Time Series Forecasting for Predicting Sustainable Performance of Companies

A final project report

Presented to project sponsor 'ClarityAI'
and Academic Faculty of OMSA
Course: CSE/ISYE/MGT 6748

by

Seungwon Lee (slee3510@gatech.edu)

Marshall Palfenier (mpalfenier3@gatech.edu)

Ramy ElGendi (relgendi3@gatech.edu)

In Partial Fulfillment
of the Requirement for the Degree
Master of Science in Analytics

Georgia Institute of Technology

July 2024

Table of Contents

Abstract

In the current trend of business operations, sustainability performance indicators of a company have become high-demand evaluation metrics for investors. This study aims to enhance the predictive accuracy of these sustainability indicators, especially scope 1 emission (direct emission) for the mid-term forecasting range. In this forecasting exercise, annual direct emissions vary heavily by industry type, a company's historical operational efficiency, and regulatory environment. The study assumes a strong relationship between a company's revenue and direct emissions as a primary influencing factor and orders sustainability features by importance using feature selection methods, Lasso Regression and Random Forest. This study offers a detailed review of the literatures and industry-specific domain knowledge to propose a qualitative assessment and its influence on emission forecasting in the future, given the demand for emission disclosures from investors and stakeholders. Company data are provided by the project sponsor ClarityAI, which includes 155 industry types, 53 countries, revenue (millions USD) with Scope 1 emission (MtCO2e), and five sustainability indicators. The model is trained on 20 years' worth of data, addressing missing years and introducing a new sustainability indicator in later years. The best results prove that Long Short-Term Memory Recurrent Neural Networks can retain mid-term dependencies in high volatility time series. LSTM obtained the lowest error results for the testing period.

Keywords: time-series forecasting; emissions forecasting; Scope 1 emission; sustainability performance indicator (SPI); feature selection; Lasso Regression; Random Forest; Exponential Smoothing; ARIMA; XGBoost; Prophet; LSTM

1 Introduction

Forecasting a company's Scope 1 emissions involves analyzing historical emission data to predict future emissions. While companies focus on reporting Scope 1 emission accurately and transparently, to set a baseline for sustainable business operations metrics, researchers and investors have been attempting to model effective CO2 emission prediction algorithms via time-series and machine-learning techniques. Sustainability performance has become a crucial indicator for investment decisions for both internal and external stakeholders of a company as it helps the investors assess the long-term risks associated with the company in terms of environmental, social, and governance risks and impact. It is believed that companies with strong sustainability practices often demonstrate better long-term operational performance as they tend to be more innovative, energy efficient and adaptable in changing market conditions. Growing demand from institutional investors for sustainable and responsible investments for pension funds and mutual funds drives greater needs to forecast sustainability performance indicators. Scope 1 direct emissions are critical to a company's environmental impact, regulatory compliance, operational efficiency, and strategic planning. By focusing on the direct emissions, companies can demonstrate.

1. accountability by directly controlling the emissions from fuel combustion and business operation processes,
2. Build stakeholder trust (investors, employees, and regulators) by planning and executing the emission reduction target, which demonstrates the company's ability to deliver and grow sustainably,
3. Reduce the cost of business operations holistically and
4. Contribute to global climate goals.

Scope 1 emission forecasting is essential for sustainable business practice and long-term success. Types of direct emission forecasting techniques fall into two main categories: traditional time-series forecasting statistics and machine learning (ML). Time-series methods try to infer the population based on the sample data and provide quantitative analysis; ML is used to make repeatable predictions by finding patterns within the data that require substantial computational resources and hyperparameter tuning. Both methods will provide deeper insights into the possibilities and limitations of the analysis and explain what the model can support with a degree of confidence in this paper.

2 Literature Review

Choosing the right time-series forecasting method depends on the characteristics of the emission data, such as the presence of trends, seasonality, cycle, and the complexity of the relationship within the data. (IBM Technology, 2022) Directionally, it is understood that the world has become more efficient over the past two decades in every industrial activity by electrifying and adopting fuel-efficient machinery that consumes less fuel/electricity per dollar value, creating lower carbon intensity in all 155 industries. Seasonality is not expected to be observed due to annual emission reporting, eliminating weather and holiday impacts. The cycle is a repeating but non-seasonal pattern, like an economic boom or financial crisis over several years or a decade, resulting in a much smoother curve than seasonality. The variation shows unpredictable irregularity in the data that cannot be explained by the data, making it challenging to capture a trend.

2.1 Statistical Models

Popular time-series forecasting methods include Exponential Smoothing, Holt-Winters (Triple Exponential Smoothing), and Autoregressive Integrated Moving Average (ARIMA)

Exponential Smoothing: The model is used to forecast a time-series that does not have clear trend or seasonality, giving more weight to recent observations and less weight to the older values, making them more responsive to changes in the data. Exponential smoothing can be expressed as:

$$s_t = ax_t + (1-a)s_{t-1} = s_{t-1} + a(x_t - s_{t-1})$$

$s_t\ =\ smoothed\ statistic$
a = smoothing factor, where $0 \leq\ a \leq 1$
x_t = the current observation

The smoothed statistic s_t (at time t) is a simple weighted average of the current observation x_t and the previous smoothed statistic $s_{t-1}$to produce a smoothed statistics as soon as two observations are available. Values of a close to 1 have less of a smoothing effect and give greater weight to recent changes in the data, while values of a closer to 0 have greater smoothing effect and are less responsive to recent changes. A benefit of Exponential Smoothing is that it does not require any minimum number of observations to produce results; however, all stages of the signal (as the theory was developed for signal processing) is required due to decay of older samples in weight exponentially. It is noted that the method of least square might be used to determine if the value of a for which sum of the quantities $(s_t - x_{t+1})^2$ is minimized.

There are several exponential methods it includes the following:

- Simple Exponential Smoothing (SES): Suitable for data without a trend or seasonality.
- Holt-Winters Seasonal Model: Also called ‘Triple Exponential Smoothing’, applies exponential smoothing three times to account for linear trends and seasonal patterns. (BYJUS, n.d.)

$s_0 = x_0$ $s_t = a\frac{x_t}{c_{t-L}} + (1-a)(s_{t-1} + b_{t-1})$ $b_t = \beta(s_t - s_{t-1}) + (1-\beta)b_{t-1}$ $c_t = \gamma\frac{x_t}{s_t} + (1-\gamma)c_{t-L}$	s_t = smoothed statistic, it is the simple weighted average of current observation x_t s_{t-1}= previous smoothed statistic a = smoothing factor of data; $0 \le a \le 1$ t = time period b_t= best estimate of a trend at time t β = trend smoothing factor; $0 \le \beta \le 1$ c_t = sequence of seasonal correction factor at time t γ = seasonal change smoothing factor; $0 \le \gamma \le 1$

ARIMA: ARIMA models are powerful and flexible, making them suitable for a wide range of time-series data. The model has three main components: Autoregressive (AR) – looks at how past values affect future values, Integrated (I) - a differencing component that accounts for trends and seasonality, and Moving Average (MA) – smooths the noise by removing non-deterministic or random movement from the time series (forecast error is a linear combination of past respective errors). ARIMA (V Kotu et al, 2019) can be expressed as:

$$y'_t = I + \alpha_1 y'_{t-1} + \alpha_2 y'_{t-2} + \cdots + \alpha_p y'_{t-p} + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + +\theta_q e_{t-q}$$

The ARIMA (p, d, q) model estimates the coefficient a and θ for a given (p, d, q) from the training data in a time-series. The equation shows the predictors are the lagged p data points for the autoregressive part and the lagged q errors are for the moving average part. The prediction is the difference y_t in the d^{th} order. Performance of ARIMA model can be evaluated by specifying (p, d, q) with the recognition of potential overfitting and redundancy and cancellation in AR-MA (R Nau, n.d.)

Machine Learning Models

Several machine learning algorithms have been gaining popularity for forecasting as the methods are often more flexible and capable of capturing complex patterns in data. Considering the volume of data for direct emissions, which are relatively low and incomplete in comparison to financial accounting data, the flexibility from ML helps to fit the pattern better. Some ML algorithms considered in this study include; Prophet, XGBoost and LSTM.

Prophet: Developed by Facebook, Prophet is an additive model that works well with time series that have strong seasonal effects and several seasons of historical data. It robustly handles missing data and shifts in the trend. It was originally developed to forecast business results taking ‘structured time series’ as the foundation. The latest Prophet model is considered in the study. (R Hyndman et al, 2021)

$$y_t = g_t + s_t + h_t + x_t + \epsilon$$

g_t = piecewise-linear trend (or ‘growth term’)
s_t = various seasonal pattern
h_t= holiday effects, events or holidays (e.g. good summer means more surf sales)
x_t = external regressors (e.g.marketing investments, bad weather)
ϵ = white noise term

XGBoost: XGBoost is a supervised machine learning method that implements gradient-boosting decision trees. It is a function that optimizes loss functions while applying several regularization techniques. Some literature says that the XGBoost method ‘wins every machine learning competition’ and tries to find out why XGBoost is better over MART (Multiple Additive Regression Tree, the tree-boosting method of choice until 2015).

XGBoost takes Newton Tree Boosting as its foundation: (D Leventis, 2018)

Algorithm 3: Newton tree boosting

Input : Data set $\mathcal{D}$.
A loss function L.
The number of iterations M.
The learning rate η.
The number of terminal nodes T

Initialize $\hat{f}^{(0)}(x) = \hat{f}_0(x) = \hat{\theta}_0 = \arg\min_{\theta} \sum_{i=1}^{n} L(y_i, \theta)$;

for m = 1,2,..,M **do**

$\hat{g}_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=\hat{f}^{(m-1)}(x)}$;

$\hat{h}_m(x_i) = \left[\frac{\partial^2 L(y_i, f(x_i))}{\partial f(x_i)^2}\right]_{f(x)=\hat{f}^{(m-1)}(x)}$;

Determine the structure $\{\hat{R}_{jm}\}_{j=1}^{T}$ by selecting splits which maximize $Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L} + \frac{G_R^2}{H_R} - \frac{G_{jm}^2}{H_{jm}}\right]$;

Determine the leaf weights $\{\hat{w}_{jm}\}_{j=1}^{T}$ for the learnt structure by $\hat{w}_{jm} = -\frac{G_{jm}}{H_{jm}}$;

$\hat{f}_m(x) = \eta \sum_{j=1}^{T} \hat{w}_{jm} \mathrm{I}(x \in \hat{R}_{jm})$;

$\hat{f}^{(m)}(x) = \hat{f}^{(m-1)}(x) + \hat{f}_m(x)$;

end

Output: $\hat{f}(x) \equiv \hat{f}^{(M)}(x) = \sum_{m=0}^{M} \hat{f}_m(x)$

Characteristics of XGBoost:

- Employs the Newton Tree Boosting to approximate the optimization problem, deploying Hessian matrix that results in better tree structure and stricter on selecting loss function (twice differentiated) to be concave.
- Randomization: Randomization on both row and column subsampling
- Missing value: learns the missing values by learning default direction and imputes by reduction on training loss.
- Penalization of complexity

LSTM: LSTM networks are a type of recurrent neural network (RNN) capable of learning long-term dependencies. They are particularly useful for time series with long-term trends and patterns and can capture non-linear relationships in the data. (M Sanjeevi, 2018)

A noticeable difference in the LSTM neural network is that it has ‘Memory’ in every ‘time-step’. A diagram of an LSTM cell at a time step t is shown below: (neural network, NN)

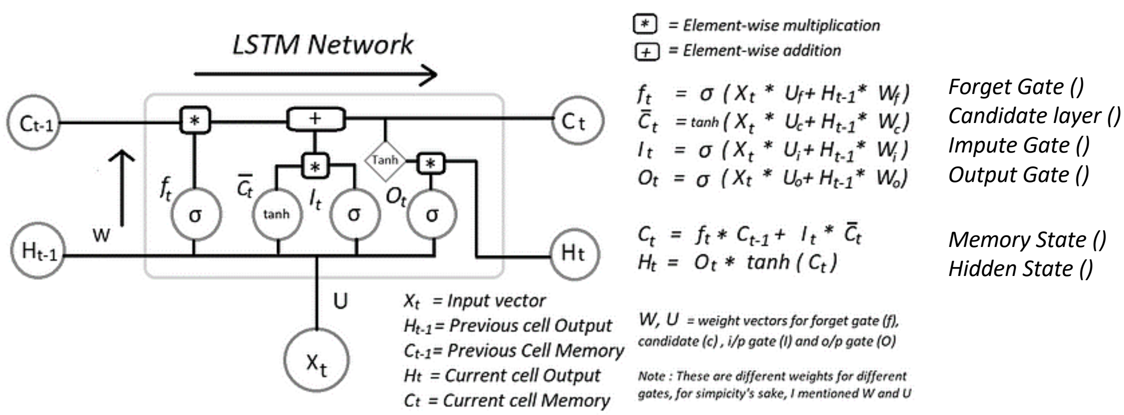


Figure. 1. LSTIM diagram at T time step

Note that the above diagram explains LSTM at one time step. This means that the equations will be recomputed for the next time step. The weight matrices, ($W_t, W_i, W_o, W_c, U_f, U_i, U_o, U_c$), and bias (not noted in the equation but have subsequent b_f, b_i, b_o, b_c), are not time dependent, meaning the calculated output of different timesteps uses the same weight matrices. (M Rastogi, 2020)

2.2 Performance Metrics

Performance of the forecasting methods will be evaluated with regression metrics (Loss Functions) to ensure the model is refined, while not overfitting the training data, to build effective and relevant forecasting solutions. Performance evaluation regression metrics include the following: (M Padhma, 2024)

- Mean Absolute Error (MAE): Simple to calculate, provides an even measure of how well the model performs, and does not penalize high errors caused by outliers. However, it is not differentiated at zero and has a scale-dependent accuracy measure.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

- Mean Squared Error (MSE): Expressed in a quadratic equation, hence a gradient descent with only one global minima. Sensitive to outliers and penalizes larger errors due to squaring. When there is a bad prediction, the sensitivity to outlier magnifies the high errors, and is scale dependent.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Root Mean Squared Error (RMSE): intuitive measure of model accuracy and easy to interpret.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

where: y_i = actual, $\hat{y}_i$ = predicted, n = sample size

3 Exploratory Data Analysis (EDA)

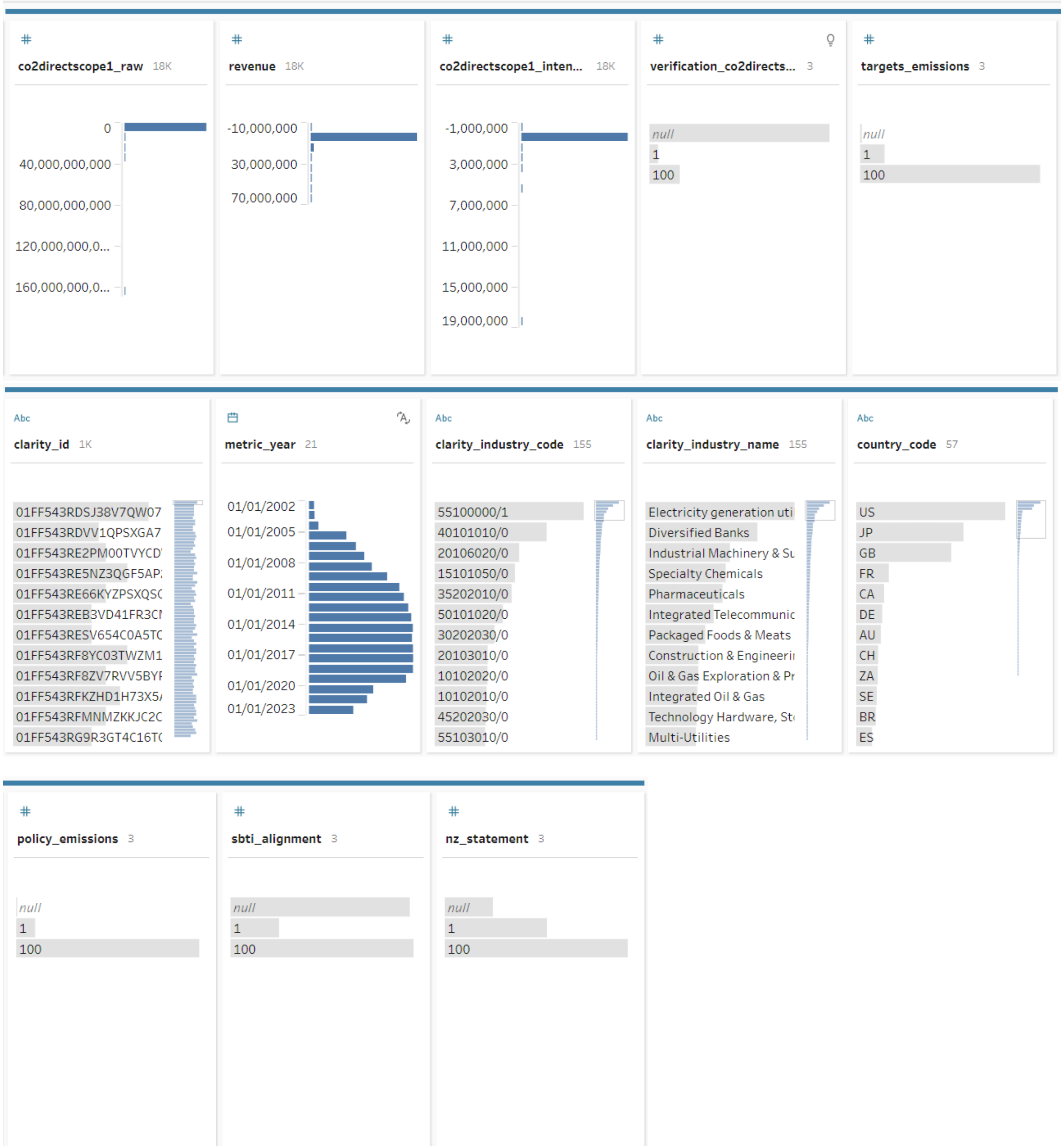
3.1 Individual Feature Assessment

The forecasting methods implemented in this paper attempt to forecast Scope 1 direct emissions from public companies who have disclosed their Scope 1 emissions from 2002 to 2022. The training data has fifteen attributes to describe each row of data. Four (4) quantitative and eight (8) categorical features are summarized in the following table along with the bar charts to show the spread of the data:

Feature	Description	Unit	Metric Type	Comment
clarity_id	ClarityAI's unique identifier of a company	N/A	N/A	1372 unique company id exists

`metric`	CO2DIRECTSCOPE1	N/A	N/A	Data has been prefiltered to Scope 1 Direct emission value only
`metric_year`	The year scope 1 emission was reported	N/A	Quantitative	Ranges from 2002 to 2023. Not all companies have complete data from 2002 to 2022. Majority data centered around 2009-2019.
`provider_code`	TR“XX”	N/A	N/A	Not relevant for forecasting
`clarity_industry_code`	Unique industry code	N/A	Categorical	155 unique industry code
`clarity_industry_name`	Industry segmentation	N/A	Categorical	155 unique industry name
`country_code`	Two-letter country codes defined in ISO 3166-1 alpha-2	N/A	Categorical	57 unique country code, country code references company’s headquarter office location
`co2directscope1_raw`	Direct emission that occur within a company’s organization boundary given year	$MtCO_2e$	Quantitative	
`revenue`	Total amount of revenue gained by the company in the given year	Million USD	Quantitative	18204 unique revenue values, 1 value in
`co2directscope1_ intensity`	Scope 1 emission that occur within a company’s organization boundary	$MtCO_2e$	Quantitative	CO2 direct scope 1_raw divided by Revenue
`verification_ co2directscope1`	Have the Scope 1 HGH emissions of a company been verified by a third party?	Yes/no (100/1)	Categorical	13.6% (2500) of population has 3rd party verification
`targets_emissions`	Has the company set targets or objectives to be achieved on emission reduction?	Yes/no (100/1)	Categorical	87% (15,966) has target emission, while 12% (2,164) does not have target emission, <1% (153) with null value
`policy_emissions`	Does the company have a policy to improve emission reduction?	Yes/no (100/1)	Categorical	91% (16,560) have policy to improve emission reduction while, 9% (1,701) does not and <1% (22) with null
`sbti_alignment`	Has the company made a commitment to the SBT Initiative?	Yes/no (100/1)	Categorical	45% (8,146) made commitment to SBTi while12% (2,160) does not with 44% (7,977) null
`nz_statement`	Indicates it the organization has any kind of target, goal, pledge towards achieving net zero	Yes/no (100/1)	Categorical	55% (10,034) made net zero statement, while 31% (5,606) did not with 14% (2,643) null

<Figure 2. Features in 'training_data.csv'>



<Figure 3. Tableau Prep: Bar chart of individual features in 'training_data.csv'>

Feature-specific commentary:

- `clarity_industry_name`: Emission reports in ‘Electricity Generation Utilities’ are most frequent, accounting for 5% (965 rows) of the population, followed by ‘Diversified Banks (4%, 721)’, ‘Industrial Machinery & Supplies & Components (3%, 544)’, ‘Specialty Chemicals (3%, 517)’ and Pharmaceuticals (3%, 496)’. Industry type could be categorized into two categories to improve the forecasting accuracy: energy intensive (production) industry and low energy intensive (service) industry.

- `nz_statement, sbti_alignment, targets_emissions & policy_emissions`: Companies tend to announce reduction target and internal policy first, then promising Net Zero or obliging to SBTi. While it is easier to set a Net Zero statement (target, goal, pledge) with varying degrees of legal enforcement, companies have a difficult time aligning to the Science Based Target Initiative (SBTi). The latter is defined by CDP, United Nations Global Compact, World Resource Institute and the World Wide Fund of Nature as 'best practices in emissions reductions and net-zero targets in line with climate science'.
- `verification_co2directscope1`: Approximately 14% of the data are third-party verified, and the earliest entries begin in 2018. A company is highly recommended to get the emission data validated by a third party assurance if SBTi commitment to identify any discrepancies and ensuring continuous improvement of the data quality. In this study we treat 'null (82%, 15034)' as the same value of 'not verified (4%, 722)'.

The visual inspection of the spread of the 'training_data.csv' shows that companies intend to reduce carbon emissions by stating emission reduction targets and establishing internal policies. However, the execution and rigor to drive Scope 1 direct emissions down to net zero by practicing SBTi's industry-specific best practices with third-party validation present slow progression toward achieving the Paris Climate Agreement of 'limit the temperature increase to 1.5˚C above pre-industrial levels'. (UNCC, n.d.) While companies may not choose to align with the SBTi method, third-party validation of emission data to track their progress is often observed, resulting in higher accuracy and transparency.

3.2 Largest Emitters (Absolute CO2)

Twenty companies account for the top two hundred largest record emissions. To understand the characteristics of large emitters, the top eight emitters were observed in detail.

- Finding 1: Despite the progress these carbon-intensive industries, such as Integrated Oil and Gas, Electricity Generation Utility, Steel Making, and Cement Producers, are making, their absolute emission values are not decreasing. This persistence of high emissions in these sectors, which we refer to as 'hard-to-abate,' is a significant factor in our forecasting efforts.
- Finding 2: The reporting of direct emissions by companies is inconsistent. Voluntary disclosures can lead to companies reporting every two years or having irregular reporting periods. For instance, a Cement Producer in China (Clarity ID: '01FF54…MBCBZ') reported yearly between 2004 and 2020 but has stopped since 2021. It is challenging to determine whether this is due to the company going bankrupt during COVID-19, not being able to afford direct emission reporting or other reasons. Economic hardship can lead to a deprioritization of direct emission reporting.
- Finding 3: Direct emissions do not reduce by the same order of magnitude simply because a company may experience financial hardship (poor revenue). Taking the Multi-Utilities company in Denmark (Clarity ID: '01FF54…..YXKF3') as an example, the company's revenue fell by 67% while the raw emission level dropped by only 10%.

A complete analysis of the top 8 largest emitters is presented in Appendix 1 in detail.

3.3 Carbon Intensity (CO2 emission to Revenue ratio)

Observing the 'Carbon Intensity' attribute in descending order identifies abnormality and pins potential outliers. This leads to two additional assessment criteria: industry peer-group assessment and peak year assessment, which are used to identify abnormality. A box plot is used to decide whether to exclude outliers.

A detailed analysis of the top 4 carbon-intensive companies is presented in Appendix 2.

3.4 Handling Missing Data

Working with real-world data often presents the problem of missing data. Scope 1 Direct Emissions are not an exception, especially since the emission reporting is voluntary. As highlighted in 'EDA 1: Individual Feature Assessment', not all companies report their emissions yearly. Additionally, with the 'Company' name represented by 'clarity_id', is it not possible to track whether the company still exists or if a business has not reported the direct emission value for that year while the 'revenue' and other categorical attributes are still present. Given the industry and country-dependent nature of carbon intensity, imputing the missing data may not significantly improve the results or be deemed more 'accurate' unless the original data owner, the company, verifies the direct emission value. For simplification of the project and working with the provided data at face value, this project does not impute missing years' emission values of a company, nor the revenue and direct emission in between missing years, while acknowledging the pros and cons of imputing data and its impact on statistical analysis. (E Raheem, 2024)

Handling missing data by imputation or omission is crucial for maintaining the validity and reliability of the modeling results. Consideration factors are:

1. Amount of missing data and understanding why the data is missing (D Cermak, 2022)
2. The mechanism behind the missing data
3. Potential impact on the analysis and how this data will be used

The literature strongly recommends discussing the extent of missing data and the limitations. It also suggests conducting a sensitivity analysis to determine if missing data significantly impacts the results. This type of analysis, often done via best-worst case analysis, provides a robust way to assess the reliability of the analysis (J Jakobsen, 2017).

Note Carbon Intensity (`co2directscope1_intensity`) is a byproduct of the direct emission divided by the revenue.

Categorical Features: All the categorical features (`verification_co2directscope1, targets_emissions, policy_emissions, sbti_alignment, nz_statement`) assume 'no' (represented by 1 in the raw data, where 100 represents 'yes') if null for the particular year's data. However, it does not impute the in between years. (W Lin et al, 2019)

3.5 Outliers & Boxplots

Visual inspection from Figure 4 highlighted two extreme outliers in the data set. These outliers were primarily detected by a bar chart of `co2directscope1emission` and `co2directscope1_intensity` attributes.

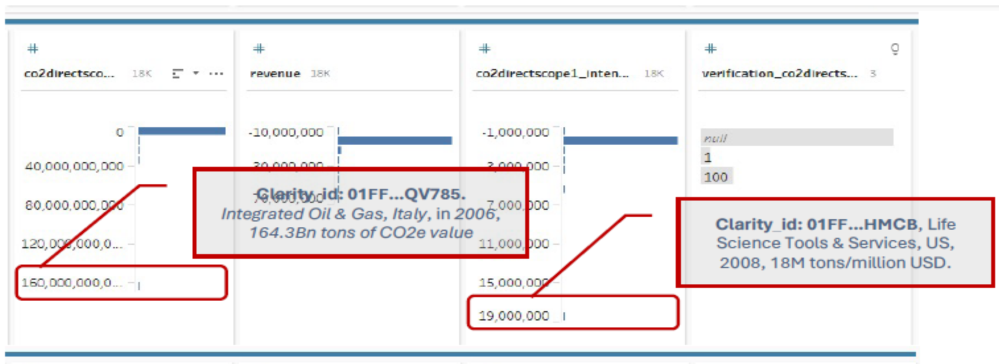


Figure 4. Outliers in Direct Emission and Carbon Intensity

Clarity_id: 01FF…QV785. Integrated Oil & Gas, Italy, in 2006, 164.3Bn tons of CO2e value
Clarity_id: 01FF…HMCB, Life Science Tools & Services, US, 2008, 18M tons/million USD (Across 2008 – 2015)

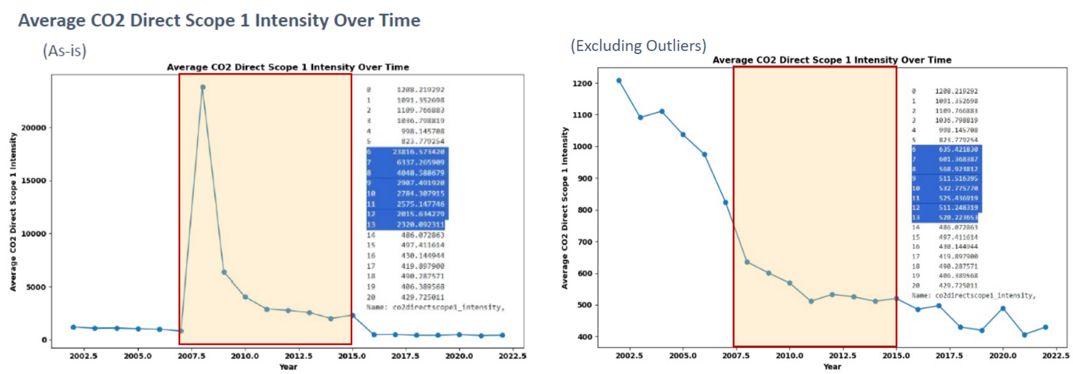


Figure 5. Average Direct Emission with Life Science company (01FF...HMCB) - Before and After exclusion

Further assessment has been conducted using Boxplots to define the outlier threshold. This study assumes data points outside of the ninety-eighth (98) percentile are excluded from method development. Skewed boxplots in both raw emission and intensity charts are shown below:

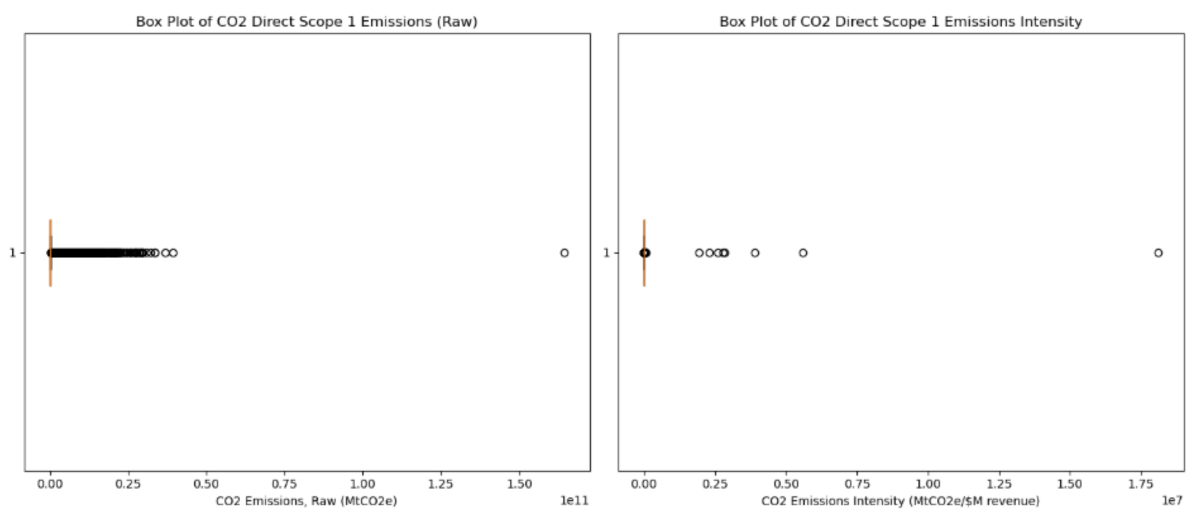


Figure 6. Outlier assessment using Boxplot

The visualization of the outlier removal and its impact on the overall Direct Scope 1 Emission (raw) and Average Carbon Intensity, which can be found in Appendix 3, is a significant aspect of our data analysis. Despite the low average intensity and total sum of direct emission values, these figures are of utmost importance in our trend analysis and forecasting results.

Supplementary Boxplot for EDA 5: Outlier Assessments are included in Appendix 4: 4.1 'Carbon Intensity by Industry type', 4.2 'Direct Emissions by Industry type', 4.3 'Revenue by Country' and 4.4 'CO2 Emission by Country'.

3.6 Feature Correlation Heatmap

The correlation heatmap indicates that more companies are getting their direct emissions verified by a third party for transparency and accuracy YoY. Net Zero Statement and SBTi commitment variables go hand in hand. However, the impact of SBTi-driven emission reduction is not indicative, resulting in the lowest correlation value of -0.098.

