TABLE OF CONTENT

Abstract

Scope 3 emissions have a significant impact on climate change due to their indirect nature and broad reach across the value chain. For most organizations, a substantial portion of their carbon emissions are attributed to activities within their supply chain and the post-sale phases, including product processing, usages, and end-of-life treatment, collectively referred to as scope 3 emissions. However, measuring scope 3 emissions is challenging many companies because of its inherent difficulty in calculation, the lack of consistent data, and the limited control over supplier's decision and customer's behaviors. In this paper, we expand the idea of implementing machine learning approach to estimate scope 3 emissions by reporting another well-performing machine learning algorithm, which is Extra Trees regressor. The model uses broadly available data including mandatory scope 1 and scope 2, financial statement data, and industry classification as input features for predicting the scope 3 emissions target. In addition, we measure the computational cost of running models in terms of execution times, power consumption, and equivalent carbon emissions generated. The selected model is the one balancing the trade-off between predictive accuracy and energy efficiency. We also describe the trend of scope 3 emissions generation during the pandemic covid-19 to examine the influence of globally unexpected events on the volume of scope 3 emissions produced. Finally, we predict scope 3 emissions across categories and sectors to evaluate the model performance by examining different prediction metrics and energy consumption coefficients.

Keywords: *Scope 3 emissions, machine learning, sustainability, cross-validation*

Acknowledge

Firstly, we would like to express our deepest gratitude to our supervisor, Alan Saied, for his valuable feedback and guidance throughout our writing process. He was very supportive and encouraged us a lot to complete our thesis from the first meeting date until the very end. This is our honor to have an opportunity to work with him on this project.

Secondly, we would like to thank BI Norwegian Business School for facilitating us during our master's program. BI also supports us in offering accounts to access and collect data for our study. Without this significant support from BI, our thesis might be incomplete.

Finally, we would like to extend our sincere thanks to our family and friends for encouraging and motivating us through our academic journey.

Trang Tran Linh Le

List of abbreviation

AdaBoost	: Adaptive Boosting
CDP	: Carbon Disclosure Project
CSR	: Corporate Social Responsibility
ESG	: Environmental, Social, Governance
GHG	: Greenhouse gases
IMF	: International Monetary Fund
KNN	: K-Nearest Neighbor
LCA	: Life Cycle Analysis
MAR	: Missing at Random
MCAR	: Missing Completely at Random
MNAR	: Missing Not at Random
ML	: Machine Learning
OLS	: Ordinary Least Square
PDF	: Probability Density Function
R^2	: R-squared
RMSE	: Root Mean Squared Error
TRBC	: The Refinitiv Business Classification

Chapter 1: Introduction

1.1. Background

Greenhouse gases (GHG) have always been a concerning problem when the world is facing more serious consequences from climate change. According to the United States Environmental Protection Agency, greenhouse gases are gases that trap heat in the atmosphere, including carbon dioxide (CO_2), methane (CH_4), Nitrous oxide (N_2O), and fluorinated gases. These gas emissions are considered the main driver of global warming leading to severe changes in the global climate. An *Emissions Gap Report 2022* conducted by United Nations Environment Program shows that while the rate of growth in GHG emissions has slowed in the past decade compared to the previous decade, average GHG emissions in the past decade were the highest on record. Specifically, between 2010 and 2019, the average annual growth was 1.1 percent per year, compared to 2.6 percent per year between 2000 and 2009. This decadal slowdown could be explained by a global reduction in new coal capacity additions (especially in China), the worldwide increasing pace of renewable energy deployment, and the steady substitution of coal for gas in the power sector of developed countries (Global Carbon Project, 2022; Lamb et al., 2021; Dhakal et al., 2022). However, between 2010 and 2019, the total global GHG emissions averaged 54.4 gigatons of CO_2 and set a record in 2019 with 56.4 gigatons. The year 2022 was also estimated to reach a similar amount, or even surpass, the 2019 level in the report. There are many causes for rising emissions globally, but one of the most significant ones is the economic development goals of countries around the world in which high demands of burning coal, oil, and gas are required for manufacturing and a huge amount of other energy consumption for operating and developing purposes. In this aspect, firms and corporations are responsible for controlling and adjusting their greenhouse gas emissions.

