Corporate Disclosures Decoded: Forecasting Real Decarbonization Rates

Leveraging Machine Learning Methods for Insight into Corporate Environmental Data by Identifying Key Features and Building Predictive Decarbonization Models

A THESIS PRESENTED
BY
FABRIZIO SERAFINI
TO
THE DEPARTMENT OF STATISTICS
AND
THE DEPARTMENT OF COMPUTER SCIENCE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
BACHELOR OF ARTS (HONORS)
IN THE SUBJECT OF
STATISTICS AND COMPUTER SCIENCE

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS

APRIL 2024

Thesis advisors: George Serafeim and Kevin Rader Fabrizio Serafini

Corporate Disclosures Decoded: Forecasting Real Decarbonization Rates

Abstract

Companies can report carbon emissions via voluntary carbon disclosure surveys, a method gaining global traction. However, the capacity of these surveys to forecast future decarbonization efforts for individual firms is largely unexplored. This study examines carbon disclosure data from the Carbon Disclosure Project (CDP) Climate Survey to forecast decarbonization rates for 3,574 firms from 2011 to 2022. Using machine learning techniques including Mixed Effect, Bayesian Ridge, and Gradient Boosting regression, we are one of the first to leverage disclosure data for predictive analysis at the firm-level. We identify a set of crucial predictors for decarbonization, including the firm's sector, country, emission targets across Scope 1 and 2, reporting of Scope 2 market-based emissions, verification of Scope 3 emissions, and the application of a Marginal Abatement Cost Curve (MACC) for emission reduction initiatives. These predictors prove consistent across modeling techniques, highlighting our analysis's reliability. CatBoost regression stands out for its predictive accuracy, while Mixed Effect regression provides a balance between interpretability and predictive capability, both serving as benchmarks for future research. Overall, our study offers critical insights for investors, policymakers, and companies on firm-level actions that indicate future decarbonization, alongside recommendations to enhance the efficacy of future disclosure surveys and the predictive validity of corporate disclosure scores.

Contents

List of Figures

List of Tables

This thesis is dedicated to My Family.

Acknowledgments

TBD. Grazie.

Purpose-driven companies have better outcomes.

George Serafeim

1 Introduction

Chapter Preview

In this chapter I will introduce the topic of **decarbonization** and the role of **corporate emissions** in driving the transition to a low-carbon economy. I will discuss the role of data and disclosures in understanding corporate emissions, focusing on the **Carbon Disclosure Project (CDP)**. I will also review the current literature and highlight the contributions of this thesis to the field. Finally, I will discuss the motivations behind this research.

1.1 The Climate Transition

The world is currently facing an unprecedented climate crisis, with global temperatures rising at an alarming rate and extreme weather events becoming more frequent

and severe. The negative externalities of human-induced greenhouse gas emissions are worrying, with the World Health Organization estimating that climate change will be responsible for over 250,000 additional deaths per year between 2030 and 2050 [57]. There is an urgent need for immediate and sustained action to reduce the negative effects of climate change. To do so, we must transition to a low-carbon economy, by increasing the adoption of renewable energy sources, improving energy efficiency, and promoting sustainable practices. This is one of the most pressing challenges of our time, and it is not an easy task: it requires a concerted effort from governments, businesses, and individuals to reduce greenhouse gas emissions in a race against time, where every second matters [40].

1.2 Companies Drive Decarbonization

The Paris Agreement sets forth ambitious objectives to combat climate change, aiming to cap the increase in global temperatures to 2°C, with an aspirational target of 1.5°C, above pre-industrial levels. This is to be achieved through a series of significant measures, including reaching net-zero greenhouse gas emissions by 2085 and a reduction of these emissions by 10% by 2030 [48]. While a great emphasis is often placed on *what* should be changed, it is not always clear *who* should be responsible for these changes. Should it be goverments designing better policies? Should it be consumers demanding more sustainable products? Should it be non-profits and NGOs advocating for change? Should it be the reader (who is very kind in reading this thesis) turning off the lights when leaving the room? While all of these actors play an important role in the fight against climate change, taking a pragmatic approach, a study from the Carbon Disclosure Project - a leading organization in the field of corporate emissions - shows how about 71% of global greenhouse gas (GHG) emissions from 1988 to 2015 are linked to just 100 companies [14]. So there we have it, we have found our culprits, and we might have also just found our solution.

The data is clear, most of GHG emissions come from companies' activities and, for this reason, firms are the ones that can make the most significant impact in the fight against climate change. Simply put, we need companies to be the driving force

behind decarbonization. Reassuringly, there is an aligment of interests as businesses are governed by executives and influenced by stakeholders who live in the same world as the rest of us, and they too are concerned about the future of the planet (albeit some more than others). As Professor Serafeim argues, there is an evolving business paradigm reflecting a collective desire for companies to positively impact the world [32]. That is,firms are increasingly playing an active role in addressing climate change by setting ambitious emission reduction targets, investing in renewable energy sources, and adopting low-carbon technologies. Companies do not do undertake this effort only because they are good citizens, but also because they see a business opportunity in the transition to a low-carbon economy. As Gallego-Alvarez et al. argue, a reduction in carbon emissions generates a positive impact on financial performance [25]. Ultimately, decarbonization is not to be seen as an inefficiency, but rather a demand for innovation and a source of new business opportunities. By transitioning to a low-carbon economy, companies can create value for their shareholders, employees, and society at large [32].

1.3 The Role of Data

Decarbonization will have a profound impact on the business world and, when it comes to sustainability, it will require significant changes in the way companies operate, creating both risks and opportunities [32]. Companies that are able to reduce their emissions will be better positioned to succeed in the future business landscape, while companies that are unable to do so will face significant risks.

Inevitably, as in any paradigm shift, there will be winners and loosers. A key question remains, how can we tell them apart? In this context, the role of data becomes crucial. Data serves an *internal function*, by helping companies understand their climate risk and identify the best strategies to reduce their emissions. At the same time, data also serves an *external function*, by helping investors allocate their resources more efficiently and policymakers design effective regulations. Such role has been recognized by the public markets, where companies' greenhouse gas emission disclosures are an indicator of performance, with stock prices reflecting estimates of

non-disclosed emissions and a significant market response to climate change information in 8-K filings [26]. Additionally, governments are increasingly interested in regulating data reporting through the implementation of mandatory disclosure requirements. For example the European Union's as of the 5th of January 2023 started enforcing the Corporate Sustainability Reporting Directive (CSRD) which requires companies to disclose their climate-related information. The directive modernises and strengthens the rules concerning the social and environmental information that companies have to report. A broader set of large companies, as well as listed SMEs, will now be required to report on sustainability. Some non-EU companies will also have to report if they generate over EUR 150 million on the EU market [24]. While the European Union is leading the way, other countries are also following suit, with the United States, Switzerland, United Kindgdom, and Canada also implementing similar regulations and with China, India, Israel, and Japan also considering similar measures [31]. In line with this global movement, key institutions such as the Carbon Disclosure Project (CDP) are playing a crucial role in promoting corporate disclosure of climate-related information, driving engagement on environmental issues, and providing a globally consistent disclosure standard for GHG emissions and information on a firm's activities to reduce carbon emissions [16].

1.4 Carbon Disclosure Project

The primary data source for this thesis is the Carbon Disclosure Project (CDP) Climate Change Questionnaire [16], which was kindly provided to me by the Climate and Sustainability Impact Lab from the Digital Design Institute at the Harvard Business School [21]. The Carbon Disclosure Project is a not-for-profit charity that runs the global disclosure system for investors, companies, cities, states and regions to manage their environmental impacts [15]. The importance of the CDP is widely recognized by the business and the academic communities. As Ban Ki Moon, former Secretary General of the United Nations, states "The work of CDP is crucial to the success of global business in the 21st century... helping persuade companies throughout the world to measure, manage, disclose and ultimately reduce their greenhouse gas

emissions. No other organization is gathering this type of corporate climate change data and providing it to the marketplace" [15]. The Carbon Disclosure Project Sustainability Questionnaire uses the Greenhouse Gas (GHG) Protocol as a reporting model for carbon-related data [7]. It is one of the largest datasets of self-reported GHG emissions and collects a wide range of information on climate change-related topics. The questionnaire provides a globally consistent disclosure standard for GHG emissions and information on a firm's activities to reduce GHG emissions. The CDP is backed by a large number of institutional investors, including banks, insurance companies, asset management companies, and pension funds holding US$100 trillion in assets (i.e., CDP signatories), which act as "norm entrepreneurs" [43]. Currently more than 23,000 companies disclose their emission data through the survey, representing two thirds of global market capitalization [15].

1.4.1 Motivations Behind Corporate Disclosure to the CDP

The Carbon Disclosure Project (CDP) questionnaire has not only gained increasing popularity among companies but has also become an important tool for investors and other stakeholders in evaluating corporate climate risks. It plays a crucial role in identifying effective strategies for emission reduction and in navigating the transition towards a low-carbon economy. The growth in the completion and publication rates of the CDP questionnaire reflects its importance, with institutional investors exerting a notable influence on climate change disclosure through corporate communication channels [19]. Consequently, the annual increase in the number of companies engaging in disclosure represents a substantial data pool, invaluable for analyzing the decarbonization process and projecting future emission trends. The rationale for companies to disclose varies: primary reasons include regulatory compliance, investor expectation alignment, reputation enhancement, peer benchmarking, emission reduction opportunity identification, and risk assessment. Furthermore, disclosing to CDP is accomplished through two independent steps: the first involves the completion of the questionnaire, while the second involves the publication of the response. The latter step is particularly significant, as it shows a company's commitment to

transparency and accountability, typically enhancing its reputation and credibility [19].

1.4.2 CDP Scores

The CDP survey assigns a score (Figure 1.4.1) that ranks the performance of companies when decarbonizing. For reference, 48% of S&P companies scored high-performance band B ratings and above in their Carbon Disclosure Project (CDP) reports in 2014 [54]. When assiging a score, CDP assesses the level of detail and comprehensiveness in a response, as well as the company's awareness of environmental issues, its management methods and progress towards environmental stewardship [13]. Additionally, specifically for climate-change scores, to recieve an A-level grade a copmany must verify at least 70% of Scope 1, Scope 2 and Scope 3 emissions [1] with a CDP-approved verification standard. Among other criteria, to score an A on Climate Change, companies must have robust governance and oversight of climate issues, rigorous risk management processes, verified scope 1 and 2 emissions and be reducing emissions across their value chain. Most Climate Change A List companies as of 2022 have well established emissions targets that have been approved by the SBTi, and evidence of targets which cover their scope 3 emissions [13]. The CDP ranking is widely accepted as a benchmark to reflect and review the corporate awareness of environmental challenges, and the best practices for risk mitigation [60].

1.4.3 Limitations

While the CDP score is a valuable and widely recognized metric, it has several limitations. First, the CDP score is assigned based on adherance to best-practices and does not provide a future outlook on the company-specific ability to reduce emissions. It signals that the company is currently adhearing to best practices, but there is no immediate way to know by how much will the company be able to reduce its emissions in the future. Second, the CDP score is based on self-reported data,

[1] For a detailed breakdown of emission scopes, see Appendix A.1

Figure 1.4.1: Caption of CDP Scoring Grades. Adapted from [13].

which can be subject to biases and errors. Third, the score does not provide an estimate of the company's future emissions, which is crucial for investors and policy makers.

1.5 Contributions to The Current Literature

While the field of decarbonization is expansive, addressing topics from economic impacts of climate change to corporate roles in a low-carbon transition, a significant gap remains: **the forecasting of real decarbonization rates at the individual company level**. This thesis bridges this critical gap by leveraging CDP data for predictive modeling, a novel approach within the field [5, 11, 27, 42]. In this review, I will focus specifically on the use of disclosure data to understand what has been done in the past and to define the key questions that remain unanswered. In the following section, I will then use those unanswered questions to motivate the hypothesis and modeling techniques that I will present in the following chapters.

1.5.1 Expanding Beyond Carbon Footprint Estimation

In terms of footprint estimation, Nguyen et al. developed a machine learning model employing advanced techniques like feature selection, boosting, and bagging to predict corporate carbon footprints for climate finance risk analyses [42]. Their innovative approach, particularly the use of the Meta-Elastic Net model, combines insights from multiple models to enhance predictive accuracy. This methodology is not only foundational for understanding present carbon liabilities but also offers a structural baseline for our work. While they concentrate on estimating current footprints using firm-year disclosure data, our research extends these methodologies to project future decarbonization rates. This shift from retrospective analysis to forward-looking forecasts allows for a more dynamic approach to evaluating corporate strategies and their impact on future emissions, setting the stage for an innovative predictive framework in decarbonization research.

1.5.2 Filling the Forecasting Void

In terms of pure forecasting models, the literature is primarily focused on country-level or industry-level forecasts, given that the company-level data is not always avaialble, less standardized, and more difficult to work with. Examples of country-level forecasting include the work of Boateng et al. who show how Artificial neural network models can be useful for aiding climate change policy decision-making as they effectively predict carbon emission intensity for Australia, Brazil, China, India, and USA with negligible forecasting errors [4]. Predictive models have also been applied to specific countries, such as China with the SSA-LSSVM model by Zhao et al. and the combined MNGM-ARIMA and MNGM-BPNN models by Wang et al. [55, 62]. In the context of industry-level forecasting, Liu et al. use a least squares support vector machine to predict CO2 emissions in China's major industries and residential consumption [51] or Yang et al. use an ARIMA model to forecast GHG emissions in Shanghai's aviation industry [58]. While these studies demonstrate the potential of predictive models in understanding and managing climate impacts, they primarily operate at a macro level, overlooking the granular details at the company

level as well as the impact of **individual** corporate actions on future emissions.

1.5.3 Understanding CDP Data in Contemporary Research

Current research widely uses the Carbon Disclosure Project (CDP) data to explore how companies report on their environmental practices and governance. For instance, Ben-Amar et al. have looked into how companies' environmental disclosures relate to the number of women on their boards [11], highlighting the link between good environmental practices and diverse leadership. Similarly, Adel et al. show that companies with better scores from CDP often have better financial results and handle their greenhouse gas information more effectively [5]. This suggests that companies paying attention to the environment might also be better managed overall.

In addition, Hassan et al. focus on Britain's top companies, studying how their environmental reporting, as assessed by CDP, impacts their overall performance [27]. Overall, these studies analyze different ways companies behave and report on environmental matters. Though, they fall short in providing detailed insights into how individual companies operate, rather than general trends in industries or countries. This aspcet is crucial for our work as it helps us understand the specific actions companies are taking towards becoming more environmentally responsible and how these actions impact their future emissions.

1.5.4 The Carbon Disclosure Project data is used for inferencial purposes

Overall, existing studies primarily utilize Carbon Disclosure Project (CDP) data and ratings to explore the correlation between management behaviors and financial outcomes [5, 11, 27, 42]. This thesis is significantly influenced by the research of Lu, Nguyen, and Serafeim, who analyze the dynamics of Incentive Diffusion and Decarbonization rates within the framework provided by CDP data [34]. My work builds on theirs, adopting their dataset as a key starting point for further investigation. Both studies concentrate on real decarbonization rates—actual decreases in emissions rather than shifts resulting from business transactions or changes in production

outputs. However, the primary point of our research diverges: while they examine the effects of incentives on present decarbonization efforts, my focus lies on forecasting decarbonization rates for the upcoming year. In essence, their work adopts an inferential stance aimed at testing specific hypotheses, whereas mine employs a predictive lens to determine the features most indicative of future decarbonization trajectories, leveraging the extensive data provided by the CDP.

1.5.5 Filling the gap: Leveraging CDP Data for Predictive Insights

This thesis takes on the task of forecasting real decarbonization rates at the company level, an area still largely unexplored in research. Filling this gap is crucial, as I believe that understanding the effects of current corporate actions on future decarbonization rates is vital for creating better disclosure methods and identifying which companies are best positioned for the transition to a low-carbon economy. The opportunity to base our analysis on CDP data, the largest dataset of self-reported greenhouse gas emissions, offers a unique chance for advancement in this field and for evaluating the variables that will have the most significant impact in terms of future results, and not just in terms of current compliance. The hopes are high, let's get started!

1.6 Motivation

This thesis aims to leverage the potential of machine learning algorithms (ML) to enhance our understanding and capabilities in forecasting corporate decarbonization efforts. By comparing various modeling techniques, we not only set a benchmark for future research but also contribute to the strategic decisions critical for environmental and economic sustainability by providing all stakeholders with novel insights on compliance data to predict future emisisons.

1.6.1 Significance and Stakeholder Impact

While decarbonization should be a topic of interest to all, the work undertaken here is particulary relevant for several key stakeholders, each impacted by and contributing to the field of decarbonization in a unique way.

Investors: They require precise, predictive insights to assess the climate risk and opportunities within their portfolios, seeking out leaders in sustainable practices for potential investment. Using compliance data to predict future emissions can provide investors with a competitive edge in identifying companies best positioned for the transition to a low-carbon economy and allocate resources more efficiently, prioritizing firms with a commitment to decarbonization that is likely to result in tangible future emissions reductions.

Companies: With an increasing need to understand and mitigate their environmental impact, companies can use this research to gauge their performance, strategize emissions reduction, and position themselves effectively for the transition. A key aspect is identifying which activities have an impact not only on same-year emissions but also on future emissions, allowing for more informed prioritization of decarbonization efforts.

Policy Makers: Detailed and predictive emissions data is vital for creating informed, effective policies that encourage corporate and sector-wide sustainability efforts. While there is ample literature on the impact of policies on emissions, and on emissions aggregated at the country or industry level, policymakers can also benefit from understanding the impact of individual corporate actions on future emissions, and understanding the strenght and weaknesses of the current disclosure systems.

The Carbon Disclosure Project Climate Survey: As a platform collecting extensive corporate environmental data, the CDP can significantly benefit from this research. Enhancements to their survey and data collection methods, informed by the findings of this study, could improve the quality and utility of the information they gather, enabling better corporate environmental accountability. Furthermore, the results of this study can provide a valuable starting point to design a new grading system that is more forward-looking, outcome-baed, and predictive of real decar-

bonization.

1.7 Research Questions and Roadmap

1. **Geography and Industry:** Are decarbonization rates uniform across countries and regions, or do they vary significantly? Are all industries decarbonizing at the same rate, or are there important differences between them? Which firms are leading the way in decarbonization and what can we learn from them?

   In **Chapter 2** we contuct Exploratory Data Analysis (EDA) to understand the distribution of decarbonization rates across countries, regions, and industries, and identify any significant variations. We then provide a comprehensive overview of the global decarbonization landscape, which for the first time is based on real decarbonization rates at the company level and not change in emission footprint.

2. **Modeling Individual Firm Actions:** How can we take into account individual firm behavior and at the same time find general trends? What model allows us to do this the best? How do corporate actions detailed in the CDP data influence future decarbonization rates, and can we derive an optimal set of predictors from these actions that best explain decarbonization efforts under a set modeling framework?

   In **Chapter 4** we develop and test iterations of Mixed-Effects models that handle individual firm data while identifying generalizable trends across the dataset, comparig the performance of each and deriving an optimal model that strikes an optimal balance between number of features, modeling individual firm behavior, and generalizing across the dataset.

3. **Enstablishing a benchmark predictive model:** Which machine learning models and features best forecast company-specific decarbonization rates? Can we set a benchmark for future research and application? Which features are most important for predicting decarbonization rates, and do they change significantly across modeling techniques?

In **Chapter 5** we develop a state-of-the-art ML model for forecasting company-specific decarbonization rates, thereby establishing a foundational benchmark for subsequent research and application. This involved identifying the most effective ML algorithms, comparing them against each other for interpreting and utilizing CDP data for predictive purposes. We also analyze the most important feautres for predicting decarbonization rates and find that they do not change significantly across models, signaling that the most important predictors are robust and generalizable across different modeling techniques.

4. **Improving the CDP Climate Survey:** Can we use our results to advise the CDP on how to improve their survey for better future emission forecasting and decarbonization strategies? Can we design a better scoring system to assign grades to companies based on their decarbonization efforts and future commitments?

   In **Chapter 6** we provide recommendations to the CDP on how to improve their survey for better future emission forecasting and decarbonization strategies, based on the results of our analysis. We furthermore propose a new scoring framework that more accurately reflects companies' actual and projected decarbonization efforts, incorporating future commitments.

2

Data Sources

Chapter Preview

We present the three primary data sources: the **CDP Climate Change Response Questionnaire**, the **Worldscope Fundamental Core Items**, and the **Global Industry Classification Standard (GICS)**. The CDP Climate Survey data contains the response surveys from all companies from 2011 to 2022, the Worldscope database provides detailed standardized financials, and the GICS offers a hierarchical classification system for industries.

2.1 Overview of Data Sources

2.1.1 Climate Change Response Questionnaire

The main data source comes from the CDP Climate Survey, it contains the response surveys from all companies from 2011 to 2022. The data was partially cleaned and processed by the Climate and Sustainability Impact Lab [21] before being shared with me. This comprehensive dataset was provided a repository that includes both raw and processed data in the form of Stata files. Organized by firm-years, each observation in the dataset corresponds to a specific firm in a given year, and is structured in a panel format, having a unique id year pair to uniquely identify each entry. The original dataset contained $34,588$ firm-years across 11 years. Since the analysis controls for financial and industry-specific predictors, I decided to focus on public companies, which represent 71% of the firm-years in the dataset. Therefore, $9,785$ firm-years were dropped from the analysis because they did not have an ISIN code, which is a unique identifier for public companies. A detailed breakdown of the data cleaning process is provided in the following section 2.2.

Important Considerations on the CDP Data

- **Reporting year lag:** The data from a given year corresponds to the financial and operational data from the previous year.This was an important consideration when merging the CDP data with other data sources, such as the Worldscope financial data.
- **Data processing:** The original data processing entailed the extraction of multiple sections from the survey, which were then systematically aligned across different years, ensuring consistency across times and adjusting the format when the questions on the CDP surveyed changed or were slightly modified. It is important to note that the fact that some questions were not asked in some years, and that the questions were not always the same across years, is a significant challenge for the analysis which is specifically focused on forecasting emissions.

2.1.2 Worldscope Fundamental Core Items

In addition to the CDP Climate Survey data, financial predictors were obtained using the Worldscope database [3] accessed through Wharton Research Data Services (WRDS) [56]. Worldscope offers detailed standardized financials, allowing for comparisons of financial information across companies from various industries worldwide. This database boasts a long history, with over 35 years of data for key developed markets dating back to 1980 and more than 25 years for emerging markets. With its extensive coverage of over 100,000 companies in more than 120 countries, including full standardized coverage of over 30 developed and emerging markets and accounting for 99% of global market capitalization, Worldscope is a comprehensive source for firm-level data. Specifically, I queried the fundamental annuals through Worldscope via WRDS, which provided key global information such as *revenue*, *total assets*, *number of employees*, and *net income*, which I then used as predictors for my analysis [3]. Data was retrieved based on the ISIN code, and resulted in 96% of the firm-years having matching financial data. Of those, 17% had missing values for *at least one* of the selected financial variables, thus the corresponding firm-years were dropped from the data-set. This choice has been made as firms with missing financial data are likely to have total assets less than 1 million, thus I removed them following a similar criteria enstablished by Serafeim et Al. [34].

2.1.3 GICS

Accessed through WRDS using Capital IQ, the Global Industry Classification Standard (GICS) provides the framework for this study's industry analysis. GICS, a collaborative creation by MSCI and S&P Dow Jones Indices, offers a hierarchical, four-tiered classification system, encompassing Sectors, Industry Groups, Industries, and Sub-Industries. This standard ensures a consistent approach to defining company activities worldwide, crucial for comparative financial analysis. The classification of a company within GICS hinges on its principal business activity, with revenue being a primary determinant. The system also considers earnings and market perception, elements that contribute to the annual refinement of the classifications to

mirror evolving market conditions. This research utilizes the 25 industry groups defined within GICS, facilitating a detailed examination of firm-level data against a backdrop of global industry standards [1, 2]. I queried the GICS data only for the firm-years that had matching financial data, resulting in 19,200 firm-years with complete financial and GICS data. GICS data was avaialble for 99% of the firm-years that had matching financial data.

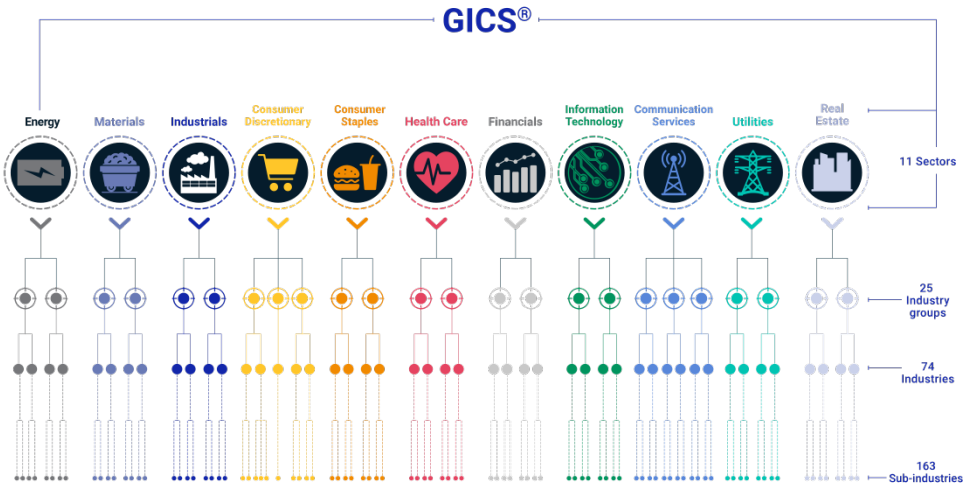


Figure 2.1.1: Global Industry Classification Standard (GICS) Structure [2].

