CLARITY AI

Add Clarity. Choose Sustainably™

SUSTAINABILITY DATA

HOW ADVANCED TECHNOLOGY CAN INCREASE DATA RELIABILITY

EXECUTIVE SUMMARY

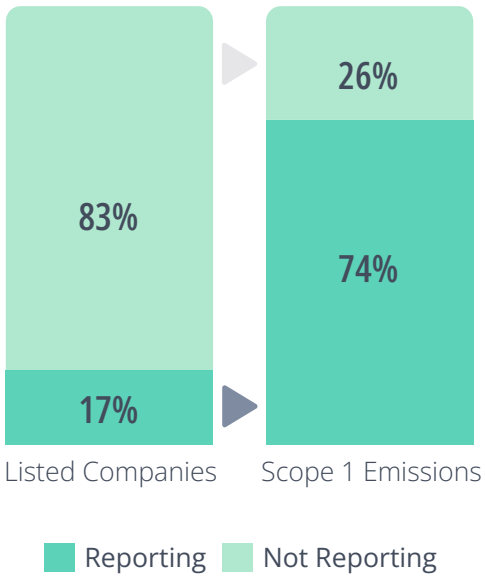
There are somewhere between 40-45,000 publicly listed companies in the world, according to the World Bank, and their direct emissions (Scope 1) account for between 20-25% of the world's GHG emissions.

Only about 17% of those publicly listed companies report their direct emissions – that's only 6,500 companies worldwide. So, yes we have a problem with reporting. As it turns out though, the problem isn't quite as big as you might think, because those 6,500 companies that do report emissions emit about 74% of the Scope 1 emissions of the full universe of public companies.[1] So while we don't have all the data we'd like to, the good news is that we have a good chunk of the data that matters in this case.

But, what about the reliability of that data? This is where we, at Clarity AI, believe there is a problem worth digging into. The old adage, "garbage in, garbage out" is centered on the concept that flawed data inputs will inevitably lead to flawed outputs. Here, sustainability data is no exception and for most asset managers and asset owners, data reliability and quality is a major point of concern in regards to sustainable investing. Investors have legitimate doubts that they won't be able to come to a sound investment conclusion if the sustainability data they are starting with is not reliable.

Reporting companies account for 3/4 of Scope 1 emissions

Figure 1: Out of 40-45,000 listed companies, ~17% report their CO2 Scope 1 emissions. Despite representing a small fraction, those ~6,500 companies account for 74% of Scope 1 emissions of the full universe of listed companies.



Clarity AI sought to understand the impact of data reliability and highlight how important your choice of provider can be on the output of your analysis, and in turn your sustainability related investment decisions. In this paper, we will review a current analysis of data provider consistency and reliability. To understand why these discrepancies exist we will examine the three most common mistakes that data providers make and provide examples to highlight how significant these mistakes can be. Next, we will explore how the Clarity AI reliability model leverages technology to increase reliability and thereby increasing the confidence investors can have on their investment decisions.

1. The 74% is calculated based on direct emissions data (Scope 1).

HOW PREVALENT ARE DATA DISCREPANCIES?

To understand the current state of data quality we needed to analyze the discrepancies that occur across data providers. Using company reported direct CO2 emissions as an example, we wanted to answer the following question: if someone were to reach out to any two of these providers and request Scope 1 emissions data from a given company, would they receive the same answer from each of them?