The GHG Protocol Corporate Standard classifies a company's GHG emissions into three scopes: scope 1, scope 2, and scope 3. Scope 1 emissions are direct emissions from company-owned and controlled resources. In other words, emissions are released into the atmosphere as a direct result of a set of activities, at a firm level. For instance, airlines produce scope 1 emissions by burning jet fuel while flying their own planes. Scope 2 emissions are indirect emissions generated from electricity, heat, steam, and cooling usage in company operations and purchased from an external utility provider. For example, supermarkets produce scope 2 emissions by using refrigeration equipment and light systems. Scope 3 emissions are indirect emissions (not included in scope 2) generated from all other activities using assets not owned by the reporting company, however, these activities are involved in the production and usage of the reporting company's product or service (Bhatia et al., 2011). An example of this is when a company purchases, uses and disposes of products from its suppliers. In fact, scope 3 emissions of one organization are scope 1 and scope 2 of another organization. This scope is often referred to as value chain emissions and is separated into 15 different categories.

Understanding where emissions are sourced from in the first place and being able to measure emissions enables the company to understand its full value chain emissions, and the impacts of its operation on climate change, hence focusing its effort on the greatest emissions reduction opportunities. According to GHG Protocol, reporting on Scope 1 and Scope 2 is mandatory. The reason is these emissions are easy to collect data on since they come from sources owned and directly influenced by the reporting organization. Meanwhile, Scope 3 reporting is optional because of its difficulty in calculation as well as its complexity in data collection. Therefore, there are not many companies reporting Scope 3 except for companies located in countries regulating compulsory Scope 3 reporting such as the United States. Out of 13,000 companies reporting GHG data, only approximately 20% disclosed their Scope 3 emissions for the 2020 fiscal year. Among companies submitting Scope 3 emissions estimation, less than 10% comprehensively and accurately measure it (Boston Consulting Group). One of the main reasons for this is poor data quality. In fact, to make estimations, firms have to rely on data shared by their supply chain partners or third-party data such as statistics released by governments, industry averages, or regulatory disclosures. Therefore, if the data and estimations provided by the supply chain partners are inaccurate, a corporation's Scope 3 emission calculations will be thrown off. In addition, many suppliers do not calculate their emissions, or they are reluctant to share because of confidentiality or contractual issues, preventing purchasing companies from drawing meaningful conclusions. Finally, there is a lack of a standard disclosure format to synchronically calculate Scope 3 correctly. Even though the CDP provides templates for disclosure and the GHG Protocol has guidelines on measuring Scope 3 emissions, there is still considerable variance in the metrics and methodologies that each company uses to measure Scope 3 emissions. Calculating corporate carbon emissions suffers the challenge of established calculation methods (such as environmental input-output, process analysis, or hybrid approaches) that requires intensive data up to the level of emissions sources, activities, raw materials, and emissions factors, most of which are not available publicly (Wiedmann, 2009). However, despite existing challenges, measuring and reporting Scope 3 emissions are strongly encouraged among businesses because of the significant benefits it brings to companies. Firstly, calculating Scope 3 helps firms to identify the largest source of emissions, hence finding opportunities to reduce energy costs. Based on data from CDP, Scope 3 emissions can make up more than 75% of a company's greenhouse gas emissions, and certain industries such as financial services could reach close to 100%. Therefore, tracking Scope 3 emissions could enable a company to figure out vulnerabilities in its value chain to changing regulations, such as carbon pricing, mandatory disclosures, and taxation. Secondly, having Scope 3 well estimated helps companies considerably cut their carbon footprint and gain the trust of customers, meet investor expectations in terms of achieving sustainable development goals, and comply with any regulations for carbon accounting. Realizing the importance of the accurate Scope 3 emissions calculation together with the challenges that companies are facing during the measuring procedure, we are inspired to develop a corporate emissions prediction model to estimate Scope 3 emissions of non-disclosing companies based on a set of multiple variables that are widely available in the financial statements of companies.