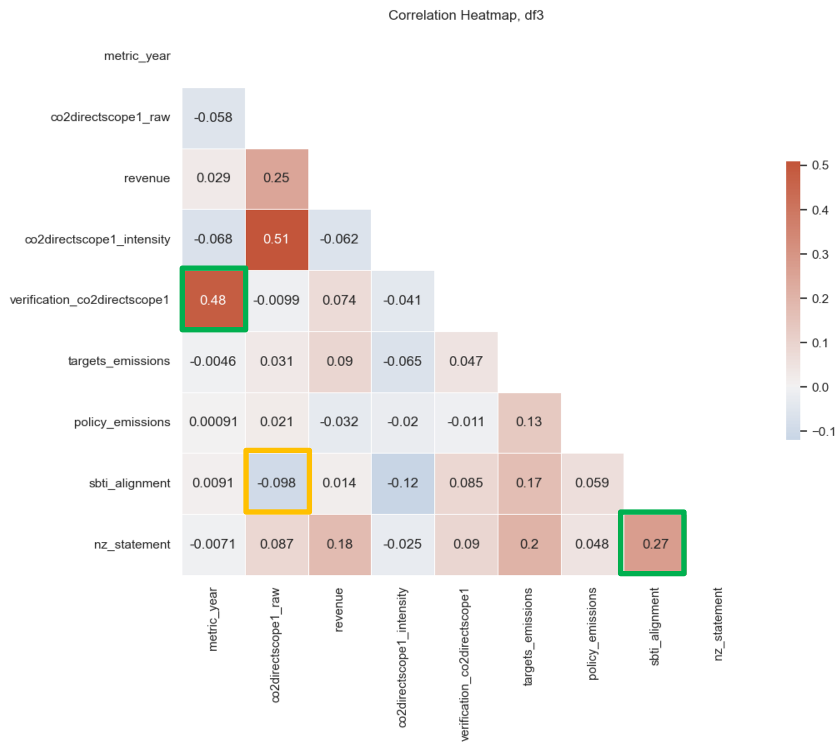


Figure 7. Correlation Matrix: Sustainability Feature Heatmap

4 Feature Selection

Feature selection helps identify the most relevant predictors, reducing dimensionality, mitigating overfitting, and enhancing model interpretability. Of the given fifteen features, eleven were considered due to the following reasons: `metric` and `provider_code` carry no values in indicating trend; `clarity_industry_code` is removed as a functional duplicate of `clarity_industry_name`; and `clarity_id` is simply an id, not a feature of the data.

Our research leverages Lasso Regression and Random Forest Regression models for feature selection. Lasso Regression, with its L1 regularization, is a key player in variable selection and model simplicity improvement (U Michelucci, 2024). On the other hand, Random Forest Regression, with its ensemble of decision trees, each on a random subset of features and data samples, evaluates the contribution of each feature to the model's predictive power via the Gini coefficient.

While these approaches are straightforward for numerical features, the categorical features require a decision among a few options: 1) assigning numeric values, 2) one-hot encoding, 3) mean replacement technique, and 4) domain knowledge. (D. Jain, 2020) One-hot encoding was selected as a simple and intuitive method, which converts a single column of categorical data into many columns – one for each category – with a binary indicator for a true or false membership of that category (maximum 1).
In our research, we used these two methods to rank the effect of each variable by its influence on intensity. Intensity is chosen because it is an interaction of the other two desired values being forecast, ultimately attempting to simultaneously capture the effects on these two measurements.

Our analysis isolates the binary variables (Figure 8 and Figure 9 below), consistently ranking verification and Net Zero statements as the most important binary indicators. Some of our models are restricted to these key factors to measure their effect. Following these are emission reduction policy, SBT Initiative commitment, and setting emission reduction targets.

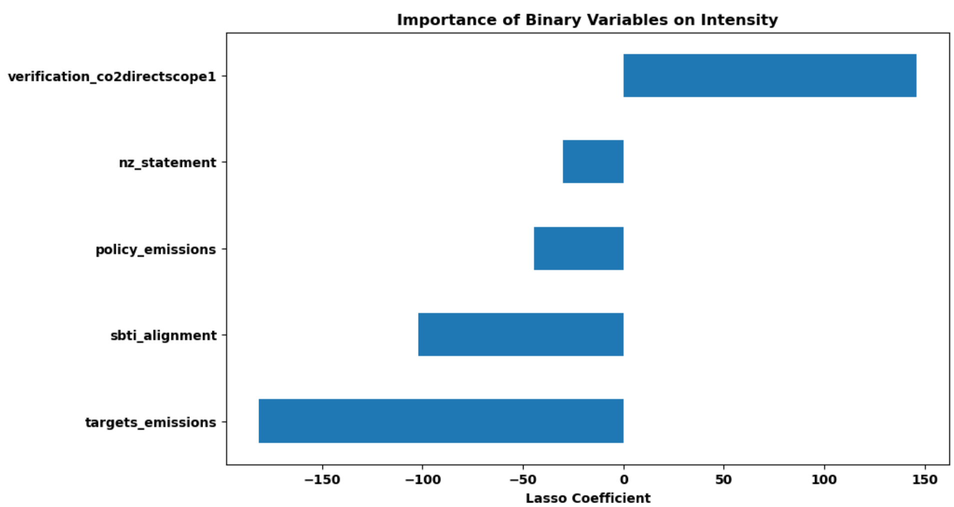


Figure 8. Lasso Regression on all binary variables. R^2 *achieved: 0.685%*

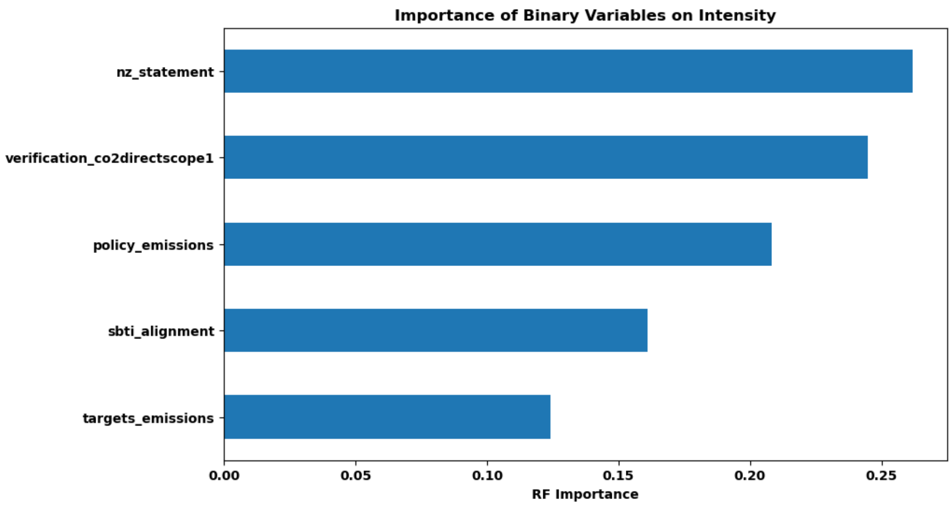


Figure 9. Random Forest Regression on all binary variables. R^2 achieved: 3.31%

Understanding the above effects is important because they are the only factors for which a company can be held accountable out of all the features available in the dataset. However, they are not the most influential factors, as seen in the comparison below against the country in which the company is headquartered and the industry ClarityAI has categorized them into.

The below approaches highlight in each chart -- and more importantly, their R^2 values -- the criticality of considering industry and country as predictive factors in intensity values; for example, knowing that a company is a Cement Producer, or that they are headquartered in India, allows one to anticipate higher intensity values compared to companies in other categories. Overall, these categories effectively mirror the influence of country-specific policies or industry-specific standards.

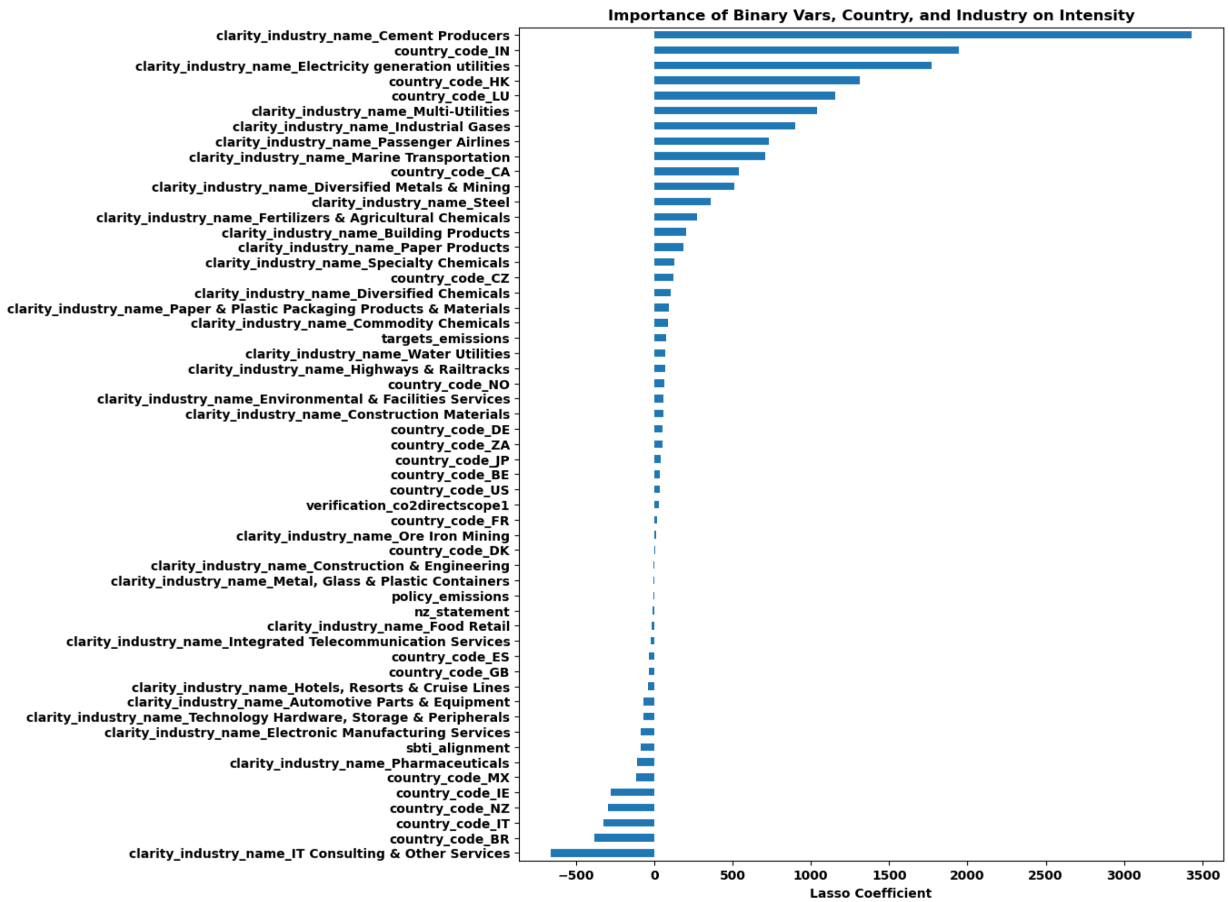


Figure 10. Lasso Regression on all binary, country, and industry variables, excluding those reduced to 0 by Lasso selection. R^2 achieved: 63.83%

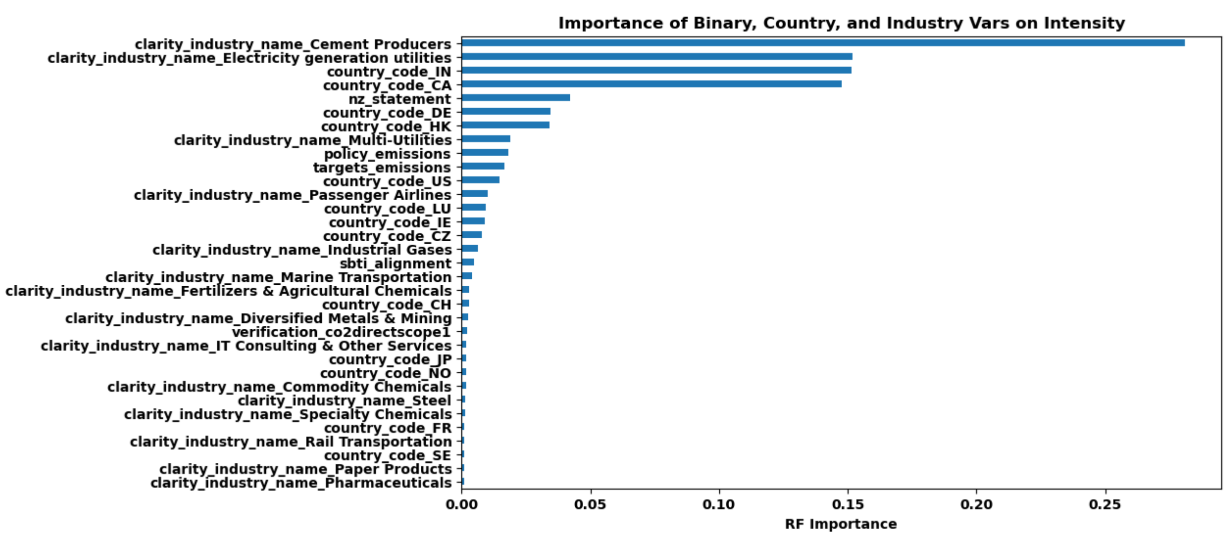


Figure 11. Random Forest Regression on all binary, country, and industry variables, excluding those reduced to less than 0.001 by RF selection. R^2 achieved: 79.12%

Influence

Our research seeks to answer these 'key questions' from stakeholders at ClarityAI: What are the key drivers and factors influencing the variability of sustainability indicators over time, and how can forecasting models effectively capture them?

Below are several plots that illustrate the effect of each binary indicator on each of the sustainability metrics, including whether a company has set emission reduction targets, implemented a policy to improve emission reduction, committed to the SBT Initiative, pledged Net Zero emissions, or verified its emissions with a third party.

The first triplet of plots displays the mean value of each sustainability metric by their target emission reduction status, as well as the mean of the dataset. The intensity plot shows a steady pattern of higher intensity from companies that have not set these targets (or have null values, which may as well be a false value). However, intensity does not tell the whole story for this metric: the plot of raw emission values shows that companies that have set these targets had consistently higher emissions for the first 20 years' worth of data, which may conflict with an intuitive expectation of this metric. This indicates that perhaps the companies that set these targets were the worst offenders in the greatest need for emission reduction. The data show a steady downward trend, ultimately achieving a mean value below companies that do not set these targets for emission reduction. Importantly, the companies that have set these targets successfully reduce their emissions without compromising their revenue, which has been generally upward since 2015.

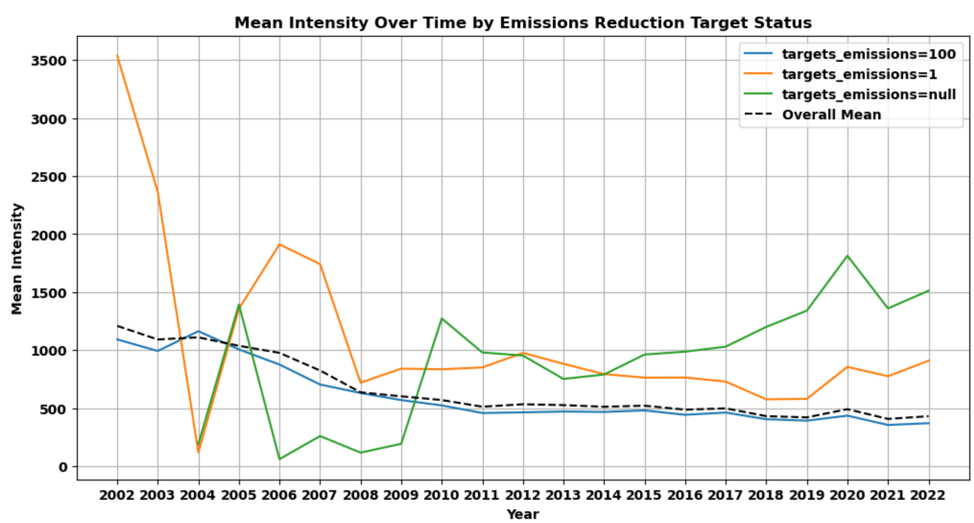


Figure 12. Annual intensity by emission reduction target status.