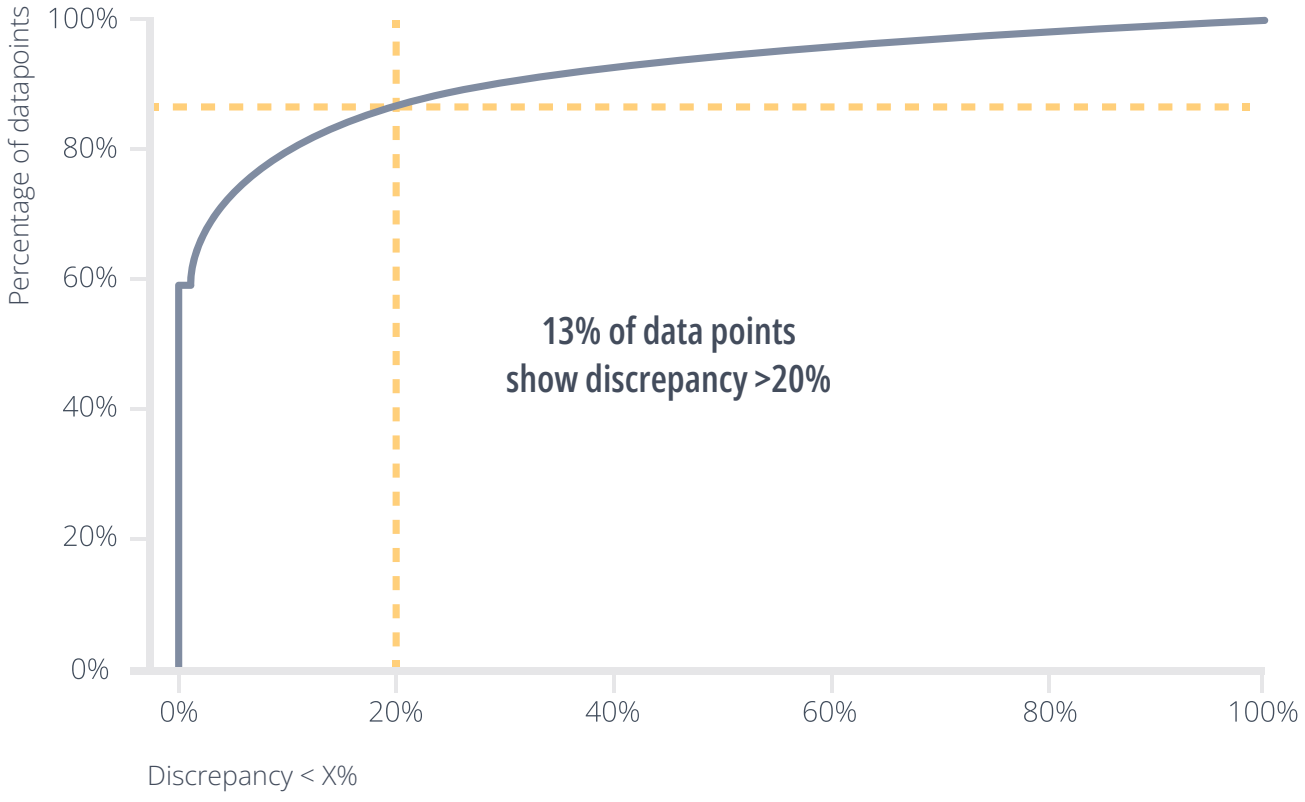
Of those 6,500 companies that do report direct emissions, we compared data points from 3 leading providers to cross-check for consistency. Across providers and across the last 5 years – for those companies that report, we have almost 30,000 data points to compare.

Among those data points, **we found that there were discrepancies (e.g., reliability issues) in 42% of the data points.** Now, when we say a discrepancy, we are referring to anything off by more than 1%, but let's amp that threshold up to at least a 20% difference. Even at that level, we still see discrepancies in more than 13% of the data – that is in more than 1 in 8 data points.

These discrepancies in 13% of the data is a significant percentage and highlights the valid trepidations that investors have when selecting a data provider.

CO2 Scope 1 discrepancies cumulative distribution

Figure 2. Cumulative distribution of discrepancies for organizations with CO2 Scope 1 reported data from two or more providers. Discrepancy Formula = (larger_value - smaller_value) *100 / larger_value



THE TOP THREE DATA ANALYSIS MISTAKES

When we discuss data reliability with clients, they are usually shocked to learn of the high degree of discrepancies that we outlined in the previous section. The first question that is usually asked is, how does this happen if this is the same company self-reported data, available to all providers? Shouldn't it be consistent across providers?

In theory, yes, but errors occur, and often, so to help illustrate how these discrepancies can occur, we wanted to highlight the most common errors that data providers make. We provide examples of each type of error, highlight the magnitude of these errors, and show how they can compromise the investors' strategic thinking, resulting in capital misallocation.

