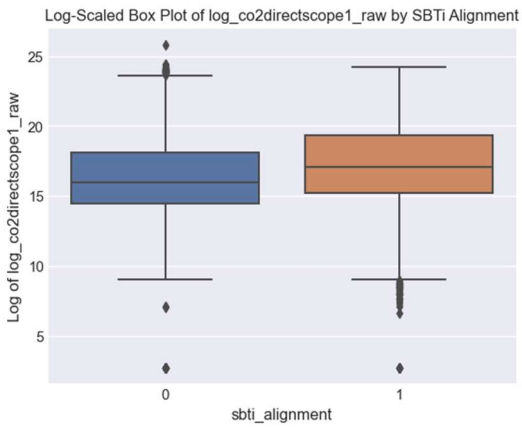
[Figure 5-1: Policy Emissions]



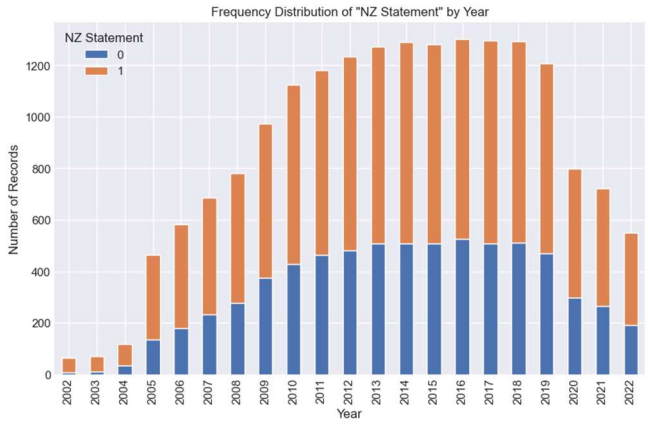
[Figure 5-2: Policy Emissions]



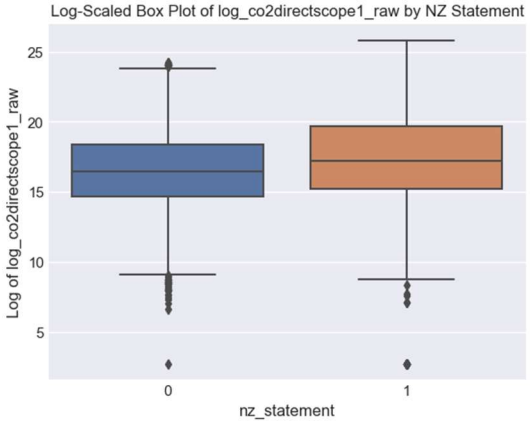
[Figure 6-1: SBTi Emissions]



[Figure 6-2: SBTi Emissions]



[Figure 7-1: NZ Statement]



[Figure 7-2: NZ Statement]

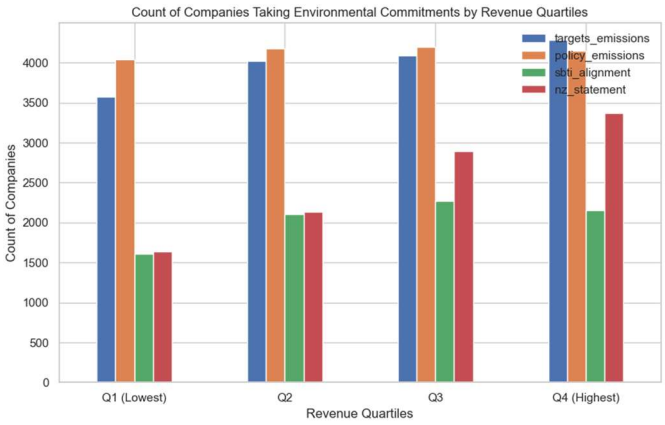
The analysis of commitment variables—"Targets Emissions," "Policy Emissions," "SBTi Alignment," and "NZ Statement"—reveals important trends in corporate CO2 emissions. Companies with commitments such as emission targets, policy emissions, and Net Zero statements generally exhibit higher median CO2 emissions compared to those without these commitments. This pattern suggests that larger or more emission-intensive companies are more likely to adopt these initiatives, possibly due to their greater environmental impact or regulatory scrutiny. However, the effectiveness of these commitments in reducing emissions varies significantly across companies, as indicated by the wide range of emission levels within these groups. In contrast, companies that align their targets with the Science-Based Targets initiative (SBTi) tend to have lower CO2 emissions.