2.2 Data Cleaning Process Flowchart

This is a visual representation of the data cleaning process described in the data sources with a specification of the number of firm-years dropped at each step:

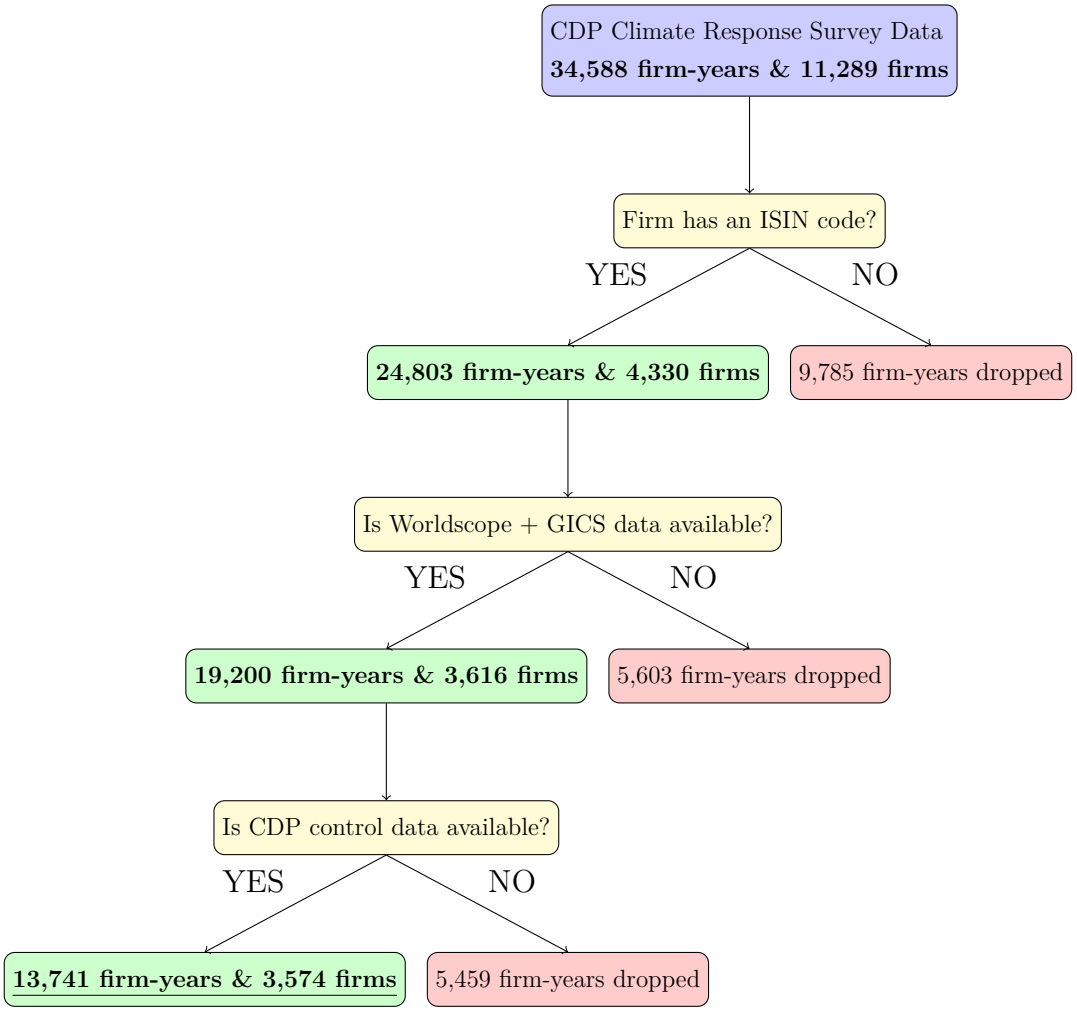


Table 2.2.1: Data Cleaning Process Flowchart

The final dataset contains **13, 741 firm-years across 3, 574 firms, with complete CDP, Worldscope, and GICS data**.

2.3 Exploratory Data Analysis

2.3.1 The Response Variable: Next-Year Real Decarbonization Rate

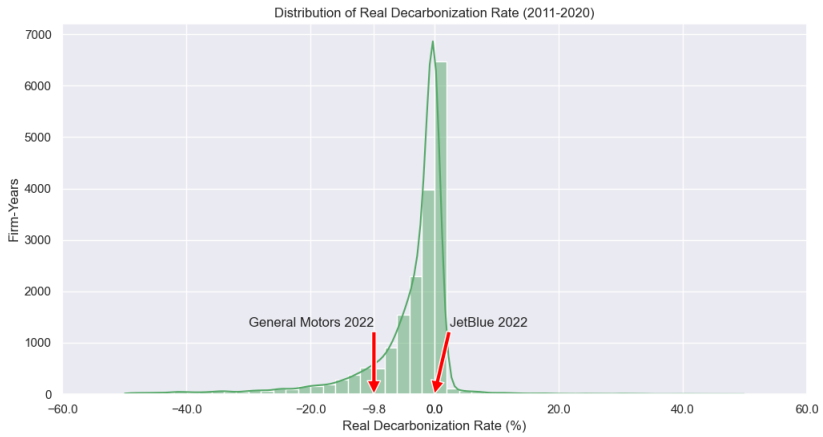


Figure 2.3.1: Real Decarbonization Rate

This is the distribution of our response variable: real decarbonization rate. As we can observe from Figure 2.3.1, the distribution has three key features (i) has a negative mean and a heavy left tail - that is on average firms are (gladly) reducting real emissions, with varying level of success (ii) the mode is zero - most firms don't change their real emissions over time, that is they do not make any change to their operations or their renewable energy stupply (iii) some frims increase their real emissions, this is a minority but can be observed in the right tail.

Why this response variable?

Out interest is in determining which firms are taking action to reduce their carbon footprint by adopting better technology, and not by operating on other indirect metrics. That is, we overall amount of emissions to go down at a global level, and therefore we are not interested if a firm reduces its emissions by divesting from a

subsidiary, or by producing less. Rather, we want to indentify and forecast which firms are taking actions that are going to enable them to emit less and operate more efficiently, and ideally we want those firms to be rewarded. In this regard, real decarboninzation rate only takes into account change in emissions due to process and renewable energy use.

2.3.2 Number of Unique Reporting Firms by Year

As can be observed from Figure 2.3.2, over time the number of unique firms reporting increases. Note that the total number of firms reporting is even higher, this is an illustration of the firms that were selected in the training set. As a requirement, the firm must have been reported for at least three years. Therefore, not only the number of firms reporting to the CDP is increasing, but also the firms who do so over time. I expect this trend to continue, and this is a positive sign since with more firms reporting, more data will be available to build better models and enhance our understanding of emissions' forecasting.

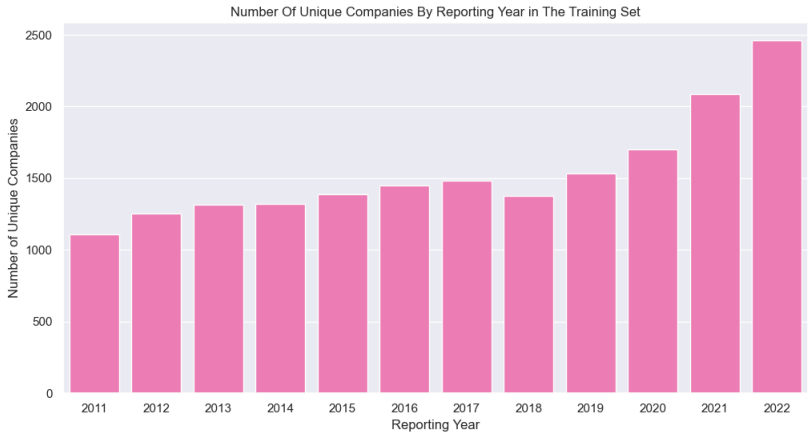


Figure 2.3.2: Number of unique reporting firms by year in the training set

2.3.3 Real Decarbonization Rate Breakdown by Continent

Figure 2.3.3 and the relative table show the mean real decarbonization rate by continent across all CDP reporting years from 2011 to 2022. As expected, there is significant class imbalance between continents, with Europe having the most number of firms, followed by North America and Asia. There is a significant difference in the mean decarbonization rate across continents, with Europe having the best mean decarbonoization rate with an average yearly decrease of -4.94% and Africa having the worst mean decarbonization rate with an average yearly decrease of -2.81%. Overall, the data suggests that operating in an environment with more incentives to report and reduce emissions, such as Europe, is associated with a higher mean decarbonization rate. This is consistent with the findings of Downar et al. [23] which shows in a UK-based study that firms with a carbon disclosure mandate reduced emissions by 8% without negatively impacting their financial operating performance. The hypothesis will be further tested in the following sections.

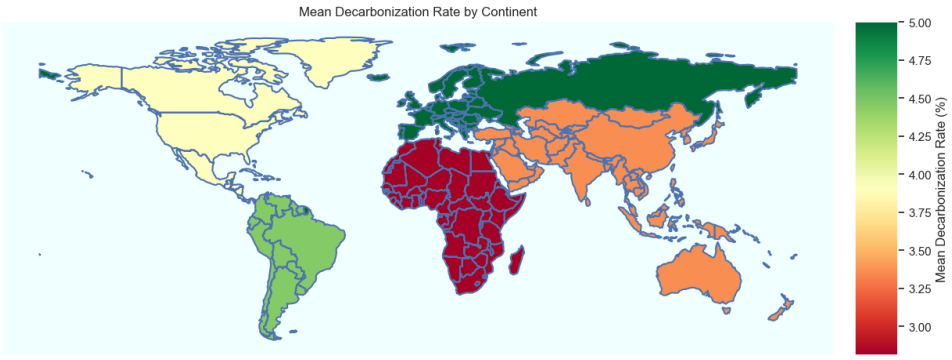


Figure 2.3.3: Mean *real decarbonization rate* by continent from 2011 to 2022

Continent	# firms	Mean	Median	Std
Africa	79	-2.8%	-0.9%	5.9%
Asia	1175	-3.1%	-1.0%	6.4%
Europe	1217	-4.9%	-1.9%	8.6%
North America	933	-3.5%	-1.0%	7.3%
Oceania	101	-3.1%	-0.6%	6.9%
South America	90	-4.0%	0.0%	10.1%

Table 2.3.1: Emissions Breakdown By Continent

2.3.4 Real Decarbonization Rate Breakdown by Sector

Figure 2.3.4 and the relative table show the mean real decarbonization rate by sector across all CDP reporting years from 2011 to 2022. The data shows that the mean decarbonization rate varies significantly across sectors, with the best mean decarbonization rate in the Software and Services sector, with an average yearly decrease of -6.67%, and the worst mean decarbonization rate in the Materials sector, with an average yearly decrease of -2.46%. Additionally, there are significant differences in the number of firms across sectors, with the Capital Goods sector having the most number of firms, 475 and the Household and Personal Products sector having the least number of firms, 43. Differences in sectors are important to consider, as they can be indicative of the difficulty of decarbonizing a given industry. For example, our data suggests that Transportation and Materials are the sectors with the worst mean decarbonization rates, which is consistent with the findings of Davis et al. [20] which suggest that difficult-to-decarbonize energy services include aviation, long-distance transport, steel and cement production.

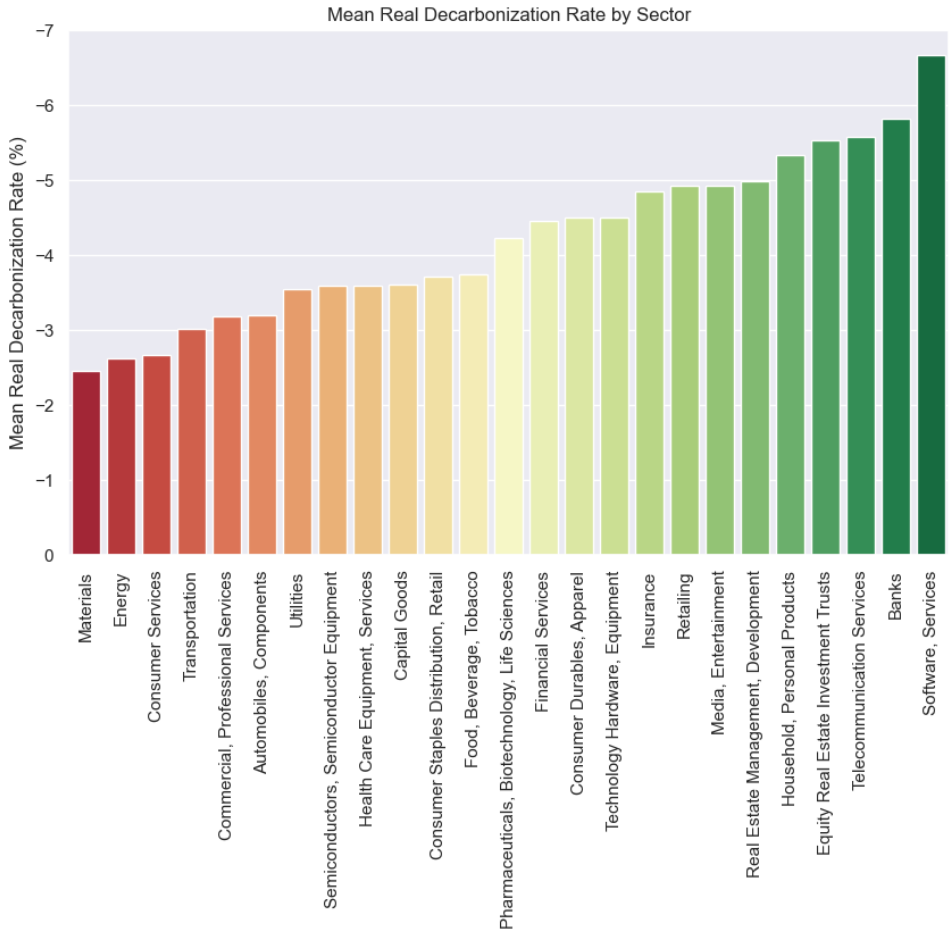


Figure 2.3.4: Mean Real Decarbonization Rate by Sector

Sector	# firms	Mean	Median	Std
Automobiles, Components	124	-3.2%	-1.91%	5.54%
Banks	166	-5.82%	-2.9%	9.46%
Capital Goods	475	-3.6%	-1.1%	6.87%

Continued on next page

Sector	# firms	Mean	Median	Std
Commercial, Professional Services	134	-3.18%	-0.04%	7.7%
Consumer Durables, Apparel	127	-4.5%	-1.4%	8.28%
Consumer Services	89	-2.67%	-0.96%	6.65%
Consumer Staples Distribution, Retail	65	-3.72%	-1.9%	6.72%
Energy	151	-2.62%	-0.15%	5.66%
Equity Real Estate Investment Trusts	96	-5.53%	-2.33%	8.9%
Financial Services	158	-4.45%	-0.6%	8.98%
Food, Beverage, Tobacco	187	-3.75%	-1.5%	6.63%
Health Care Equipment, Services	100	-3.6%	-0.9%	7.45%
Household, Personal Products	43	-5.33%	-2.4%	8.09%
Insurance	96	-4.85%	-2.0%	8.39%
Materials	420	-2.46%	-0.6%	5.78%
Media, Entertainment	70	-4.93%	-0.54%	8.25%
Pharmaceuticals, Biotechnology, Life Sciences	97	-4.23%	-1.8%	7.63%
Real Estate Management, Development	53	-4.99%	-1.1%	8.86%
Retailing	115	-4.93%	-1.3%	9.43%
Semiconductors, Semiconductor Equipment	79	-3.6%	-0.5%	8.15%
Software, Services	140	-6.67%	-2.85%	10.21%
Technology Hardware, Equipment	185	-4.5%	-1.8%	8.76%
Telecommunication Services	77	-5.58%	-2.34%	9.26%
Transportation	155	-3.02%	-1.0%	6.3%
Utilities	173	-3.54%	-0.1%	8.08%

Table 2.3.2: Real decarbonization rate breakdown by industry

2.3.5 Real Decarbonization Rate Breakdown by Country

Figure 2.3.5 shows the mean real decarbonization rate by country across all CDP reporting years from 2011 to 2022. There are significant differences both in the number of firms and in the mean decarbonization rate across countries. Table 2.3.3 shows summary statistics for the worst 10 performing countries with nonzero mean real decarbonization rates. For a complete list of countries, see appendix table.

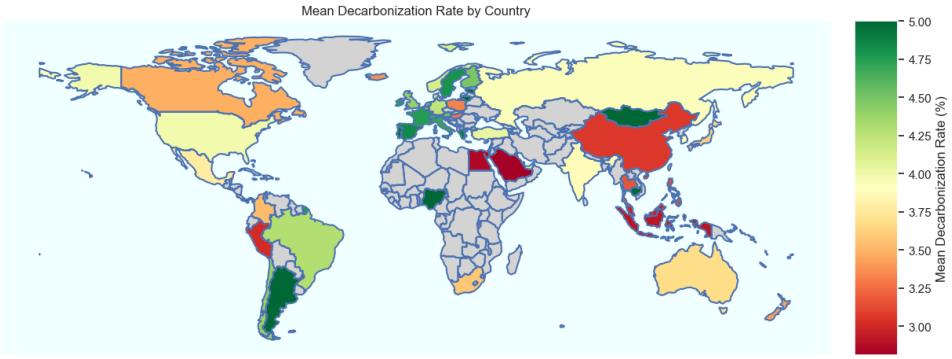


Figure 2.3.5: Mean Decarbonization Rate by Country

Country	# firms	Mean	Median	Std
Saudi Arabia	1	-0.6%	-0.6%	nan%
Egypt	2	-1.46%	0.0%	2.85%
Indonesia	10	-1.53%	0.0%	2.76%
Malaysia	13	-1.57%	0.0%	6.19%
Cayman Islands	2	-1.62%	0.0%	7.54%
Peru	1	-1.67%	0.0%	4.53%
China	78	-1.76%	0.0%	5.85%
Hong Kong	35	-1.7%	-0.27%	7.43%

Continued on next page

Country	# firms	Mean	Median	Std
Philippines	12	-1.83%	0.0%	4.57%
Thailand	19	-1.85%	0.0%	5.8%

Table 2.3.3: Emission Breakdown by Country

2.3.6 Real Decarbonization Rate Breakdown by Year

Figure 2.3.6 shows the mean and median real decarbonization rate by year across all CDP reporting years from 2011 to 2022. The data shows that the mean and median decarbonization rates have been (assuringly) decreasing over time.

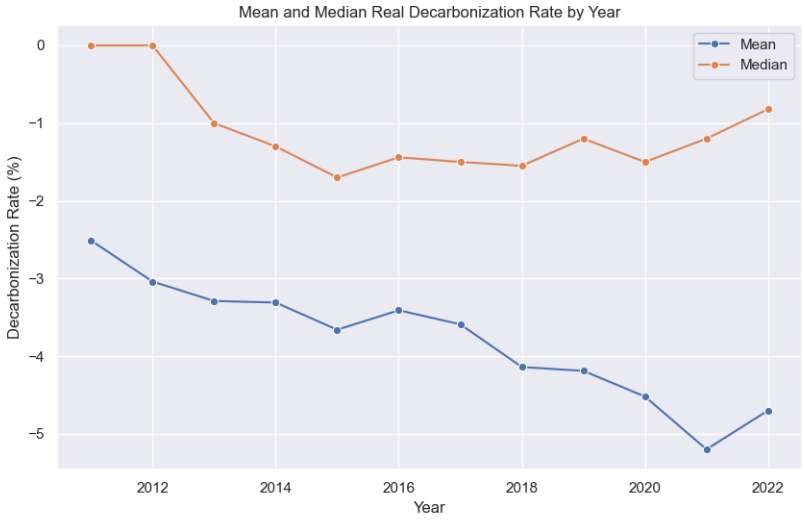


Figure 2.3.6: Mean and Median Real Decarbonization Rate by Year

Year	Count	Mean	Median	Std
2011	1109	-2.51%	0.0%	6.09%
2012	1252	-3.04%	0.0%	6.03%
2013	1317	-3.29%	-1.0%	5.85%
2014	1322	-3.31%	-1.3%	6.01%
2015	1388	-3.66%	-1.7%	6.1%
2016	1451	-3.41%	-1.44%	5.97%
2017	1483	-3.59%	-1.5%	6.68%
2018	1374	-4.14%	-1.55%	7.69%
2019	1530	-4.19%	-1.2%	8.08%
2020	1701	-4.52%	-1.5%	8.59%
2021	2088	-5.2%	-1.2%	9.75%
2022	2461	-4.7%	-0.82%	9.79%

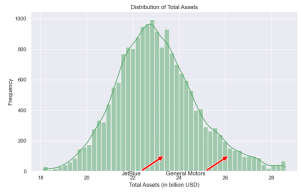
Table 2.3.4: Real decarbonization rate by year

2.4 Financial Predictors Feature Engineering

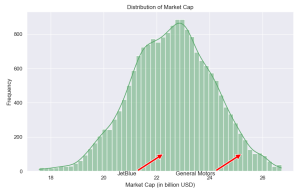
Figure 2.4.1 shows the distribution of the financial predictors used in the analysis. This is a list of each predictor along with a brief description of how it was derived:

- **Total Assets 2.4.1a**: The total assets of the firm, which is a measure of the firm's size and the scale of its operations. Directly obtained from the Worldscope database and transformed using the natural logarithm $log(1 +$ Total Assets$)$.
- **Market Capitalization 2.4.1b**: The market capitalization of the firm, which is a measure of the firm's size and the scale of its operations. Directly obtained from the Worldscope database and transformed using the natural logarithm $log(1 +$ Market Cap$)$.

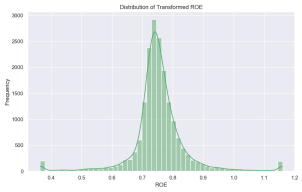
- **Return on Equity 2.4.1c**: The return on equity of the firm, which is a measure of the firm's profitability. Since the return on equity is a percentage which can be negative, the following transformation was used: $log(1 + \frac{\text{ROE}}{100})$.
- **Revenue 2.4.1d**: The total revenue of the firm, which is a measure of the firm's size and the scale of its operations. Directly obtained from the Worldscope database and transformed using the natural logarithm $log(1 + \text{Revenue})$.
- **Net Income 2.4.1e**: The net income of the firm, which is a measure of the firm's profitability. Directly obtained from the Worldscope database and transformed using the natural logarithm $log(1 + \text{Net Income})$.
- **Employees 2.4.1f**: The total number of employees of the firm, which is a measure of the firm's size and the scale of its operations. Directly obtained from the Worldscope database and transformed using the natural logarithm $log(1 + \text{Employees})$.
- **Total Assets 1yr Growth 2.4.1g**: The one year growth of the total assets of the firm, which is a measure of the firm's growth. Directly obtained from the Worldscope database and since the growth can be negative, the following transformation was used: $log(1 + \frac{\text{Total Assets 1yr Growth}}{100})$.
- **Employees 1yr Growth 2.4.1h**: The one year growth of the total number of employees of the firm, which is a measure of the firm's growth. Directly obtained from the Worldscope database and since the growth can be negative, the following transformation was used: $log(1 + \frac{\text{Employees 1yr Growth}}{100})$.
- **Net Income over Assets 2.4.1e**: The net income of the firm over its total assets, which is a measure of the firm's profitability. The feature was calculated with the following formula $log(1 + \frac{\text{Net Income}}{\text{Total Assets}})$.



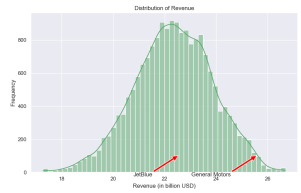
(a) Total Assets



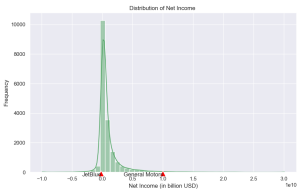
(b) Market Capitalization



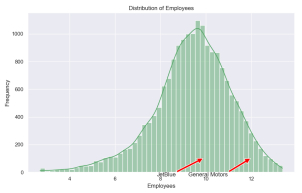
(c) Return on Equity



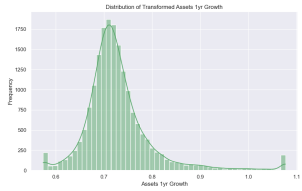
(d) Revenue



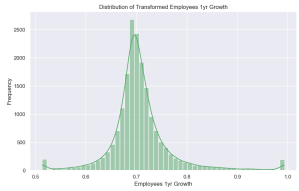
(e) Net Income



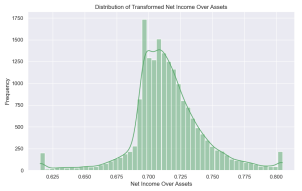
(f) Employees



(g) Tot. Assets 1yr Growth



(h) Employees 1yr Growth



(i) Net Income Over Assets

Figure 2.4.1: Financial Predictors

3 Methods

Chapter Preview

We describe the three primary modeling techniques used in this study: **Linear Mixed-Effects regression**, **Bayesian Ridge regression**, and **CatBoost regression**. We chose to explore these models because they are **well-suited for the analysis of longitudinal data**, they provide a good balance between model complexity and generalizability, and they are **particularly effective for handling categorical variables with a high number of levels**.

3.1 Data Considerations

Throughout our analysis, our primary focus is on the Next-Year Decarbonization Rate, a continuous variable. Our goal is to identify key predictors and assess their

impact on this response within a multiple timeseries framework, where each firm's annual reports represent individual yearly observations. With an average data span of just 6 years, the dataset is too small for the application of deep learning or traditional single-firm timeseries models. Therefore, our strategy involves employing models that allow us to borrow strength across firms and industries while ensuring the outcomes remain interpretable. Moreover, we must navigate the high dimensionality of our data, which includes 130 predictors, among which are categorical variables with a vast number of levels.

I have chosen Mixed-Effects models as they are exceptionally suitable for our longitudinal data analysis. This approach enables us to introduce random intercepts for categorical variables with extensive levels, specifically *Id*, *Country*, *Industry*, and *Continent*. By doing so, we can effectively borrow strength across these dimensions. Mixed-Effects models offer a balance between interpretability and the necessity for manual tuning. In our process of determining the most significant predictors, we will, in Chapter 4, proceed to develop models of increasing complexity. We start by establishing the optimal panel structure, then methodically incorporate groups of CDP predictors by their significance, including Emission Figures, Investment, Incentives, Risks, and Opportunities. Each group will be added to the model iteratively, thus building increasingly complex models.

This step-by-step model building approach allows us to analyze key relationships to gather insights from CDP data and to ensure that those reletionsihps are both logical and intuitive. As you will see, we will find surprisignly coherent and consistent relationships, indicating that the CDP data can be highly informative of future decarbonization efforts. At every step, we will advance the most relevant predictors to subsequent models to prevent overfitting. Our model selection and refinement process is guided by the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), which serve as tools to compare the statistical quality of our models.

3.2 Mixed Effects Models

The main statistical method used in this study is the linear mixed-effects model (LMM) implemented via the R package lme4 [10]. LMMs are a generalization of linear regression models that allow for the inclusion of both fixed and random effects. Fixed effects are the parameters of interest, while random effects are used to account for the correlation between observations within the same group. Mixed-effects models are particularly useful for longitudinal data, where repeated measurements are taken on the same subjects over time. In our case, the repeadted measurements are the yearly company CDP reports, given that the same companies report their emissions over multiple years. Furthermore, we use random effects for categorical variables that have a high number of levels, such as countries, industries, companies and continent. This approach allows the model to capture the variability in the data due to these categorical variables, while also reducing the risk of overfitting and improving the generalizability of the model. As Bates explains, mixed-effects models are also known as *multilevel* models becauase the random effects represent levels of variation in addition to the pre-observation noise term that is incorporated in common statistical models such as linear regression models, generalized linear models, and nonlinear regression models [9]. Furthermore, there have been many successful applications of mixed-effects models in problems with similar settings, such as the work of Maruotti et al. who use bivariate bidimensional mixed-effects regression model and effectively captures heterogeneity in CO2 emissions and growth across OECD countries from 1990 to 2018 [38], or the work of Zanin et al. who assessed the functional relationship between CO2 emissions and economic development using an additive mixed model approach [61].