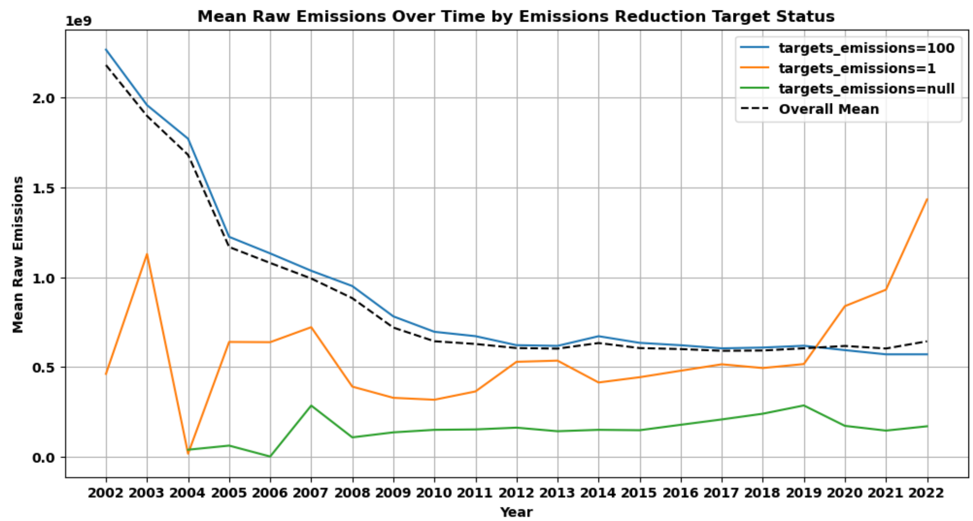


Figure 13. Annual emissions by emission reduction target status.

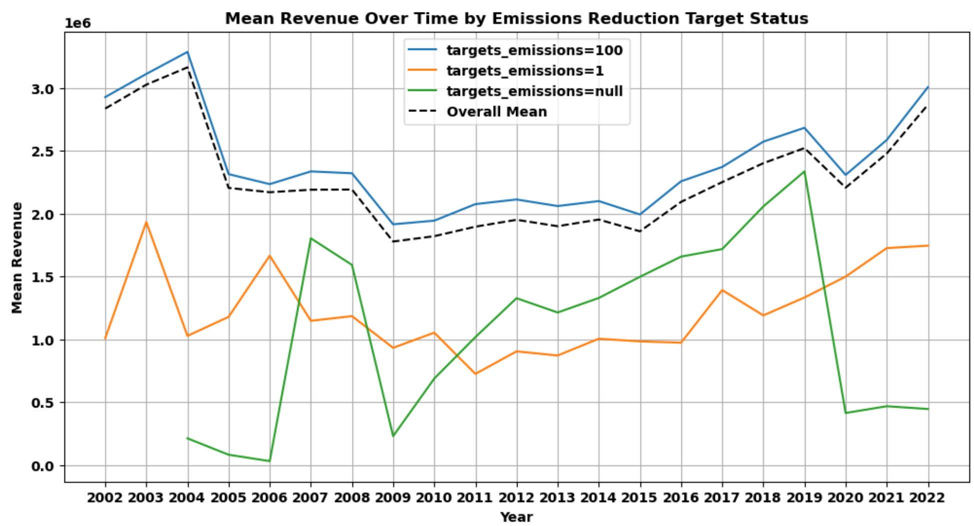


Figure 14. Annual revenue by emission reduction target status.

The next triplet of plots is less informative due to the spikiness in the charts, presenting difficulty in any conclusions around long-term trends. It is shown, however, that companies with a policy in place steadily decrease their intensity in small increments over time, while companies without this policy are less predictable in their intensity, and predictability is a valuable trait for forecasting. The trends are even more dire for companies with null values. A potential conclusion to be drawn from this is that, instead of using this as a predictive variable, it could be used as a minimum requirement for the dataset in order to reduce variance.

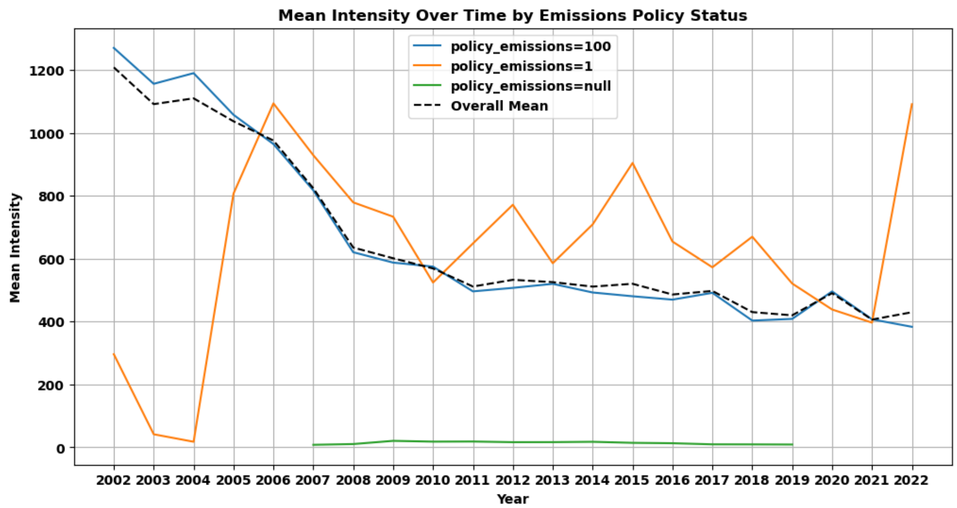


Figure 15. Annual intensity by emission reduction policy status.

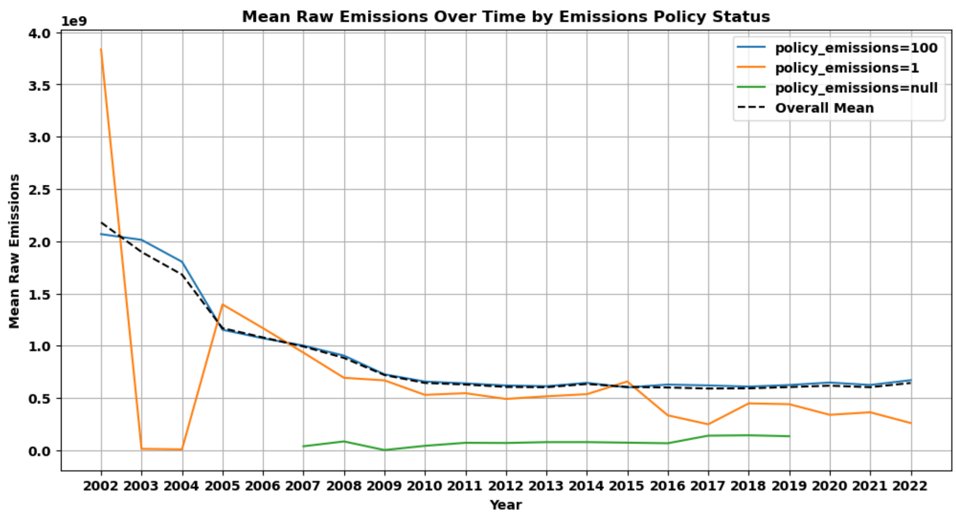


Figure 16. Annual emissions by emission reduction policy status.

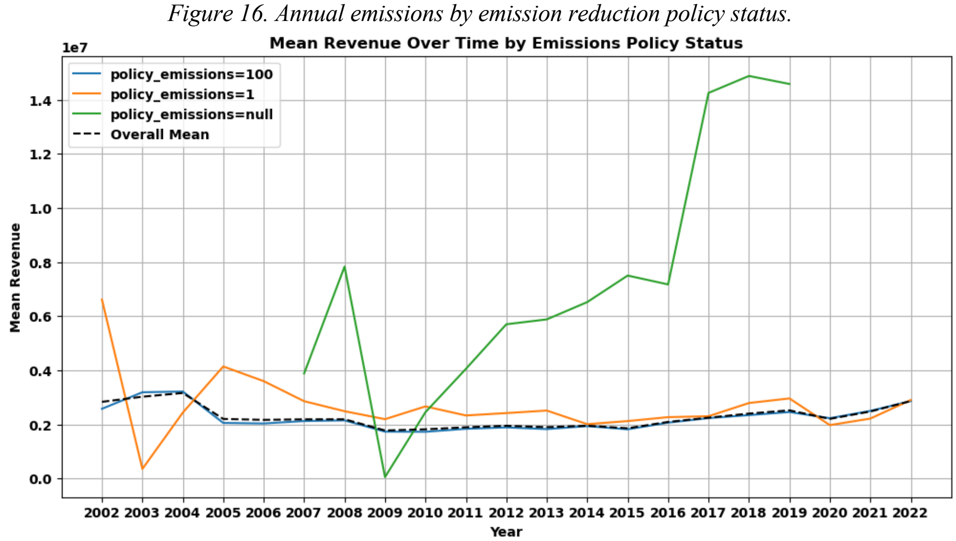


Figure 17. Annual revenue by emission reduction policy status.

The following triplet of plots show roughly the same trends regardless of commitment to the SBT Initiative. Across all statuses, intensity and emissions trended down from the peaks of the early 2000s, and revenue has been steadily increasing since 2015. Intuitively, the lack of differentiation between these labels aligns with the feature selection algorithms, which show that commitment to this initiative is not a very influential factor in predicting sustainability metrics; this is illustrated by a larger delta between the null values and the false values than the false values and the true values.

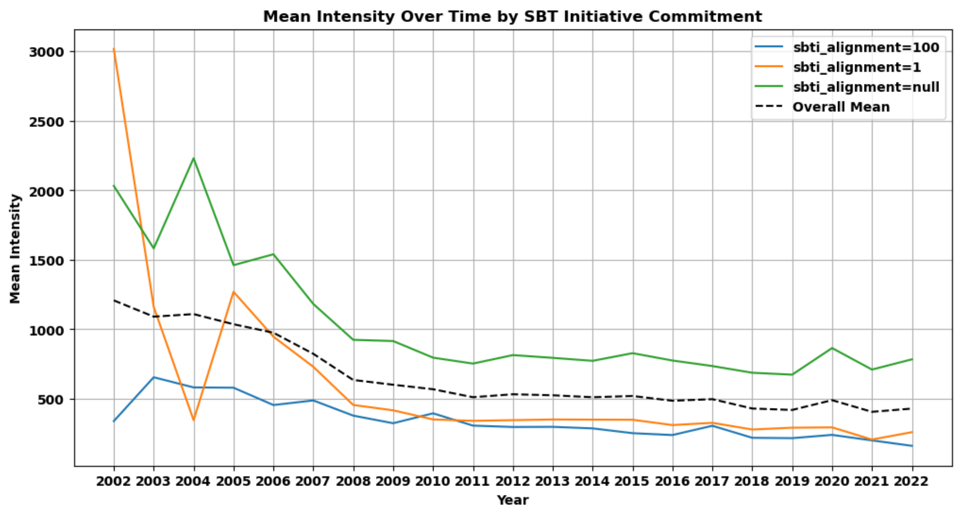


Figure 18. Annual intensity by SBT Initiative commitment status.

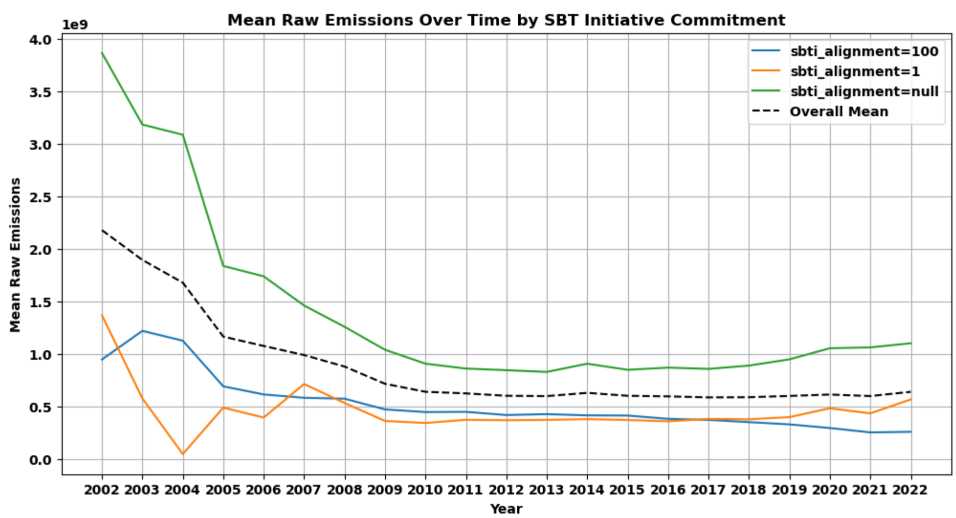


Figure 19. Annual emissions by SBT Initiative commitment status.

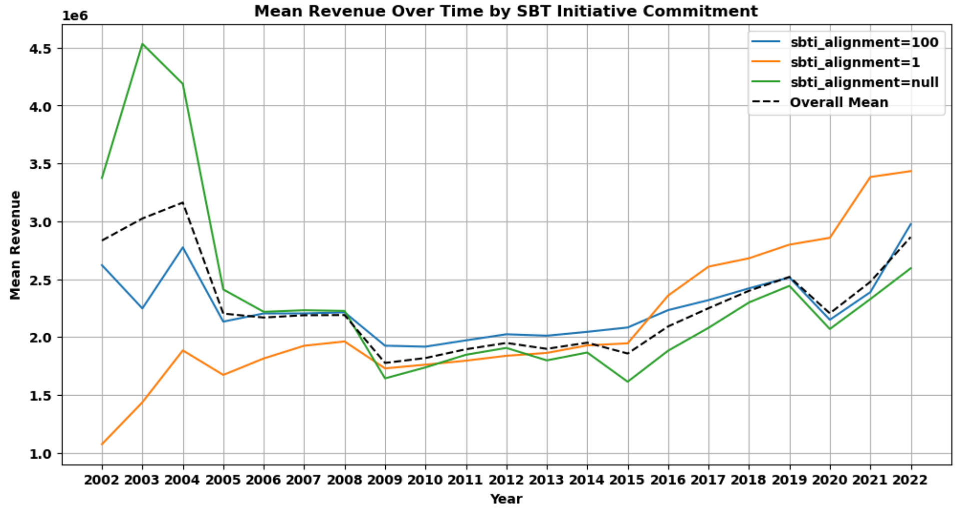


Figure 20. Annual revenue by SBT Initiative commitment status.

The triplet plots for Net Zero pledge status are the most interesting among this exercise, particularly in the raw emissions chart. It shows a clear pattern: companies that have pledged to achieve global net zero emissions consistently emit more Scope 1 emissions than companies that have not made this pledge. This reinforces a possibility presented in the first triplet of emission reduction target status plots: the companies that set these goals are the largest culprits. While there has been a steady downward trend into relative stationarity for the last decade, the gap hardly appears to be closing between companies that have pledged and those that have not. The intensity remains tightly coupled between the two categories. However, critically, Net Zero Emissions is an absolute goal -- not a ratio like intensity -- and companies that have committed to absolute zero are further from it than those that have not.

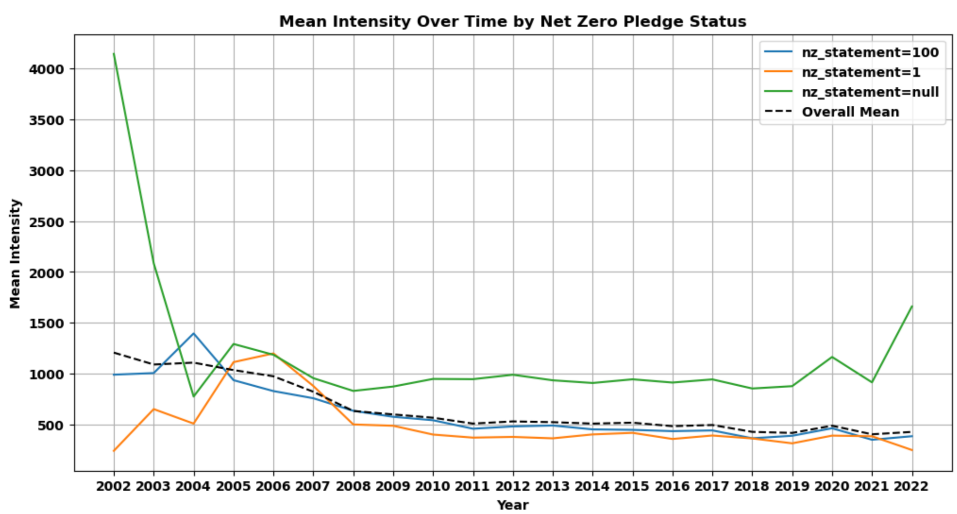


Figure 21. Annual intensity by Net Zero pledge status.