In fact, many models have been used to estimate corporate carbon footprints, however, most of those have certain limitations. For example, using carbon projections from commercial data providers such as MSGI ESG Carbon Metrics or Thomson Reuters ESG, the estimation models use a sequence of naïve models that extrapolate emissions from hand-collected energy information, peer group benchmarks, or historical emissions (Q. Nguyen et al., 2021). The target of these models is the aggregated figure, total emissions, scope 1 and scope 2. A sequence of three models generates the final prediction. The purpose of these models is to fill in the data gap for companies not disclosing their carbon emissions in any single year. The advantage of these naïve models is the simplicity of calculation and easy interpretation while the limitation is no attempt to measure the out-of-sample predictive performance. Another approach is to build conventional regression models (such as Ordinary Least Square (OLS) or Gamma Generalized Linear model) as in the case of (Global Carbon Project, 2022; Griffin et al., 2017; CDP, 2016). These models are also easy to interpret, however, they have limitations in restricted samples and are evaluated mainly by in-sample goodness-of-fit (D. K. Nguyen et al., 2021). Therefore, the prediction accuracy of these approaches is not validated in an out-of-sample setting. Recognizing these drawbacks, in recent research, there are gradually more machine learning models developed to predict corporate carbon emissions in general and Scope 3 emissions specifically with significant improvement in accuracy performance. The estimation of greenhouse gas emissions could benefit from the application of machine learning algorithms (Gorge and Gladys 2022). However, the amount of research with regard to applying machine learning algorithms to predict Scope 3 carbon emissions is not much. Therefore, there are still existing gaps for further study. For these reasons, we choose to carry out research in which we develop different machine learning models to optimally solve the problem of Scope 3 emissions estimations of non-disclosing companies.

1.2. Research questions

Within the scope of this paper, the limitations in measuring and reporting scope 3 emissions could be addressed by answering the main following questions:

Question 1: Are scope 3 emissions affected by Covid - 19?

Question 2: Can machine learning methods be used to improve prediction accuracy of scope 3 emissions calculation for non-disclosing firms?

Sub-question 2.1: Does the prediction accuracy improve if the total scope 3 emissions target is classified into 16 categories?

Sub-question 2.1: Does the prediction accuracy improve if the total scope 3 emissions target is broken down into sectors?

***Question 3*:** Is there any trade-off between prediction accuracy and energy consumption across machine learning models?

1.3. Contribution of the thesis

This paper provides an overview of the reporting Scope 3 emissions status by firms and corporations in different countries. From that, this study develops a sustainable machine learning model to solve a sustainable problem, which is predicting Scope 3 emissions for non-disclosing companies based on the available data from financial statements of companies.

Specifically, there are two main objectives in this master thesis. First, we are using a machine learning approach to solve a sustainable problem. In detail, we are figuring out the answer to the question of whether machine learning would be able to predict Scope 3 emissions of non-disclosing firms well. By attaining a well-performing prediction model, companies could take these models as a reference for their estimation of Scope 3 emissions, hence gaining benefits from the accurate measurement of Scope 3 emissions as mentioned above. In addition, we could provide a bigger forecasting picture of how the global carbon emissions trend is going. The other goal is to achieve the sustainability of machine learning models. Specifically, we consider both predictive accuracy and computational complexity (power consumption, execution time, equivalent carbon emissions) when training machine learning models. Sustainable machine learning is known as a subset of sustainable artificial intelligence (AI). The definition of sustainable AI is in its infancy. Sustainable AI is a field of research that applies to the technology of AI (the hardware powering AI, the method to train AI, and the actual processing of data by AI) and the application of AI while addressing issues of AI sustainability and/or sustainable development (Aimee, 2021). Based on this definition, sustainable machine learning could be defined as an efficient and sustainable model that balances its good performance and its resource consumption. The selected model is the one providing the best predictive performance in alignment with a relatively efficient computational cost and low carbon emissions.