3.2.1 Model Specification

Let's consider a simple linear mixed-effects model having next year decarbonization rate as a response with an intercept, a single fixed effect, and a single random intercept. In our case, the fixed effect will be the year of the report, and the random effect will be the company. Let y_{ij} be the response variable at the i-th year for the j-th

company, let x_{ij} be the fixed effect at the i-th year for the j-th company, and let z_{ij} be the random effect at the i-th year for the j-th company. The linear mixed-effects model can be written as [53]:

$$Y_{ij}|\alpha_j = \alpha_j + \beta_0 + \beta_1 X_{ij} + \epsilon_{ij} \tag{3.1}$$

$$\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2) \tag{3.2}$$

$$\epsilon_{ij} \sim N(0, \sigma^2) \tag{3.3}$$

$$i = 1, 2, ..., n_j \tag{3.4}$$

$$j = 1, 2, ..., J \tag{3.5}$$

Where J is the number of companies, n_j is the number of years for the j-th company, α_j is the random intercept for the j-th company, μ_α is the mean of the random intercepts, σ_α^2 is the variance of the random intercepts, β_0 is the fixed intercept, β_1 is the fixed effect, and ϵ_{ij} is the error term. Note how the corresponding OLS model would require to control for the company fixed effects by including a dummy variable D_j for each company:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \sum_{j=1}^{J} \alpha_j D_j + \epsilon_{ij} \tag{3.6}$$

In the ordinary least squares (OLS) model, we introduce a separate parameter for each of the J groups which leads to the estimation of J additional coefficients. This can be inefficient, particularly when J is large. In contrast, in the mixed-effects model we introduce only two additional parameters for the random effects: σ_α^2, which represents the variance of the random intercepts across groups, and σ_e^2, which represents the residual variance within groups. These parameters are estimated from the data using the Restricted Maximum Likelihood (REML) method [9]. REML is particularly effective because it corrects for the potential bias introduced by estimating fixed effects alongside the variance components. This approach enhances model efficiency and reduces the risk of overfitting, especially in scenarios with a large number of groups. Once the variance components σ_α^2 and σ_e^2 are estimated, the model employs

a process known as *shrinkage* to compute the individual random effects (intercepts) for each group. Shrinkage is a regularization technique that adjusts the individual group estimates, pulling them towards a central value, typically towards the overall population mean, which in the context of many mixed models is assumed to be zero [9]. This shrinkage effect is governed by the relative sizes of σ_α^2 and σ_e^2. Specifically:

- If σ_α^2 is large compared to σ_e^2, it indicates significant variability among the groups. Thus, the random intercepts for each group are allowed to deviate more from the overall mean, reflecting the distinct characteristics of each group.
- Conversely, if σ_α^2 is small relative to σ_e^2, it suggests that the groups are not markedly different from each other. As a result, the random intercepts are more heavily shrunk towards the central value, reducing the differences among group intercepts.

This approach allows the mixed-effects model to balance between capturing the unique attributes of each group and maintaining model parsimony and generalizability, making it particularly useful when dealing with a large number of groups.

3.2.2 Extension to Multiple Fixed Effects and Random Slopes

The linear mixed-effects model can be extended to include multiple fixed effects and random slopes. Let X be a $n \times p$ matrix of fixed effects, where n is the number of observations and p is the number of fixed effects. Let Z be a $n \times q$ matrix of random effects, where q is the number of random effects. The linear mixed-effects model can be written as [53]:

$$Y|\gamma = X\beta + Z\gamma + \epsilon \tag{3.7}$$

$$\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I) \tag{3.8}$$

$$\gamma \sim \mathcal{N}(\mathbf{0}, \sigma^2 \Sigma) \tag{3.9}$$

$$\gamma \perp \epsilon \tag{3.10}$$

We followed the notation of Baayen et al. [8], where Y is a $n \times 1$ vector of response variables, in our case *Next Year Decarbonization Rate*, β is a $p \times 1$ vector of all the CDP predictors that we test in various iterations, γ is a $q \times 1$ vector of random effects, in our case *Id*, *Country*, *Continent*, and *Industry*, ϵ is a $n \times 1$ vector of error terms, Σ is a $q \times q$ covariance matrix of random effects, and I is the identity matrix. The covariance matrix Σ is estimated from the data using the REML method in the lmer package [9].

3.3 Bayesian Ridge Regression

3.3.1 Considerations

The choice to deploy a Bayesian Ridge regression model specifically addressess the main shortcoming of the mixed-effects model, which is the **lack of regularization and variable selection**. While I manually selected optimal variables for Mixed-Effecs models, ensuring that the model was not overfitting, the Bayesian Ridge regression model automatically selects the most relevant variables and regularizes the coefficients. This is particularly useful in the context of our study, where we have a large number of CDP predictors and we are interested in identifying the most important ones. Furthermore, the Bayesian Ridge regression model provides confidence intervals for the coefficients, which can be highly informative and lead to interesting interpretations. Bayesian Ridge regression has been successfully applied in various fields, such as Forecasting COVID-19 outbreak progression by Saqib [49] or to predict Overall Equipment Effectiveness performance in corporate setting by Imane et al. [30]

Bayesian techniques can be used to include regularization parameters which is not set in a hard sense but tuned to the data. This can be done by introducing uninformative priors over the hyper parameters of the model [36, 46]. The l_2 regularization parameter used in Ridge regression is equivalent to finding a maximum a posetiori estimation under a Gaussian prior over the coefficients w with precision α^{-1} [46]. Effectively, we can use a Bayesian method to replicate regularization, with

the advantage of being able to build confidence intervals for the coefficients. Using this model makes sense given our objectives, as we have a large number CDP of features and we are interested in finding the most relevant ones, while at the same time avoiding overfitting and building an overly complex model. Furthermore, a bayesian approach is a great complement to the mixed-effects model, as it provides confidence intervals for the coefficients, which can be highly informative in the context of our study. We will report the specification presented by Tipping [52] which is implemented in the scikit-learn library [46].

3.3.2 Model Specification

Consider a data set of input-target pairs $\{\mathbf{X_n}, t_n\}_{n=1}^{N}$ we follow the standard probabilistic formulation and assume that targets are samples from the model with additive noise:

$$t_n = y(\mathbf{x}_n, \mathbf{w}) + \epsilon_n \tag{3.11}$$

where ϵ_n iare independent samples from some noise process which is assumed to be mean-zero Gaussian with variance σ^2 [52]. Therefore we have:

$$p(t_n|\mathbf{x}) = \mathcal{N}(t_n|y(\mathbf{x}_n), \sigma^2) \tag{3.12}$$

the notation specifies a Gaussian distribution over the target variable t_n with mean $y(\mathbf{x}_n, \mathbf{w})$ and variance σ^2. Assuming independence of t_n, we can write the likelihood of the complete data set as [52]:

$$p(\mathbf{t}|\mathbf{w}, \sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}||\mathbf{t} - \Phi\mathbf{w}||^2\right\} \tag{3.13}$$

$$t = (t_1, t_2, ..., t_N)^T \tag{3.14}$$

$$w = (w_1, w_2, ..., w_N)^T \tag{3.15}$$

Then, we encode a preference of smoother functions by using a zero-mean Gaussian

prior over the parameters $\mathbf{w}$, with precision α [52]:

$$p(\mathbf{w}|\alpha) = \prod_{i=0}^{N} \mathcal{N}(w_i|0, \alpha^{-1}) \tag{3.16}$$

$$\tag{3.17}$$

where α is a vector of $N+1$ hyperparameters. Note that there is an individual hyperparameter associated independently with every weight. To complete the specification of this hierarchical model, we define hyperparameters over α as well as the final remaining hyperparameter σ^2, we do so by using Gamma priors[52]:

$$p(\alpha_i) = \text{Gam}(\alpha_i|a, b) \tag{3.18}$$

$$p(\beta) = \text{Gam}(\beta|c, d) \tag{3.19}$$

$$\beta \equiv \frac{1}{\sigma^2} \tag{3.20}$$

$$\text{Gam}(\alpha|a, b) = \Gamma(a)^{-1} b^a \alpha^{a-1} \exp(-b\alpha) \tag{3.21}$$

3.3.3 Non informative Priors

The key aspect of the model is that we intentionally make these priors non-informative by fixing their parameters to small values e.g. $a = b = c = d = 10^{-4}$ [52]. The incredibly interesting aspect of such approach is that this formulation of priors is a type of *automatic relevance determination* [37] where a broad prior over the hyperparameters allows the posterior to concentrate at very large values of some of those α values, with the advantage that the associated posterior weights probability will be concentratd around zero, effectively eliminating the corresponding inputs and deeming them as irrelevant, thus providing a form of automatic feature selection [52].

3.4 CatBoost Regression

To achieve maximum predictive accuracy, and address the high dimensionality of the data, I chose to use the CatBoost algorithm. CatBoost is a state-of-the-art machine learning algorithm developed by Yandex, specifically designed to efficiently handle categorical variables [47]. This algorithm is particularly useful in our case as we have multiple important categorical predcitors, namely *Id*, *Country*, *Industry*, and *Continent* which are crucial to the analysis but at the same time have a high number of levels. If we were to one-hot encode these categorical variables, we would end up with a large number of features, which would make the model more prone to overfitting and would require a larger dataset to train effectively, which we do not have. This is one of the key reasons of why we previously used mixed-effects models Therefore, the CatBoost algorithm is particularly well suited for forecasting the decarbonization rate of companies based on the CDP data as it allows to use the whole dataset for training and to handle categorical features efficiently. Additionally, there is ample literature demostrating that CatBoost is a state-of-the-art algorithm for multi-source data, which is precisely the case of our study. As Pan et al. explain, the CatBoost model is particularly suited for multi-source heterogeneous data, outperforming popular machine learning algorithms like random forest and gradient boosting decision tree [44]. We find similar results in our own study, as we will show in chapter 5. Additionally, CatBoost has been successfully used in multiple emission forecasting settings, such as in the research of Marco et al. to forcast carbon dioxide emissions of light-duty vehicles [41] or in the work of Bai to predict emission characteristics in low carbon biofuel-hydrogen dual fuel engines [50].

Handling Categorical Features

CatBoost transforms categorical features into numerical values using an efficient strategy which reduces overfitting and allows to use the whole dataset for training [22]. In particular, a random permutation of the dataset is performed and for each example the algorithm computes the average label value for the example with the same category value placed before the given one in the permutation [22]. Let

$\sigma = (\sigma_1, \sigma_2, ..., \sigma_n)$ be the permutation of the dataset, then $x_{\sigma_p,k}$, the predictor corresponding to observation k in partition σ_p is substituted with:

$$\frac{\sum_{i=1}^{p-1}[x_{\sigma_i,k} = x_{\sigma_p,k}] \cdot y_{\sigma_j} + a \cdot P}{\sum_{i=1}^{p-1}[x_{\sigma_i,k} = x_{\sigma_p,k}] + a} \tag{3.22}$$

finally, the algorithm uses the transformed variable to build the decision tree [22].

Gradient Boosting Framework

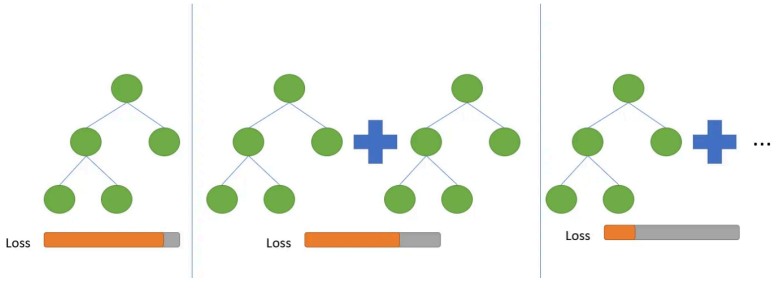


Figure 3.4.1: Gradient boosting. Adapted from: [59]

In this section I will briefly highlight Catboost's unique features. The CatBoost algorithm is based upon oblivious decision trees, which are trees that make decisions based on only one feature at a time [22], this design ensures that the threes are balanced and less prone to overfitting. At the same time, using oblivious trees allows for a more efficient implementation and faster execution speed. In Ordered boosting mode [22], Catboost usues supporting models $M_{r,j}$ to make predicitons based on peromutations of training samples. Note that the random permutation σ_r chosen for each tree-building iteration are the same used to calculate the Target Statistics for categorical features embeddings. Finally, the implementation of Catboost I will be using in the following chapters uses *l2* regularization on the leaf values of the trees to prevent overfitting [22, 47]. For a general overview of gradient boosting, the literature is vast and I suggest to go over the Survey from He et al. [28] or the overview of XGBoost, a popular tree bosting system, by Chen et al. [17].

4

Parametric Modeling Results

4.1 Introduction

In this chapter I will present the results of parametric modeling in forecasting next-year decarbonization rates. The chapter presents a sequence of models in increasing order of complexity that allow us to:

1. Test key hypothesis on the relationship between the predictors and next year decarbonization rates. In particular, I followed the CDP structure in testing multiple sets of predictors groped based on focus areas, such as carbon credits, initiaties and incentives, risks and opportunites, to understand what role each area plays in next year decarbonization rate and which predictors are useful from each set.

2. Build a state-of-the-art model that strikes an optimal balance between number

of features and prediction accuracy. To achieve this task, models were scored based on the Akaike information criterion (AIC), which is a method for selecting models that balance goodness of fit and model complexity, aiming to identify how new data might behave [33]

For each presented model, I will first present the key questions and hypothesis, then a brief model descriptions, the results, and a discussion. At each step, I will select the most relevant predictors and carry them to the next model.

4.1.1 Response Variable

Unless otherwise specified, most models are forecasting *Next Year Decarbonization Rate* (distribution and further information are provided in the previous chapter corresponding to Figure 2.3.1), which is the year on year change in Real emissions for a given company espressed as a percentage.

4.2 Geography, Industry, and Financial Predictors

4.2.1 Model I: Year and Current Decarbonization Rate

Table 4.2.1: Model Comparison: Fixed Effects Only vs. Random Intercept for Firm Id

	Dependent variable:	
	Next Year Decarbonization Rate	
	OLS	*linear mixed-effects*
	(1)	(2)
Year	-0.228^{***}	-0.257^{***}
	(0.021)	(0.021)
Ghg.Change.Real	0.295^{***}	0.210^{***}
	(0.009)	(0.009)
Constant	-1.985^{***}	-2.141^{***}
	(0.129)	(0.138)
Random Effects:		
Number of Firms		1870
sd(Firms)		2.142
Akaike Inf. Crit.	94134.242	94018.786
Bayesian Inf. Crit.	94164.354	94056.427

Note: *p<0.1; **p<0.05; ***p<0.01

Model 1: Simple Ordinary Least Squares regression with intercept using Year and Same-Year Decarbonization Rate as predictors. Model 2: Linear Mixed Effect Model with the same predictors as model (1) but with an added random intercept for the unique firm identifier.

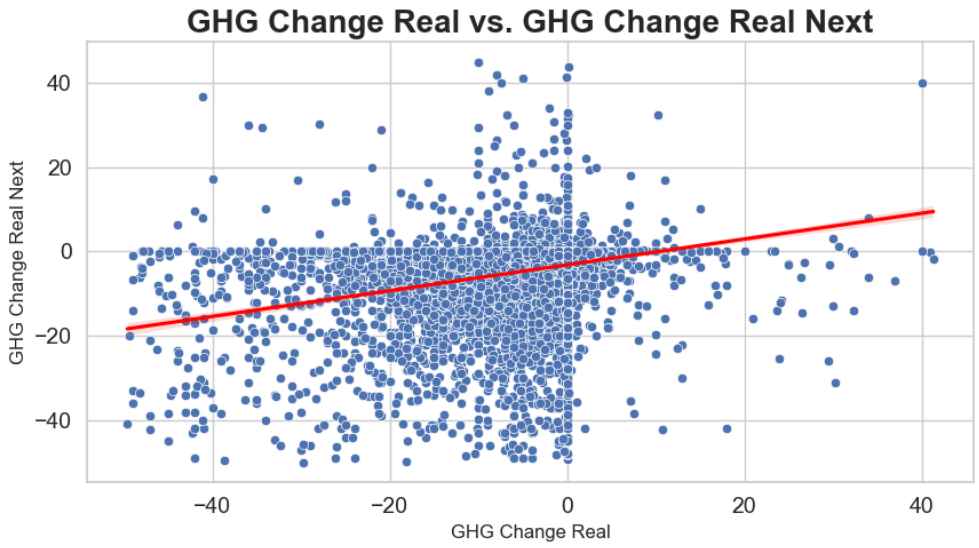


Figure 4.2.1: Current vs. Next Year Decarbonization Rate

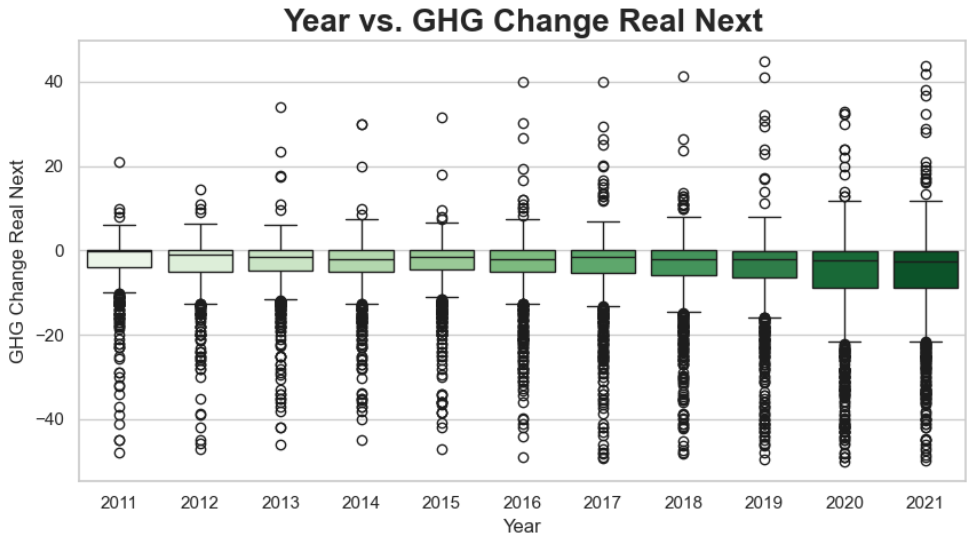


Figure 4.2.2: Next Year Decarbonization Rate Over Time

Discussion

In model (1) from Table 4.2.1, we start by predicting *Next-Year Decarbonization Rate* using year and Same-Year Decarbonization Rate. We observe how on average we predict a 22% decrease in real decarbonization rate, as well as a positive and significant correlation between previous year and next year decarbonization rate. In particular, controlling for year, an increase in decarbonization rate from the previous year corresponds to 0.3 predicted next year increase in decarbonization rate. The marginal relationship between Same-Year Decarbonization Rate and Next-Year decarboninzation rate can be observed in Figure 4.2.1, where we observe a positive correlation. Analogously, the relationship between year and the response variable is presented in Figure 4.2.2 where we observe that over time the variance of decarboninzation rates increases, while the mean slightly decreases.

In model (2) from Table 4.2.1, we add our first random effect: a random intercept for the firm unique identifier code. Mixed effect models are particularly suited when we have repeated measures on the same individuals, in our case, the same firms. In particular, including an effect for each individual firm allows to take into account the correlation between each firm's timeseries without introducing an excessively high number of parameters. When adding Id as a random effect, the model's AIC decreases from 94134.242 to 94018.79, signaling a significant improvement in the model's fit. Furthermore, the signs of the coefficient yeear and ghg change real remain the same, and the coefficient for year remains significant.

Key Finding: Adding a random effect significantly improves the model's fit, and we will therefore iterate to find the optimal combination of random effects to then identify a comprehensive set of fixed effects that best predict decarbonization rates.

4.2.2 Model II: Analyzing GICS Industry Sector

Table 4.2.2: Impact of GICS Industry on Next-Year Real Decarbonization Rate

	Dependent variable:	
	Next Year Decarbonization Rate	
	linear mixed-effects	
	(3)	(4)
Year	−0.260*** (0.021)	−0.259*** (0.021)
Ghg.Change.Real	0.206*** (0.009)	0.209*** (0.009)
IndustryAutomobiles, Components	3.851*** (0.631)	
IndustryBanks	1.870*** (0.568)	
IndustryCapital Goods	3.347*** (0.519)	
IndustryCommercial, Professional Services	4.121*** (0.645)	
IndustryConsumer Durables, Apparel	2.356*** (0.631)	
IndustryConsumer Services	3.528*** (0.715)	
IndustryConsumer Staples Distribution, Retail	3.216*** (0.698)	
IndustryEnergy	4.069*** (0.592)	
IndustryEquity Real Estate Investment Trusts	1.207* (0.656)	
IndustryFinancial Services	2.591*** (0.607)	
IndustryFood, Beverage, Tobacco	3.452*** (0.574)	
IndustryHealth Care Equipment, Services	3.062*** (0.653)	
IndustryHousehold, Personal Products	1.674** (0.797)	
IndustryInsurance	2.519*** (0.617)	
IndustryMaterials	4.475*** (0.522)	
IndustryMedia, Entertainment	2.226*** (0.779)	
IndustryPharmaceuticals, Biotechnology, Life Sciences	2.721*** (0.626)	
IndustryReal Estate Management, Development	2.227** (0.874)	
IndustryRetailing	1.527** (0.683)	
IndustrySemiconductors, Semiconductor Equipment	3.585*** (0.702)	
IndustryTechnology Hardware, Equipment	2.381*** (0.614)	
IndustryTelecommunication Services	1.770*** (0.657)	
IndustryTransportation	3.754*** (0.605)	
IndustryUtilities	3.487*** (0.562)	
Constant	−5.196*** (0.484)	−2.395*** (0.241)
Random Effects:		
Number of Firms	1870	1870
Number of Industries		25
sd(Firms)	1.953	1.93
sd(Industry)		0.972
Akaike Inf. Crit.	93891.439	93919.569
Bayesian Inf. Crit.	94109.755	93964.738

Note: *p<0.1; **p<0.05; ***p<0.01

Model (3): Linear Mixed Effect Model incorporating firm Id as a random intercept and including Industry as a fixed effect. Model (4): Linear Mixed Effect Model incorporating Industry as a random effect nested within firm ID, along with the same fixed effects as model (3).

Figures

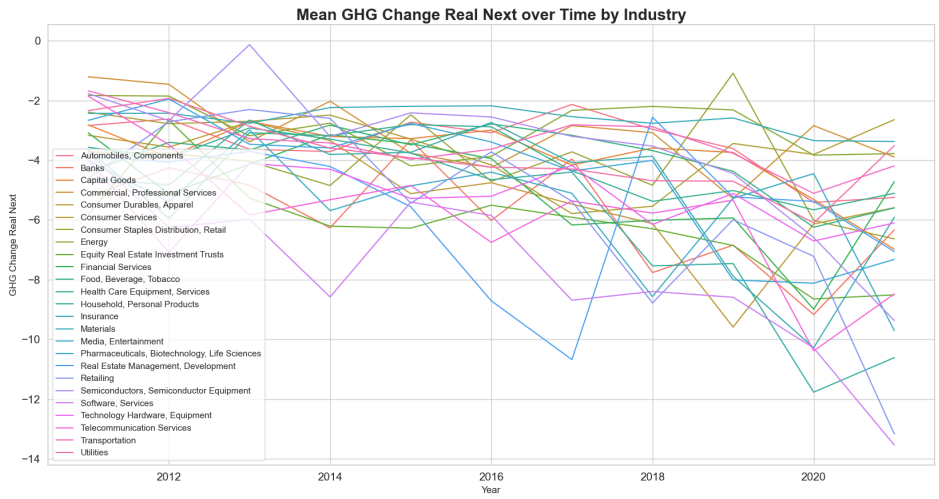


Figure 4.2.3: Next Year Decarbonization Rate by Industry

Discussion

In model (3), the inclusion of industry sector as a categorical predictor provides a clear indication that the decarbonization rate is significantly influenced by the sector in which a firm operates. The coefficients from the model suggest that, compared to the Software, Services industry (reference category), firms in sectors like Energy, Materials, and Transportation have significantly higher decarbonization rates, aligning with the expectation due to their carbon-intensive nature.

Model (4) incorporates industry sector as a random effect nested within firm IDs, accounting for the non-independence of observations within the same sector and firm. This model variation reflects the hierarchical structure of data. Despite an increase in AIC, indicating a more complex model, the significant predictors remain consistent, suggesting that industry can be effectively used as a random effect.

The mean decarboninzation trends by industry are presented in Figure 4.2.3. We can observe that over time mean decarboninzation rates are improving (decreasing)

and that some industries tend to decarbonize more than others as expected.

Key Finding: The significant effect of the industry sector on decarbonization rates underscores the importance of considering sector-specific factors when evaluating environmental performance. The persistence of significant coefficients for industry sectors across both models indicates that certain industries face inherent challenges in reducing carbon emissions. This could be due to technological barriers, regulatory differences, economic constraints, or the fundamental carbon intensity of their operations. Overall, the addition of industry as a random effect nested within firm IDs in Model (4) helps to better capture the intra-sector variability and the unique characteristics of firms. While this increases model complexity, as evidenced by the higher AIC, it provides a more accurate representation of the real-world scenario where firms within the same industry may follow different decarbonization trajectories due to various factors such as size, location, and management practices.

4.2.3 Model III: Analyzing Geographical Location

Table 4.2.3: Impact of Country and Continent on Next-Year Real Decarbonization Rate

	Dependent variable:	
	Next Year Decarbonization Rate	
	(5)	(6)
Year	-0.267^{***} (0.021)	-0.267^{***} (0.021)
Ghg.Change.Real	0.207^{***} (0.009)	0.207^{***} (0.009)
ContinentAfrica	1.895^{***} (0.409)	
ContinentAsia	1.714^{***} (0.202)	
ContinentNorth America	1.377^{***} (0.185)	
ContinentOceania	1.585^{***} (0.519)	
ContinentSouth America	1.058^{**} (0.524)	
Constant	-3.282^{***} (0.258)	-2.027^{***} (0.413)
Random Effects:		
Number of Firms	1871	1871
Number of Industries	25	25
Number of Continents	6	6
Number of Countries		48
sd(Firms:Industry)	1.789	1.766
sd(Industry)	0.975	0.97
sd(Continent)		0.74
sd(Country:Continent)		0.349
Akaike Inf. Crit.	93831.892	93834.542
Bayesian Inf. Crit.	93914.702	93894.767

Note: *p<0.1; **p<0.05; ***p<0.01

Model (6): Linear Mixed Effect Model incorporating Year, GHG Change Real, and Continent as fixed effects, and firm ID and industry as random intercepts. Model (6): Linear Mixed Effect Model adding Country as a random effect nested within Continent, in addition to the fixed and random effects included in model (5).