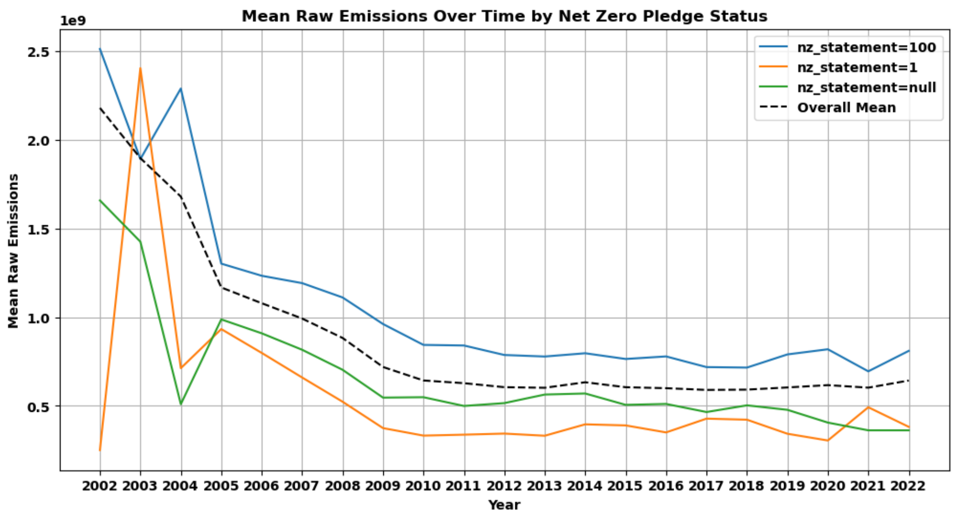


Figure 22. Annual emissions by Net Zero pledge status.

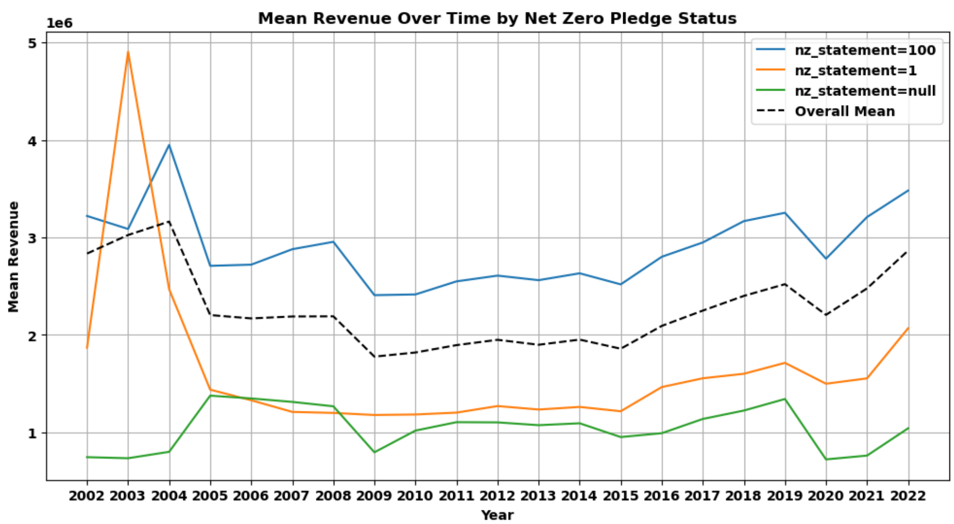


Figure 23. Annual revenue by Net Zero pledge status.

Lastly, the triplet of third-party verification status plots provides more unintuitive evidence that the companies that are most visibly committed to controlling (or, in this case, monitoring) emissions are the largest contributors. While there are fewer years of data for verification status (which are only present from 2018 onward), verified companies are setting a trend of higher intensity and higher emissions than the unverified.

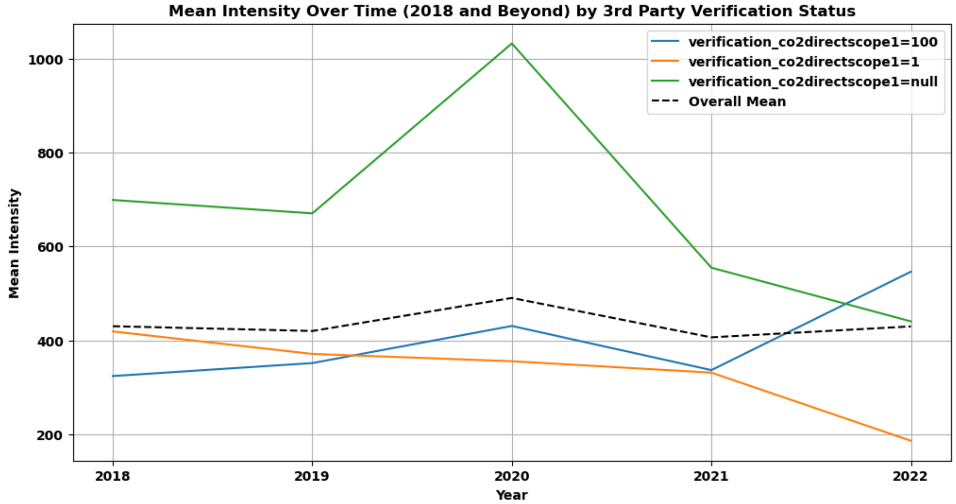


Figure 24. Annual intensity by third party verification status.

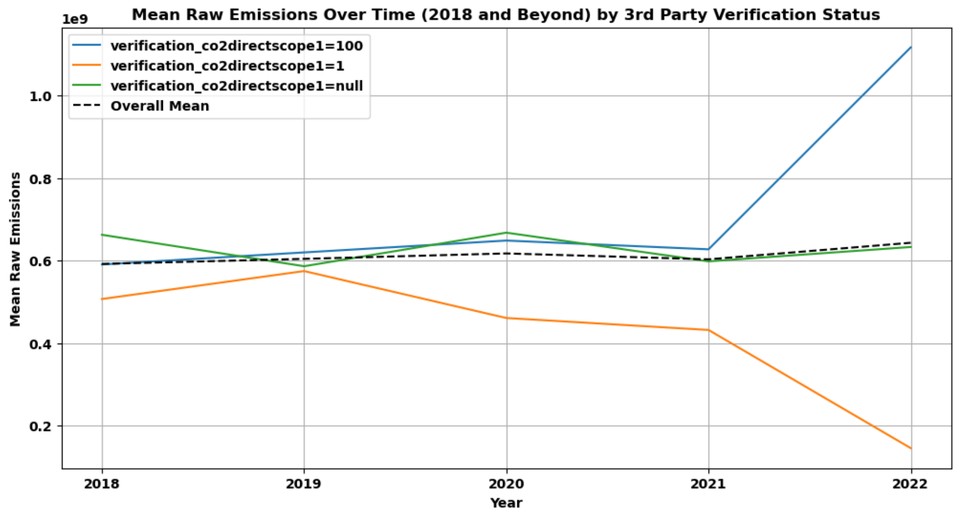


Figure 25. Annual emissions by third party verification status.

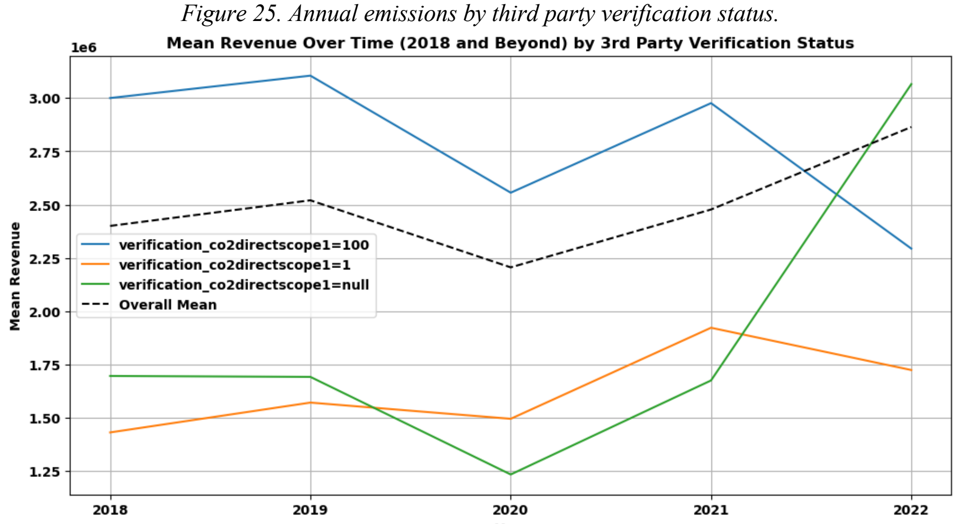


Figure 26. Annual revenue by third party verification status.

5 Methods

For the implementation, TensorFlow was used for deep learning applications. For the data preparation and visualization of the results, Scikit-learn, Numpy, Seaborn, and Matplotlib were used.

From all the algorithms implemented in this study, variations of LSTM Multivariate and Exponential Smoothing Univariate, offered the best results for forecasting.

5.1 ARIMA

The ARIMA model was implemented using Python's `statsmodels` library. Key steps included making the data stationary using differencing techniques, selecting the best parameters (p, d, q) using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), and fitting the model to the historical CO2 emissions data. The model was then used to forecast future emissions.

5.2 Exponential Smoothing / Holt-Winters Smoothing

The Exponential Smoothing model was implemented using Python's `statsmodels` library. Key steps included ensuring the data exhibited trends and seasonal patterns, selecting the Holt-Winters method to capture level, trend, and seasonality, and fitting the model to the historical CO2 emissions data. The model was then used to generate future forecasts.

5.3 Prophet

The Prophet model was implemented using Python's `prophet` library. Key variables included historical Scope 1 emissions and revenue. The model parameters were tuned to account for seasonal patterns and holiday effects relevant to the dataset. Prophet's strength lies in its ability to handle missing data and adapt to trend shifts, making it suitable for the volatile nature of emission data.

5.4 XGBoost

The XGBoost model was implemented using Python's `xgboost` library. Key steps included preparing the data with historical CO2 emissions and revenue, fine-tuning parameters such as the number of trees, learning rate, and max depth using cross-validation, and fitting the model to the historical data. The model was then used to generate future forecasts.

5.5 LSTM

Long short-term memory, or LSTM, is a recurrent neural network, or RNN, that is more effective at forecasting long-term trends compared to other RNNs. However, the same attributes that enable LSTM to retain long-term trends also demand a relatively longer sequential data set (a minimum of 10) than other methods.

The LSTM model was implemented using Python’s TensorFlow library. Pre-processing steps included identification of variables that meet the minimum sequence requirements and tuning the number of neural net layers, neurons per layer, and other factors that generally scale favorably with model performance, but at the cost of computation demands.

6 Results

The goal of this study is to enhance the predictive accuracy of sustainability indicators, in particular scope 1 emission (direct emission) for mid-range forecasting. Five time-series and machine learning algorithms were analyzed and implemented (using the tools mentioned in section 3) to have a solid comparison.

UNIVARIATE MODELS

6.1 ARIMA

The results indicate that ARIMA provides reliable forecasts for Scope 1 emissions. The model's performance was evaluated using MAE, RMSE, and MAPE.

Key results from the ARIMA model:

- **MAE (CO2)**: 85,630,948
- **RMSE (CO2)**: 612,495,824.50
- **MAPE (Accuracy) (CO2)**: 82%
- **MAE (Revenue)**: 312,499.45
- **RMSE (Revenue)**: 895,632.78
- **MAPE (Accuracy) (Revenue)**: 80.25%

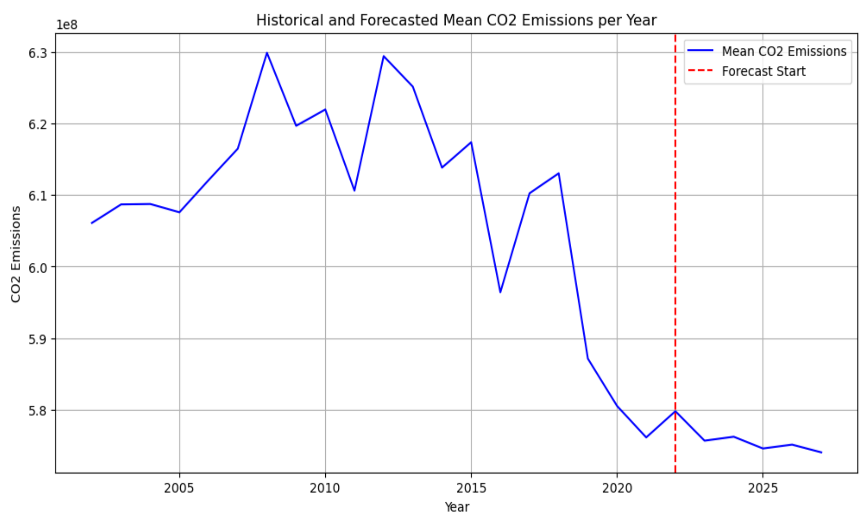


Figure 27: Historical and Forecasted Mean CO2 Emissions per Year (ARIMA)

The graph illustrates the historical and forecasted mean CO2 emissions per year using the ARIMA model. The blue line represents the actual CO2 emissions data from previous years, while the red dashed line marks the start of the forecast period. From 2005 to around 2015, CO2 emissions exhibit an upward trend with notable fluctuations, indicating periods of increase and some decrease. Following 2015, there is a noticeable decline in CO2 emissions, which continues until the forecast start point in 2020.

After 2020, the forecasted emissions display a slight downward trend, suggesting a continued decrease in mean CO2 emissions. This forecast maintains the pattern of decline observed in the recent historical data, reflecting the model's ability to capture the underlying trend accurately. The ARIMA model appears to have effectively captured the overall trend and fluctuations in the historical data, which is crucial for generating reliable forecasts. The smooth transition from historical data to the forecasted period indicates that the model has managed to align the forecast closely with the observed data patterns.

6.2 Exponential Smoothing / Holt-Winters Smoothing

The results indicated that smoothing provided the best forecasts for Scope 1 emissions. The model's performance was evaluated using MAE, RMSE, and MAPE.

Key results from the Exponential Smoothing model:

- **MAE (CO2)**: 54,282,066
- **RMSE (CO2)**: 455,284,123.30
- **MAPE (Accuracy) (CO2)**: 87.5%
- **MAE (Revenue)**: 178,954.32
- **RMSE (Revenue)**: 657,832.10
- **MAPE (Accuracy) (Revenue)**: 85.62%

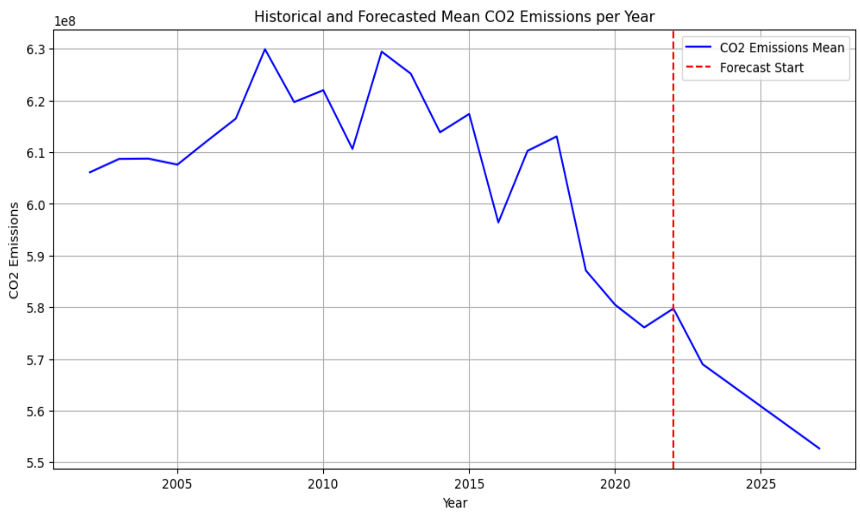


Figure 28: Historical and Forecasted Mean CO2 Emissions per Year (Smoothing)

The graph illustrates the historical and forecasted mean CO2 emissions per year using the Exponential Smoothing (Holt-Winters) model. The blue line represents the actual CO2 emissions data from previous years, while the red dashed line marks the start of the forecast period. From 2005 to around 2015, CO2 emissions show an overall upward trend with noticeable fluctuations, indicating periods of increase and decrease. Following 2015, there is a significant decline in CO2 emissions, continuing until the forecast start point in 2020.

After 2020, the forecasted emissions exhibit a continuous downward trend, suggesting a further decrease in mean CO2 emissions. This forecast mirrors the pattern of decline observed in the recent historical data, reflecting the model's capability to capture underlying trends and seasonal variations effectively. The Exponential Smoothing model has managed to align the forecast closely with the observed data patterns, demonstrating its strength in handling stable trends and seasonal effects. However, abrupt changes in trends can pose a challenge, which might impact the accuracy for certain datasets.