1. Human Reporting Errors

Human errors account for more than 80% of the errors found.[2] These vary in nature but some examples include: incorrect addition of category values (which is the most common), misinterpretation of report details (such as misreporting years or the type of emissions scope), and inaccurate unit measurements (such as misreporting Tons (t) vs Gigaton (Gt)).

Example: Leading Aerospace and Defense Firm

A leading data provider inaccurately reported Scope 1 emissions data due to a human error in reading the publicly available Global Emissions Report. The error occurred by misreading a table, taking the location based Scope 2 value instead of the global total Scope 1 value.

Discrepancy data for 2019
CO2 emissions Scope 1

Correct value

613,000

Tons

Leading Provider

1,208,000

Tons

Data Discrepancy Impact

The discrepancy of roughly 600,000 tons is comparable to the yearly emissions of Puerto Rico (an unincorporated U.S. territory). This would make the company move from the ~30th to the ~60th percentile of its industry when compared to the peers in terms of GHG emissions per Million USD Revenues.

2. We analyzed in depth a sample of ~150 data points to understand the origin of these discrepancies and identified three common types of errors.

2. Inconsistent Reporting Boundaries

Data providers use boundaries (i.e., rules to decide which entities from the group to include or not, what to do with joint ventures, investments, etc.) for emission reporting inconsistently.

Example: Leading organization for the Mining and Refining of Nickel

A leading data provider only included emissions data from two of this leading organization's main business lines and left out its joint venture. This error occurs by inaccurately creating data boundaries that exclude joint ventures from the total CO2 emissions figures. According to the GHG Protocol, joint ventures should be proportionally consolidated based on equity share if applying the financial control rules (which was the case in this example). Excluding the joint venture means that the data provider wasn't disclosing the complete volume of emissions for the company.

Discrepancy data for 2020
CO2 emissions Scope 1

Correct value

1,998,000
Tons

Leading Provider

1,396,000
Tons

Data Discrepancy Impact

When joint ventures are not included, the amount of CO2 that investors believe this company has would be considerably lower than its industry peers. In fact, the CO2 emissions of joint ventures for this company are equivalent to the building and running of 2 gas fired power plants for an entire year.

3. Incomplete Disclosures

Companies publish incomplete disclosures that miss relevant emissions (Scope 3 categories, regions/offices, business lines).

Example: Multinational Oil & Gas company

A leading provider includes only 45% of the company's Scope 1 and 2 emissions in its dataset (direct emissions + indirect emissions of the energy consumed). The error occurred by not including all the subsidiaries of the company. The result is that this company seems to be an average emitter when compared with peers but the reality is that the company is a below average performer in terms of intensity of emissions.

Discrepancy data for 2020
CO_2 emissions Scope 1 and 2

Correct value

220,000,000

Tons

Leading Provider

103,000,000

Tons

Data Discrepancy Impact

Based on actual emissions this company is one of the worst performers. If we use the incorrect data, their ESG score increases by 50 points (on a 100 point scale), meaning that you could be making the wrong investment decision or including large errors in your regulatory reporting.

THE IMPACT OF DATA DISCREPANCIES

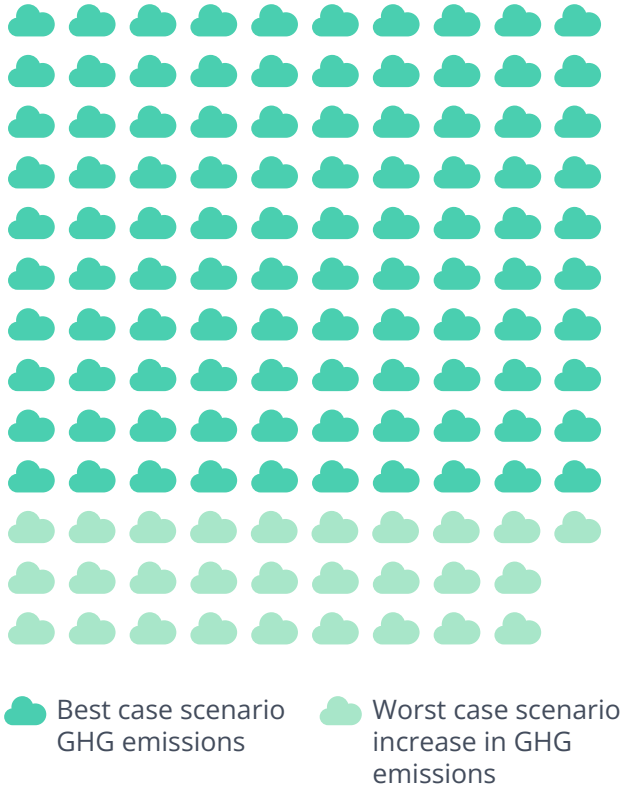
It is obvious to see the potential impact that a significant data discrepancy could have on your investment strategy but to further illustrate the relevance of these discrepancies for investors we examined their impact on the calculation of the GHG emissions of a climate focused fund.