Figures

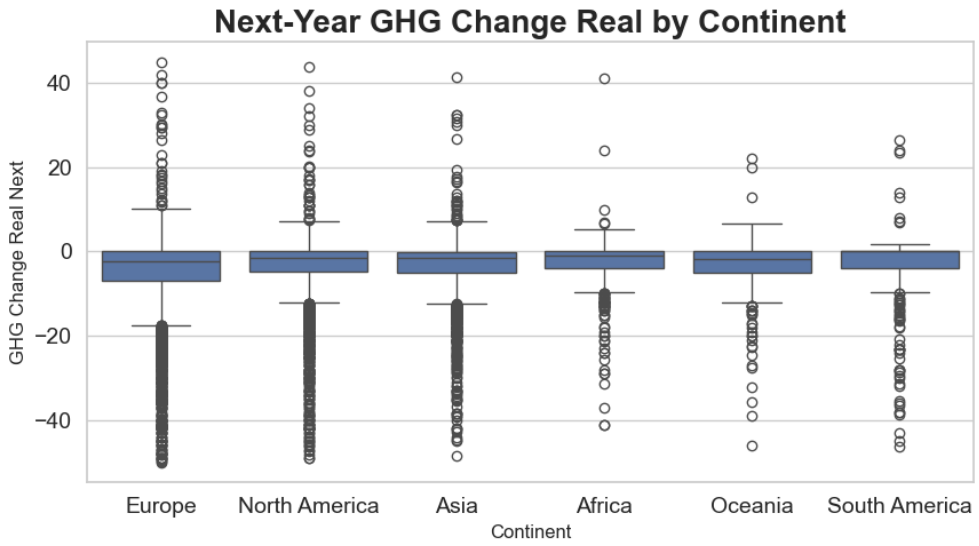


Figure 4.2.4: Continent vs. Next Year Decarbonization Rate

Discussion

Model (5) introduces the effect of geographical location at the continental level as a fixed effect while controlling for year and GHG change, and using firm ID and Industry as random intercepts. The significant coefficients for continents compared to Europe (the reference category) suggests that geographical location is correlated with a significant change firm's decarbonization rate, in particular we observe how Oceania is the continent associated with the least decarbonization. That is, compared to Europe, according to the model a firm in Oceana is predicted to decarbonize by an average of 1.585% less compared to Europe. This finding aligns with expectations, reflecting variations in regulatory environments, access to clean technologies, and economic conditions across continents. A comparison between continents is also provided in Figure 4.2.4, where we observe how Europe's distribution is centered around a lower mean compared to other continents.

Model (6) incorporates country as a random effect nested within continents. This

allows to take into account the effect of geographical location in future models without increasing the degrees of freedom. Although the AIC slightly increases, indicating a more complex model, the consistency of significant coefficients for Year and GHG Change Real across both models suggests that these are robust predictors of decarbonization rates.

Key Finding: The analysis demonstrates that the continent and country where a firm is located significantly impact its decarbonization rate, highlighting the importance of geographical factors in environmental strategies. Moving forward, the inclusion of country and continent as significant variables will be included in all subsequent models. Form now on, the random effects will always be the same: that is random intercepts for firm Id nested wihtin Industry, and random intercepts for Country nested wihtin Continent.

4.2.4 Model IV: Analyzing Financial Predictors

Table 4.2.4: Impact of Financial Predictors on Next Year Real Decarbonization Rate

	Dependent variable:		
	Next Year Decarbonization Rate		
	(7)	(8)	(9)
Year	−0.258*** (0.021)	−0.256*** (0.021)	−0.256*** (0.021)
Ghg.Change.Real	0.204*** (0.009)	0.205*** (0.009)	0.205*** (0.009)
Market.Cap	−0.405*** (0.115)	−0.520*** (0.088)	−0.529*** (0.089)
Employees	0.047 (0.093)		
Revenue	0.234 (0.156)	0.154* (0.091)	0.162* (0.092)
Employees.1Y.Gr	−0.024 (1.222)	0.489 (1.070)	
Assets.1Y.Gr	1.020 (1.099)		0.791 (0.954)
Tot.Assets	−0.225 (0.138)		
Net.Income.Over.Assets	−2.597 (4.659)		
Roe	−0.662 (1.239)		
Constant	8.158** (3.222)	5.854*** (1.605)	5.618*** (1.573)
Random Effects:			
Number of Firms	1871	1871	1871
Number of Industries	25	25	25
Number of Continents	6	6	6
Number of Countries	48	48	48
sd(Firms:Industry)	1.733	1.729	1.729
sd(Industry)	0.898	0.899	0.898
sd(Continent)	0.74	0.746	0.745
sd(Country:Continent)	0.275	0.274	0.277
Akaike Inf. Crit.	93831.892	93834.542	93793.357
Bayesian Inf. Crit.	93914.702	93894.767	93913.807

Note: *p<0.1; **p<0.05; ***p<0.01

Model (7): Linear Mixed Effect Model evaluating the impact of Year, GHG Change Real, Market Cap, and Revenue on Next-Year Decarbonization Rate, with Firm ID and Industry as random intercepts. Model (8): Extension of Model (7) adding Assets Growth as an additional predictor. Model (9): Further extension of Model (8) adding Employees Growth as an additional predictor.

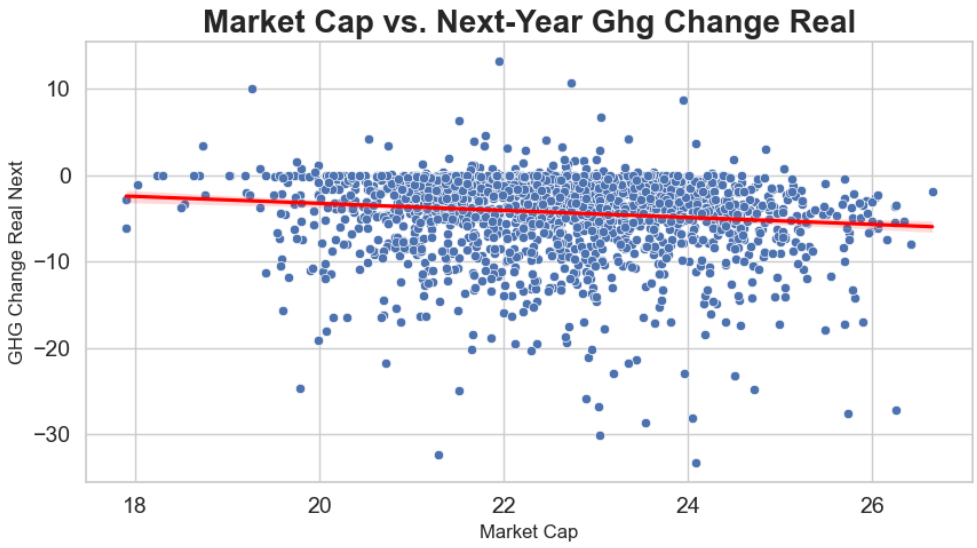


Figure 4.2.5: Market Capitalization vs. Next Year Decarbonization Rate

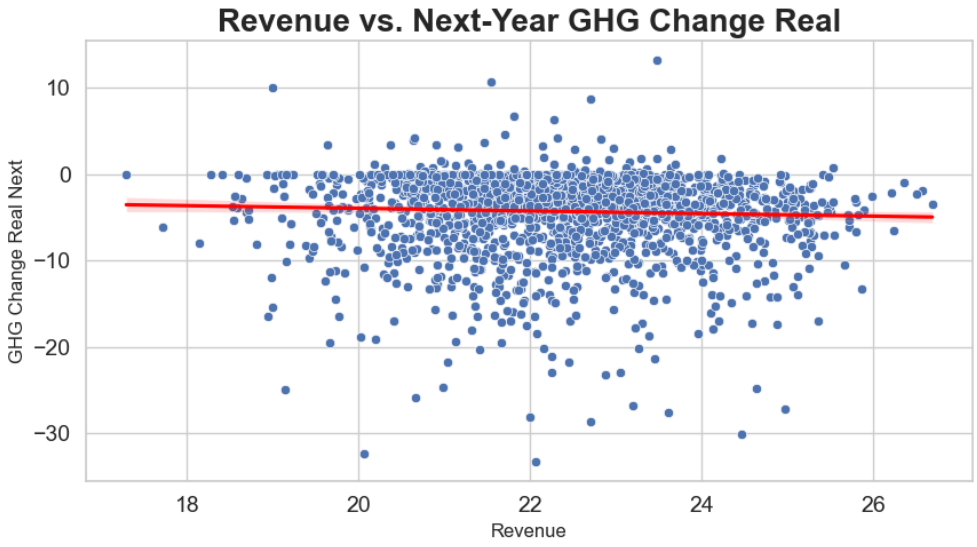


Figure 4.2.6: Revenue vs. Next Year Decarbonization Rate

Discussion

In Model (7), significant predictors include Year, GHG Change Real, and Market Cap, with Revenue showing significance at the 10% level. The negative coefficient for Market Cap suggests that larger firms have a lower, which in this context means better, decarbonization rate. The marginal relationship is presented in Figure 4.2.5 where we can clearly observe a negative correlation between market cap and decarbonization rate. A similar, though less pronounced marginal relationship also applies to revenuue, as can be observed in Figure 4.2.6.

Model (8) focuses on Assets Growth as a variable, controlling for Market Cap and Revenue and removing other non-significant predictors. The positive but non-significant coefficient for Assets Growth suggests an expected correlation relationship where firms expanding their assets might experience a worse rate of decarbonization, potentially due to increased operations or capital investment not directly tied to reducing emissions.

Model (9) includes Employees Growth, which also shows a non-significant result. Similar to Assets Growth, the positive direction of the Employees Growth coefficient hints at a conceivable relationship where firms increasing their workforce may not proportionally enhance their decarbonization efforts. Yet, this remains speculative without statistical significance.

Key Finding: Financial indicators, particularly Market Cap and, to a lesser extent, Revenue, serve as relevant predictors for a firm's decarbonization rate. The consistent significance and direction of these coefficients across models suggest a potential link between a firm's financial scale and its environmental performance. Financial metrics will continue to be used as control variables in future models to refine our understanding of their impact on decarbonization rates. The positive but non significant coefficients of growth indicators in my opinion hint at the fact that it is challenging to grow and decarbonize at the same time.

4.3 Emission Scopes, Incentives, Targets, Risks and Opportunities

4.3.1 Model V: Analyzing GHG Emissions and Verification

Table 4.3.1: Impact of GHG and Verification on Decarbonization

	Dependent variable:			
	Next Year Decarbonization Rate			
	(10)	(11)	(12)	(13)
Year	−0.118*** (0.028)	−0.211*** (0.022)	−0.111*** (0.028)	−0.111*** (0.027)
Ghg.Change.Real	0.198*** (0.009)	0.200*** (0.009)	0.196*** (0.009)	0.196*** (0.009)
Market.Cap	−0.438*** (0.087)	−0.447*** (0.087)	−0.403*** (0.086)	−0.409*** (0.086)
Revenue	0.183* (0.099)	0.246*** (0.090)	0.204** (0.098)	0.176* (0.095)
Ghg1	0.107*** (0.037)		0.126*** (0.037)	0.115*** (0.035)
Ghg2Location	−0.087** (0.042)		−0.088** (0.042)	−0.066*** (0.025)
Ghg2Market	−0.016 (0.031)		−0.022 (0.031)	
Ghg3.Total	−0.032 (0.020)		−0.015 (0.020)	
Ghg3.Count	−0.086*** (0.030)		−0.058* (0.030)	−0.073*** (0.024)
Ghg1.Na	1.732*** (0.583)		1.845*** (0.586)	1.682*** (0.507)
Ghg2Location.Na	−0.324 (0.580)		−0.396 (0.578)	
Ghg2Market.Na	0.886** (0.358)		0.767** (0.358)	0.991*** (0.178)
Ghg3.Total.Na	0.039 (0.437)		−0.010 (0.438)	
Methane.Emissions	0.078*** (0.022)		0.079*** (0.022)	0.077*** (0.022)
Type.Scope1Limited/Moderate		0.320 (0.225)	0.367 (0.224)	0.360 (0.221)
Type.Scope1N.A		0.334 (0.450)	0.629 (0.457)	0.958*** (0.263)
Type.Scope1Third.Party.Underway		0.858*** (0.299)	0.807*** (0.300)	0.850*** (0.295)
Ghg.Verification.Scope1.Yes		−0.444 (0.517)	−0.344 (0.528)	
Ghg.Verification.Scope2.Yes		−0.317 (0.389)	−0.047 (0.391)	
Ghg.Verification.Scope3.Yes		−0.610*** (0.185)	−0.410** (0.190)	−0.458** (0.181)
Constant	2.645* (1.543)	2.732* (1.499)	1.035 (1.602)	1.000 (1.523)
Random Effects:				
Number of Firms	1871	1871	1871	1871
Number of Industries	25	25	25	25
Number of Continents	6	6	6	6
Number of Countries	48	48	48	48
sd(Firms:Industry)	1.617	1.643	1.588	1.58
sd(Industry)	0.704	0.857	0.649	0.648
sd(Continent)	0.671	0.689	0.644	0.637
sd(Country:Continent)	0.291	0.242	0.208	0.197
Akaike Inf. Crit.	93723.94	93738.237	93703.092	93683.497
Bayesian Inf. Crit.	93874.502	93858.688	93898.824	93834.06

Note: *p<0.1; **p<0.05; ***p<0.01

Model (10): Includes all predictors in the set and analyzes all hypothesis: how different greenhouse gas (GHG) emissions, whether they're checked (verified), and other basic company details like size and earnings, affect how much a company can reduce its emissions in a year. Model (11): Just focuses on the types of GHG emissions to see which ones are most important for reducing emissions, still considering basic company details. Model (12): Only looks at whether companies check (verify) their emissions data and how that influences their ability to cut down emissions, keeping the company details the same. Model (13): Picks out and uses only the most important factors from the first three models, those predictors will be passed on in the next models.

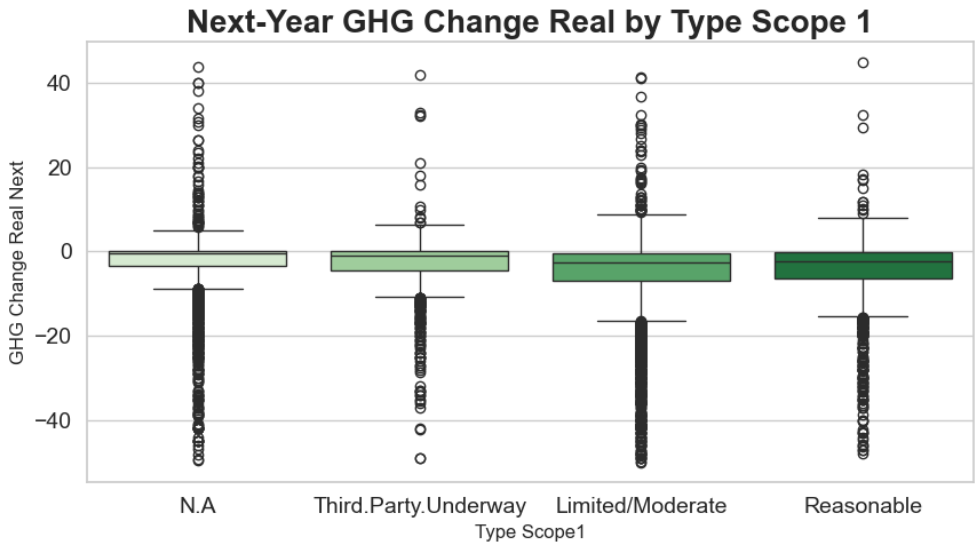


Figure 4.3.1: GHG Emission Scope 1 Verification Type vs. Next Year Decarbonization Rate

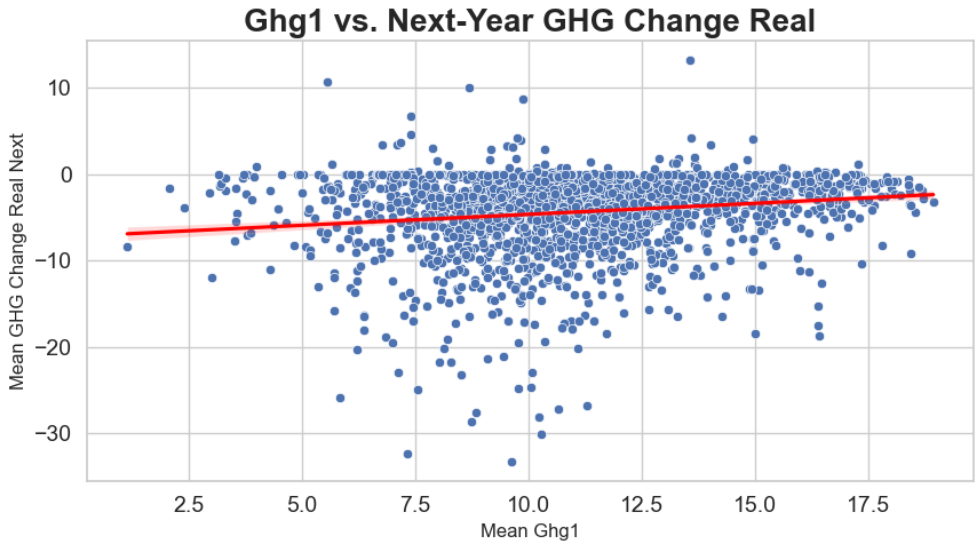


Figure 4.3.2: GHG Emission Scope 1 vs. Next Year Decarbonization Rate

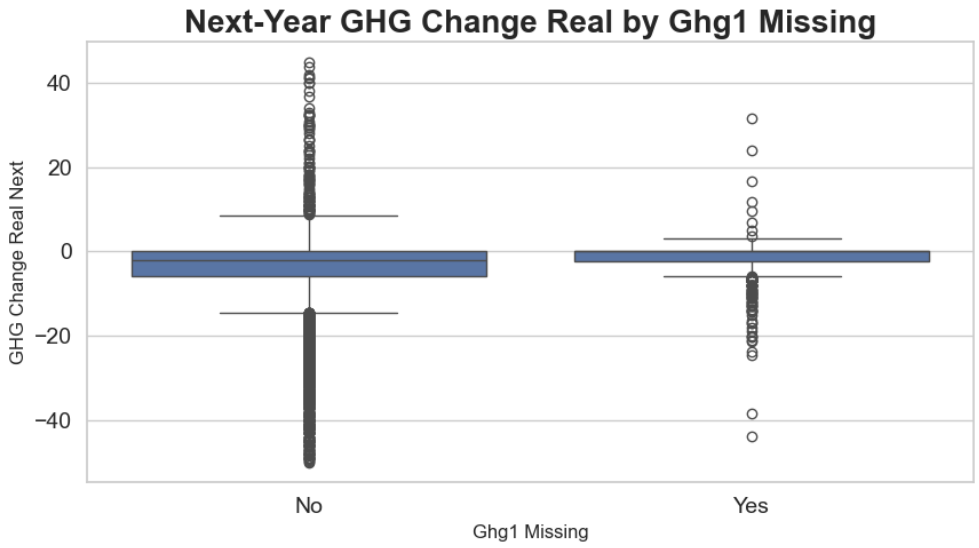


Figure 4.3.3: GHG Emission Scope 1 Missing vs. Next Year Decarbonization Rate

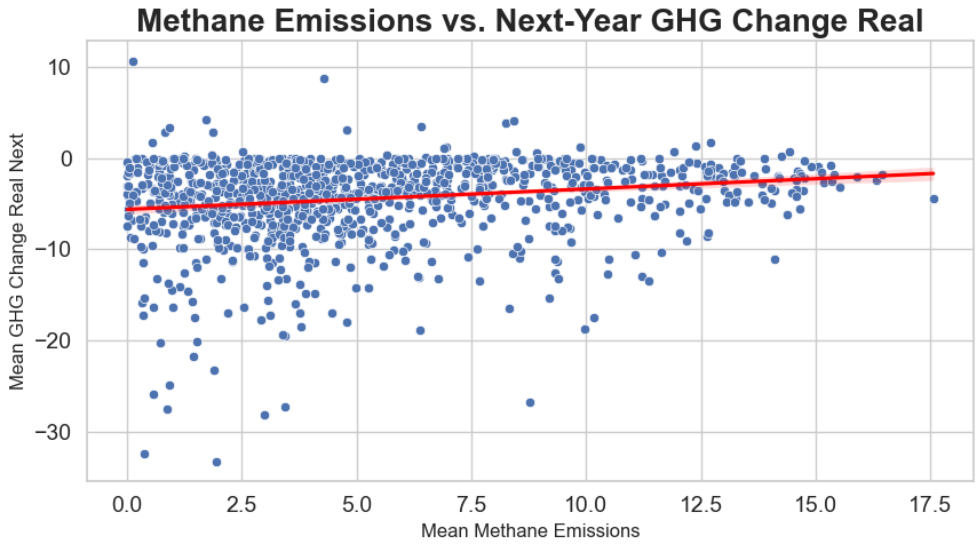


Figure 4.3.4: Methane Emissions vs. Next Year Decarbonization Rate

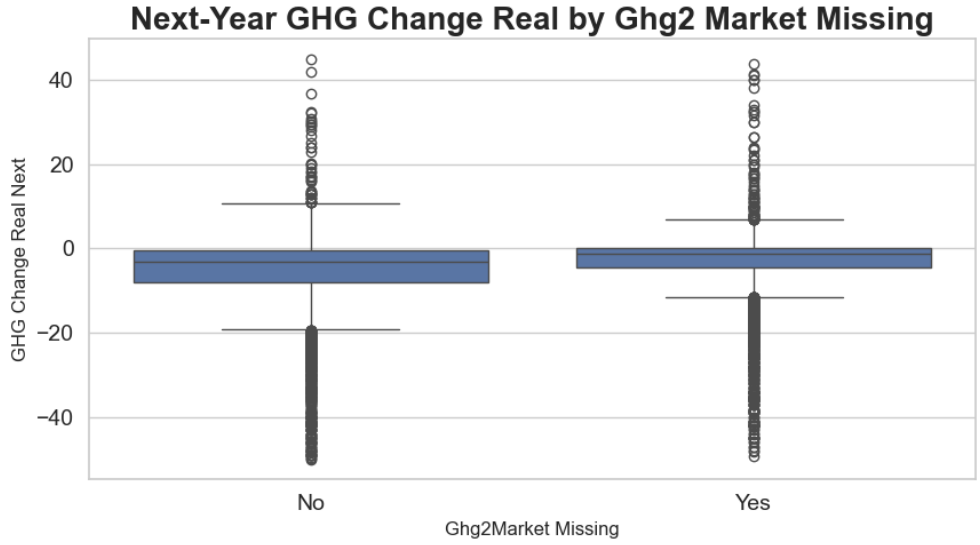


Figure 4.3.5: GHG Emission Scope 2 Market Missing vs. Next Year Decarbonization Rate

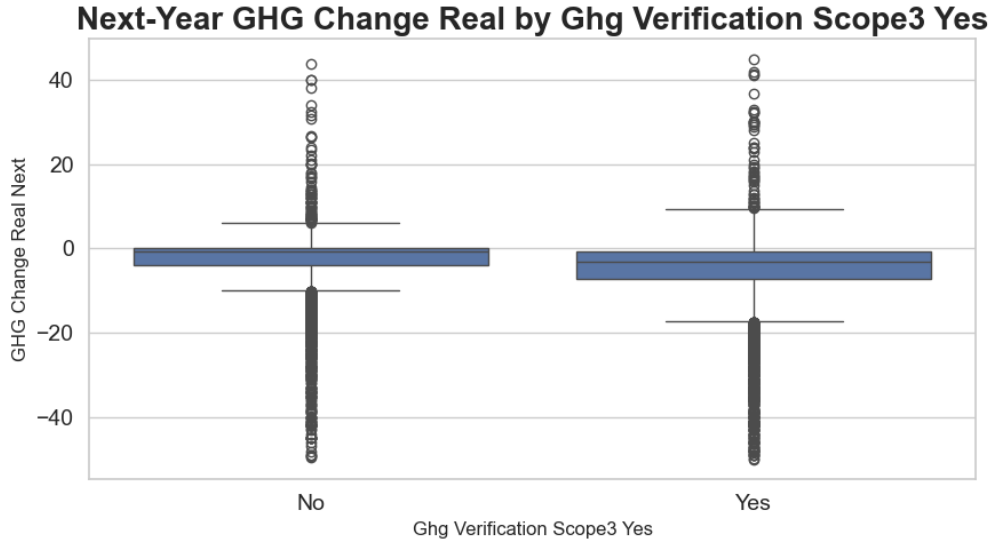


Figure 4.3.6: GHG Emission Scope 3 Verification vs. Next Year Decarbonization Rate

Discussion

The analysis shows that Scope 1 emissions (Ghg1) significantly influence a firm's decarbonization efforts. A positive coefficient for Ghg1 (0.107, significant at the 1% level) indicates that higher Scope 1 emissions are correlated with a poorer decarbonization rate for the next year. Furthermore, when data for Scope 1 emissions is missing (Ghg1.Na), there's a notable jump in the decarbonization rate, with a coefficient increase to 1.732. Specifically, firms that do not report their Scope 1 emissions are predicted to have a 0.577% worse decarbonization rate compared to those that do report. This gap emphasizes the importance of accurate reporting; not reporting is linked to poorer environmental performance. The marginal relationship between Scope 1 emssions and Next-Year Decarbonization Rate is presented in Figure 4.3.2 where we observe that higer emissions are correlated with worse decarbonization rates. The relationship between reporting Scope 1 emissions is presented in the boxplot Figure 4.3.3.

Regarding market emissions, the positive coefficient for not reporting market-based emissions (Ghg2Market.Na) is correlated with a predicted worse next-year decarbonizaiton rate by an average of 0.850%. This makes sense since market-based reporting is more complex and firms that invest the effort likely prioritize decarbonization more strongly. The marginal relationship between reporting market-based emissions and the response is presented in Figure 4.3.5 where the firms that don't report have wrose decarboninzation rates on average.

Methane emissions also play a significant role in the analysis. The presence of Methane Emissions with a positive coefficient (0.078, significant at the 1% level) suggests that firms with higher reported methane emissions are expected to have a worse next-year decarbonization rate. This underlines the importance of focusing on specific pollutants like methane, which have a high impact on global warming and can significantly affect a firm's overall decarbonization performance. The marginal relationship between methane emissions and next-year decarboninzation rate is presented in Figure 4.3.4 where we observe a clear trend between firms with higher methane emissions having worse next-year decarboninzation rates.

The role of emissions verification is also central in our findings, particularly the positive impact of verifying Scope 3 emissions. Firms that verify their Scope 3 emissions (Ghg.Verification.Scope3.Yes) are associated with a better decarbonization rate, with a negative coefficient of -0.610 (significant at the 1% level in the second model and -0.458, significant at the 5% level in the final model). The marginal relationship is presented in Figure 4.3.6. This suggests that firms taking steps to validate their broadest category of emissions are likely more committed to comprehensive decarbonization efforts, as Scope 3 encompasses indirect emissions not produced by the firm directly but related to their value chain. Hence, verifying these emissions can be seen as an indication of a firm's comprhensive approach to understanding and mitigating its environmental impact.

The type of Scope 1 verification process is a significant predictor. The model indicates that having a third-party verification of Scope 1 emissions underway (Type.Scope1Third.Party.Und is associated with a worse decarbonization rate compared to other methods, evidenced by a positive coefficient (0.858, significant at the 1% level). The refernce category in this case is reasonable Sccope1 verification. Marginal relationships are presented in Figure 4.3.1. This could suggest that firms only beginning to engage with third-party verification may have previously neglected deeper decarbonization efforts, or it may reflect relatively lower decarbonization rate due to more accurate reporting or transitional operational changes.

Additionally, the negative impact of not having any Scope 1 verification (Type.Scope1N.A) with a significant worse predicted decarbonization rate (0.958, significant at the 1% level in the final model) suggests the importance of not just reporting emissions but also verifying them.

Overall, our findings suggest that while the verification of emissions, especially for Scope 3, is a positive step towards better decarbonization (see Figure 4.3.6), the initial stages of Scope 1 verification might reflect a period of adjustment where firms are just starting to confront and accurately report their emissions. This phase may not immediately reflect in improved decarbonization rates but is essential for transparent and effective environmental management in the long run.