6.3 XGBoost

The results from the XGBoost model were not as expected. It did not show a good performance in handling large volumes of data and capturing complex patterns in the emissions data. The model's accuracy was assessed using MAE, RMSE, and Mean Absolute Percentage Error (MAPE). XGBoost provided the lowest MAPE, indicating its effectiveness in forecasting Scope 1 emissions with high accuracy.

Key results from the XGBoost model:

- **MAE (CO2)**: 107,503,466.75
- **RMSE (CO2)**: 489,531,865.20
- **MAPE (Accuracy) (CO2)**: 59.11%
- **MAE (Revenue)**: 562,625.22
- **RMSE (Revenue)**: 1,799,433.94
- **MAPE (Accuracy) (Revenue)**: 72.47%

Several graphs generated during the analysis illustrated the model's performance and accuracy in forecasting Scope 1 emissions. These graphs visually demonstrated the model's capability to handle large datasets and complex patterns effectively.

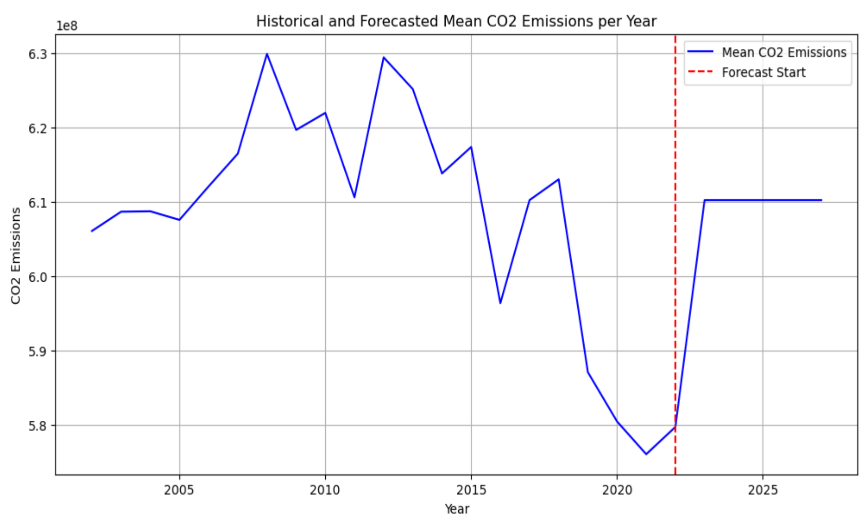


Figure 29. Historical and Forecasted Mean Revenue per Year (XGBoost)

This graph illustrates the historical and forecasted mean CO2 emissions per year using the XGBoost model. The red dashed line represents the forecast start. The model forecasts a significant change in emissions trend after the forecast starts.

6.4 Prophet

These results indicate that Prophet provides reliable mid-term forecasts for Scope 1 emissions. The model's performance was evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Prophet demonstrated robust performance and maintained a high level of accuracy, even with incomplete data.

Key results from the Prophet model:

- **MAE (CO2)**: 65,281,774.50
- **RMSE (CO2)**: 342,944,318.10
- **MAPE (Accuracy) (CO2)**: 84.28%
- **MAE (Revenue)**: 251,090.78
- **RMSE (Revenue)**: 775,086.12
- **MAPE (Accuracy) (Revenue)**: 82.53%

Additionally, several graphs generated during the analysis illustrated the seasonal effects and trends captured by the Prophet model. These graphs provided a visual representation of the model's ability to forecast Scope 1 emissions accurately over time.

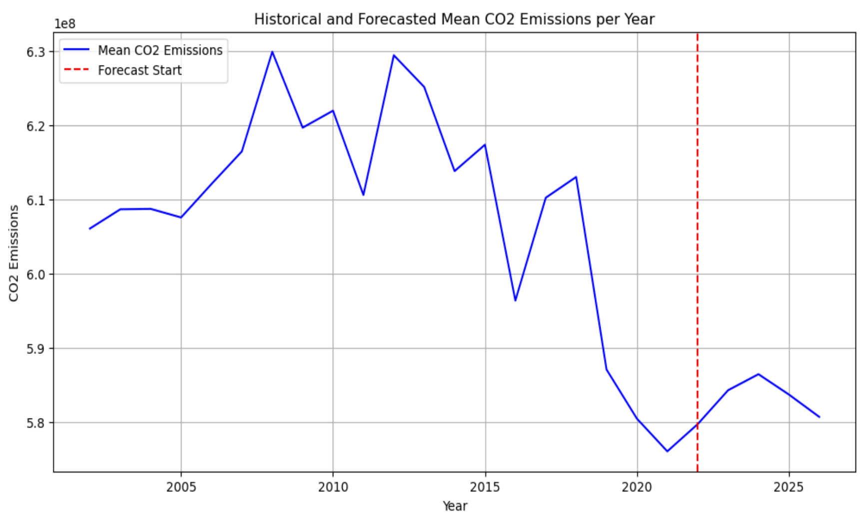


Figure 30: Historical and Forecasted Mean Revenue per Year (Prophet)

The graph illustrates the historical and forecasted mean CO2 emissions per year using the Prophet model. The blue line represents the actual CO2 emissions data from previous years, while the red dashed line marks the start of the forecast period. From 2005 to around 2015, CO2 emissions show an upward trend with noticeable fluctuations, indicating periods of both increase and decrease. Following 2015, there is a decline in CO2 emissions, which continues until the forecast start point in 2020.

After 2020, the forecasted emissions exhibit a mixed trend with some fluctuations, suggesting periods of both increase and decrease in mean CO2 emissions. This forecast reflects the pattern of recent historical data, indicating the model's ability to capture underlying trends and seasonal variations effectively. The Prophet model has managed to align the forecast closely with the observed data patterns, demonstrating its strength in handling strong seasonal effects and trend shifts. However, the model's handling of extreme data volatility and irregular patterns can pose challenges, impacting accuracy for certain datasets.

MULTIVARIATE MODELS

6.5 Multivariate ARIMA

In the multivariate ARIMA model, given the additional insights from null value imputation in categorical value and outlier removal, the model considers the following points to down-select features. 1) More companies are getting their Direct Emissions verified by third-party for transparency and accuracy YoY. 2) Net Zero Statement and SBTi commitment go hand in hand. 3) Impact of SBTi driven emission reduction is not clear, given lowest correlation value to directs emission value.

As a result, the features considered in the multivariate ARIMA model include 1) revenue (USD million), 2) 'Industry Carbon Intensiveness', which are newly defined post' average carbon intensity peak assessment' by industry types, and 3) countries.

Note that a new feature has been introduced called 'industry carbon intensiveness' to resolve the sparse matrix nature of the one-hot encoding of 155 industries. After evaluating the industry trend in its emission intensity, the companies have been re-categorized into two types: production and services. Above >0.75 quantile from boxplot assessment, 39 carbon-intensive industries resulted. (refer to Appendix 3 and Appendix 5. Carbon Intensive Industry categorization)

Parameter tuning; ARIMA(p,d,q)

p = auto-regressive, d= non-seasonal different, q = lagged forecast errors

ARIMA	AIC	BIC	MSE
(1,1,1)	835828316	836304.920	1293847.025
(2,1,0)	840688.205	841164.809	1330306.454
(3,0,2)	835510.154	836018.015	1282544.029
(3,0,3)	835439.305	835954.978	1175894.884
(3,0,4)	835503.206	836026.693	1289602.606

The best results for multivariate ARIMA were obtained with (3,0,3) yielding the lowest AIC, BIC, and MSE values. We recognize that MSE is more sensitive to outliers, resulting in high values, as shown above. This suggests additional Mean Absolute Percentage Error (MAPE) calculation to avoid scale-dependent loss function. Alternatively, two independent models, the carbon-intensive and non-carbon-intensive industry ARIMA models, could be further developed.

6.6 Multivariate LSTM

In our application, a minimum of 10 points is required, which means dropping the verification_co2directscope1 indicator, which only has data back to 2018 at the earliest. This is a concern due to the consistent evaluation of verification as an important feature per our feature selection analysis. However, the results are still promising, and we can expect the model to only improve once ten years of verification status data are available.

Four LSTM models are presented to illustrate the effects of different parameters. The constants across each of these four models include:

- A single Dense Layer at the end of the model architecture connects each input node to each output node to combine features learned from the previous layer
- Use of the Adam optimizer (at a rate of 0.001) ,which combines two other stochastic gradient descent optimizations (Adaptive Gradient Algorithm also known as AdaGrad, and Root Mean Square Propagation also known as RMSProp). It adjusts the weights of the neural network to minimize the loss function.
- The loss function is the Mean Squared Error.
- A batch size of 32 training samples was used in each forward and backward pass of the network. Batch sizes mostly affect the efficiency of computation.
- A training and validation split of 80% and 20%, respectively.

The table below outlines the parameters of four differently tuned LSTM models and their results.

Model #	LSTM Layers	Neurons/Layer	Epochs	Dropout Rate	Early Stopping Patience
1	1	50	50	N/A	N/A
2	1	100	100	0.2	N/A
3	1	100	100	0.2	10
4	2	100/50	100	0.2	10

The table below displays the resulting Mean Squared Error for the dataset described in each column header.

Model #	NANs and Verification dropped	NANs filled, All Predictors	NANs filled, Verification removed	NANs filled, 2 most important features
1	297875.6	1009239.98	1003741.67	1130760.486
2	439493.88	1032879.54	1030971.39	1122788.98
3	335263.69	993423.62	1010136.04	1090162.83
4	276372.45	988510.72	926258.16	1099769.97

The first model is the simplest implementation, with a single LSTM layer and no dropout function. It achieves a moderate MSE result, but risks overfitting due to the lack of a dropout function to forget some amount of traits specific to the training data.

The second model increases the number of epochs, which allows the model to learn progressively more from the data at the cost of additional computation. It also implements the dropout function, which results in a higher MSE result, but is less likely to be overfit to one specific dataset.

The third model adds an early stopping method to the dropout function, which helps to prevent overfitting by halting the training process if the model's performance on a validation set does not improve for a specified number of epochs. The MSE improves (reduces) from the second model, demonstrating the value of the early stopping method, but is still meaningfully higher than the (likely overfitted) first model.

The fourth model adds an additional LSTM layer. This model tends to have a similar MSE to the first model, but does not consistently perform better or worse. The additional LSTM layer is a step towards overfitting, but unlike the first model, balances that with the dropout function. More layers will capture more complex patterns and hierarchical features, translating to better handling of long-term dependencies, at the cost of computation and additional overfitting risk.

Consistency is a concern when using the Adam optimizer, which was used for all LSTM models. Even in Python environments with a set seed for randomization, the TensorFlow Adam optimizer is known to be non-deterministic in its current state.

The model clearly favors the dataset that drops any rows with NAN values, instead only using the 9792 rows of complete data. Models that overwrite NANs with False (zero) values suffer in their performance, especially when applied to the Verification indicator, which has entirely False values up until verification begins in 2018.

Evaluating the highest quality model (the fourth model), the Root Mean Squared Error (RMSE) of 525.71 is slightly lower than the mean of 551.74 from the actual Y values; on average, the model's predictions are relatively close to the mean value of the actual data. The RMSE is also much smaller than the standard deviation of 1840.59, indicating the model's predictions are within a reasonable range of variability. Considering the high variability across the full range of data, the model appears to be performing reasonably well.

Finally, visualizing the results of the highest quality model shows that the LSTM predictions mostly follow the same trends as the actual data. Naturally, the sharpest declines in the actual data are not quite as pronounced in the predicted values, but the neural network adapts quickly to these trend changes. The green point on the chart represents the predicted point for 2023 which successfully mimics the downward trend in intensity from 2018 to 2022.

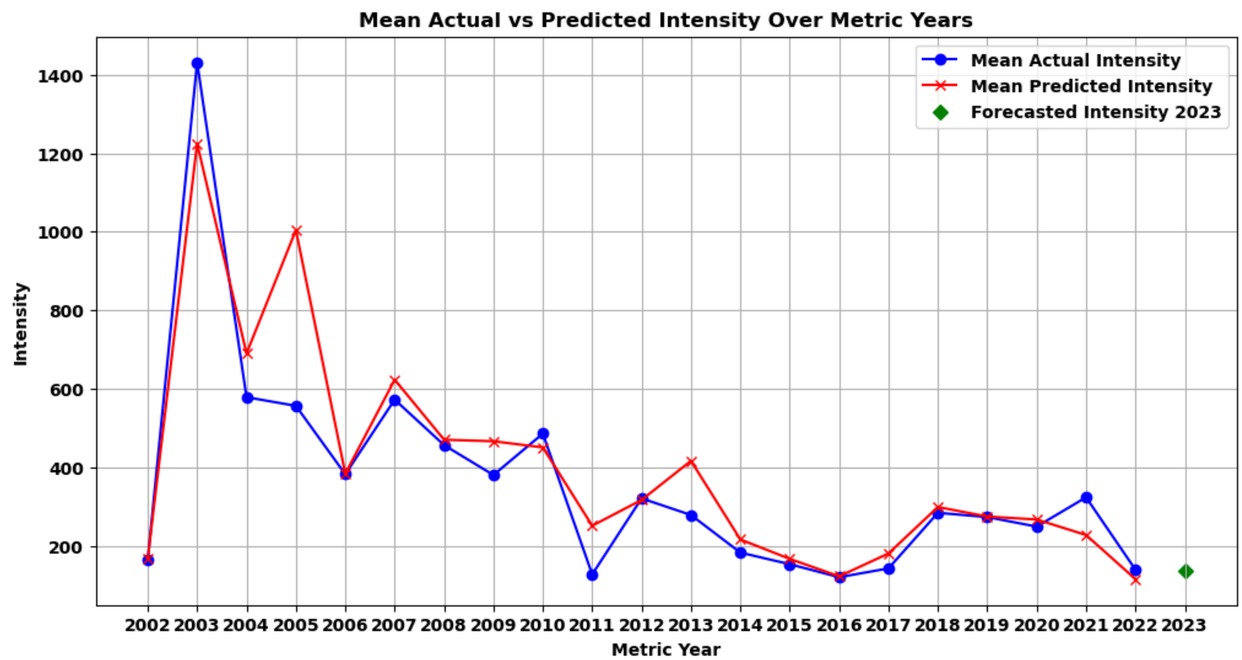


Figure 31: Actual and Predicted Mean Intensity per Year (LSTM)

7 Discussion

The following discussion points addresses the 'additional research questions' to explain the impact of explicit sustainability commitment, challenges and limitation faced during model creation, and actionable insights for investors and stakeholders.

<u>7.1) What are the challenges associated with relying on forecasting techniques to predict the future sustainability performance of companies?</u>

Inconsistent feature reporting poses challenges for fully leveraging forecasting techniques. Detailed reviews at the company level reveal that variability in emission reporting years and intermittent sustainability commitments reduce accuracy.

Data imputation can address these inconsistencies by setting new standards:

a) For missing emission reports, *interpolate expected emissions using revenue* data from SEC 10-K or equivalent reports.

b) For categorical features like sustainability commitments, use a conservative approach by assuming the previous year's status for missing data. For example, if a company commits to net zero emissions in 2016 but not 2014, assume no commitment for 2015 (n = n-1).

c) Assign weights to each feature and apply penalties for high imputation levels to improve forecasting performance.

<u>7.2) How would the model perform when limited historical data is available for a given company and what is the minimal historical depth required to apply any meaningful forecasting technique?</u>

In LSTM: performance suffers greatly when the model accounts for all feature due to lack of historical dataset for ‘verification_scope1emission’ data being less than 10 time-steps (in this case years).

7. 3) How do you anticipate that the integration of the forecasting technique could contribute to providing investors and stakeholders with actions insights regarding the future sustainability performance of companies?

Hard to abate sector: Industries such as integrated oil and gas, electricity generation, steelmaking, and cement production are among the largest polluters. Despite their progress, absolute emission values are not decreasing, classifying them as 'hard-to-abate' sectors. Forming an alliance to invest in 'breakthrough technology' for absolute emission reduction could be considered a more sustainable company. (reference to 'Net Zero Teeside' Carbon Capture)

Classification of company's existence: Companies' emission reporting is often inconsistent due to voluntary disclosures, with some reporting every two years or irregularly (e.g., a Chinese cement producer reported annually from 2004 to 2020 but stopped in 2021). Determining whether a company has ceased operations or simply discontinued reporting due to financial constraints is essential. Identifying the company's status helps decide whether to impute data for missing years or rely solely on historical data for model training.

Sunk cost like Sunk emission: Poor financial performance does not proportionally reduce emissions. For example, a Danish (Clarity ID: '01FF54…..YXKF3') multi-utility company's revenue fell by 67%, but emissions dropped only by 10%. Characterizing carbon-intensive industries and establishing a baseline for minimum emissions can help investors and stakeholders calculate actual emission reductions versus purchasing carbon credits. This translates into operational costs impacting profit margins and investor returns.