Commitment Variable	**Kruskal-Wallis Statistic**	**P-value**
Targets Emissions	176.637	2.63E-40
Policy Emissions	35.287	2.85E-9
SBTi Alignment	253.139	5.37E-57
NZ Statement	324.175	1.78E-72

[Table 2: Kruskal-Wallis Test Results]

The Kruskal-Wallis tests for each variable show statistically significant differences in CO2 emissions between companies with and without these commitments, with extremely low p-values indicating that these differences are not due to chance (Table 2). This suggests that these commitments are closely linked to distinct CO2 emission patterns, making them useful predictors of emission levels. However, the wide variability in emissions among companies that have adopted these commitments indicates that the effectiveness of these initiatives can vary greatly, depending on how rigorously they are implemented. Overall, the results highlight the significant role these commitments play in influencing CO2 emissions, while also emphasizing the diverse outcomes that can result from different approaches to implementation.

6.5 Environmental Commitments and Revenue Quartiles



[Figure 8: Counts of Companies Taking Environmental Commitments by Revenue Quartiles]

The plot suggests that companies in different revenue quartiles show varying levels of commitment to environmental initiatives, with those in the highest revenue quartile (Q4) generally being more likely to adopt environmental commitments, except for SBTi alignment (Fig. 8). This indicates that revenue is correlated with the likelihood of a company making environmental commitments, which could imply that companies with higher revenues have more resources to invest in such initiatives.

Given this observation, revenue appears to be a significant factor in determining a company's environmental behavior. Therefore, it would make sense to include revenue as a feature in a tree-based model, as it might help the model capture relationships between a company's financial capacity and its environmental actions, potentially improving the accuracy of CO2 emission predictions.

7. Data Transformation Strategies

With the exploratory data analysis and preprocessing complete, we turned our attention to the modeling and prediction process. Before proceeding with the models, it was essential to address the transformation of categorical variables with varying levels of cardinality. To handle the diverse range of categorical variables in the dataset, we employed two distinct data transformation strategies. The first approach involved pure one-hot encoding, which provided a straightforward representation for all categorical variables. The second approach was a hybrid encoding strategy, where we applied one-hot encoding to low-cardinality features and target encoding to those with high cardinality. These strategies were designed to optimize the modeling process by effectively managing the complexity and dimensionality of the categorical data.

7.1 Approach 1: Pure One-Hot Encoding:

All categorical variables were transformed using one-hot encoding, resulting in a high-dimensional feature space of 1616 variables. This method was chosen to simplify the initial model training process by providing a straightforward categorical representation.

7.2 Approach 2: Hybrid Encoding and Feature Engineering:

For low-cardinality features, one-hot encoding was applied, while high-cardinality categories underwent target encoding. This hybrid approach was employed to reduce the dimensionality and improve model interpretability.

We engineered additional features to effectively capture temporal variations in CO2 emissions and revenue within each company. This approach allowed us to leverage the inherent time series aspect of our data, fully utilizing the temporal dynamics unique to each company. These features were essential for analyzing and understanding year-over-year changes and trends, which would have been difficult to discern without such feature engineering.

8. Modeling and Evaluation: Pure One-Hot Encoding

8.1 Model Training and Setup

Various ensemble models including Random Forest, LightGBM, XGBoost, and CatBoost were trained on the one-hot encoded dataset. We split the data into training (80%) and test sets (20%) to ensure an unbiased evaluation of model performance.