Key Findings: When looking at a firm decarboninzation strategies, the most

important factors to consider are methane emissions, GHG1 and relative verification, whether the frims reports market Scope 2 emissions, whether the firm has scope 3 verification, and if the type of scope 1 emissions is either limited, moderate, or underway. This combination of features offers a comprehensive set when it comes to forecasting next-year decarbonization rate and shows that the effect of those firm actions are correlated not only with a same-year effect, but also with a next-year effect.

4.3.2 Model VI: Analyzing Incentives

Table 4.3.2: Impact of Incentives on Next-Year Real Decarbonization Rate

	Dependent variable:			
	Next Year Decarbonization Rate			
	(14)	(15)	(16)	(17)
Year	−0.102*** (0.028)	−0.113*** (0.027)	−0.105*** (0.028)	−0.108*** (0.027)
Ghg.Change.Real	0.195*** (0.009)	0.196*** (0.009)	0.195*** (0.009)	0.195*** (0.009)
Market.Cap	−0.406*** (0.086)	−0.394*** (0.086)	−0.391*** (0.086)	−0.391*** (0.086)
Revenue	0.182* (0.095)	0.172* (0.095)	0.178* (0.095)	0.177* (0.095)
Type.Scope1Limited/Moderate	0.346 (0.221)	0.329 (0.222)	0.315 (0.222)	0.318 (0.222)
Type.Scope1N.A	0.864*** (0.264)	0.933*** (0.263)	0.842*** (0.264)	0.847*** (0.264)
Type.Scope1Third.Party.Underway	0.789*** (0.295)	0.824*** (0.295)	0.765*** (0.295)	0.772*** (0.295)
Ghg.Verification.Scope3.Yes	−0.419** (0.182)	−0.458** (0.181)	−0.420** (0.181)	−0.423** (0.181)
Ghg1	0.122*** (0.035)	0.115*** (0.035)	0.123*** (0.035)	0.122*** (0.035)
Ghg2Location	−0.063** (0.025)	−0.065*** (0.025)	−0.062** (0.025)	−0.062** (0.025)
Ghg3.Count	−0.065*** (0.024)	−0.071*** (0.024)	−0.063*** (0.024)	−0.064*** (0.024)
Ghg1.Na	1.701*** (0.507)	1.694*** (0.507)	1.711*** (0.507)	1.721*** (0.507)
Ghg2Market.Na	0.990*** (0.178)	0.993*** (0.178)	0.993*** (0.178)	0.991*** (0.178)
Methane.Emissions	0.077*** (0.022)	0.079*** (0.022)	0.079*** (0.022)	0.079*** (0.022)
Cdp.Boardoversight.I	−0.161 (0.215)		−0.166 (0.215)	
Cdp.Incentivebinary.I	−0.561*** (0.192)		−0.546*** (0.192)	−0.565*** (0.190)
Method.IndInternal Incentives		−1.776*** (0.576)	−1.751*** (0.576)	−1.746*** (0.576)
Method.IndMacc		−0.807** (0.367)	−0.793** (0.367)	−0.794** (0.367)
Constant	1.227 (1.527)	0.790 (1.523)	1.017 (1.528)	0.935 (1.522)
Random Effects:				
Number of Firms	1871	1871	1871	1871
Number of Industries	25	25	25	25
Number of Continents	6	6	6	6
Number of Countries	48	48	48	48
sd(Firms:Industry)	1.563	1.573	1.558	1.558
sd(Industry)	0.644	0.656	0.652	0.65
sd(Continent)	0.649	0.637	0.648	0.648
sd(Country:Continent)	0.215	0.189	0.208	0.197
Akaike Inf. Crit.	93680.41	93672.856	93670.175	93667.531
Bayesian Inf. Crit.	93846.029	93838.475	93850.851	93840.678

Note: *p<0.1; **p<0.05; ***p<0.01

Model (14): This one checks if having any incentives (binary yes/no) and board oversight affects how a company reduces its emissions, along with other usual details like year, GHG changes, and company size. Model (15): This one looks only at the specific types of incentives used by the company, comparing them to companies that use "other" types of incentives not categorized, to see which incentives might be better at helping reduce emissions. Model (16): This combines everything – it looks at whether having incentives, the types of incentives, and board oversight, along with all the other usual factors, influences how much a company can reduce its emissions. Model (17): This focuses only on the most important factors from the third model. It keeps the binary yes/no on having incentives and the specific types of incentives but drops the board oversight, to see what really matters most for reducing emissions.

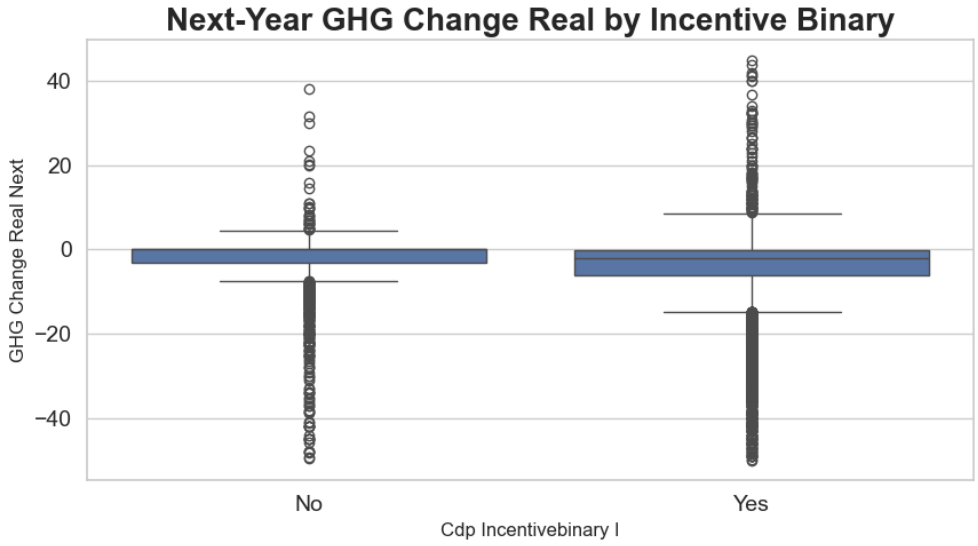


Figure 4.3.7: Incentive Binary vs. Next Year Decarbonization Rate

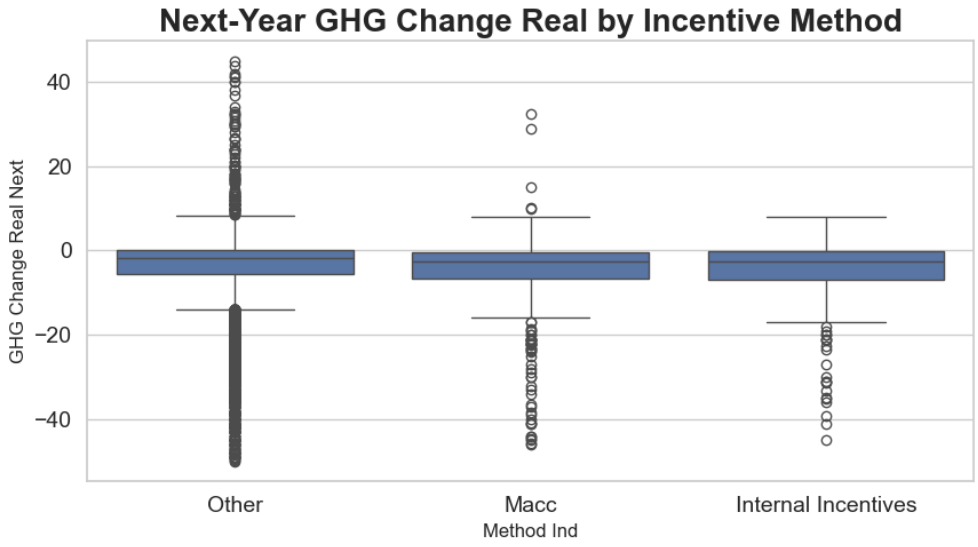


Figure 4.3.8: Incentive Method vs. Next Year Decarbonization Rate

Discussion on Incentives

The analysis from the first model indicates that simply having incentives in place (Cdp.Incentivebinary.I) is significantly associated with a decarbonization rate, with a negative coefficient (-0.561, significant at the 1% level). The marginal relationship can be observed in Figure 4.3.7. This suggests that firms with any form of incentives aimed at reducing emissions are correlated with relatively better decarbonization efforts compared to those without as expected.

When analyzing the types of incentives in the second model, we observe how specific incentive methods have better outcomes than general incentives. In particular, the Method.IndInternal Incentives (internal incentives for decarbonization) show a very strong negative relationship with the decarbonization rate (-1.776, significant at the 1% level), indicating that internal corporate incentives are particularly effective compare to the "other" baseline. Similarly, Method.IndMacc (Marginal Abatement Cost Curve) also show a negative coefficient (-0.807, significant at the 5% level), suggesting that firms who use a marginal abatement cost curve to reward decarbonization efforts are correlated with a better decarboninzation performance. For more details on the Marginal Abatement Cost Curve, see Appendix A.2. Additionally, the marginal relationship between inective methods and next-year decarbonization rates is presented in Figure 4.3.8 wehre we observe how internal incetives and using the Marginal Abatement Cost Curve are both correlated with better average decarbonization rates compared to other forms of incentives.

The consistent negative coefficients for incentives across models demonstrate that well-structured incentive programs, especially those embedded within the company's internal operations, can be useufl for achieving meaningful decarbonization not only on a same year basis, but also when considering next-year decarbonizaiton rate. This could imply that incentives not only drive immediate actions but also foster a culture of sustainability and long-term commitment to reducing emissions.

Interestingly, board oversight did not emerge as a significant factor in the final model, suggesting that while governance is important, the direct impact on decarbonization rates may be more strongly influenced by tangible, operational incentive

mechanisms rather than solely by high-level oversight.

Key Findings: Our study highlights the significant role of targeted incentives in enhancing a firm's decarbonization efforts. Specifically, internal incentives and the use of Marginal Abatement Cost Curves are notably effective. This suggests that detailed, well-implemented incentive schemes are key to reducing emissions, more so than general executive endorsements or board oversight.

4.3.3 Model VII: Analyzing Targets, Risks, and Opportunities

Table 4.3.3: Impact of GHG, Verification, Incentives, Targets, Risks and Opportunities on Next-Year Real Decarbonization Rate

	Dependent variable:			
	Next Year Decarbonization Rate			
	(18)	(19)	(20)	(21)
Year	−0.087*** (0.028)	−0.057* (0.031)	−0.040 (0.031)	−0.047 (0.031)
Ghg.Change.Real	0.191*** (0.009)	0.192*** (0.009)	0.188*** (0.009)	0.188*** (0.009)
Market.Cap	−0.403*** (0.084)	−0.400*** (0.086)	−0.410*** (0.085)	−0.409*** (0.085)
Revenue	0.241** (0.094)	0.187** (0.095)	0.247*** (0.094)	0.249*** (0.094)
Type.Scope1Limited/Moderate	0.344 (0.220)	0.347 (0.222)	0.369* (0.220)	0.370* (0.220)
Type.Scope1N.A	0.746*** (0.263)	0.867*** (0.264)	0.769*** (0.263)	0.796*** (0.262)
Type.Scope1Third.Party.Underway	0.738** (0.294)	0.809*** (0.296)	0.774*** (0.295)	0.792*** (0.294)
Ghg.Verification.Scope3.Yes	−0.325* (0.181)	−0.401** (0.181)	−0.309* (0.181)	−0.317* (0.181)
Ghg1	0.123*** (0.034)	0.121*** (0.035)	0.122*** (0.035)	0.122*** (0.034)
Ghg2Location	−0.059** (0.025)	−0.054** (0.025)	−0.052** (0.025)	−0.052** (0.025)
Ghg3.Count	−0.033 (0.025)	−0.060** (0.024)	−0.031 (0.025)	−0.032 (0.024)
Ghg1.Na	1.673*** (0.503)	1.676*** (0.507)	1.637*** (0.503)	1.657*** (0.502)
Ghg2Market.Na	0.897*** (0.178)	1.005*** (0.178)	0.915*** (0.178)	0.911*** (0.178)
Methane.Emissions	0.078*** (0.022)	0.082*** (0.022)	0.081*** (0.022)	0.083*** (0.022)
Method.IndInternal Incentives	−1.681*** (0.572)	−1.740*** (0.576)	−1.678*** (0.572)	−1.700*** (0.572)
Method.IndMacc	−0.775** (0.364)	−0.736** (0.367)	−0.723** (0.364)	−0.725** (0.364)
Cdp.Incentivebinary.I	−0.356* (0.192)	−0.491** (0.193)	−0.300 (0.195)	
Cdp.Baseyearemission.Mean	0.016 (0.016)		0.016 (0.016)	
Cdp.Targetscope.Percent.Mean	0.001 (0.002)		0.001 (0.002)	
Cdp.Targetamount.Mean	−0.433*** (0.073)		−0.420*** (0.073)	−0.368*** (0.054)
Cdp.Targettype.Absolute	−0.194*** (0.066)		−0.181*** (0.066)	−0.165*** (0.058)
Cdp.Targettype.Intensity	0.088 (0.069)		0.086 (0.069)	
Cdp.Aggregated.Risk		0.615*** (0.178)	0.565*** (0.177)	0.553*** (0.176)
Cdp.Aggregated.Opp		−1.283*** (0.238)	−1.151*** (0.237)	−1.184*** (0.236)
Constant	0.193 (1.512)	1.110 (1.543)	0.362 (1.534)	0.232 (1.525)
Random Effects:				
Number of Firms	1871	1871	1871	1871
Number of Industries	25	25	25	25
Number of Continents	6	6	6	6
Number of Countries	48	48	48	48
sd(Firms:Industry)	1.47	1.58	1.495	1.49
sd(Industry)	0.588	0.653	0.593	0.605
sd(Continent)	0.614	0.651	0.619	0.617
sd(Country:Continent)	0.219	0.197	0.218	0.219
Akaike Inf. Crit.	93634.714	93641.746	93614.903	93589.986
Bayesian Inf. Crit.	93845.502	93829.949	93840.747	93785.717

Note: *p<0.1; **p<0.05; ***p<0.01

Model (18): We analyze the role that targets play in forecasting next-year decarbonization. In particular, we look at both absolute figures (such as target amount) and target types. Model (19): This model adds in how companies perceive their environmental risks and opportunities. It checks if recognizing more risks or seeing more green opportunities changes how well they can reduce emissions, alongside their financial metrics and targets. Model (20): third model combines all predictors from targets, risk, and opportunities together. Model (21): The final model narrows down excluding variables that are not significant at the 5% level

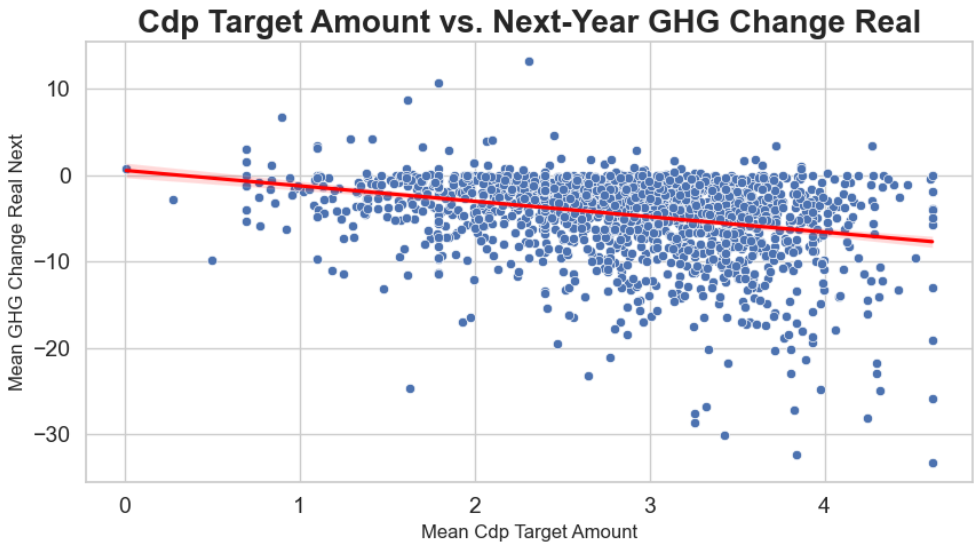


Figure 4.3.9: Cdp Target Amount vs. Next Year Decarbonization Rate

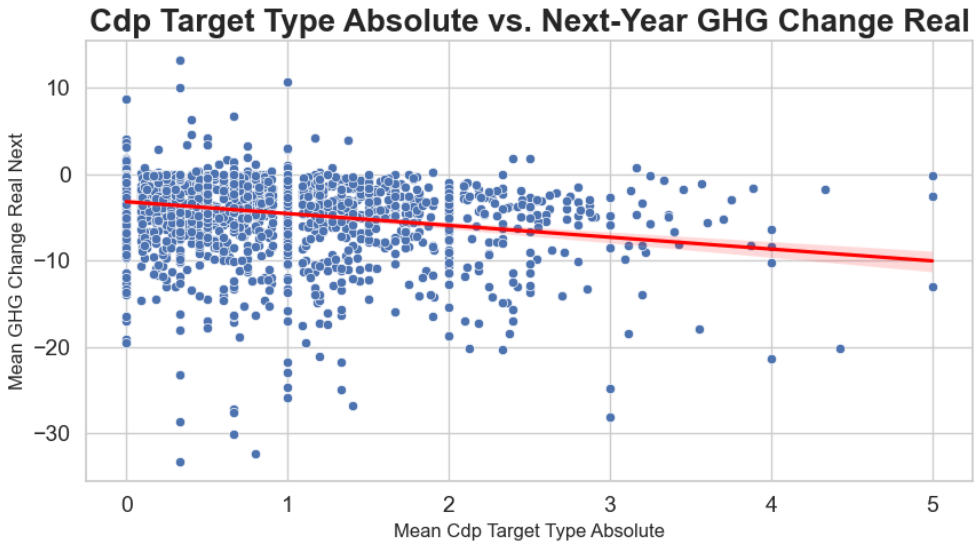


Figure 4.3.10: Cdp Target Type Absolute vs. Next Year Decarbonization Rate

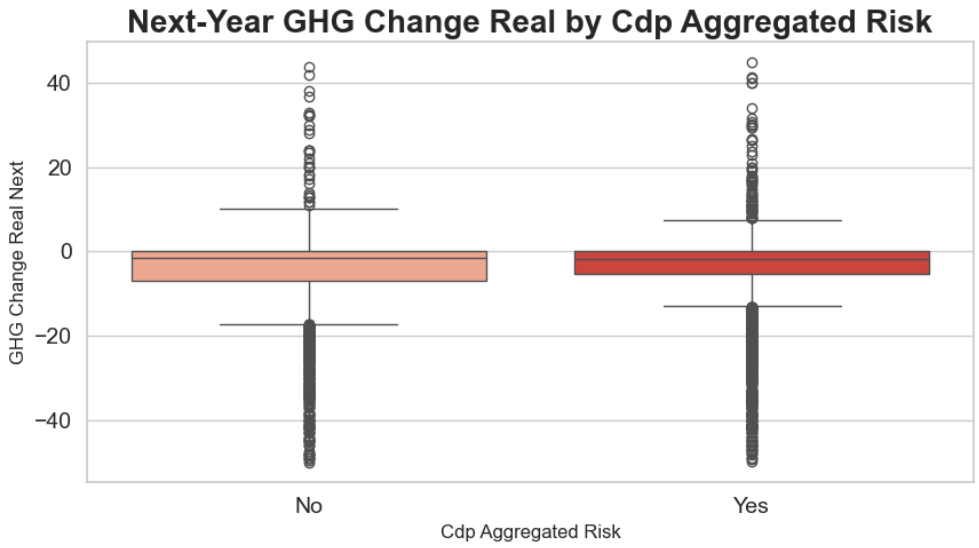


Figure 4.3.11: Cdp Aggregated Risk vs. Next Year Decarbonization Rate

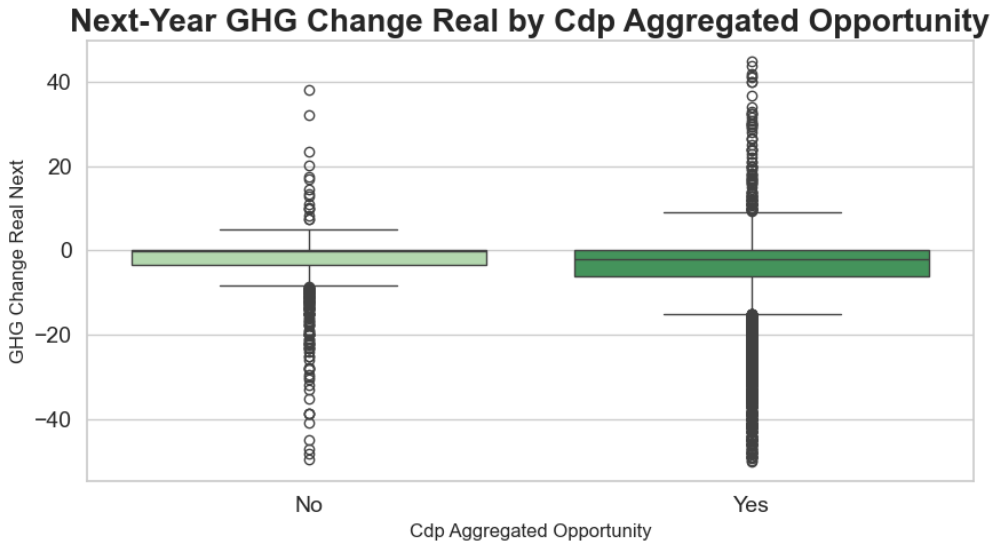


Figure 4.3.12: Cdp Aggregated Opportunity vs. Next Year Decarbonization Rate

Discussion

The results suggest that setting specific emission reduction targets (Cdp.Targetamount.Mean) has a significant negative impact on the next year's decarbonization rate, with coefficients like -0.433 (significant at the 1% level). This indicates that firms with more ambitious targets are correlated with a better decarboninzation rate. This can also be seen in the Figure 4.3.9, where we observe a negative relationship between the target amount and the decarbonization rate. This suggests that firms with higher target amounts are expected to have a better next-year decarbonization rate.

The type of target set by firms also matters. Companies with absolute reduction targets (Cdp.Targettype.Absolute) show a notable improvement in decarbonization efforts, as indicated by a negative coefficient of -0.194 (significant at the 1% level) this relationship is also showin in Figure 4.3.10 where we observe a slight negative trend between absolute targets and decarbonization rate. This contrasts with intensity-based targets (Cdp.Targettype.Intensity), which did not show significant effects, suggesting that absolute targets might be more impactful.

Assessing risks (Cdp.Aggregated.Risk) and identifying opportunities (Cdp.Aggregated.Opp) associated with climate change also appear to influence decarbonization rates significantly. Firms recognizing risks show a negative impact on decarbonization rate with a coefficient of 0.615 (significant at the 1% level), indicating that firms that identify risks are correlated with lower decarboninzation, this relationship can also be observed in the boxplot in Figure 4.3.11 where we observe a positive relationship between the aggregated risk and the decarbonization rate. Conversely, recognizing opportunities related to climate change is linked with better decarbonization outcomes, as shown by a negative coefficient of -1.283 (significant at the 1% level), suggesting that firms identifying decarboninzation opportunities tend to reduce their emissions over the next year. The relationship between opportunities and next year real decarbonization rate is also shown in Figure 4.3.12 where we observe that the distribution of firms who identify an opportunity corresponds to a relatively lower next-year decarbonization rate.

Key findings: Targets play an important role in enhancing decarbonization rates

over the next year, and an understanding of risks and opportunities, is equally important. Additionally, to gain a comprehensive understanding, it is most valuable to observe the mean amount target set by a firm, and whether that firm has identified risks and opportunites.

4.4 Investments, Initiatives, Carbon Credits, and Intensity Figures

4.4.1 Model VIII: Analyzing Investments and Initiatives

Table 4.4.1: Impact of Investments and Initiatives on Next-Year Real Decarbonization Rate

	Dependent variable:			
	Next Year Decarbonization Rate			
	(22)	(23)	(24)	(25)
Year	−0.049 (0.034)	−0.038 (0.031)	−0.059* (0.034)	−0.068** (0.032)
Ghg.Change.Real	0.188*** (0.009)	0.186*** (0.009)	0.185*** (0.009)	0.186*** (0.009)
Market.Cap	−0.406*** (0.085)	−0.397*** (0.084)	−0.400*** (0.085)	−0.403*** (0.084)
Revenue	0.270*** (0.094)	0.252*** (0.094)	0.269*** (0.094)	0.268*** (0.094)
Type.Scope1Limited/Moderate	0.388* (0.220)	0.352 (0.220)	0.373* (0.220)	0.370* (0.220)
Type.Scope1N.A	0.771*** (0.261)	0.728*** (0.262)	0.730*** (0.262)	0.723*** (0.262)
Type.Scope1Third.Party.Underway	0.770*** (0.294)	0.717** (0.295)	0.731** (0.295)	0.724** (0.294)
Ghg.Verification.Scope3.Yes	−0.307* (0.180)	−0.261 (0.181)	−0.264 (0.181)	−0.261 (0.181)
Ghg1	0.109*** (0.035)	0.123*** (0.034)	0.108*** (0.035)	0.112*** (0.035)
Ghg2Location	−0.048* (0.025)	−0.047* (0.025)	−0.045* (0.025)	−0.045* (0.025)
Ghg3.Count	−0.027 (0.025)	−0.013 (0.025)	−0.014 (0.025)	−0.012 (0.025)
Ghg1.Na	1.437*** (0.507)	1.668*** (0.502)	1.423*** (0.508)	1.465*** (0.506)
Ghg2Market.Na	0.888*** (0.179)	0.913*** (0.178)	0.904*** (0.179)	0.916*** (0.178)
Methane.Emissions	0.079*** (0.022)	0.083*** (0.022)	0.080*** (0.022)	0.081*** (0.022)
Method.IndInternal Incentives	−1.667*** (0.571)	−1.752*** (0.571)	−1.714*** (0.571)	−1.705*** (0.571)
Method.IndMacc	−0.710* (0.364)	−0.716** (0.363)	−0.704* (0.363)	−0.705* (0.363)
Cdp.Targetamount.Mean	−0.349*** (0.055)	−0.345*** (0.055)	−0.335*** (0.055)	−0.332*** (0.055)
Cdp.Targettype.Absolute	−0.164*** (0.058)	−0.153*** (0.058)	−0.156*** (0.058)	−0.160*** (0.058)
Cdp.Aggregated.Risk	0.584*** (0.180)	0.617*** (0.178)	0.584*** (0.181)	0.574*** (0.178)
Cdp.Aggregated.Opp	−1.109*** (0.237)	−1.132*** (0.236)	−1.073*** (0.237)	−1.063*** (0.237)
Initiative.Scope1	0.258** (0.124)		0.311** (0.125)	0.275** (0.117)
Initiative.Scope2	−0.153 (0.113)		−0.072 (0.115)	
Initiative.Scope3	−0.149 (0.165)		−0.040 (0.170)	
Absent.Cdp.Initiative.Firm.Year.Processed.Csv	0.855*** (0.269)		0.725** (0.283)	0.752*** (0.269)
Co2.Counter		−0.135 (0.084)	−0.106 (0.088)	
Msaving.Counter		0.183* (0.105)	0.177* (0.105)	
Investment.Counter		−0.313*** (0.121)	−0.358*** (0.122)	−0.278*** (0.064)
Investment.Total.Log1P		−0.002 (0.015)	0.007 (0.015)	
Constant	−0.364 (1.531)	−0.072 (1.522)	−0.368 (1.529)	−0.345 (1.532)
Random Effects:				
Number of Firms	1871	1871	1871	1871
Number of Industries	25	25	25	25
Number of Continents	6	6	6	6
Number of Countries	48	48	48	48
sd(Firms:Industry)	1.474	1.467	1.461	1.468
sd(Industry)	0.584	0.608	0.59	0.593
sd(Continent)	0.631	0.612	0.63	0.639
sd(Country:Continent)	0.207	0.2	0.2	0.231
Akaike Inf. Crit.	93589.449	93590.17	93593.274	93570.684
Bayesian Inf. Crit.	93815.293	93816.014	93849.23	93789

Note: *p<0.1; **p<0.05; ***p<0.01

Model (22): We analyze the predictors related to initiatives, in particual whether the company has initiatives to reduce carbon emissions and which scopes have a related initiative. Model (23): we analyze investments that companies make to reduce carbon emissions by looking at the number of invesmtents, the potential GHG savings, and the estimated capital requirement to execute the investments. Model (24): we combine all predicors together to then perform feature selection. Model (25): The final model revomes non-significant predcitors to enhance the AIC score and strike an optimal balance between number of features and forecasting ability.