Using as a case study the BNP Paribas Easy Low Carbon 100 Europe PAB, classified as an Article 9 fund under SFDR, we find that there are discrepancies larger than 20% for 18% of the 100 companies included.[3] We then calculate the GHG emissions following two different approaches, either selecting the lowest or the highest value whenever we face a data discrepancy. The GHG emissions resulting from both calculations differ by 28% (see Figure 3).

Noticeably, the required annual decarbonisation rate derived by the EU Technical Expert Group (TEG) and the Intergovernmental Panel for Climate Change (IPCC) to reach the 1.5°C Paris target is 7%. If the error in the GHG emissions of a financial product due to data reliability issues is up to 4 times the annual carbon reduction rate that it would have to hit to be aligned with the Paris Agreement, it is easy to understand why investors should care about data quality issues.

Degree of variation based on potential data discrepancies

Figure 3. The GHG Scope 1 emissions from a climate fund can vary up to a 28% depending on the reported data chosen for the calculation. This is four times the 7% annual carbon reduction rate that a product would need to hit to be aligned with the Paris Agreement.



3. From the 100 companies included in this case study Fund, we have data for the Scope 1 emissions (2019) from two or more providers in 95+ cases which cover +97% of the fund's holdings. Calculated as:

$$\sum_{n}^{i}\left(\frac{\text{current value of investment}_i}{\text{investee company's enterprise value}_i} \times \text{investee company's Scope (x) GHG emissions}_i\right)$$

HOW DOES CLARITY AI WORK TO SELECT RELIABLE DATA?

At Clarity AI we rely on technology and data experts to solve this problem. First, we curated a robust dataset of sustainability data points, which have gone through rigorous quality checks. Then, we trained, calibrated, and validated a supervised machine learning model to select the most reliable data points and filter out non-reliable data. The model leverages more than 90 features for each data point which provide it with enough context to understand its plausibility. The type of features we use are similar to the information a human expert would require to make the same decision. Some examples are:

- If the company has reported this metric for multiple years, is this new ESG data point for the organization consistent with its own reporting history?
- Is it consistent with its industry, given the size and other factors of the reporting company?
- Is the datapoint consistent with other sources for the same company-metric-year?

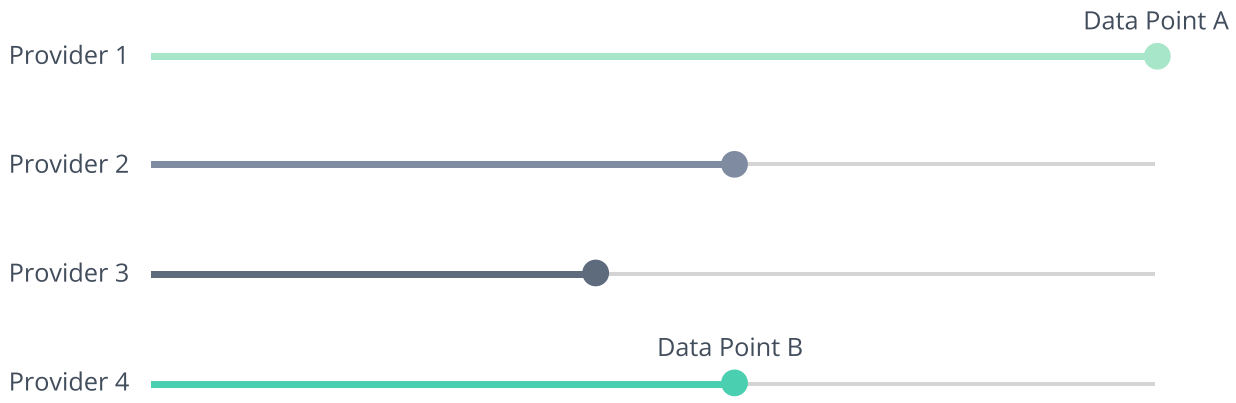
At Clarity AI we created machine learning solutions that combine all of the features mentioned into one single model, which are superior to other simpler methods like outlier filtering.