8.2 Performance Metrics

We implemented a custom function to compute key performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2). These metrics were calculated for both the training and test sets, providing a thorough assessment of each model.

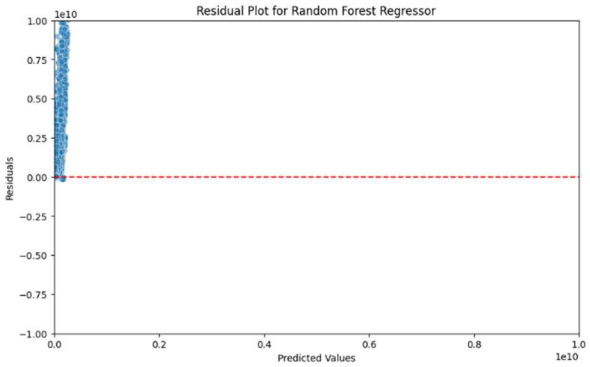
Hyperparameter tuning was performed using RandomizedSearchCV to enhance model performance. We defined parameter distributions for each model, and the search process was conducted over 50 iterations with 5-fold cross-validation.

8.3 Model Evaluation and Results

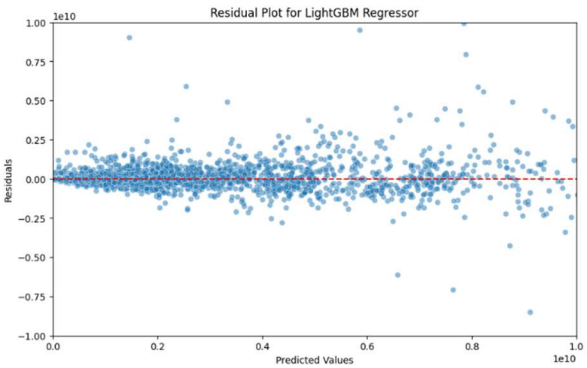
	Training Error				Test Error			
	MAE	MSE	RMSE	R-sq	MAE	MSE	RMSE	R-sq
Random Forest Regressor	6.8E+08	7.6E+18	2.8E+09	-0.0401	6.6E+08	6.1E+18	2.5E+09	-0.0478
LightGBM Regressor	2.8E+08	3.2E+18	1.8E+09	0.56633	3.2E+08	1.9E+18	1.4E+09	0.67344
XGBoost Regressor	1.8E+08	2.1E+18	1.5E+09	0.70752	2E+08	7.4E+17	8.6E+08	0.87218
CatBoost Regressor	2.9E+08	3.1E+18	1.8E+09	0.57083	2.9E+08	1.6E+18	1.3E+09	0.7283

[Table 3: Training vs. Test Errors for Different Models]

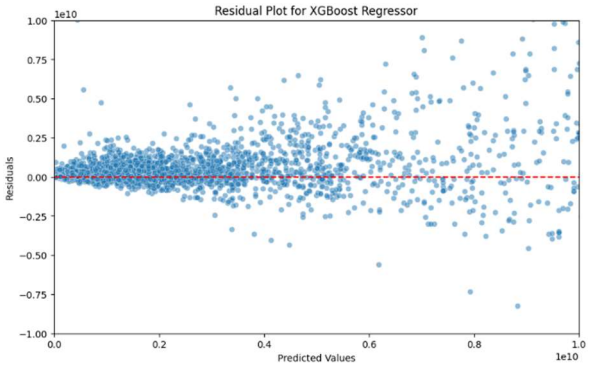
The XGBoost and CatBoost regressors performed the best, showing low error metrics and high R-squared values on both training and test sets, indicating strong generalization and no overfitting. LightGBM also performed well with balanced errors, while the Random Forest Regressor struggled with high errors and negative R^2, suggesting poor model fit and an inability to capture the data patterns effectively.