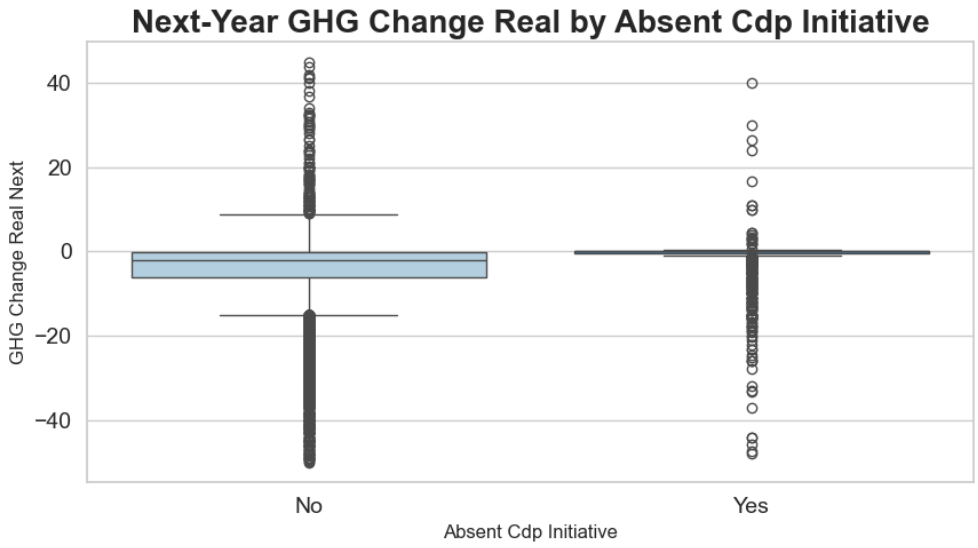


Figure 4.4.1: Absent Cdp Initiative vs. Next Year Decarbonization Rate

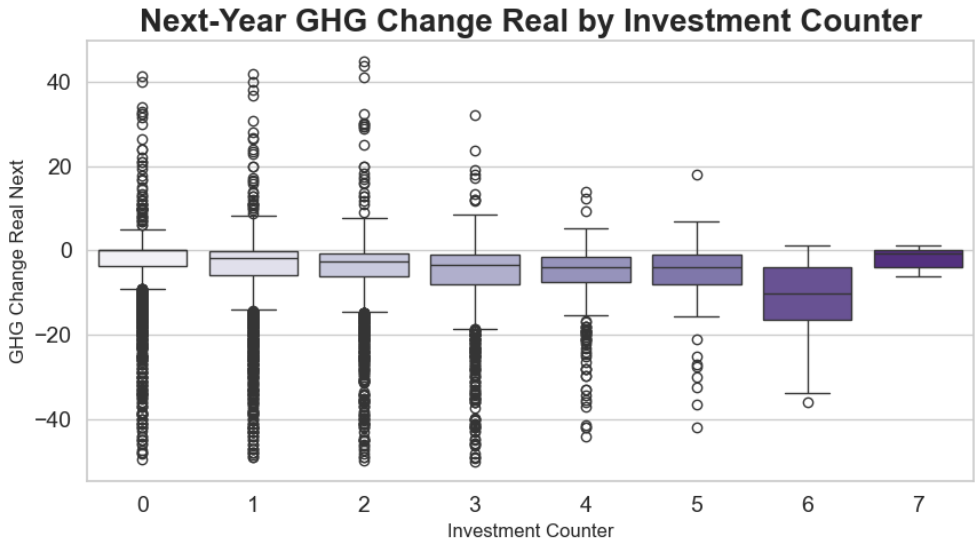


Figure 4.4.2: Investment Counter vs. Next Year Decarbonization Rate

Discussion

The only significant initiative metrics are whether the company does not have any initiatives, associated with a worse decarboninzation of 0.855, and whether the company has initiatives on scope 1, which is associated with worse decarbonization rate of 0.258%. The relationship between the absence of initiatives and next year decarbonization rate is also shown in Figure 4.4.1, where we observe that firms who don't report initiatives also have a decarbonization rate of zero, that is either they don't decarbonize, or they don't accurately report their decarbonization efforts. The Scope 1 coefficient makes sense given the context of the model as it likely acts as an indicator of whether a firm that has incentives has them in Scope 1 emissions or in either Scope 2 or Scope 3. With having incentives in Scope 2/3 being associated with better decarboninzation rates.

The absolute number of investments is significant and associated with a better decarboninzation rate of 2.78% per investment in model (4). Other invesmtent metrics are not significant at the 5% level and were not included in the final model. The relationship between the number of investments and the next year decarbonization rate is also shown in Figure 4.4.2.

Key Insights: Investments and Initiatives are important components of carbon emission reduction. Based on this analysis, which is correlational, they have an effect on decarboninzation that goes past the same year and affects the following year as well. This makes intutitive sense as we expect the investment that a firm makes along with economic incentives to have an effect over time as in both cases positive outcomes result in systemic change has a positive impact over time.

4.4.2 Model IX: Analyzing Carbon Credits and Intensity Figures

Table 4.4.2: Impact of Carbon Credits and Intensity Figures on Next-Year Real Decarbonization Rate

	Dependent variable:		
	Ghg.Change.Real.Next	Ghg.Change.Real	Ghg.Change.Real.Next
	(26)	(27)	(28)
Ghg.Change.Real	0.187*** (0.009)		0.186*** (0.009)
Ghg.Change.Real.Prev		0.114*** (0.008)	
Year	−0.038 (0.031)	0.016 (0.029)	−0.045 (0.031)
Market.Cap	−0.396*** (0.084)	−0.204** (0.080)	−0.396*** (0.085)
Revenue	0.260*** (0.094)	0.177** (0.088)	0.262*** (0.094)
Type.Scope1Limited/Moderate	0.362* (0.220)	0.258 (0.207)	0.347 (0.220)
Type.Scope1N.A	0.730*** (0.262)	0.491** (0.246)	0.702*** (0.262)
Type.Scope1Third.Party.Underway	0.730** (0.295)	0.527* (0.275)	0.696** (0.294)
Ghg.Verification.Scope3.Yes	−0.259 (0.181)	−0.569*** (0.169)	−0.252 (0.181)
Ghg1	0.124*** (0.034)	0.149*** (0.031)	0.125*** (0.034)
Ghg2Location	−0.049** (0.025)	−0.017 (0.023)	−0.046* (0.025)
Ghg3.Count	−0.012 (0.025)	−0.067*** (0.023)	−0.010 (0.025)
Ghg1.Na	1.633*** (0.503)	1.512*** (0.462)	1.612*** (0.502)
Ghg2Market.Na	0.912*** (0.178)	0.951*** (0.165)	0.904*** (0.178)
Methane.Emissions	0.082*** (0.022)	0.029 (0.020)	0.082*** (0.022)
Method.IndInternal Incentives	−1.722*** (0.571)	−0.560 (0.534)	−1.710*** (0.571)
Method.IndMacc	−0.707* (0.363)	−0.744** (0.342)	−0.698* (0.363)
Cdp.Targetamount.Mean	−0.340*** (0.055)	−0.455*** (0.051)	−0.330*** (0.055)
Cdp.Targettype.Absolute	−0.153*** (0.058)	−0.122** (0.055)	−0.152*** (0.058)
Cdp.Aggregated.Risk	0.623*** (0.177)	0.472*** (0.163)	0.623*** (0.177)
Cdp.Aggregated.Opp	−1.098*** (0.237)	−0.796*** (0.220)	−1.081*** (0.237)
Absent.Cdp.Initiative.Firm.Year.Processed.Csv	0.733*** (0.270)	0.574** (0.250)	0.678** (0.267)
Investment.Counter	−0.254*** (0.063)	−0.377*** (0.058)	−0.247*** (0.062)
Ghg.Int.Figure.Na	−2.861 (1.888)	−0.402 (1.739)	
Ghg.Int.Figure	0.021 (0.034)	0.028 (0.031)	
Ghg.Int.Change	−0.001 (0.003)	0.059*** (0.003)	
Absent.Cdp.Ghg.Int.Processed.Csv	−0.266 (0.201)	0.069 (0.186)	
Cdp.Num.Credits.Clean.Count			0.011 (0.043)
Cdp.Orig.Or.Purchase.Clean.Credit.Origination			0.006 (0.062)
Cdp.Purpose.Clean.Voluntary.Offsetting			−0.048 (0.067)
Absent.Cdp.Carbon.Credits.Full.Processed.Csv			0.067 (0.196)
Constant	−0.423 (1.533)	−2.738* (1.459)	−0.578 (1.559)
Random Effects:			
Number of Firms	1871	1871	1871
Number of Industries	25	25	25
Number of Continents	6	6	6
Number of Countries	48	48	48
sd(Firms:Industry)	1.466	1.604	1.468
sd(Industry)	0.605	0.425	0.602
sd(Continent)	0.625	0.53	0.625
sd(Country:Continent)	0.213	0.446	0.227
Akaike Inf. Crit.	93587.623	91292.459	93593.136
Bayesian Inf. Crit.	93828.524	91533.357	91533.357

Note: *p<0.1; **p<0.05; ***p<0.01

Model (26): We analyze the predictors related to intensity figures and intensity change controlling for all other variables carried forward from previous models. Model (27): similar to model (26), but we perform inference on the same-year by using as response Ghg.Change.Real and controlling for Ghg.Change.Real.Prev (lag-1 variable of Ghg.Change.Real). Model (28): we analyze carbon credits and the impact that they have on next year decarboninzation rate, still controlling for all other variable carried forward from previous models.

Discussion

Intensity is a variable that is important when assessing decarbonization and in model (27) intensity change is significant and positive as expected. Though, in model (26) the predictor is not significant and the coefficient value is almost 0, this seems to suggest that while intensity can explain same-year emissions, given the other control varaibles present in the model, it has little predictive power on next year emisssions.

Carbon Credits show no rpedictive power for next year emissions, this makes sense as carbon credits have no impact on *Real* Decarbonization Rate which can be achieved only through utilization of a greater share of renewable energy or by enhacning internal processes.

Key Insights: Neither Intensity figures nor Carbon Credits are useufl predictors when assessing next year real decarbonization rate. Therefore, I will exculde them from the final model.

4.5 Finalized Models

Table 4.5.1: Impact of Selected Predictors on Next-Year Real Decarbonization Rate

	Dependent variable:		
	Next Year Decarbonization Rate		Ghg.Change.Real
	(29)	(30)	(31)
Year		−0.048 (0.030)	−0.013 (0.028)
Ghg.Change.Real	0.188*** (0.009)	0.186*** (0.009)	
Ghg.Change.Real.Lag1			0.117*** (0.008)
Market.Cap	−0.414*** (0.084)	−0.398*** (0.084)	−0.253*** (0.080)
Revenue	0.245*** (0.091)	0.257*** (0.094)	0.163* (0.089)
Type.Scope1Limited/Moderate	0.361* (0.218)	0.358 (0.220)	0.235 (0.209)
Type.Scope1N.A	0.914*** (0.242)	0.720*** (0.261)	0.534** (0.249)
Type.Scope1Third.Party.Underway	0.833*** (0.291)	0.713** (0.294)	0.535* (0.277)
Ghg.Verification.Scope3.Yes		−0.272 (0.178)	−0.667*** (0.169)
Ghg1	0.115*** (0.033)	0.125*** (0.034)	0.179*** (0.032)
Ghg2Location		−0.046* (0.025)	−0.022 (0.023)
Ghg1.Na	1.954*** (0.472)	1.614*** (0.502)	1.845*** (0.466)
Ghg2Market.Na	1.026*** (0.149)	0.917*** (0.177)	1.030*** (0.166)
Methane.Emissions	0.075*** (0.022)	0.082*** (0.022)	0.030 (0.021)
Method.IndInternal Incentives	−1.752*** (0.570)	−1.722*** (0.571)	−0.438 (0.541)
Method.IndMacc	−0.679* (0.362)	−0.712** (0.363)	−0.827** (0.346)
Cdp.Targetamount.Mean	−0.353*** (0.054)	−0.334*** (0.055)	−0.506*** (0.052)
Cdp.Targettype.Absolute	−0.152*** (0.058)	−0.158*** (0.058)	−0.154*** (0.055)
Cdp.Aggregated.Risk	0.774*** (0.156)	0.625*** (0.177)	0.437*** (0.165)
Cdp.Aggregated.Opp	−1.182*** (0.233)	−1.085*** (0.236)	−0.828*** (0.223)
Absent.Cdp.Initiative.Firm.Year.Processed.Csv	0.686** (0.267)	0.685** (0.267)	0.676*** (0.250)
Investment.Counter	−0.274*** (0.061)	−0.254*** (0.062)	−0.409*** (0.058)
Constant	−0.621 (1.443)	−0.395 (1.526)	−1.687 (1.464)
Random Effects:			
Number of Firms	1871	1871	1871
Number of Industries	6	6	6
Number of Continents	6	6	6
Number of Countries	25	25	25
sd(Firms:Industry)	1.463	1.464	1.617
sd(Industry)	0.613	0.602	0.414
sd(Continent)	0.625	0.623	0.509
sd(Country:Continent)		0.221	0.427
Akaike Inf. Crit.	93553.665	93564.395	91631.244
Bayesian Inf. Crit.	93726.812	93767.655	91834.502
R-squared	0.145	0.146	0.155

Note: *p<0.1; **p<0.05; ***p<0.01

Model (29): result of applying backward stepwise regression to eliminate non-significant control variables and random effects to generate a final model with the best AIC. Model (30): adding back some predicors that I believe are important and that were analysed before, the model is very similar to model (29), with the only difference that we don't remove the coefficient for year and the random intercept for country. Model (31): testing model (30) on same-year decarbonization rate controlling for Ghg.Change.Real.Lag1 to check which predictors are significant in predicting same-year decarboninzation rates and pefrom a final comparison

Relevant Figures

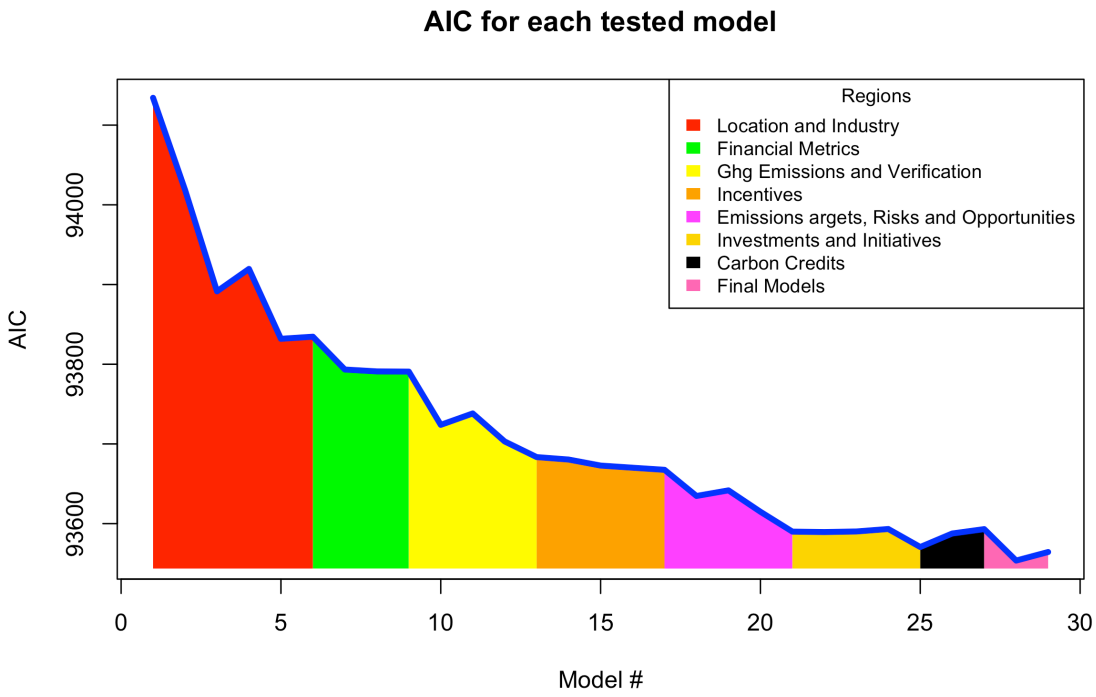


Figure 4.5.1: AIC of all tested models by model iteration

Discussion

We observe how methane emissions is significant at the 1% level when forecasting next year decarboninzation rate, while it is not significan when the response variable is same year decarbonization rate. This is an interesting result and seems to suggest that methane emissions are correlated more with future struggle to decarbonize compared to present decarbonizaion.

Overall, in model (30) nearly all the predictos are significant, with most being significant at the 1% level. Additionally Figure 4.5.1 shows how in all model iterations performed in this chapter, we manage to reduce the AIC significantly with the addiiton of significant predictors and the inclusion of relevant random effects.

In the final stages, the AIC plateaus, signaling that successive models will only be marginally better than our current final one which strikes a good balance between number of features, interpretability, and forecasting accuracy.

Next Steps: In the following chapter, we will test the prediction accuracy of the final model and use it as a benchmark to develop nonparamatric and more advanced modeling techiques, prioritizing prediction accuracy over interpretability.

5

Finding The Best Model

Chapter Preview

Overall, we find that while the **Mixed Effects Model** is a strong contender, but **CatBoost** is the best model for this dataset. Most importantly, we find that **different models identify similar sets of important features**, confirming the hypothesis that from the CDP survey we can identify a group of key variables that together can best forecast next year decarbonization rate.

5.1 Modeling Objectives

In this chaper, we will tune, benchmark, and report the results of various models to forecast next year's real decarbonization rate. The primary objectives are to identify the best model to forecast next year's decarbonization rate with the current dataset

to inform industry stakeholders on the current predictive capabilities of the survey. At the same time, we will determine the most important features for predicting the decarbonization rates and we will compare the results with the previous chapter to ensure consistency across models. Finally, we will use our results to inform the CDP on the most important variables to focus on when designing the survey and assigning future grades to firms based on their projected real decarbonization rate. Initially, we will be re-training the final mixed-effects model and comparing it against more flexible, data-driven non-parametric approaches such as decision trees and ensemble methods. Then, we will be evaluating all the models based on Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared value [29]. In summary, the models that are evaluated in this chapter are the Mixed Effects Model from Chapter 5, Bayesian Ridge Regressor , Catboost Regressor, Ridge Regression, Orthogonal Matching Pursuit, Elastic Net, Lasso Regression, Lasso Least Angle Regression, Gradient Boosting, Random Forest, Light Gradient Boosting Machine, Extra Trees, Dummy Regressor, Extreme Gradient Boosting, K Neighbors Regressor, Huber Regressor, Decision Tree, AdaBoost, Passive Aggressive Regressor, and Ordinary Least Squares. The models are all benchmarked using the Pycaret library [6]. The body of the chapter is dedicated to exploring and refining the three models introduced in the Methods chapter, which I believe to be the most promising: the Mixed Effects Model, Bayesian Ridge, and Catboost Regressor. For a detailed methods description of all other models used for benchmarking, please refer to Pycaret's and Sklearn's documentation [6, 46].

5.1.1 Data Preprocessing

The predictors in this chapter are the same as the ones analyzed across all the models in chapter 5, but the data has been preprocessed and split into two sets for training and testing purposes. The test set contains the year 2021, which is the most recent year for which we have data. The training set contains all years from 2011 to 2020. Additionally, the selected models are tuned using grid search and cross-validation to find the best hyperparameters. The folds are created using the *TimeSeriesSplit*

method from the scikit-learn library [46] to ensure that the data is split in a time series fashion and to prevent future leakage, an overfitting condition that occurs when data that would not be available at the time of predction is inadvertedly used to train the model, this happens when cross-validation is not done in a time series fashion and can lead to overly optimistic performance metrics. The models are evaluated based on the RMSE, MAE, and R-squared values.

Train and Test Set Summary Statistics

This is a summary of the training and testing sets used in this chapter. The training set contains all years from 2011 to 2020, and the testing set contains 2021. The year 2022 has been excluded from the analysis as we don't have the next year's decarbonization rate to compare the predictions against at the time of writing this thesis. Note that the number of features includes one-hot encoded variables, the actual number of predictors is the same as the previous chapter. All categorical features have been one-hot encoded except for the CatBoost model, which automatically handles categorical variables. Furthermore, the numerical predictors have been scaled using the Standardscaler from the scikit-learn library [46].

Table 5.1.1: Summary Statistics for Training and Testing Data

Dataset	Train	Test
Number of Observations	12411	1330
Number of Features	130	130
Number of Unique Firms	1870	1330
Mean Next Year Decarbonization Rate	-4.19	-5.98
Standard Deviation Next Year Decarbonization Rate	7.47	10.13

% of Total Observations	90.32%	9.68%

5.2 Baseline Metrics

The baseline metrics for the test set are calculated according to the following methods:

1. Using previous year decarbonization rate to predict next year's decarbonization rate
2. Using mean decarbonization rate for each firm across all reported years
3. Guessing zero for all firms as the next year's decarbonization rate
4. Using the mean for all firms for each year as the prediction for the next year's decarbonization rate

Table 5.2.1: Baseline Metrics for Test Set

	Method	MSE	RMSE	MAE	R2
1	Current Year Rate	148.06	12.17	7.20	-0.44
2	Previous Mean For Each Firm	109.16	10.45	6.41	-0.06
3	Guessing Zero for All Firms	138.31	11.76	6.99	-0.35
4	Previous Year Mean for All Firms	102.52	10.13	7.03	-0.00

As we can observe from the table, by guessing the previous year's decarbonization rate for each firm, we get a RMSE of 10.45. Similarly, by guessing the mean decarbonization rate for each firm, we get a RMSE of 10.13. We will use these metrics as a baseline to evaluate the performance of the models in this chapter.

5.3 Mixed Effects Regression Model (Chapter 5)

In this seciton we evaluate the Mixed Effects Model from the previous chapter, in particular the *Final Model (2)* presented in Table 4.5.1 which has been selected based on the AIC values and feature significance and justification. In this case, the model represents our final selection of features and random effects structure after a thorough analysis performed chapter 5. The model has been trained on the whole training set and evaluated on the test set.

5.3.1 Table of Model Performance Metrics

Table 5.3.1: Model Performance Metrics for Training and Test Sets

Set	R^2	RMSE	MSE	MAE
Training	0.15	6.71	45.05	0.00
Test	0.10	9.58	91.77	6.47

As we can observe from the table, the Mixed Effects Model has a RMSE of 9.58 on the test set and the R^2 value is 0.10, which means that the model explains 10% of the variance in the test set.

5.3.2 Residuals Plot

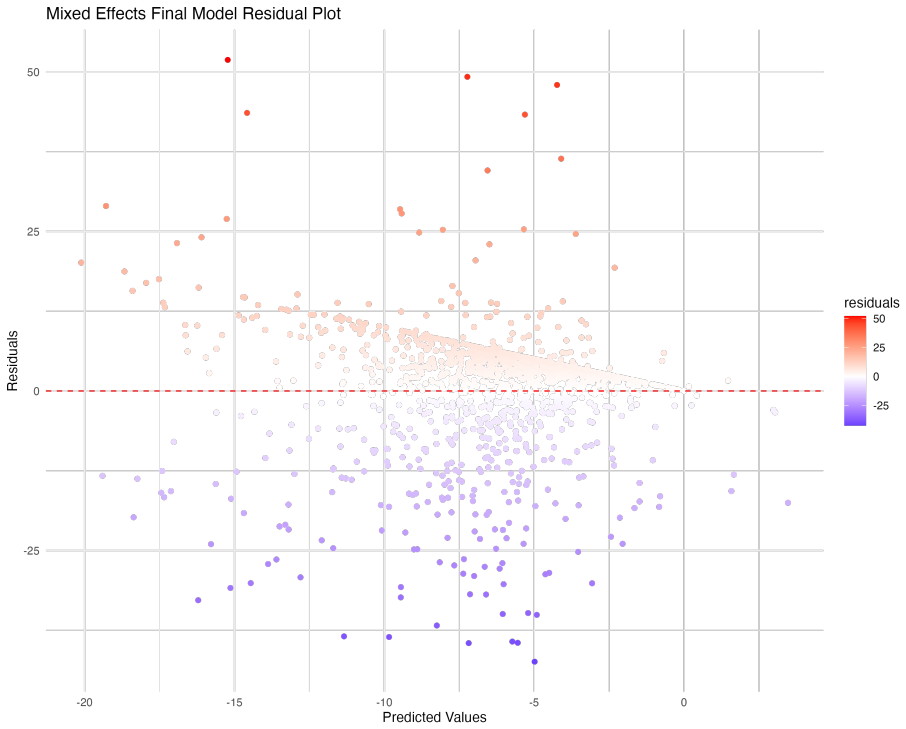


Figure 5.3.1: Mixed Effects Model Residuals Plot

The residuals plot is centered around zero, which is a good sign, but there are significant outliers, especially for firms with positive change in decarbonization rate and for firms with an absolute change of more than 20%. This suggests that the model is not performing well for these firms and it makes intuitive sense as firms who have a decarbonization rate change of more than 20% are likely to be outliers and difficult to predict.

5.4 Model Selection and Benchmarking

Using Pycaret [6], a Machine Learning library, we will be comparing the performance of various models on the dataset. The models will be evaluated based on the RMSE,

MAE, and R-squared values. The best model will be selected based on the RMSE value. We used timeseries cross-validation to ensure that the data is split in a time series fashion. As explained in the introductoiun, the folds are created using the TimeSeriesSplit method from the scikit-learn library [46]. We are not tuning the hyperparameters for the models in this section, as we will be doing that in the next section only for the best model. This analysis is useful in identifying which candidate models perform best on the dataset, and consequently which models are worth tuning.