Value in qualitative assessment: This approach will remain essential until mandatory emission disclosure and standard work for each industry type with a validation process are established.

8 Reflection

The following section describes the points of consideration for ClarityAI to improve forecasting accuracy.

Reflection during EDA:

- Which industry reports more diligently: consecutive yearly reporting, fulfillment of the last five years of emission reporting, third-party data validation adoption rate (to put more weight on prime data set for forecasting accuracy).
- Revenue fluctuation and its impact on direct emission by industry, to identify hyperparameter tuning effectiveness by industry types.
- Compare the company against its peer industry group with Carbon Intensity parameters to understand the climate impact of the business. This comparison helps identify potential reporting mistakes and remove the outliers/abnormalities in the data set when training the forecasting model. This information is also valuable for investors interested in the company's sustainability. Lower Carbon Intensity against its peer industry group will be favored considering the

introduction of carbon tax schemes around the world per \$/MtCO2e, which will eventually reduce the company's profit and yield a lower return to the shareholders.

- Feature selection using multicollinearity assessment to methodically eliminate features to reduce dimensionality. Computational efficiency can be anticipated, lowering training processing time on low value-adding features.

9 Conclusions

This study examined the growing importance of direct emission forecasting to investors for sustainable investing. Generations of forecasting algorithms have been implemented to serve the needs of stakeholders (including ClarityAI, the project sponsor) and forecast future emissions with greater confidence and accuracy. Given the data, the best-performing forecasting was determined to be LSTM among the models developed.

References

BYJUS, n.d., ‘Exponential Smoothing’, BYJUS, viewed 2024/06/14, URL< https://byjus.com/maths/exponential-smoothing/>

D Cermak, 2022, ‘3 Steps to Consider BEFORE Deciding to Impute Missing Data’, Medium, viewed 2024/06/14, URL <https://medium.com/@dcermak/3-steps-to-consider-before-deciding-to-impute-missing-data-692d57b76c4f>

D Leventis, 2018, ‘XGBoost Mathematics Explained’. Medium, viewed 2024/06/28, URL <https://dimleve.medium.com/xgboost-mathematics-explained-58262530904a>

D Jain, 2020, ‘Feature Handling: Categorical and Numerical’, Medium, viewed 2024/06/028, URL <https://towardsdatascience.com/feature-handling-3f14c12ecbb8>

E Raheem, 2024, ‘Missing Data Imputation: A Practical Guide’, Springer Link, viewed 2024/06/16, URL< https://link.springer.com/chapter/10.1007/978-3-031-41784-9_18>

IBM Technology, 2022, ‘What is Monte Carlo Simulation?’, Youtube, viewed 2024/06/12, URL < https://www.youtube.com/watch?v=7TqhmX92P6U >

J Jakobsen et al, 2017, ‘When and how should multiple imputation be used for handling missing data in randomized clinical trials – a practical guide with flowcharts’, BMC Medical Research Methodology, viewed 2024/06/14, URL< https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-017-0442-1>

M Padhma, 2024, ‘A Comprehensive Introduction to Evaluating Regression Models’, AnalyticsVidyha, viewed 2024/06/14, URL<https://www.analyticsvidhya.com/blog/2021/10/evaluation-metric-for-regression-models/>

M Rastogi, 2020, ‘Tutorial on LSTMs: A Computational Perspective’, Medium, viewed 2024/07/01, URL <https://towardsdatascience.com/tutorial-on-lstm-a-computational-perspective-f3417442c2cd#da46>

M Sanjeevi, 2018, ‘Chapter 10.1: DeepNLP – LSTM (Long Short Term Memory) Network with Math’, Medium, viewed 2024/07/01, URL<https://medium.com/deep-math-machine-learning-ai/chapter-10-1-deepnlp-lstm-long-short-term-memory-networks-with-math-21477f8e4235>

R Hyndman et al, 2021, ‘Forecasting: Principles and Practice’, 3rd edition, OTexts, Melbourne, Australia viewed 2024/06/15, URL < https://otexts.com/fpp3/prophet.html>

R Nau, n.d., ‘the Mathematical Structure of ARIMA models’, Duke, viewed 2024/06/13, URL < https://people.duke.edu/~rnau/Mathematical_structure_of_ARIMA_models--Robert_Nau.pdf>

Science Based Targets, n.d., ‘About Us’, Science Base Targets, viewed 2024/06/14, URL<https://sciencebasedtargets.org/about-us>

UNCC, n.d., ‘the Paris Agreement’, viewed 2024/06/13, URL < https://unfccc.int/process-and-meetings/the-paris-agreement>

U Michelucci, 2024, ‘Fundamental Mathematical Concepts for Machine Learning in Scienc’, viewed 2024/6/13, URL< https://doi.org/10.1007/978-3-031-56431-4 >

V Kotu et al, 2019, ‘Autoregressive Moving Average’, Science Direct, viewed 2024/06/13, URL < https://www.sciencedirect.com/topics/mathematics/autoregressive-integrated-moving-average>

W Lin et al, 2019, ‘Missing Value Imputation: a review and analysis of the literature (2006-2017)’, Springer Link, viewed 2024/06/14, URL < https://link.springer.com/article/10.1007/s10462-019-09709-4>

Appendix

Appendix 1: Largest Emitter (Absolute CO2)

Twenty companies account for top two-hundred largest record emissions. Top eight emitters were observed in detail to understand the characteristics of large emitters. Full analysis of the largest emitter in detail are presented in Appendix 1. Following companies are in descending order: *(Could turn this into table)*

1) Company ('01FF543RRCCWNHSF9RQKZQV785'), an Integrated Oil & Gas company in Italy (IT), emitting 2.723e+11 MtCO2e between 2004 to 2021. Their average carbon intensity (ACI) is 1337.13 while the total revenue is 205.6 trillion USD over the past 17 years. The company has both emission reduction target with internal policy in place, followed by net zero commitment. Although they have not committed to SBTi, the company have been validating scope 1 emission over the past 4 years, between 2018 and 2021. The company has not reported on their 2022 emission, this could be driven by the Sustainability Report publishing timing. Noticeable year for this company is in 2006, where the emission has increased by 2100% in 2006 from prior year, while the revenue grew only by 16%. Given such unusual variance, catastrophic failure or oil spill accident could be presumed and requires additional due diligence in categorizing this particular data point as well as entire company in training the forecasting model.

2) Company '01FF543WBWAWJR3F4V1R7RKV3K' an Electricity generation utility company in India (IN), emitted 3.257e+11 MtCO2e between 2012 to 2022. Their average carbon intensity is 20047.8 while the total revenue is 16.7 trillion USD over the past 10 years. The company has internal policy to reduce emission, without specific emission reduction target, no comment with respect to SBTi alignment while third part verification of Scope 1 emission data in 2019, 2020 and 2022. They have been making steady progress, reducing emission by 3~8% YoY.

3) Company '01FF543TXQTTWTTQBTS6H9S1Q4' is an Integrated Oil and Gas company in the US, emitted 2.748e+11 MtCO2e between 2002 and 2019. Note that the company did not report out on the emission between 2003 and 2004. Since 2006 they company has reported Scope 1 emission annually. Their average carbon intensity is 479.9 while the total revenue is 632.3 trillion USD over the past 15 years. The company has target emission reduction with net zero goal however do not have a set internal policy in place. In 2016 the company has signaled policy measure and had drawn back since 2018. No record of SBTi alignment or emission validated by third party.

4) Company '01FF543VY7KMG0Z16ZGVVW7VCH' is an Integrated Oil and Gas Company in Russia (RU), emitted 3.818e+11 MtCO2e between 2005 to 2019. Their average carbon intensity is 4752, while the total revenue is 114.5 trillion USD over the past 14 years. The company has both emission reduction target and internal policy in place while not committing to net zero goal. The company have been certifying emission value in the past 2 years in 2018 and 2019. The company have made drastic improvement in revenue increase while keeping the emission level stable. The carbon intensity in 2005 was 14531 while in 2019 the company managed to drop the intensity almost by 90%, 1921.4 in 14 years.

5) Company '01FF543SQXXR65BMYJWYT5AXYG' is a Steel making company in Luxembourg (LU), emitted 3.363e+11 MtCO2e between 2007 to 2022. Their average carbon intensity is

2214.8 while the total revenue is 156.9 trillion USD over the past 13 years. The company has all three measures, emission reduction target, policy in place with net zero statement but not committing to SBTi target. The company has been validating Scope 1 emission in the recent 4 years between 2018 to 2021, except for the most recent year 2022. There has been a gradual emission reduction improvement observed since the scope 1 emission validation practice has been introduced (2018).

6) Company '01FF543T9EBDX98HFKDA91NGK2' is an Electricity generation utilities in South Korea (KR) emitted 2.58e+11 MtCO2e between 2005 and 2021. Their average carbon intensity is 5458.4 while the total revenue is 55.4 trillion USD over the past 14 years. The company has significantly reduced their carbon intensity by half, between 2005 and 2011, the ACI was 7931 while 2014 to 2021 average is 3843. This excludes the 2 years, 2011 and 2012 where their business performance would have experienced a heavy turmoil with 99% reduction in operation or could have been a wrong unit reported in that particular year. It is still a valid assumption to say that business operation has significantly changed between 2010 and 2011, e.g. replacement of equipment. The company has emission reduction target, with policy in place with net zero statement. In 2021 the company has started to audit their scope 1 emission reporting values.

7) Company '01FF543SFRGWRP7C6NPF7YXKF3' is a Multi-Utilities company in Denmark (DE) emitted 3.253e+11 MtCO2e between 2004 and 2019 with the exclusion of 2003. Their average carbon intensity is 3688.4 while the total revenue is 96.1 trillion USD over the past 16 years. The company's average carbon intensity double in the recent 2 years, 2002-2017 was 3250 while 2018-2019 was 6975. While the company has strong commitment to emission reduction given all four commitments checked, the company has only validated the 2018 and 2019 emission data. Sharp increase in average intensity is due to large fall in revenue by 67% while emitting similar amount of emission (only -10% in raw emission level). Such behaviour could lead to multiple scenarios where the company have high sunk cost and business operation structure hence despite the weak business performance the emission production is unchangeable, or the emission reporting validation has identified missing scope 1 emission values.

8) Company '01FF543V7FPZ55M0JB4DTMBCBZ' is a Cement Producer in China (CH) emitted 2.36e+11 MtCO2e between 2004 and 2020. Their average carbon intensity is 4482.9 relatively stable across the years with total revenue of 52.7 trillion USD the past 17 years. The company has strong commitment to reduce carbon emission with SBTi alignment in place and have been validating their data by third party in the last 3 years. The company has not reported emission since 2021.

Appendix 2 – Carbon Intensity (Carbon Emission to Revenue ratio)

High Carbon Intensity companies were observed in detail to understand their characteristics and seek for potential outliers:

1) Company '01FF543T2JTRB6CHHBWXNNHMCB' is a Life Science Tools and Services company in US, emitted 21.7e+9 MtCO2e between 2008 and 2015. Their average carbon intensity is 6006083 while the total revenue is 6,716.2 Million USD over 8 years. The company has committed to SBTi without any target reduction, policy in place nor net zero target. Given such high scope 1 emission value with lowest revenue when compared to its industry peer group, Life Science Tools and Services, there is high likelihood this company is an outlier. Boxplot will be used to validate this data point and overall data population with this company included and excluded.

2) Company '01FF543RJ1GKQM4WA89CB9M6W9' is an Electricity generation utility company in Hongkong (HK) emitted 1.714e+11 MtCO2e between 2006 and 2022 with an exclusion of 2008-2010 and 2014-2016. Their average carbon intensity is 19641.3 while the total revenue is 11.2 trillion USD over the 10 years. The company has grown by 10 folds while reducing their emission by 74%. Significant improvement has been made between 2007 and 2011 and again in 2020 to 2021. In comparison to its' peer industry, this company emits about 5 times more than the average of similar revenue.

3) Company '01FF543VHJJC6WC3YSQAS6CMN3' is a Cement Producer in Great Britain (GR), emitted 4.17e+9 MtCO2e in 2016, 2017 and 2020. Their average carbon intensity value ranges from 5445.9 to 45121.7. Given the absolute CO2 emission variation of 12~16% while the revenue of ~600%, one would be suspicious of the 2017 revenue data's accuracy in terms of its unit.

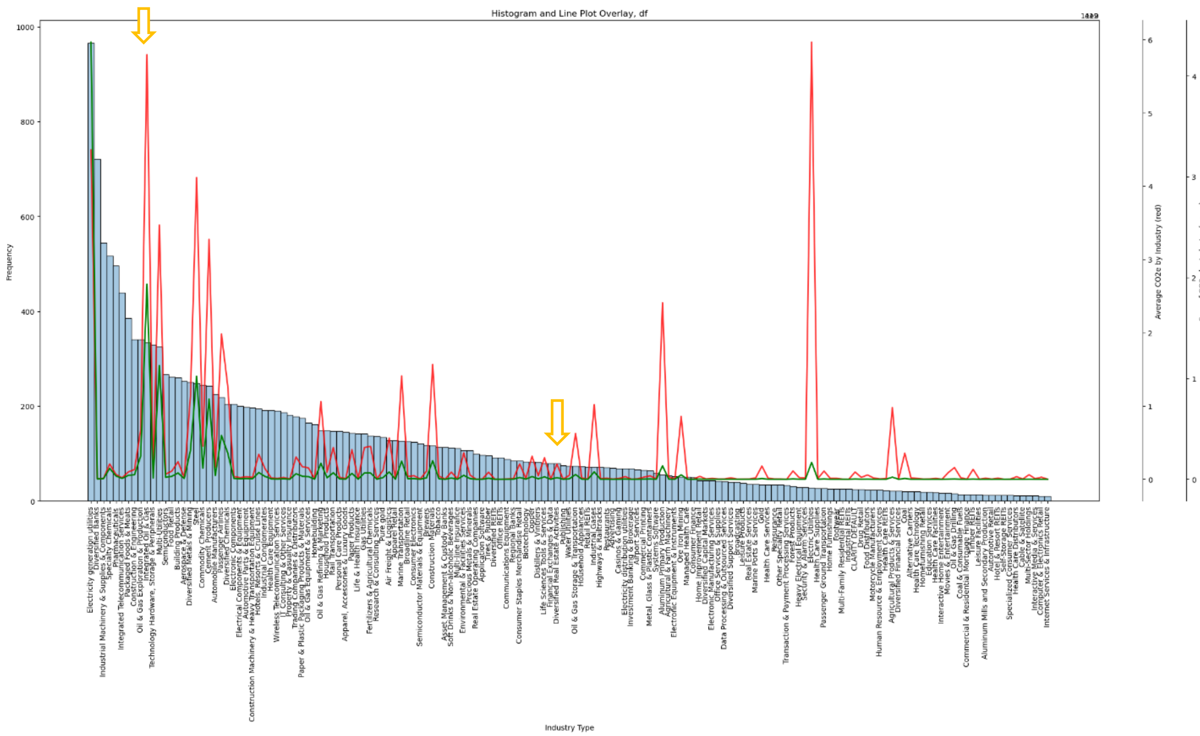
4) Company '01FF543WQ0MR056SYMVY3EHE7W' is a Cement Producer in India (IN), emitted 40.15e+9 MtCO2 between 2010 and 2019. Their average carbon intensity value is 12232, with a step change observed in 2010 vs 2012 (avg. 12881) and 2015 and 2016 (8889.9). Their revenue grew by 82% while the emission increased by 41%. Amongst the peer industry group, where the Cement Producer carbon intensity is 5536.9 (excludes company '01FF…CMN3' in GR) this company in India is approximately produces 2x more direct emission with the direction to cleaner Cement Production.

'Carbon Intensity' attribute in descending order is a useful method in identifying abnormality and potential outliers.