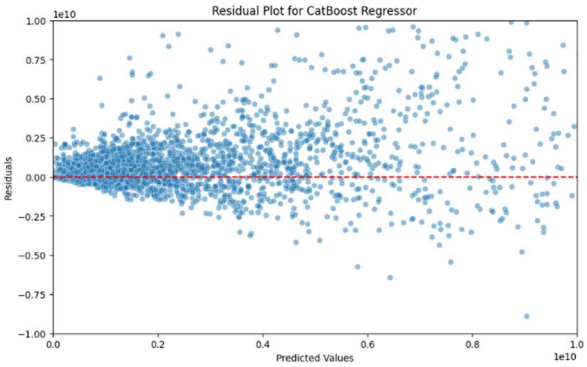
[Figure 8-1: Random Forest]



[Figure 8-2: LightGBM]



[Figure 8-3: XGBoost]



[Figure 8-4: CatBoost]

Looking at the residual plots for these models. There are three major observations:

First, the residual plot of random forest indicates that the Random Forest Regressor is not performing well on this dataset. The model may be overfitting to a particular subset of the data or failing to capture important relationships, leading to poor generalization and significant prediction errors. This issue could be related to how high-cardinality categorical variables were treated. Second, there is a common trend of underprediction in these

models, which is indicated by more residuals being positive. Third, there is increasing residual variance with higher predicted values. However, the XGBoost model shows the lowest variance among the three models.

Based on the performance metrics and the plot, XGBoost is the best model.

9. Data Transformation: Hybrid Encoding and Additional Feature Engineering

To address the limitations of one-hot encoding, we implemented a hybrid encoding strategy. For categorical features exhibiting fewer categories, such as binary variables and metric year variables, one-hot encoding remains our method of choice. However, for high-cardinality categories, we applied target encoding at the company, industry, and country levels, replacing categorical identifiers with more meaningful statistical representations, such as means and standard deviations of CO2 emissions and revenue. This approach not only reduced the overall dimensionality of the dataset but also improved model interpretability.

To prevent data leakage, target encoding of mean and standard deviation was applied only to the training data to ensure that the test data does not inadvertently affect the feature values derived during training, and these mean and standard deviations from the training data were merged into the test data.

In addition to using hybrid encoding, we enhanced our dataset by engineering features that capture temporal variations in CO2 emissions and revenue for each company. This involved calculating yearly percent changes to make the most of the time series components within each company. Given the panel format of the data, with varying years across companies, this approach was necessary to effectively utilize the available time series data.

However, due to inconsistencies and gaps in the data across some companies, these features had many missing values. To address these gaps, we employed a random forest model specifically designed to impute the missing percent change values.

10. Random Forest Imputation of Missing Yearly Changes Post-Feature Engineering

After our feature engineering phase, we identified that 2648 rows had missing values (NaNs) in the percent change features for CO2 emissions (co2_percent_change) and revenue (revenue_percent_change). These gaps arose because the data for many companies were not continuous year over year, leading to incomplete records for these yearly change calculations.

To address this issue, we constructed two random forest models: one focused on imputing missing values for co2_percent_change and the other for revenue_percent_change. We chose not to use backfill or forward fill methods because there were many gaps in the years, and these methods are inadequate for handling such extensive gaps. Additionally, we

needed a model to impute the validation dataset, which also have numerous gaps that requires accurate predictions to fill the NaNs in these key features.

10.1 Model Setup and Data Preparation

For the input to these random forest models, we utilized target-encoded features to represent high-cardinality categories and removed the original categorical variables. We also excluded features like co2directscope1_raw and co2directscope1_intensity from our model training because they were not present in the validation dataset, ensuring consistency across data applications.

We employed a strategic data splitting method where the dataset was divided based on the availability of the co2_percent_change feature into two subsets: one with known values and one with unknown (NaN) values. The subset with known values was further split into training and testing groups to facilitate thorough model evaluation. The best models were selected using RandomizedSearchCV, optimizing for accuracy and reliability.