1.4. Structure of the thesis

The thesis is structured into 08 chapters. Chapter 1 provides an introduction to the background, research questions, and the contribution of the thesis. In chapter 2, we elaborate on the theory framework for measuring and reporting corporate scope 3 emissions. We also discuss the reality of reporting scope 3 emissions. Chapter 3 presents the literature review that provides an overview of current knowledge, relevant theories, and existing research gaps. In Chapter 4, we explain the methodology used in the paper by specifying main machine learning models and important metrics to evaluate the model performance. Chapter 5 demonstrates in detail data collection, data source, feature selection, data overview, and data processing. The results of the research are thoroughly interpreted in Chapter 6 by reporting the key performance metrics, pointing out the important features, and evaluating the data overfitting of the selected model. Further discussions about the limitations and possibility for future research are conducted in Chapter 7. Finally, conclusions about the main finding, answers to research questions are drawn in Chapter 8.

Chapter 2: Theoretical background

2.1. Scope 3 emissions prediction and GHG protocol

According to climate scientists, global carbon dioxide emissions must be cut by as much as 85 percent below 2000 levels by 2050 to limit the global mean temperature increase to 2 degrees Celsius above pre-industrial levels. Rising temperatures above this threshold will result in increasingly unpredictable and dangerous consequences for people and ecosystems. For this reason, the need to accelerate efforts to reduce GHG emissions is increasingly urgent. As pointed out earlier, Scope 3 emissions account for two thirds of the total GHG emissions of a company. Therefore, by understanding and being able to measure GHG emissions generally or Scope 3 emissions specifically, companies not only can identify opportunities to bolster their bottom line, reduce risks, and discover competitive advantages but also make huge contributions to global environmental protection. To help companies accurately understand and calculate their Scope 3 emissions, the GHG Protocol Corporate Value Chain (Scope 3) Accounting and Reporting Standard was introduced and internationally accepted as a common framework for measuring companies' value chains.

2.1.1 Greenhouse gas protocol

The Greenhouse Gas Protocol (GHG Protocol) is a multi-stakeholder partnership of businesses, governments, nongovernmental organizations (NGOs), and others convened by the World Business Council for Sustainable Development (WBCSD) and the World Resources Institute (WRI). First launched in 1998, the mission of the GHG Protocol is to develop internationally accepted greenhouse gas (GHG) accounting and reporting standards and tools and to promote their adoption in order to achieve a low-emissions economy worldwide. *The GHG Protocol Scope 3 Standard* is a supplement to the *GHG Protocol Corporate Accounting and Reporting Standard, Revised Edition (2004)*. It provides guidance and requirements for organizations to prepare and publicly report a GHG emissions inventory that contains indirect emissions resulting from value chain activities (i.e., Scope 3 emissions). The primary goal of this standard is to provide a standardized step-by-step approach to help organizations understand their full value chain emissions impact to focus company efforts on the greatest GHG reduction opportunities, leading to more sustainable decisions about companies' activities and the products they buy, sell, and produce.

The standard was developed with the following objectives: (1) To help companies develop effective strategies for managing and cutting down their scope 3 emissions through an understanding of value chain emissions and associated risks and opportunities, (2) To help companies prepare a true and fair scope 3 GHG inventory in a cost-effective manner, through the use of standardized approaches and principles, (3) To support consistent and transparent public reporting of corporate value chain emissions according to a standardized set of reporting requirements. Companies of all sizes and in all economic sectors could use this standard for their GHG emissions calculation. Within the scope of this paper, definitions related to greenhouse gases and Scope 3 emissions will be cited based on *The GHG Protocol Scope 3 Standard.*

2.1.2 Scope 3 emissions categories

According to *GHG Protocol Corporate Standard,* a company's emissions are divided into three scopes: scope 1, scope 2, and scope 3. By definition, scope 3 emissions occur from sources owned or controlled by other entities in the value chain such as third-party logistics providers, travel suppliers, employees, customers, material suppliers, or waste management suppliers.