Table 5.4.1: Cross Validation Results for All Tested Models

Model	MAE	MSE	RMSE	R2
Bayesian Ridge	4.15	48.40	**6.91**	0.09
CatBoost Regressor	4.34	48.77	6.94	0.09
Ridge Regression	4.22	48.74	6.94	0.08
Orthogonal Matching Pursuit	4.29	49.09	6.96	0.08
Elastic Net	4.25	49.35	6.97	0.08
Lasso Regression	4.23	49.54	6.99	0.07
Lasso Least Angle Regression	4.23	49.54	6.99	0.07
Gradient Boosting Regressor	4.34	49.57	7.00	0.07
Light Gradient Boosting Machine	4.35	49.71	7.01	0.07
Random Forest Regressor	4.56	50.25	7.04	0.06
Extra Trees Regressor	4.51	50.90	7.08	0.05
Dummy Regressor	4.45	54.09	7.30	-0.01
Extreme Gradient Boosting	4.85	57.20	7.51	-0.07
K Neighbors Regressor	4.80	60.37	7.71	-0.13
Huber Regressor	4.38	67.05	8.13	-0.25
Decision Tree Regressor	5.89	96.97	9.80	-0.84
AdaBoost Regressor	9.13	115.14	10.60	-1.12
Passive Aggressive Regressor	6.84	187.64	12.92	-2.66

Continued on next page

Table 5.4.1: Cross Validation Results for All Tested Models

Model	MAE	MSE	RMSE	R2
Linear Regression	31.33	3024.39	35.90	-42.10

Note how Bayesian Ridge and CatBoost have the lowest RMSE values. We will be exploring these models further in the next section. In general though, no model is significantly better than the others, which suggests that there is significant unexplained variance in the data. This is to be expected, as the CDP survey data is a first step in understanding the decarbonization rate, but there are many other factors that determine the decarbonization rate, especially in the long term. Additionally, there is significant noise in the data due to inconsistent reporting, which makes it difficult to predict the decarbonization rate accurately. That is, firms may not report their decarbonization rate accurately, or they may not report it at all, which makes it difficult to predict. In this exercise, I preferred

5.5 Bayesian Ridge Model

The Bayesian Ridge model is a linear regression model that uses a Bayesian approach to estimate the coefficients of the model. The model is regularized using a prior distribution on the coefficients, which helps to prevent overfitting. The model is evaluated based on the RMSE, MAE, and R-squared values. The hyperparameters that we tuned are the alpha parameter, which controls the strength of the regularization, and the lambda parameter, which controls the strength of the prior distribution on the coefficients. The best model was selected based on the RMSE value. Train and test results are reported in the Table 5.5.1 below.

Table 5.5.1: Bayesian Ridge Model Performance Metrics for Training and Test Sets

Set	R^2	RMSE	MSE	MAE
Training	0.9	6.91	48.4	4.15
Test	0.10	9.59	91.9	6.46

The Bayesian Ridge model has a RMSE of 9.59 on the test set, which is very similar to the Mixed Effects Model. This suggests that the Bayesian Ridge model is not significantly better than the Mixed Effects Model.

5.5.1 Feature Importance Plot

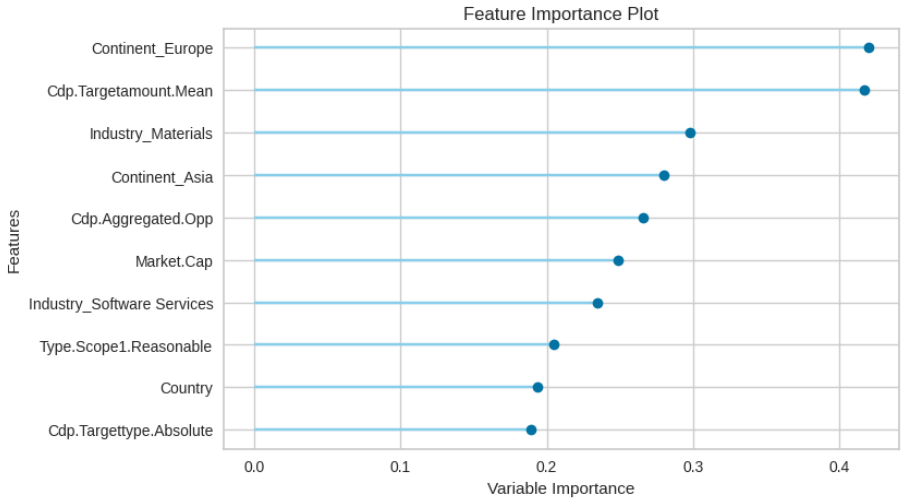


Figure 5.5.1: Bayesian Ridge Model Feature Importance Plot

The feature importance plot 5.5.1 shows that the most important features are continent Europe, the Target Amount Mean, the industry Materials, whether the firm identified an opportunity to reduce emissions. All those features are consistent with the Mixed Effects Model and the Exploratory Data Analysis we performed in Chapter 2. Consistency across model when it comes to feature importance is a good sign, as it suggests that the features are indeed important for predicting the decarbonization rate and that they are not just artifacts of the model.

5.5.2 Residuals Plot

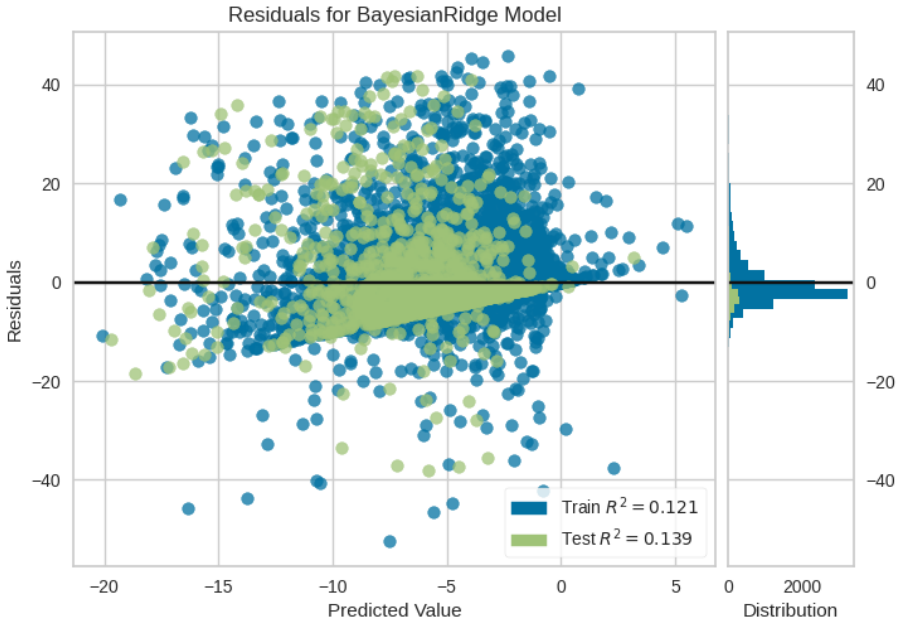


Figure 5.5.2: Bayesian Ridge Model Residuals Plot

The residuals plot is smilar to the Mixed Effects Model, centered around zero, but with significant outliers, especially for firms with positive change in decarbonization rate and for firms with an absolute change of more than 20%. This suggests that both models have similar performance and are not performing well for these firms. This makes sense and is consistent with the fact that the data is noisy and that with a greater focus on enhancing reporting quality for relevant variables, the models could be greatly improved.

5.6 Catboost Regressor Model

To tune the Catboost Regressor model, we used grid search and cross-validation to find the best hyperparameters for the model. The hyperparameters that we tuned are the learning rate, the depth of the tree, the number of trees, and the l2 regularization

parameter. We used the TimeSeriesSplit method from the scikit-learn library to ensure that the data is split in a time series fashion with number of folds $cv = 3$. The model was evaluated based on the RMSE, MAE, and R-squared values. The best model was selected based on the RMSE value. The hyperparameters that we tuned are as follows:

Table 5.6.1: Hyperparameters for Catboost Regressor

Parameter	Grid of Values	Selected Best Value
Depth	4, 6, 8	**6**
Iterations	500, 1000	**1000**
Learning Rate	0.01, 0.02, 0.03	**0.02**
L2 Leaf Reg	1, 3	**1**

5.6.1 Model Performance Metrics

Table 5.6.2: Catboost Regressor Tuned Model Performance Metrics

Set	R^2	RMSE	MSE	MAE
Training	0.33	5.73	35.91	3.47
Test	0.11	9.54	91.13	6.25

The Catboost Regressor model has a RMSE of 9.54 on the test set, and the R^2 value is 0.11, which is the best performance so far. This suggests that the Catboost Regressor model is the best model for this dataset. It does not come as a surprise, as Catboost is known for its ability to handle categorical variables and its robustness to overfitting. Therefore, model performance is in line with theoretical expectations and the results from the previous chapter with the main takeaway being that the model is able to generalize well to the test set, but the noise in the response variable significantly limits predictive performance.

5.6.2 Feature Importance Plot

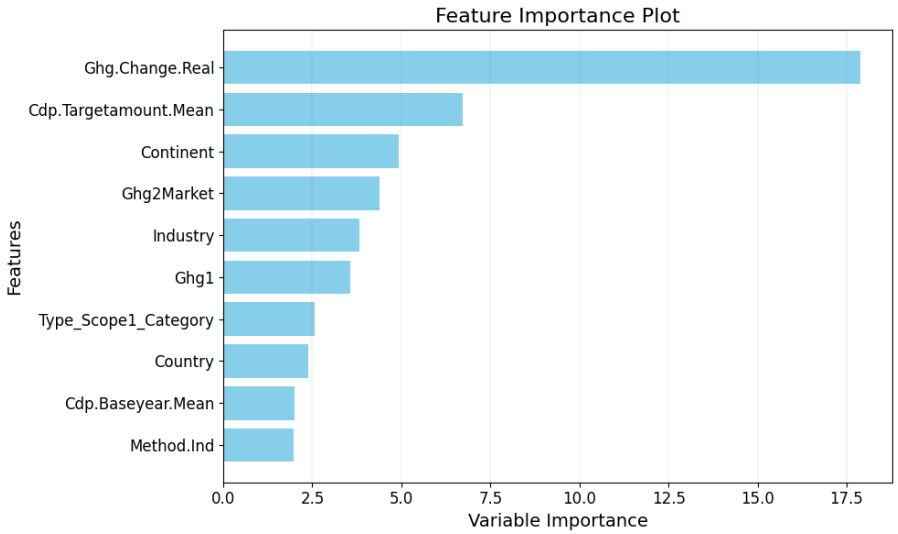


Figure 5.6.1: Catboost Regressor Feature Importance Plot

Catboost has a built-in feature importance plot, which is shown in Figure 5.6.1. The most important features are the Current Decarbonization Rate Change, Target Amount Mean, Continent, Whether the firm reports market based emisisons, and the industry. Note how reassuringly all the selected features by a tree-based boosting method, which is very different from the linear models, are consistent with the Mixed Effects Model and the Bayesian Ridge Model. We therefore learn that there is a consensus across models that the most important features are the same, thus to advance our ability to predict the decarbonization rate, we should focus on these features.

5.6.3 Shapley Beeswarm Values Plot

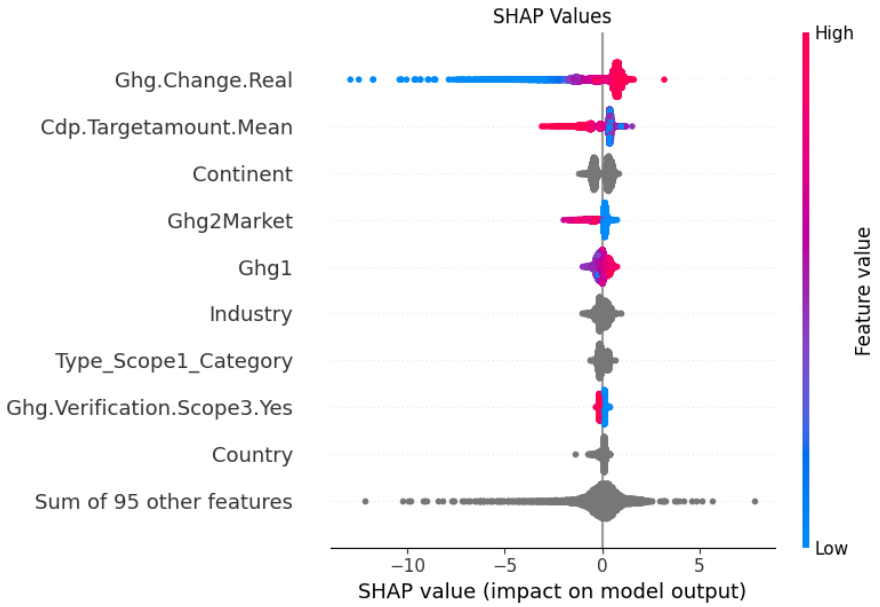


Figure 5.6.2: Catboost Regressor Shapley Beeswarm Values Plot

For the Catboost Regressor model, we used the Shapley values to explain the predictions of the model. The Shapley values are a measure of feature importance that is based on game theory [35]. The Shapley values are calculated for each feature and are used to explain the predictions of the model. The Shapley values are plotted in Figure 5.6.2. The beeswarm plot shows the distribution of the Shapley values for each feature. The plot offers an interpretation of the effect that each feature has on the model's prediction, with a high Shapley value off the x-axis indicating a strong effect on the model's prediction in that direction. Therefore, the plot is useful in understanding the relationship between the features and the model's predictions akin to interpreting coefficients in a linear model. In this regard, the effect of current decarbonization rate change is important and positive, that is firms with a positive change in decarbonization rate are likely to have a higher decarbonization rate next year. Similarly, the effect of the target amount mean is important and negative, that

is firms with a higher target amount mean are likely to have a better decarbonization rate next year. The effect of Scope 2 Market Emissions is also similiar to what we found in the Mixed Effects Model: firms that report market based emissions are likely to have a better decarbonization rate next year.

6 Conclusion

6.1 Selected Individual Firm Actions

All the modeling techniques we employed were targeted at identifying key relationships between firm-level actions and decarbonization rates. Through our analysis, we had the opportunity to explore various important actions, finding both expected and surprising results. In this section I will highlight the most important relationships, and then I will argue why theese relationships can be useful to develop a better grading system for CDP reports. The results are presented in Table 6.1.1 where we show the coefficients of the Mixed Effect final model, as well as whether the variable was selected as important by the Bayesian Ridge and CatBoost models. We will primarily focus on variables that were selected as top 10 in importance by all three models.

6.1.1 Background Predictors

Our analysis identified several core predictors essential for forecasting next-year's decarbonization rates. First among them is the **current-year decarbonization rates**, indicating a strong positive correlation with next-year's outcomes. This trend underscores a continuity in firms' environmental efforts, where entities engaged in emission reduction are likely to persist in their endeavors. Equally central to our analysis is the **industry** in which a firm operates. Our findings reveal that firms within sectors considered difficult to abate, such as **industrials** or **construction**, typically exhibit poorer decarbonization performance. Conversely, those in the **technology** and **consumer goods** sectors are more prone to reducing emissions, highlighting the impact of industry characteristics on decarbonization efforts and the need for tailored strategies across sectors.

Another significant predictor is the **geographical location**—both the **country** and **continent**—of a firm's operations. Our analysis points to a faster decarbonization pace within the **European Union**, as opposed to slower rates observed in **Asia**. This geographic variance is important to take into consideration when forecasting firm-level decarbonization rates, alongside an observed disparity in disclosure practices, with a predominance of reporting from firms in the **United States** and **Europe**.

Overall, the initial set of important variables for evaluating a company's future decarbonization trajectory include its **sector**, **country of operation**, and **current decarbonization rate**. Understanding these factors is an important starting point for assessing environmental strategies within the broader context of industry and geographic dynamics.

6.1.2 Most Important CDP Survey Metrics

In our analysis of CDP-specific metrics, certain predictors emerged as consistently significant across all models. These include the **average amount of emission targets** for Scope 1 and 2 emissions, the **verification method** for Scope 1 emissions, and the adoption of either a marginal abatement cost curve (MACC) or internal

incentives to guide **emission reduction initiatives**.

Particularly important is the role of **incentive type** in predicting next-year's decarbonization rates, which according to our analysis is more significant than its impact on same-year rates. According to Table 6.1.1, both **internal incentives** and the **MACC** approach show a negative correlation with decarbonization rates. This implies that firms employing these strategies tend to achieve greater year-on-year reductions in emissions. Such findings highlight the importance of a proactive approach to incentives, differentiating between short-term impacts and long-term sustainability in decarbonization efforts. Therefore, we find that all key stakeholders should analyze which incentive structures are in place to determine the future effectiveness of a firm's environmental strategy.

Scope 1 emission verification also stands out as a crucial predictor. Firms that undertake comprehensive verification of their Scope 1 emissions are generally more successful in decarbonizing. This contrasts with outcomes associated with limited, moderate, unavailable, or third-party verifications underway. This pattern emphasizes the value of external validation in emission reporting, suggesting that transparency and accountability in reporting, alongside consistent verification processes, are key to superior decarbonization performance. Firms that are open and accountable about their emissions tend to be more committed to environmental goals, reflecting a more structured and reliable strategy that is more likely to lead to decarbonization in the future. Therefore, the extent and rigor of emissions disclosure and verification serve as strong indicators of a firm's decarbonization commitment and effectiveness.

Lastly, the **average amount of emission targets** set across Scope 1 and 2 emissions is an essential indicator of future decarbonization performance. We find that firms that establish more ambitious targets are likely to demonstrate better decarbonization outcomes. This consistency underscores the notion that ambition in setting environmental targets correlates with a firm's dedication to its environmental objectives and its capability to execute a sustained and effective strategy. Overall, the level of ambition in emission targets is a direct indicator of a firm's potential to achieve significant decarbonization in the future.

Table 6.1.1: Mixed Effects Model Coefficients, with Indicator of Bayesian Ridge and CatBoost Top 10 Feature Importances

Predictor	**MixedEffect**	**BayesianR**	**CatBoost**
Year	-0.048	✗	✓
Ghg.Change.Real	0.186^{***}	✓	✓
Market.Cap	-0.398^{***}	✗	✗
Revenue	0.257^{***}	✗	✗
Type.S1.Limited/Moderate	0.358	✓	✓
Type.S1.N.A	0.720^{***}	✓	✓
Type.S1.Third.Party.Underway	0.713^{**}	✓	✓
Ghg.Verification.Scope3.Yes	-0.272	✗	✗
Ghg1	0.125^{***}	✗	✓
Ghg2Location	-0.046^{*}	✗	✗
Ghg1.Na	1.614^{***}	✗	✓
Ghg2Market.Na	0.917^{***}	✗	✓
Methane.Emissions	0.082^{***}	✗	✗
Method.IndInternal Incentives	-1.722^{***}	✓	✓
Method.IndMacc	-0.712^{**}	✓	✓
Cdp.Targetamount.Mean	-0.334^{***}	✓	✓
Cdp.Targettype.Absolute	-0.158^{***}	✓	✗
Cdp.Aggregated.Risk	0.625^{***}	✗	✗
Cdp.Aggregated.Opp	-1.085^{***}	✓	✗
Absent.Cdp.Initiative	0.685^{**}	✗	✗
Investment.Counter	-0.254^{***}	✗	✗

*Notes: For the Mixed Effects Model, results are reported from model (30) in Chapter 5 4.5.1, where green indicates a positive impact and red indicates a negative impact on decarbonization. Significance levels are denoted as *, **, and *** for p<0.1, p<0.05, and p<0.01, respectively. For the Bayesian Ridge and CatBoost models, the displayed feature importances are from the tuned models. A green check (✓) signifies that the feature is among the top 10 based on feature importance, whereas a red cross (✗) indicates it is not.*

6.2 Forecasting Results and Data Considerations

We found that, with our data, **building a universal model that accurately forecasts decarbonization rates for any specific firm through disclosure and financial data is not achivable**. Due to irregularities in disclosure, such as missing values, non-accurate figures, and inherent variability not captured by disclosure data, there is not enough forecasting power to accurately predict next year decarbonization for a specific firm. To illustrate this point, it is sufficient to observe the various figures presented in Chapter 4 where it is possible to observe a significant level of noise in our data.

Though, we were successful in achieving our to major objectives. As explained in the introduction, the focus of predictive modeling is twofold: first, we want to understand which models work best with disclosure data when it comes to forecasting decarbonization rates at the firm-level. Second, through a predictive exercise, our aim is to identify which predictors are significant across various modeling techniques and what is their correlation with future decarbonization.

Addressing the first question, we found that more interpretable modeling techniques, such as Mixed Effects and Bayesian Regression, are the best choice. Alternatively, models that are particulary capable at leveraging categorical adata are also valid choices, with tree-based model and CatBoost yielding the best results. **We found that any model that either requires a relatively large number of data-points per firm to train (Long-Short Term Memory, autoregressive timeseries models) or that does not perform regularization and is not able to handle categorical data with high number of categories (Ordinary Least Squares) should be avoided**. The first group's disatvantage is obvious: we do not gain better predictive results compared to Mixed Effects as on average each firm does not have more than six data-points, and we do not have interpretability. Arguing why we should not use simple linear model which are standard in the field is more nuanced: first, we though our analysis we found that it is important to place each firm in its own category and recognize the profound difference in operations between firms that operate in different industries, countries, and continents. Yet, if

we one-hot-encode all this information we risk overfitting to the data. Second, if we instead chose not to use all those categorical information to prevent overfitting, then we bias some of the important relationships we are trying to uncover as we do not control for the fact that each firm has its own unique characteristics, and we build models that are not representative of the structure of our data thus we encounter an important omitted variables problem.

The most important key takeaway of predictive modeling is that **when it comes to disclosure data, taking a forward-looking predictive apporach allows to uncover firm-level disclosure actions that correlate with future decarbonization**. Crucially, as explained in the introduction, we need individual firms to decarbonize, thus it is important to suggest to all stakeholders which actions are most likely to lead to decarbonization **at the firm level**. Therefore, unlike most literature in the field **we should not aggregate the data**, as although we would likely get a better prediction of the aggregated emissions as is typiciallly the case when aggregating, this will not actually be useful in practice for suggesting individual firm actions as aggregating can lead to the data showing a different relationship than the individual data due to Simpson's paradox [45].

6.3 Improving the CDP Survey

Discussion on the residuals -> missing data is the problem -> we can use this as a feature to enhance disclosure. Simple surveys with pinpointed questions.

6.3.1 CDP Survey Design: emphasis on key quantitative metrics

Todo

6.3.2 CDP Scores: more transparency

Todo

6.4 Future Work

Todo

A

Appendix

A.1 Emission Scopes

The thesis will assume familiarity with the concept of Scopes 1, 2, and 3 emissions which I am going to introduce in this section. In carbon-accounting and emissions reporting, it is very important to distinguish between three types of emissions: Scope 1, Scope 2, and Scope 3 emissions. Each category represents a different level of emissions associated with an organization's activities as shown in Figure A.1.1.

- **Scope 1** emissions are direct emissions from owned or controlled sources. This includes emissions from company vehicles, and emissions from chemical processes or combustion in owned or controlled boilers, furnaces, etc.
- **Scope 2** emissions are indirect emissions from the generation of purchased electricity, steam, heating, and cooling consumed by the reporting company.

These emissions occur at the facility where the energy is generated, not at the point of consumption.

- **Scope 3** emissions are all indirect emissions (not included in Scope 2) that occur in the value chain of the reporting company. This includes both upstream and downstream emissions, encompassing a wide range of activities such as the extraction and production of purchased materials, transportation of purchased fuels, and use of sold products and services.

Understanding these scopes is critical for organizations aiming to fully assess and manage their carbon footprint.

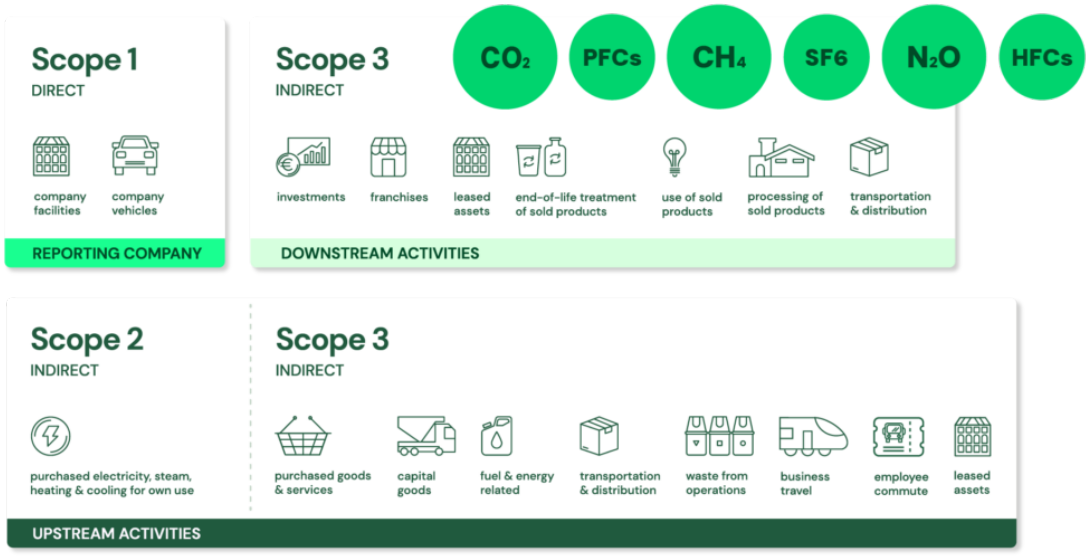
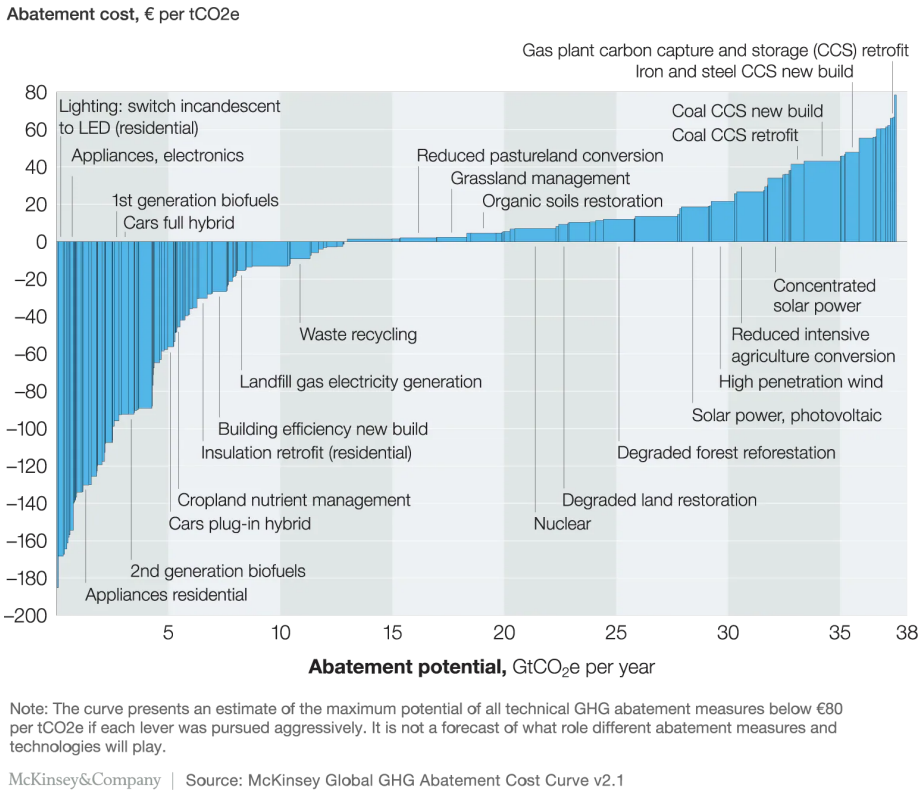


Figure A.1.1: Emissions Scopes 1, 2, and 3 . Adapted from [12].