The fact that a data point is an outlier with respect to its industry or its own history does not mean that it is wrong. It should certainly raise suspicions, but not necessarily mean anything else. And the opposite is also true.

A data point that is not an outlier by any benchmark is not necessarily right. The flexibility of a machine learning model allows for much more complex relationships between the reliability of a datapoint and its features. It analyzes the data from all angles at every level of granularity, which boosts its performance.

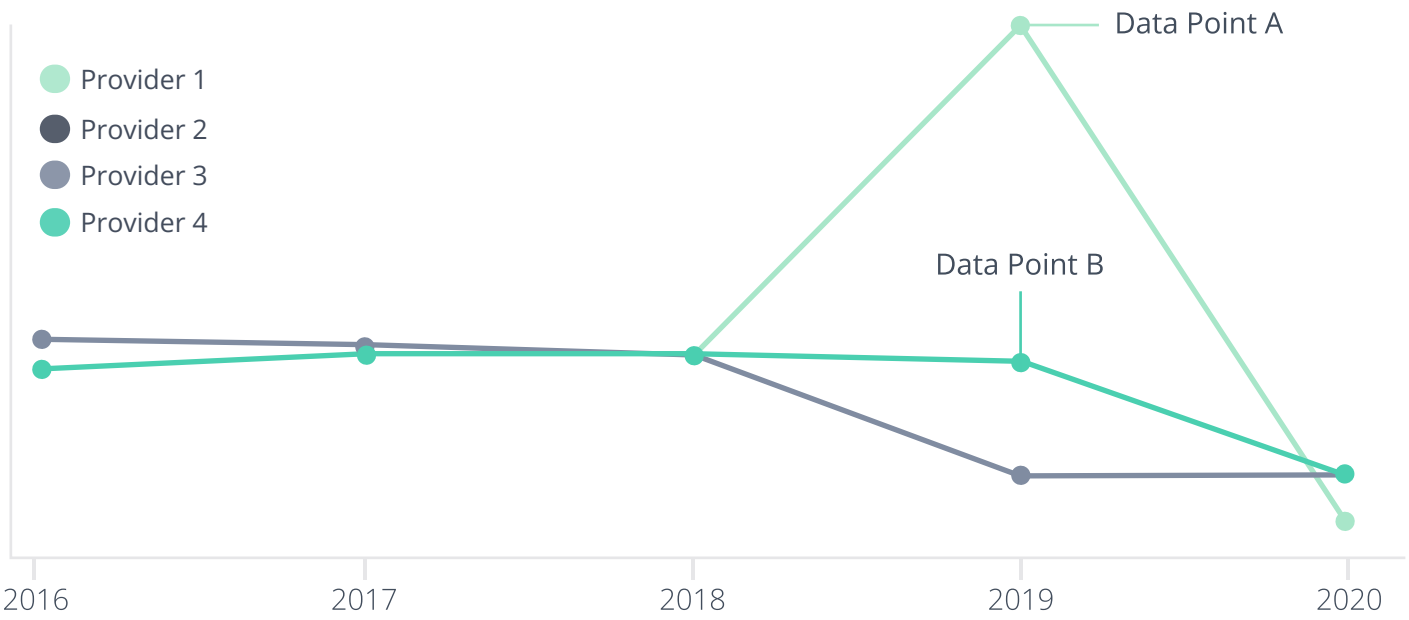
Reliability model features: Consistency with other sources