10.2 Model Performance and Evaluation

The performance metrics from the test data showed promising results. For CO2 percent change, we achieved an MSE of 0.0098, an RMSE of 0.099, and an MAE of 0.0196. A similar process was applied for the revenue_percent_change feature, yielding an MSE of 0.00054, an RMSE of 0.0233, and an MAE of 0.01256. We were satisfied with the efficacy of our models in predicting and imputing missing values accurately.

10.3 Final Adjustments and Data Readiness

After addressing the NaN values in the percent change features, only four rows remained with NaNs in the company-level standard deviation features, which were due to companies having only a single recorded observation. This lack of variance in the data led to NaN values for company_co2_std and company_revenue_std. We resolved this by setting these NaNs to zero, reflecting the absence of variance in these cases.

With all NaN values accounted for, post-target encoding and feature engineering for percent changes, our dataset was fully prepared. We then proceeded with training the ensemble models on this enriched and more complete dataset.

11. Modeling and Evaluation: Hybrid Encoding and Feature Engineering

We trained ensemble models on datasets enhanced with a combination of one-hot encoding, target encoding, and additional feature engineering. To ensure consistency in our modeling approach and the evaluation process, we applied the same parameter

distributions and key performance metrics that had been used in our previous models, which were trained exclusively on one-hot encoded data.

11.1 Model Evaluation and Results

The results, summarized in the table below, highlight the performance metrics for each model:

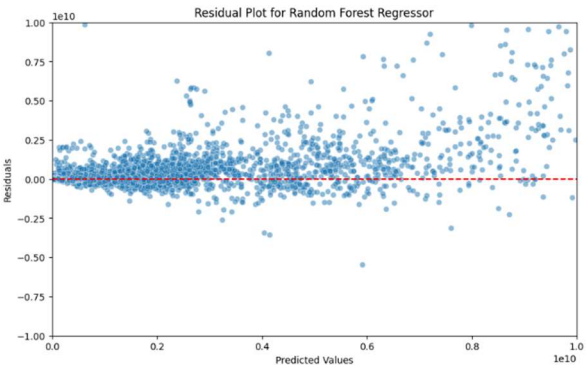
	Training Error				Test Error			
	MAE	MSE	RMSE	R-sq	MAE	MSE	RMSE	R-sq
Random Forest Regressor	1.7E+08	1.8E+18	1.3E+09	0.75409	2.1E+08	9.2E+17	9.6E+08	0.83879
LightGBM Regressor	1.1E+08	1.1E+18	1.1E+09	0.84466	1.4E+08	3.7E+17	6.1E+08	0.93444
XGBoost Regressor	4.4E+07	5.2E+16	2.3E+08	0.99292	1.2E+08	2.6E+17	5.1E+08	0.95358
CatBoost Regressor	1.1E+08	7.6E+17	8.7E+08	0.8964	1.4E+08	3.6E+17	6E+08	0.93738

[Table 4: Training vs. Test Errors for Different Models]

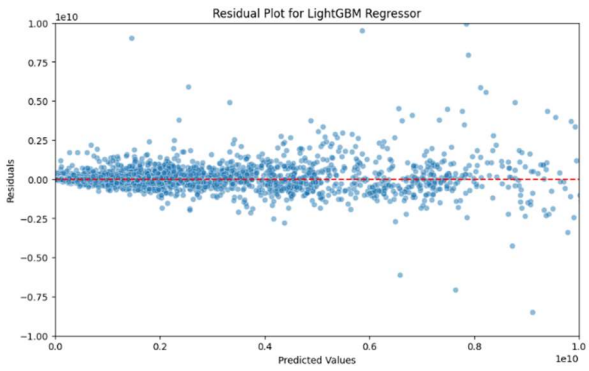
The ensemble models that incorporated hybrid encoding and feature engineering significantly outperformed those based on one-hot encoding alone. The Random Forest Regressor, which previously struggled with high errors and poor R^2 values, showed marked improvement, indicating better model fit and generalization. LightGBM, XGBoost, and CatBoost also displayed enhanced performance, with XGBoost achieving the highest R-squared and lowest error metrics, confirming its robustness and predictive power.