Within scope 3 emissions, it is further divided into 15 different categories. The categories are intended to provide companies with a systematic framework to organize, understand, and report on the diversity of scope 3 activities within a corporate value chain. The categories are designed to be mutually exclusive, such that, for any reporting company, there is no double counting of emissions between categories (*GHG Protocol Corporate Standard).* The following table provides the description of each scope 3 emissions category defined by the *GHG Protocol Corporate Standard.*

Category	**Description**
1. Purchased goods and services	Extraction, production, and transportation of goods and services purchased or acquired by the reporting company in the reporting year, not otherwise included in Categories 2 - 8

2. Capital goods	Extraction, production, and transportation of capital goods purchased or acquired by the reporting company in the reporting year
3. Fuel- and energy-related activities (not included in scope 1 or scope 2)	Extraction, production, and transportation of fuels and energy purchased or acquired by the reporting company in the reporting year, not already accounted for in scope 1 or scope 2, including: a. Upstream emissions of purchased fuels (extraction, production, and transportation of fuels consumed by the reporting company) b. Upstream emissions of purchased electricity (extraction, production, and transportation of fuels consumed in the generation of electricity, steam, heating, and cooling consumed by the reporting company) c. Transmission and distribution (T&D) losses (generation of electricity, steam, heating, and cooling that is consumed (i.e., lost) in a T&D system) – reported by the end-user. d. Generation of purchased electricity that is sold to end users (generation of electricity, steam, heating, and cooling that is purchased by the reporting company and sold to end users) – reported by the utility company or energy retailer only
4. Upstream transportation and distribution	• Transportation and distribution of products purchased by the reporting company in the reporting year between a company’s tier 1 suppliers and its own operations (in vehicles and facilities not owned or controlled by the reporting company) • Transportation and distribution services purchased by the reporting company in the reporting year, including inbound logistics, outbound logistics

	(e.g., of sold products), and transportation and distribution between a company’s own facilities (in vehicles and facilities not owned or controlled by the reporting company)
5. Waste generated in operations	Disposal and treatment of waste generated in the reporting company’s operations in the reporting year (in facilities not owned or controlled by the reporting company)
6. Business travel	Transportation of employees for business-related activities during the reporting year (in vehicles not owned or operated by the reporting company)
7. Employee commuting	Transportation of employees between their homes and their worksites during the reporting year (in vehicles not owned or operated by the reporting company)
8. Upstream leased assets	Operation of assets leased by the reporting company (lessee) in the reporting year and not included in scope 1 and scope 2 – reported by lessee
9. Downstream transportation and distribution	Transportation and distribution of products sold by the reporting company in the reporting year between the reporting company’s operations and the end consumer (if not paid for by the reporting company), including retail and storage (in vehicles and facilities not owned or controlled by the reporting company
10. Processing of sold products	Processing of intermediate products sold in the reporting year by downstream companies (e.g., manufacturers)
11. Use of sold products	The end use of goods and services sold by the reporting company in the reporting year
12. End-of-life treatment of sold products	Waste disposal and treatment of products sold by the reporting company (in the reporting year) at the end of their life