Figure 4



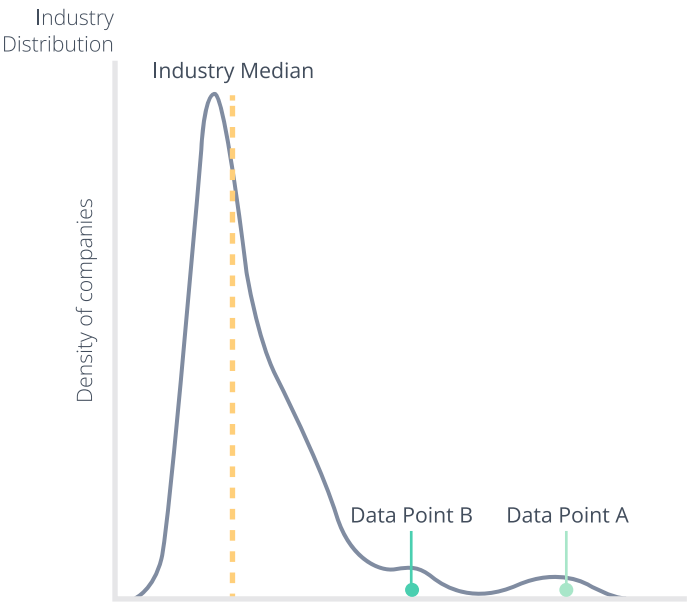
Reliability model features: Consistency with reporting history

Figure 5



Reliability model features: Consistency within industry

Figure 6



Reliability model features: Consistency with company fundamentals

Figure 7

Metric	Year	Value
Revenue	2019	$x
Employees	2019	x
Market Cap	2019	$x

CASE STUDY

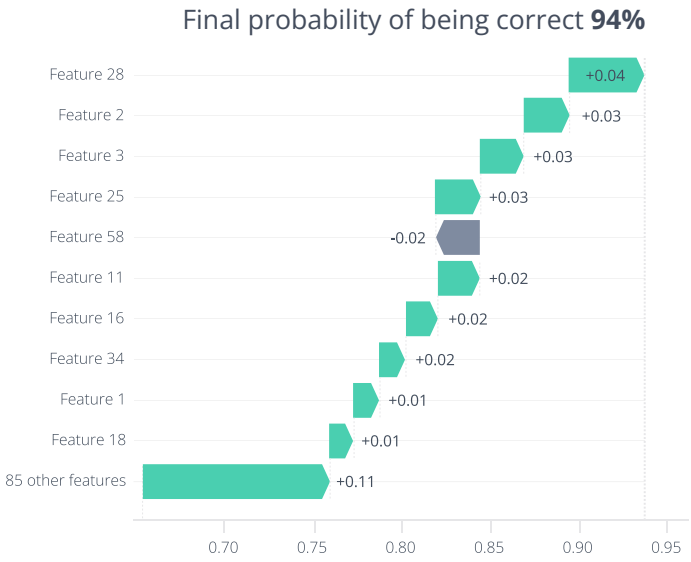
To better understand how our reliability model operates, we will use the case study of a leading software company. In this example, the value for CO2 Scope 1 emissions provided by Leading Provider 1 is almost 60% higher than the one chosen by Clarity AI. How did the model select the right one? According to our Reliability model, every data point starts with the same average probability of being correct.

Then, depending on its features' values, this probability increases (green arrows) or decreases (grey arrows) until it reaches its final predicted value. To illustrate how each component contributes to the probability of being correct, we can focus on the company's capex (Feature 1) and the number of employees (Feature 4), for example.

In the case of the Clarity AI data point, the company's capex adds 1% to the probability of this emissions data point being correct. Our model looks at the feature presented, in this case, capex, and it takes that value in relation to the rest of the information it has about the company. If the capex is consistent with the emission value (given all the company's additional features), it will increase the probability that the data point is correct.