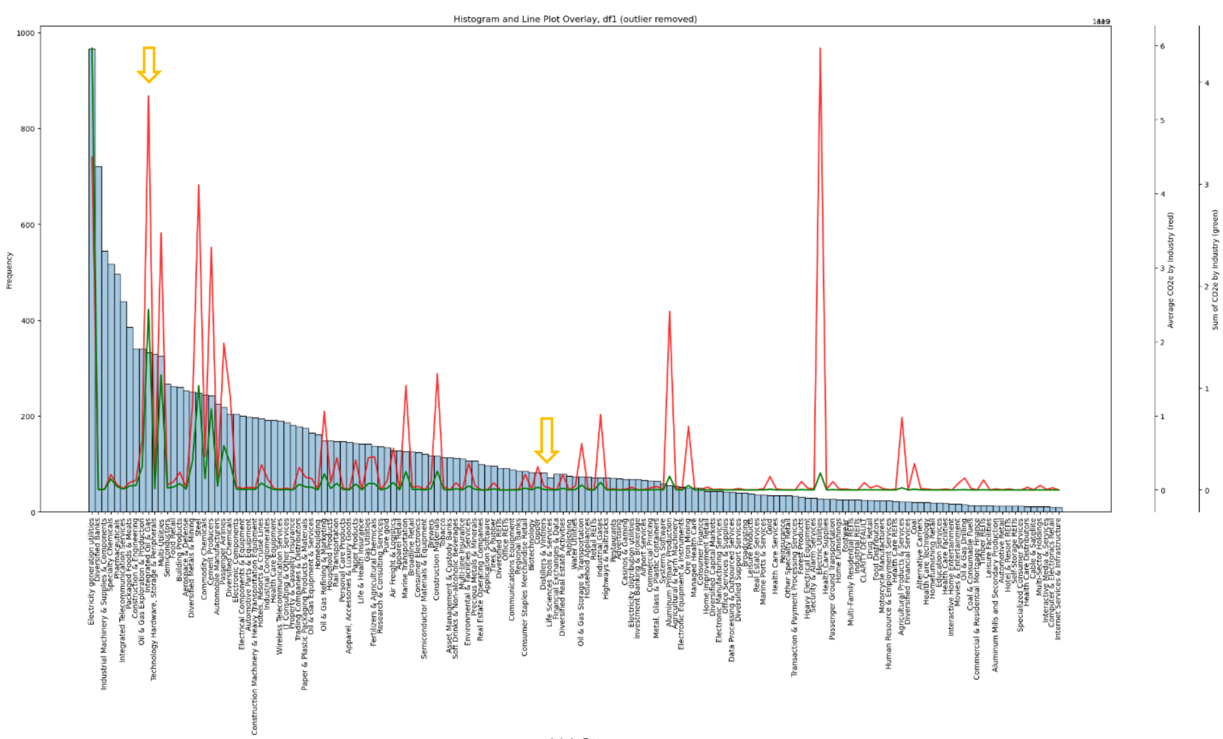
Appendix 3 – Impact of Outlier Removal on overall Direct Scope 1 Emission (raw) and Average Carbon Intensity

(Arrows indicate the outlier position and the industry)

Before:

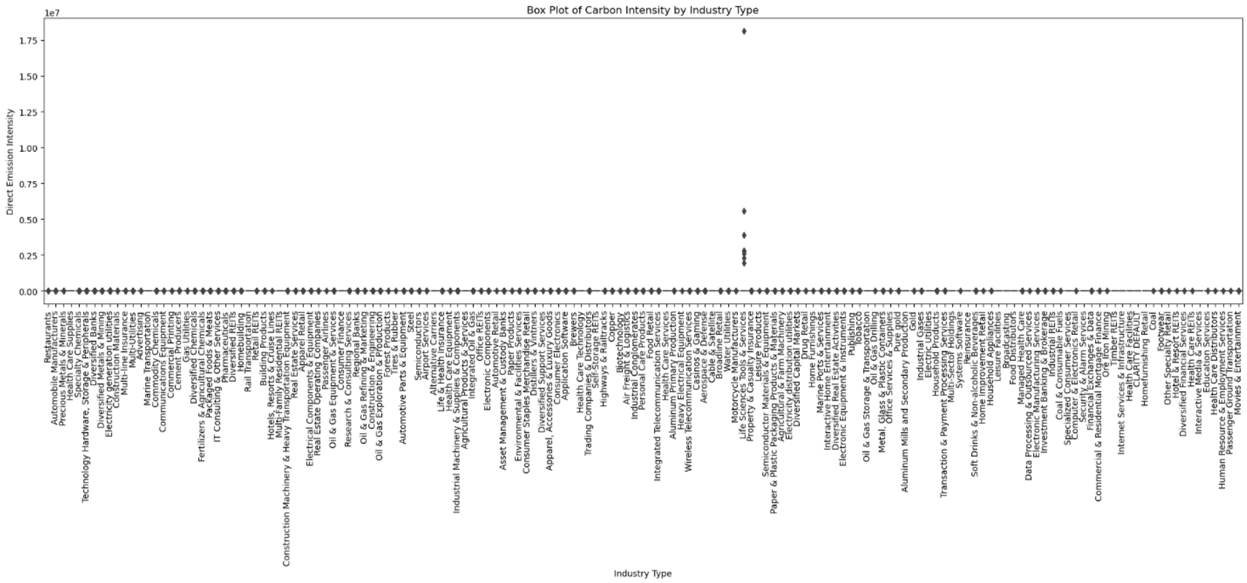


After:

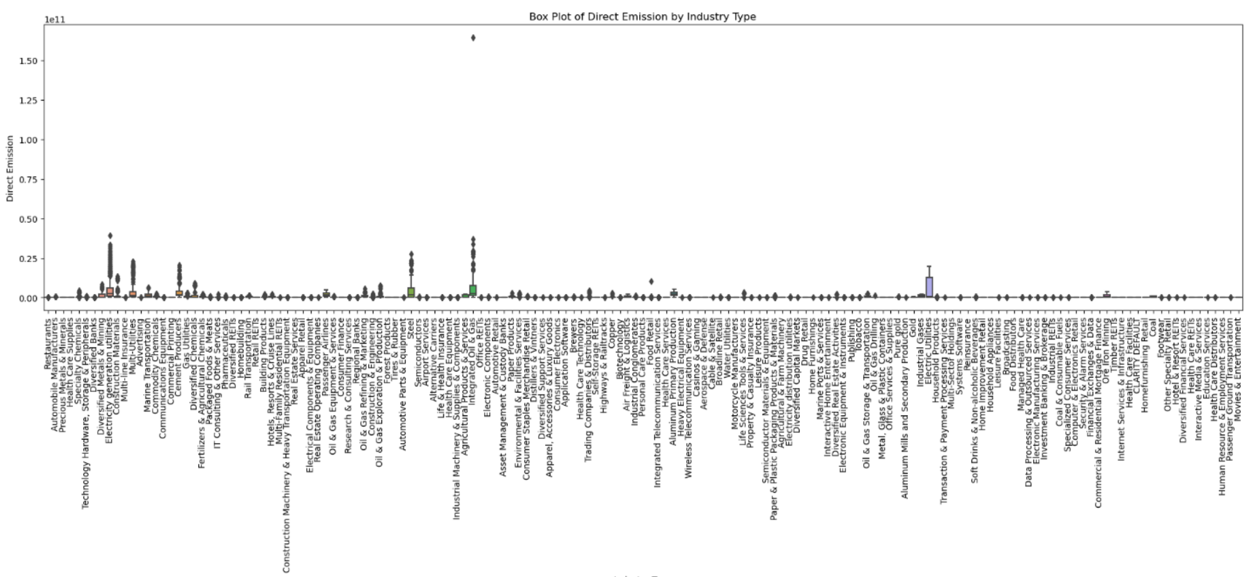


Appendix 4 – Supplementary Boxplots for EDA 5: Outlier Assessment

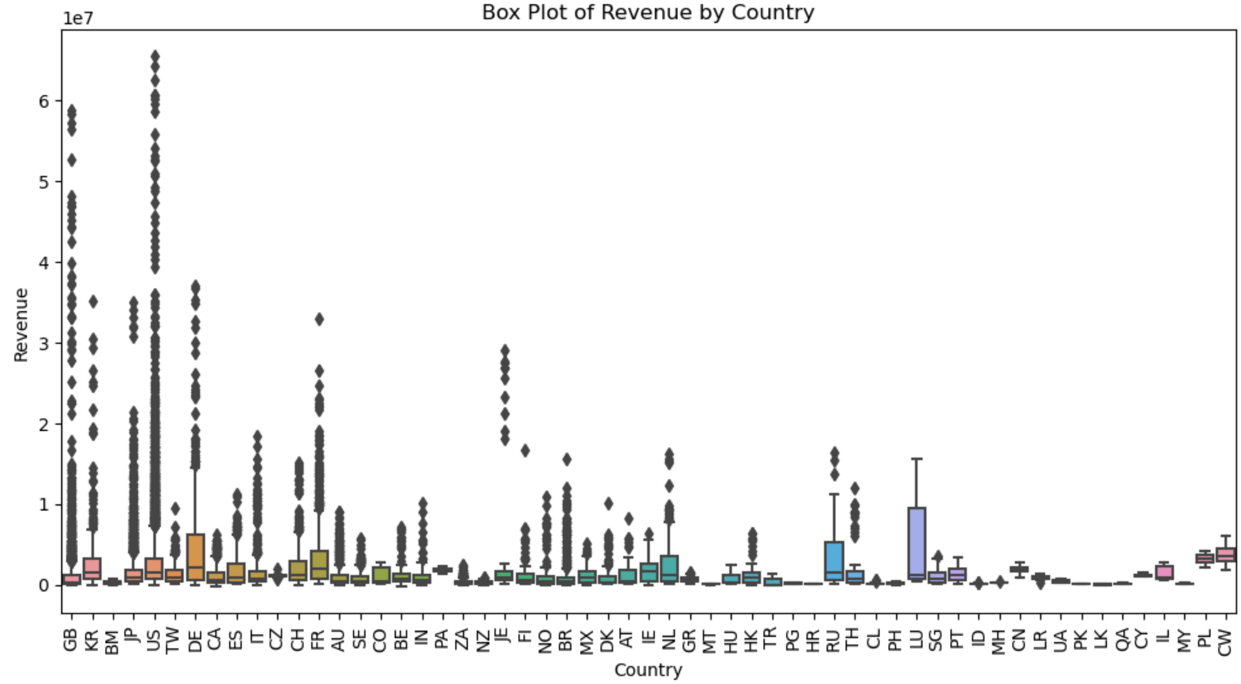
4.1 Carbon Intensity by Industry Type:



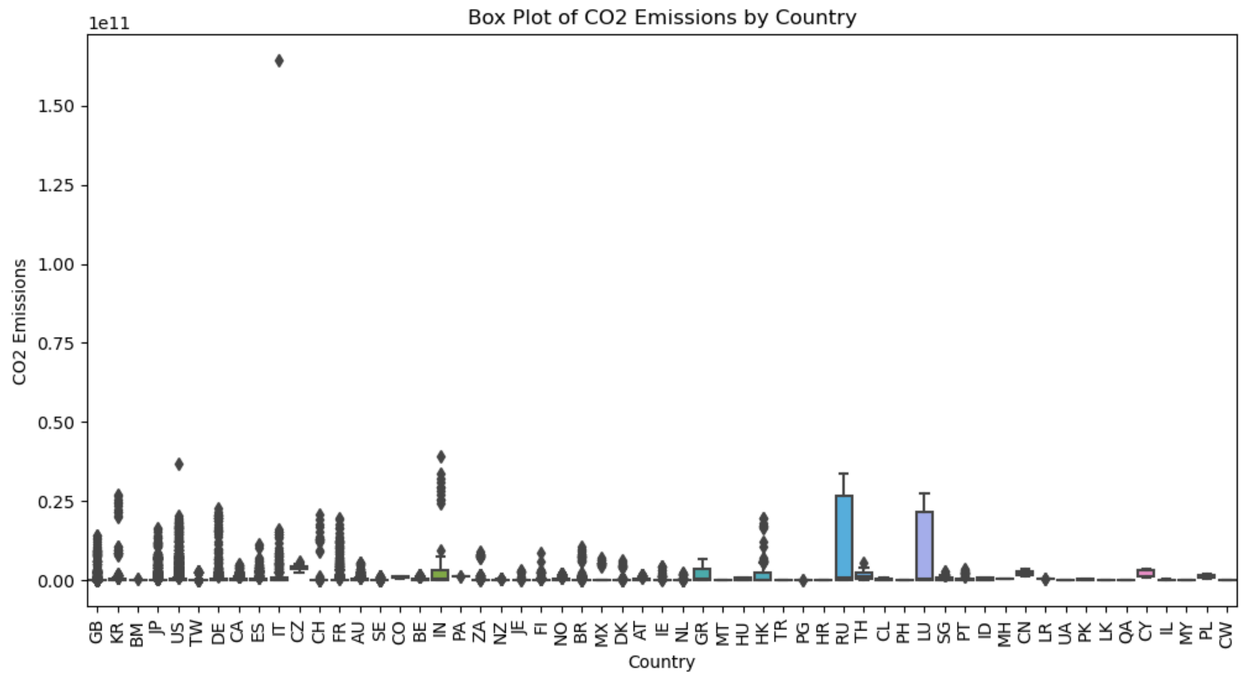
4.2 Direct Emission by Industry Type:



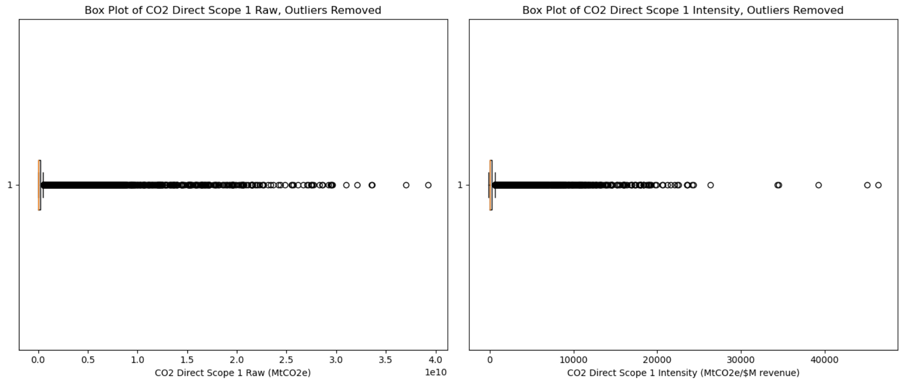
4.3 Revenue by Country:



4.4 CO2 Emission by Country:



Appendix 5. Carbon Intensive Industry Verification Methodology




```
Quantile of Carbon Intensity by sum of Industry Direct Emission:
0.00    3.924251e+04
0.25    3.866708e+06
0.50    2.325659e+07
0.75    1.396308e+08
1.00    5.963271e+09
Name: co2directscope1_raw, dtype: float64

Count of Total Industry: 155
Count of Carbon Intensive Industry (threshold qunatile > 0.75): 39
List of Carbon Intensive Industry (threshold qunatile > 0.75):
                              clarity_industry_name  co2directscope1_raw
116                               Electric Utilities         5.963271e+09
9                               Integrated Oil & Gas         5.317700e+09
0                   Electricity generation utilities         4.491816e+09
17                                             Steel         4.116252e+09
11                                    Multi-Utilities         3.467529e+09
19                                  Cement Producers         3.272365e+09
92                       Aluminum Primary Production         2.408800e+09
21                                Passenger Airlines         1.981064e+09
55                            Construction Materials         1.566266e+09
50                             Marine Transportation         1.409461e+09
22                             Diversified Chemicals         1.275370e+09
16                       Diversified Metals & Mining         1.129776e+09
38                      Oil & Gas Refining & Marketing         1.059483e+09
82                                  Industrial Gases         1.019096e+09
130                   Agricultural Products & Services         9.778026e+08
95                                  Ore Iron Mining         8.563732e+08
8                   Oil & Gas Exploration & Production         6.810246e+08
77                  Oil & Gas Storage & Transportation         6.248234e+08
48                          Air Freight & Logistics         5.622325e+08
18                              Commodity Chemicals         4.519049e+08
45              Fertilizers & Agricultural Chemicals         4.475222e+08
44                                    Gas Utilities         4.329744e+08
40                              Rail Transportation         4.326167e+08
42                                   Paper Products         4.029191e+08
60                 Environmental & Facilities Services         3.596631e+08
131                                             Coal         3.548688e+08
27                      Hotels, Resorts & Cruise Lines         3.405991e+08
70                                           Copper         3.130221e+08
33                    Trading Companies & Distributors         3.037209e+08
14                                Building Products         2.388840e+08
68                 Consumer Staples Merchandise Retail         2.089359e+08
3                               Specialty Chemicals         2.081323e+08
74                   Diversified Real Estate Activities         2.055184e+08
108                                             Gold         1.791540e+08
34     Paper & Plastic Packaging Products & Materials         1.718355e+08
20                           Automobile Manufacturers         1.633757e+08
139                                 Oil & Gas Drilling         1.588878e+08
35                      Oil & Gas Equipment & Services         1.542203e+08
28                           Industrial Conglomerates         1.447203e+08
```