A.2 Marginal Abatement Cost Curve

A Marginal Abatement Cost Curve (MACC) is a graphical representation of the cost of reducing GhG emissions. It is a tool used to identify the most cost-effective

measures to reduce emissions. The MACC is a plot of the cost of abating one additional unit of emissions against the quantity of emissions abated. The curve is typically upward-sloping, indicating that the cost of abating additional emissions increases as more emissions are reduced. The MACC is a useful tool for policymakers and businesses to identify the most cost-effective measures to reduce emissions [39]. It can also be used to compare the cost-effectiveness of different emission reduction measures and to identify the most cost-effective measures to achieve a set emissions reduction target. Figure A.2.1 shows an example of a MACC.



Note: The curve presents an estimate of the maximum potential of all technical GHG abatement measures below €80 per tCO2e if each lever was pursued aggressively. It is not a forecast of what role different abatement measures and technologies will play.

McKinsey&Company | Source: McKinsey Global GHG Abatement Cost Curve v2.1

Figure A.2.1: Marginal Abatement Cost Curve. Adopted From [39]

As we can observe from the graph, the MACC is upward-sloping, on the x-axis we have the quantity of emissions abated, and on the y-axis, we have the cost of abating one additional unit of emissions. Different industries and sectors will have different

MACCs, and the shape of the MACC will depend on the cost of abatement measures and the quantity of emissions abated.

A.3 Neural Network Model

A.3.1 Model Architecture

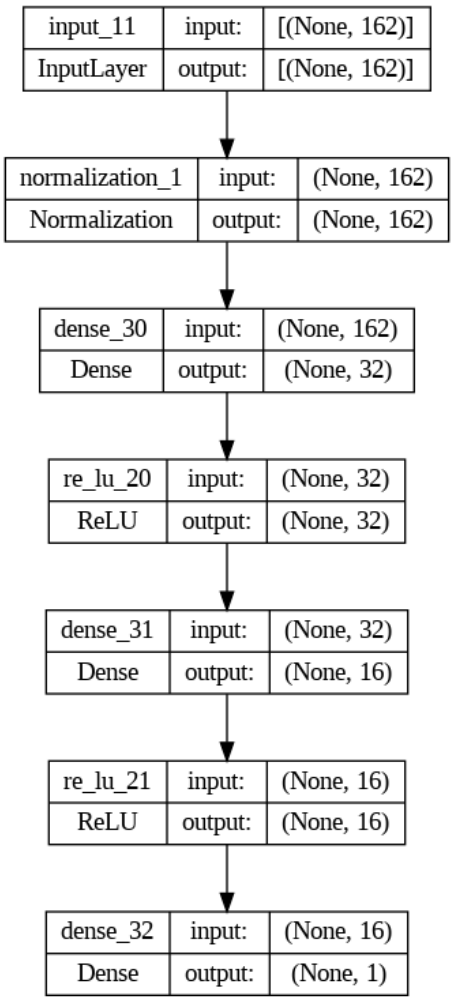


Figure A.3.1: Neural Networks Model Architecture

A.3.2 Training and Validation Loss Plot

Figure A.3.2: Neural Networks Model Training and Validation Loss Plot

A.3.3 Model Performance Metrics

Table A.3.1: Neural Networks Model Performance Metrics

Set	R^2	RMSE	MSE	MAE
Training	0.21	6.23	38.79	3.77
Test	0.09	9.65	93.26	6.32

A.4 Variable Dictionary

Table A.4.1 provides an overivew of the predictors used in the analysis, including their type, description, and source. Predictors are divided into three primary categories:

- **Firm Information:** Variables that describe the firm's characteristics, such as its unique identifier, reporting year, headquarters country, headquarters continent, and industry sector.

- **Financial Predictors:** Variables that capture the firm's financial performance, including total revenue, total assets, total employees, net income, and market capitalization.
- **CDP Predictors:** Variables derived from the CDP Climate Survey, such as the firm's emissions, energy consumption, and climate-related targets and initiatives.

Variable	Type	Description	Source
Firm Information			
ID	cat	unique firm identifier	CDP
Year	num	reporting year	CDP
Country	cat	headquarters country	CDP
Continent	cat	headquarters continent derived from country	CDP
Industry	cat	Global Industry Classification Standard 25 industry sectors	GICS
Financial Predictors			
log(Revenue)	num	natural logarithm of total revenue	WS
log(Assets)	num	natural logarithm of total assets	WS
log(Assets 1yr gr.)	num	natural logarithm of total assets growth	WS
log(Employees)	num	natural logarithm of total employees	WS
log(Empl. 1y gr.)	num	natural logarithm of total employees growth	WS
log(Net Income)	num	natural logarithm of net income	WS
log(Market Cap)	num	market capitalization	WS

Continued on next page

Table A.4.1 – *Continued from previous page*

Variable	Type	Description	Source
log(Roe)	num	natural logarithm of return on equity	WS
log(Revenue)	num	natural logarithm of total revenue	WS
CDP Predictors			
Ghg.Change.Real.Next	num	Change in real GHG emissions from previous year	CDP
Proportion.Verified.Scope1	num	Proportion of Scope 1 GHG emissions that are verified	CDP
Ghg1	num	Total direct (Scope 1) GHG emissions	CDP
Ghg2Location	num	Location-based Scope 2 GHG emissions	CDP
Ghg2Market	num	Market-based Scope 2 GHG emissions	CDP
Ghg3.Total	num	Total Scope 3 GHG emissions	CDP
Ghg3.Count	num	Count of Scope 3 GHG emissions sources reported	CDP
Ghg1.Na	cat	Indicator if Scope 1 GHG data is not available	CDP
Ghg2Location.Na	cat	Indicator if location-based Scope 2 data is not available	CDP
Ghg2Market.Na	cat	Indicator if market-based Scope 2 data is not available	CDP
Ghg3.Total.Na	cat	Indicator if Scope 3 data is not available	CDP
Methane.Emissions	num	Total methane GHG emissions	CDP

Continued on next page

Table A.4.1 – *Continued from previous page*

Variable	Type	Description	Source
Type.Scope1	cat	Type of Scope 1 emissions verification	CDP
Ghg.Verification.Scope1.Yes	cat	Indicator if Scope 1 emissions are verified	CDP
Ghg.Verification.Scope2.Yes	cat	Indicator if Scope 2 emissions are verified	CDP
Ghg.Verification.Scope3.Yes	cat	Indicator if Scope 3 emissions are verified	CDP
Method.Ind	cat	Indicator of the incentive method used	CDP
Cdp.Boardoversight.I	num	Board oversight on climate-related issues	CDP
Cdp.Incentivebinary.I	cat	Presence of incentives for climate-related performance	CDP
Cdp.Baseyearemission.Mean	num	Average of baseline year emissions data	CDP
Cdp.Targetscope.Mean	num	Average percentage of emissions scopes covered by targets	CDP
Cdp.Targetamount.Mean	num	Mean target emission reduction amount	CDP
Cdp.Targettype.Absolute	cat	Presence of absolute emission reduction targets	CDP
Cdp.Targettype.Intensity	cat	Presence of intensity-based emission reduction targets	CDP
Cdp.Aggregated.Risk	num	Aggregated measure of climate-related risks	CDP

Continued on next page

Table A.4.1 – *Continued from previous page*

Variable	Type	Description	Source
Cdp.Aggregated.Opp	num	Aggregated measure of climate-related opportunities	CDP
Initiative.Scope1	cat	Initiative related to Scope 1 emissions	CDP
Initiative.Scope2	cat	Initiative related to Scope 2 emissions	CDP
Initiative.Scope3	cat	Initiative related to Scope 3 emissions	CDP
Absent.Cdp.Initiative	cat	Indicator if firm-year data is absent in CDP initiative processed CSV	CDP
Co2.Counter	num	Count of CO2 reduction initiatives	CDP
Msaving.Counter	num	Count of money-saving initiatives due to emission reductions	CDP
Investment.Counter	num	Number of investments in emission reduction initiatives	CDP
Investment.Total.Log1P	num	Log-transformed total investment in emission reduction (log1p)	CDP
Cdp.Num.Credits.Clean.Count	num	Count of clean energy credits	CDP
Clean.Credit.Origination	cat	Origin of clean energy credits (original or purchased)	CDP
Cdp.PurposeVoluntary.Offsetting	cat	Purpose of clean energy credits for voluntary offsetting	CDP
Absent.Cdp.Carbon.Credits	cat	Indicator if carbon credits data is absent in processed CSV	CDP

Table A.4.1: Variable Dictionary

A.5 A Case Study on Two CDP Reports

To begin the discussion on exploratory data analysis, I must first address the complexities of emission accounting and reporting within the framework of CDP data, highlighting its distinct nature compared to the more standardized field of financial reporting. That is, despite ongoing improvements, emission reporting still falls short of the robust standards established in financial accounting. I will highlight this by analyzing the 2022 CDP reports from two markedly different companies: General Motors (GM) in the automotive sector and Jet Blue in the airline industry. These reports underscore the highly company-specific nature of emission data, with significant variances stemming from diverse operational practices, especially across different industries. This results in a substantial reliance on text-based and free-form answers within the CDP reports, presenting unique challenges for data analysis. To navigate this complexity, a critical starting point is to analyze the inherent differences in these reports, which will inform and refine our modeling approach. By understanding and accommodating these industry-specific nuances, we aim to develop a more accurate and representative model of emission reporting and reduction strategies as well as identifying potential areas of improvement.

A.5.1 General Motors 2022 CDP Report

General Motors Company (GM), a global leader in the automotive industry, is headquartered in Detroit, Michigan, USA. Renowned for its ownership and production of the Chevrolet, GMC, Cadillac, and Buick brands, GM was the largest automaker in the United States by sales in 2022. GM's commitment to sustainability is evident in its strategic approach to reducing Scope 1, Scope 2, and Scope 3 greenhouse gas (GHG) emissions, with comprehensive governance and ambitious environmental targets.

Scope 1 and Scope 2 Emissions

GM has set forth aggressive targets to reduce its Scope 1 and Scope 2 emissions by 71.4% by 2035, relative to its 2018 baseline. In 2018, GM reported Scope 1 emissions of 1,763,555 metric tons CO2e and Scope 2 emissions of 3,924,338 metric tons CO2e. By the reporting year 2022, GM achieved a reduction to 1,252,906 metric tons CO2e for Scope 1 and 2,150,694 metric tons CO2e for Scope 2, marking significant progress towards its goal. This reduction aligns with the 1.5 degrees Celsius strategy set by the Paris Agreement, underscoring GM's commitment to global climate initiatives [18, 63].

GM's strategy includes enhancing energy efficiency across its manufacturing operations and increasing the use of renewable energy. In 2021, GM implemented over 300 energy efficiency improvements, such as upgrading to more efficient equipment and increasing renewable electricity use from 23% to 25%, contributing to GHG reductions in Scope 2 emissions.

Scope 3 Emissions

Addressing Scope 3 emissions, GM has set a target to achieve a 50.4% reduction in its vehicle use emissions, from a baseline of 0.0002466 metric tons of CO2 per kilometer to 0.0001223136. GM's strategy to meet this target includes transitioning to an all-electric vehicle (EV) future, with plans to introduce 30 new EV models by 2025 and aspirations to be fully electric by 2035. Partnerships to increase renewable energy generation and deploy EV chargers, in collaboration with EvGo, further exemplify GM's holistic approach to reducing its carbon footprint across the value chain.

Key Takeaways

- **Strategic Emissions Reduction:** GM's targeted reductions in Scope 1 and Scope 2 emissions demonstrate a strong commitment to environmental stewardship, leveraging technological advancements and renewable energy.
- **Leadership in Electric Vehicles:** GM's aggressive transition to an all-EV

lineup by 2035 highlights its leadership role in transforming the automotive industry towards sustainability.

- **Comprehensive Approach to Sustainability:** Through its Scope 3 emissions reduction target, GM addresses the broader environmental impact of its products, emphasizing the importance of a comprehensive strategy that extends beyond direct emissions.

GM's sustainability efforts showcase a deep commitment to reducing its environmental impact and leaading the automotive industry towards a more sustainable future. By strategically targeting Scope 1, Scope 2, and Scope 3 emissions, GM is not only adhering to global climate agreements but also setting a precedent for corporate responsibility in addressing climate change.

A.5.2 JetBlue Airways Corporation 2022 CDP Report

JetBlue Airways Corporation has been steadfast in its commitment to environmental stewardship, focusing on reducing Scope 1, Scope 2, and Scope 3 greenhouse gas (GHG) emissions across its operations. The airline's governance structure emphasizes sustainability, with strategic initiatives overseen by its board and executive team, underscoring a comprehensive approach to addressing climate change.

Scope 1 and Scope 2 Emissions

In the reporting year 2022, JetBlue's Scope 1 emissions totaled 6,853,927 metric tons CO2e, predominantly from jet fuel combustion, a primary challenge within the airline industry. Scope 2 emissions amounted to 25,945 metric tons CO2e, reflecting the emissions from electricity consumption. These figures demonstrate JetBlue's significant environmental footprint, necessitating aggressive measures for reduction. JetBlue's strategies to mitigate these emissions include modernizing its fleet with more fuel-efficient aircraft, such as the Airbus A220 and A321neo, and investing in sustainable aviation fuel (SAF) to reduce lifecycle GHG emissions associated with jet

fuel. Additionally, the airline is transitioning its ground service equipment to electric power, aligning with its commitment to lower Scope 1 and Scope 2 emissions.

Scope 3 Emissions

JetBlue's Scope 3 emissions are a crucial component of its sustainability strategy, addressing emissions from purchased goods and services, capital goods, and fuel-and-energy-related activities not included in Scope 1 or 2. In 2022, the emissions reported were as follows:

- Purchased Goods and Services: 44,922 metric tons CO2e, estimated for catered food and onboard product.
- Capital Goods: 485,629 metric tons CO2e, associated with aircraft ground equipment and spare parts.
- Fuel-and-Energy-Related Activities: 1,391,126 metric tons CO2e, highlighting the broader impact of JetBlue's operational energy use.

These figures were calculated using the Quantis Scope 3 tool, demonstrating JetBlue's reliance on standardized methodologies to quantify and manage its indirect emissions. The airline's commitment to understanding and reducing its Scope 3 emissions is evident through its detailed reporting and targeted reduction strategies, including investments in SAF and efficiency improvements across its value chain.

Key Takeaways

- **Comprehensive Climate Strategy:** JetBlue's efforts to reduce Scope 1, Scope 2, and Scope 3 emissions underscore a holistic approach to sustainability, addressing both direct and indirect sources of GHG emissions.
- **Innovation and Efficiency:** Through fleet modernization, SAF investments, and operational efficiencies, JetBlue is actively working towards reducing its environmental impact, despite the inherent challenges of the airline industry.

- **Scope 3 Emissions Challenge:** JetBlue's detailed reporting on Scope 3 emissions highlights the complexity of addressing indirect emissions. The airline's engagement with its supply chain and investment in sustainable practices exemplify a forward-thinking approach to environmental responsibility.

JetBlue's sustainability efforts reflect a deep commitment to reducing its carbon footprint and contributing to the global fight against climate change. By addressing Scope 1, Scope 2, and Scope 3 emissions with targeted strategies and investments, JetBlue is paving the way for a more sustainable future in aviation.

A.6 Code Repository

The code repository for this thesis project is available on GitHub at the following link: https://github.com/fabrisera/thesis_project. The repository contains the code for data preprocessing, exploratory data analysis, feature engineering, model development, evaluation and the tex files for the thesis document. The data is not included in the repository due to confidentiality reasons, but the code is structured to work with the standard CSV data provided by the CDP.

Bibliography

[1] Global industry classification standard. https://en.wikipedia.org/wiki/Global_Industry_Classification_Standard, 2024. Accessed: February 9, 2024.

[2] Global industry classification standard (gics). https://www.msci.com/our-solutions/indexes/gics, 2024. Accessed: February 9, 2024.

[3] Worldscope fundamentals. https://www.lseg.com/en/data-analytics/financial-data/company-data/fundamentals-data/worldscope-fundamentals#feature-and-benefits, 2024. Accessed: February 9, 2024.

[4] A. Acheampong and E. B. Boateng. Modelling carbon emission intensity: Application of artificial neural network. *Journal of Cleaner Production*, 2019.

[5] Maryam Al-Qahtani and Adel Elgharbawy. The effect of board diversity on disclosure and management of greenhouse gas information: evidence from the united kingdom. *Journal of enterprise information management*, 33(6):1557–1579, 2020.

[6] Moez Ali. *PyCaret: An open source, low-code machine learning library in Python*, April 2020. PyCaret version 1.0.

[7] J. Andrew and C. Cortese. Accounting for climate change and the self-regulation of carbon disclosures. *Accounting Forum*, 35:130 – 138, 2011.

[8] R.H. Baayen, D.J. Davidson, and D.M. Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412, 2008. Special Issue: Emerging Data Analysis.

[9] Douglas Bates. Computational methods for mixed models. 07 2010.

[10] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[11] Walid Ben-Amar, Millicent Chang, and Philip McIlkenny. Board Gender Diversity and Corporate Response to Sustainability Initiatives: Evidence from the Carbon Disclosure Project. *Journal of Business Ethics*, 142(2):369–383, May 2017.

[12] Tara Bernoville. What are scopes 1, 2 and 3 of carbon emissions? https://plana.earth/academy/what-are-scope-1-2-3-emissions, June 2022. Accessed: 2024-01-30.

[13] Carbon Disclosure Project. CDP Scoring 2022: Short Explainer. https://cdn.cdp.net/cdp-production/comfy/cms/files/files/000/006/703/original/Scoring_2022_-_short_explainer.pdf, 2022. Accessed: 2024-01-30.

[14] CDP. The Carbon Majors Database: CDP Carbon Majors Report 2017. https://cdn.cdp.net/cdp-production/cms/reports/documents/000/002/327/original/Carbon-Majors-Report-2017.pdf?1501833772, 2017. Accessed: [Your access date here].

[15] CDP. Cdp - environmental reporting for companies, cities, states, and regions. https://www.cdp.net/en, 2024. [Online; accessed 30-January-2024].

[16] CDP. Guidance for companies. https://www.cdp.net/en/guidance/guidance-for-companies, 2024. [Online; accessed 30-January-2024].

[17] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. ACM, August 2016.

[18] Wikipedia contributors. General motors — wikipedia, the free encyclopedia, 2024. [Online; accessed 1-February-2024].

[19] Julie Cotter and Muftah M. Najah. Institutional investor influence on global climate change disclosure practices. *Australian Journal of Management*, 37:169 – 187, 2012.

[20] S. Davis, N. Lewis, Matthew Shaner, Sonia Aggarwal, D. Arent, I. Azevedo, S. Benson, Thomas H. Bradley, J. Brouwer, Y. Chiang, C. Clack, Armond Cohen, S. Doig, J. Edmonds, P. Fennell, C. Field, B. Hannegan, B. Hodge, M. Hoffert, Eric Ingersoll, P. Jaramillo, K. Lackner, K. Mach, M. Mastrandrea, J. Ogden, P. Peterson, D. Sanchez, D. Sperling, J. Stagner, J. Trancik, Chi-Jen Yang, and K. Caldeira. Net-zero emissions energy systems. *Science*, 360, 2018.

[21] Digital Data Design Institute at Harvard. Climate and sustainability impact lab. https://d3.harvard.edu/labs/climate-and-sustainability-impact-lab/, 2024. [Online; accessed 30-January-2024].

[22] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. Catboost: gradient boosting with categorical features support, 2018.

[23] Benedikt Downar, J. Ernstberger, S. Reichelstein, Sebastian Schwenen, and A. Zaklan. The impact of carbon disclosure mandates on emissions and financial operating performance. *Review of Accounting Studies*, 26:1137 – 1175, 2020.

[24] European Commission. Corporate sustainability reporting, 2023. Accessed: March 22, 2024.

[25] Isabel Gallego-Álvarez, L. Segura, and Jennifer Martínez-Ferrero. Carbon emission reduction: the impact on the financial and operational performance of international companies. *Journal of Cleaner Production*, 103:149–159, 2015.

[26] P. Griffin, D. Lont, and Estelle Sun. The relevance to investors of greenhouse gas emission disclosures. *Capital Markets: Market Efficiency eJournal*, 2012.

[27] A. Hassan, Andrew Wright, and J. Struthers. Carbon disclosure project (cdp) scores and the level of disclosure on climate change related activities: an empirical investigation of the ftse 100 companies. *International Journal of Sustainable Economy*, 5:36–52, 2013.

[28] Zhiyuan He, Danchen Lin, Thomas Lau, and Mike Wu. Gradient boosting machine: A survey, 2019.

[29] T. O. Hodson. Root-mean-square error (rmse) or mean absolute error (mae): when to use them or not. *Geoscientific Model Development*, 15(14):5481–5487, 2022.

[30] Mjimer Imane, Es-Saadia Aoula, and El Hassan Achouyab. Using bayesian ridge regression to predict the overall equipment effectiveness performance. In *2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, pages 1–4, 2022.

[31] Emma Jonson. The complete guide to national climate-related disclosures, 2022. Accessed: March 22, 2024.

[32] Douglas J Lamdin. George serafeim: Purpose + profit: how business can lift up the world. *Bus. Econ.*, October 2023.

[33] P. Larrañaga and C. Bielza. Akaike information criterion. *Dictionary of Bioinformatics and Computational Biology*, 2014.

[34] Shirley Lu, Trang Nguyen, and George Serafeim. Incentive diffusion and decarbonization rates. Early draft – please do not share – comments are welcome, August 2023.

[35] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[36] David J. C. MacKay. Bayesian Interpolation. *Neural Computation*, 4(3):415–447, May 1992. _eprint: https://direct.mit.edu/neco/article-pdf/4/3/415/812340/neco.1992.4.3.415.pdf.

[37] David J. C. MacKay. *Bayesian Non-Linear Modeling for the Prediction Competition*, pages 221–234. Springer Netherlands, Dordrecht, 1996.

[38] A. Maruotti and Pierfrancesco Alaimo Di Loro. Co2 emissions and growth: A bivariate bidimensional mean-variance random effects model. *Environmetrics*, 34, 2023.

[39] McKinsey & Company. A revolutionary tool for cutting emissions, ten years on, 2017. Accessed: 20 March 2024.

[40] Muryani Muryani, K. Nisa', M. A. Esquivias, and S. H. Zulkarnain. Strategies to control industrial emissions: An analytical network process approach in east java, indonesia. *Sustainability*, 2023.

[41] Yuvaraj Natarajan, Gitanjali Wadhwa, K. R. Sri Preethaa, and Anand Paul. Forecasting carbon dioxide emissions of light-duty vehicles with different machine learning algorithms. *Electronics*, 12(10), 2023.

[42] Quyen Nguyen, Ivan Diaz-Rainey, and Duminda Kuruppuarachchi. Predicting corporate carbon footprints for climate finance risk analyses: A machine learning approach. *Energy Economics*, 95:105129, 2021.

[43] Christian Ott, Frank Schiemann, and Thomas Günther. Disentangling the determinants of the response and the publication decisions: The case of the carbon disclosure project. *Journal of Accounting and Public Policy*, 36(1):14–33, 2017.

[44] Yue Pan and Limao Zhang. Data-driven estimation of building energy consumption with multi-source heterogeneous data. *Applied Energy*, 268:114965, 2020.

[45] Judea Pearl. Understanding simpson's paradox. Social Science Research Network, 9 2013. Available at SSRN: https://ssrn.com/abstract=2343788 or http://dx.doi.org/10.2139/ssrn.2343788.

[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[47] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018.

[48] B. Sanderson, B. O'Neill, and C. Tebaldi. What would it take to achieve the paris temperature targets? *Geophysical Research Letters*, 43:7133 – 7142, 2016.

[49] Mohd Saqib. Forecasting covid-19 outbreak progression using hybrid polynomial-bayesian ridge regression model. *Applied Intelligence*, 51(5):2703–2713, 2021.

[50] Femilda Josephin Joseph Shobana Bai. A machine learning approach for carbon di oxide and other emissions characteristics prediction in a low carbon biofuel-hydrogen dual fuel engine. *Fuel*, 341:127578, 2023.

[51] Wei Sun and Mohan Liu. Prediction and analysis of the three major industries and residential consumption co2 emissions based on least squares support vector machine in china. *Journal of Cleaner Production*, 122:144–153, 2016.

[52] Michael E. Tipping. Sparse bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.*, 1:211–244, sep 2001.

[53] Unknown. Lecture 5: Linear mixed models. https://www2.stat.duke.edu/courses/Fall03/sta216/lecture5.pdf, September 2003. Course materials for STA216.

[54] Dr. Aditya Upadhyay. Improving band ratings in carbon disclosure project reports. *International Journal for Research in Applied Science and Engineering Technology*, 2022.

[55] Qiang Wang, Shuyu Li, and Zhanna Pisarenko. Modeling carbon emission trajectory of china, us and india. *Journal of Cleaner Production*, 2020.

[56] Wharton Research Data Services. Wharton research data services. https://wrds-www.wharton.upenn.edu/. Accessed: 2024-2-10.

[57] World Health Organization. Climate change and health. https://www.who.int/news-room/fact-sheets/detail/climate-change-and-health, 2023. Accessed: 2024-03-24.

[58] Huijuan Yang and John F. O'connell. Short-term carbon emissions forecast for aviation industry in shanghai. *Journal of Cleaner Production*, 275:122734, 2020.

[59] Okan Yenigun. Smart aspects of catboost algorithm: Introduction and implementation of yandex's catboost ml model. *Python in Plain English*, Sep 2022.

[60] Baojun Yu, Changming Li, Nawazish Mirza, and Muhammad Umar. Forecasting credit ratings of decarbonized firms: Comparative assessment of machine learning models. *Technological Forecasting and Social Change*, 174(C), 2022.

[61] Luca Zanin and Giampiero Marra. Assessing the functional relationship between co2 emissions and economic development using an additive mixed model approach. *Economic Modelling*, 29(4):1328–1337, 2012.

[62] Huiru Zhao, Guo Huang, and Ning Yan. Forecasting energy-related co 2 emissions employing a novel ssa-lssvm model: Considering structural factors in china. *Energies*, 11:1–21, 2018.

[63] Wenji Zhou, D. McCollum, Oliver Fricko, S. Fujimori, M. Gidden, Fei Guo, T. Hasegawa, Han Huang, D. Huppmann, V. Krey, Chang-Yi Liu, S. Parkinson, K. Riahi, P. Rafaj, W. Schoepp, Fang Yang, and Yuanbing Zhou. Decarbonization pathways and energy investment needs for developing asia in line with 'well below' 2°c. *Climate Policy*, 20:234 – 245, 2020.