These results suggest that this approach offers a more meaningful representation of the data, enhancing accuracy and reliability across all tested models. We believe that the significant improvements in the performance metric s are due to this approach addressing the challenges of complexity and overfitting associated with the high-dimensional data from one-hot encoding, leading to better generalization and performance.

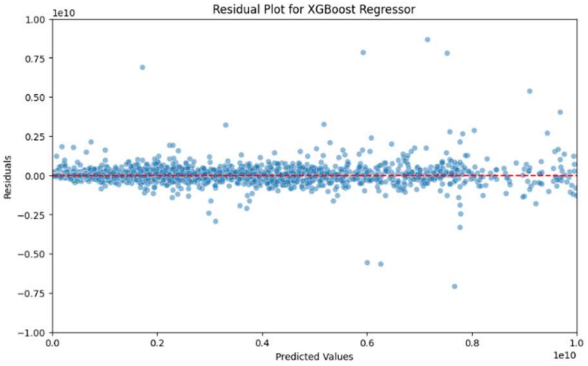
The residual plots reinforce this conclusion:



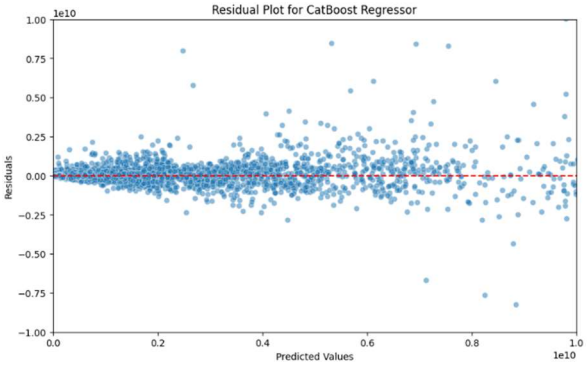
[Figure 9-1: Random Forest]



[Figure 9-2: LightGBM]



[Figure 9-3: XGBoost]



[Figure 9-4: CatBoost]

The residual plots for the models using hybrid encoding show that residuals are more tightly clustered around zero compared to models trained on one-hot encoded data. Although the variance still tends to be higher for larger predictions, it is significantly smaller than before. Lastly, there is no observable tendency for underprediction or overprediction, indicating that hybrid encoding, and feature engineering have improved the models' ability to learn and capture patterns.

The conclusion we made from this approach is that: among the models, XGBoost stands out as the best overall model.

12. Backfilling the Validation Dataset: Methods and Outcomes

To ensure that the validation dataset mirrored the completeness of the training set, we enriched it with company, industry, and country-level features. This enrichment was accomplished by merging these detailed features using specific identifiers such as clarity_id, clarity_industry_name, and country_code. This step was crucial for aligning the validation data with the enriched training dataset, enabling our trained models to effectively make predictions on the validation data.

In the process of enriching the dataset, we addressed a few missing values in the revenue column by replacing them with the company_revenue_mean. Additionally, we addressed missing values in company_co2_std and company_revenue_std that resulted from entries with only a single data point. For these, we set the standard deviations to zero, as no variance calculation was possible from a single observation.

We also identified a few missing values in the co2_percent_change and revenue_percent_change features. To fill these gaps, we utilized two previously built random forest models to imputing missing values for these specific features. This approach ensured that all necessary data points were present, allowing for model predictions.

Once the dataset was fully prepared and complete, we proceeded with the predictions. We used eight models developed from two different approaches to generate forecasts for the validation set. For each model, we added the predictions as new columns to the dataset. After generating the predictions, we reverted them to their original scale. We successfully backfilled the validation data and saved it in a CSV file.