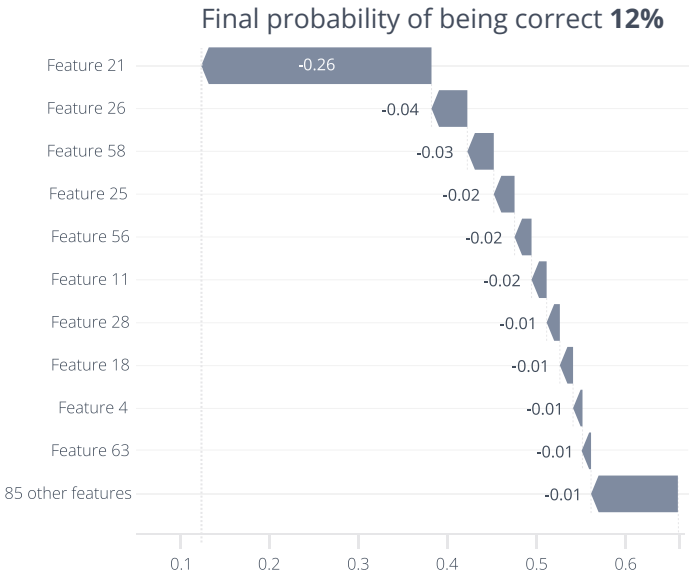
Reliability model case study: Leading software company

Figure 8. Clarity AI data point



In the case of Leading Provider 1, however, the feature related to the number of employees reduces its probability by 1%, meaning that the emissions reported by this provider are slightly inconsistent for a software company having the precise number of employees this company has.

Figure 9. Leading provider data point



Considering all features and their interactions, the Clarity AI data point shows the highest final probability of being correct, around 94%. In contrast, the Leading Provider 1 data point ends up with a 12% probability of being right. Therefore, our Reliability model selects the Clarity AI data point as the most reliable one.

The workflow described above highlights specific features within the model. Still, it is essential to remark that the model looks at each sustainability data point while considering the entire context of the company, much as a human expert would do after years of training. Is it consistent with industry standards? Is it aligned with the emissions of similar companies? Is it plausible for a company of a given size and growth pattern? And so on.

The great advantage of our technological solution is that it is transparent and scalable. This scalability means that we can get the most reliable data to our customers in the fastest way possible.

THE PATH FORWARD

The ESG investment landscape is rapidly changing, along with the regulations that monitor these activities. Investors need the most up to date data to work to stay compliant with their associated regulatory bodies, and more importantly, to meet client demand. As we illustrated in the three common mistakes and the company case study, details can be missed or misunderstood, which could have implications for strategy and investment decisions.

Clarity AI believes the only way to prioritize data reliability at scale is to leverage advanced technology, including machine learning. We create scientific and evidence-based methodologies, leverage the research and data science expertise of our global team, and continuously innovate, create, deploy and maintain our tools and scores.

We trained state-of-the-art Machine Learning algorithms leveraging the input from sustainability experts. When a data point is detected as non-reliable it is sent for external review and corrected if necessary — and this data will enter back the system for training the model. This advanced technology is the only way to create clean, reliable data that investors can rely on. And remember that only 17% of listed companies report their emissions. If we want to address the measurement and plans for emissions reduction, it is critical that we use high quality, reliable data.

AUTHORS

Ron Potok
Head of Data Science at Clarity AI

Ignacio Tamarit
Lead Data Scientist at Clarity AI

Bruna Correa
Data Scientist at Clarity AI

Patricia Pina
Head of Product Research & Innovation at Clarity AI

For more information, reach out to
insights@clarity.ai

ABOUT CLARITY AI

Clarity AI is a sustainability technology platform that uses machine learning and big data to deliver environmental and social insights to investors, organizations, and consumers. As of October 2022, Clarity AI's platform analyzes more than 50,000 companies, 320,000 funds, 198 countries and 188 local governments – 2-13 times more than any other player in the market – and delivers data and analytics for investing, corporate research, benchmarking, consumer e-commerce and reporting. Clarity AI has offices in North America, Europe and the Middle East, and its client network manages tens of trillions in assets.

CLARITY AI

Add Clarity. Choose Sustainably™