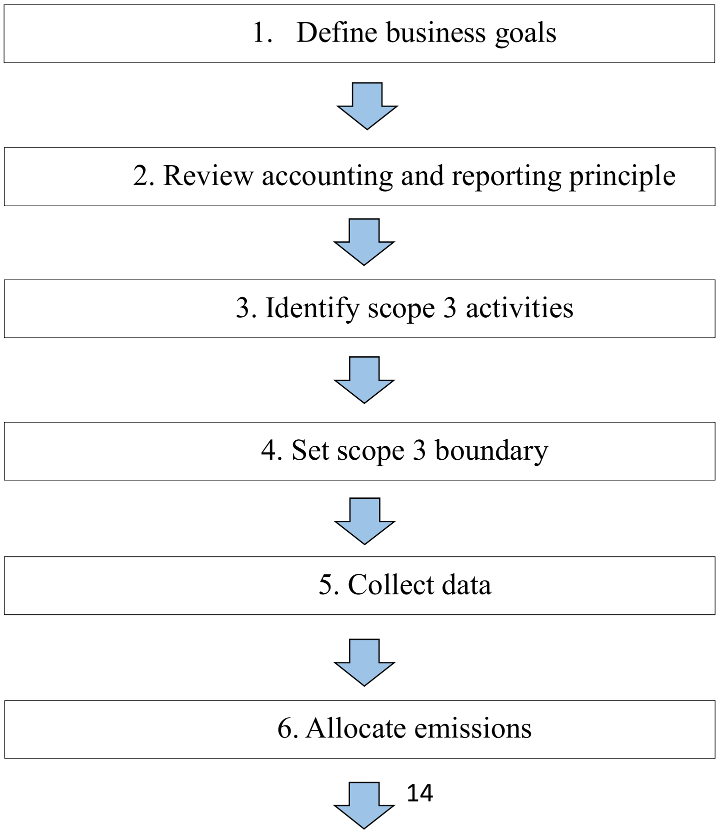
13. Downstream leased assets	Operation of assets owned by the reporting company (lessor) and leased to other entities in the reporting year, not included in scope 1 and scope 2 – reported by lessor
14. Franchises	Operation of franchises in the reporting year, not included in scope 1 and scope 2 – reported by the franchisor
15. Investments	Operation of investments (including equity and debt investments and project finance) in the reporting year, not included in scope 1 or scope 2

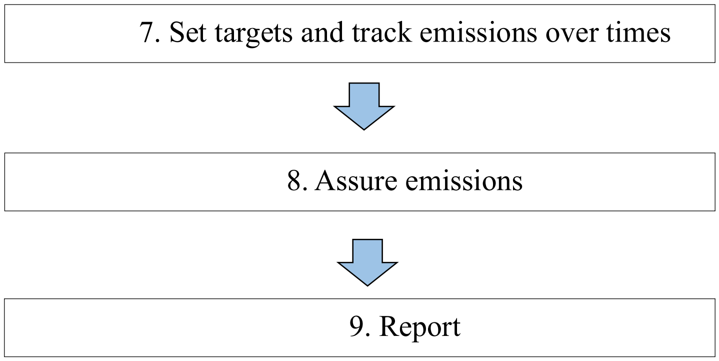
Table 1: Explanation of Categories in Scope 3 Carbon Emissions. Source: GHG Protocol Corporate Standard

2.1.3 Scope 3 accounting and reporting by GHG Protocol

Generally, to have a complete scope 3 accounting and reporting, companies should follow steps that are clearly specified and organized by *the GHG Protocol Corporate Standard*. Accordingly, there are several requirements for each step that companies need to meet to have a true and fair scope 3 emissions calculation.

The following flow chart presents an overview of steps in scope 3 accounting and reporting introduced in *the GHG Protocol Corporate Standard.*





2.1.4 Scope 3 emissions reporting reality

Despite having specific and detailed guidance from GHG Protocol for measuring and reporting scope 3 emissions, there remain several obstacles that deter companies from undertaking this task. The objections to reporting scope 3 emissions primarily revolve around challenges related to data collection and accounting. These include dependencies on industry average data, the lack of primary data, and the possibility of double-counting emissions among reporting organizations. In addition, the inability to control the actions of value chain partners is considered one of the biggest challenges. On the contrary, counterarguments express that scope 3 emissions measurement is too important to be omitted. The paramount importance of calculating Scope 3 emissions is emphasized in understanding climate-related financial risks, preventing organizations from claiming lower emissions and related liabilities by outsourcing carbon-intensive activities (i.e., transferring emissions from Scope 1 or Scope 2 to Scope 3), facilitating actual emissions reductions within the organization’s value chain, and preventing companies from skirting responsibilities to be transparent to their shareholders about their overall risk exposure, which is especially relevant for the industry with a majority of their emissions classified as scope 3 (WRI). However, in spite of data challenges, thousands of companies publicly disclose scope 3 emissions estimates. According to CDP, the number of disclosing companies has increased from 936 companies in 2010 to 3,317 companies in 2021. The following chart presents the number of companies publicly disclosing scope 3 emissions.

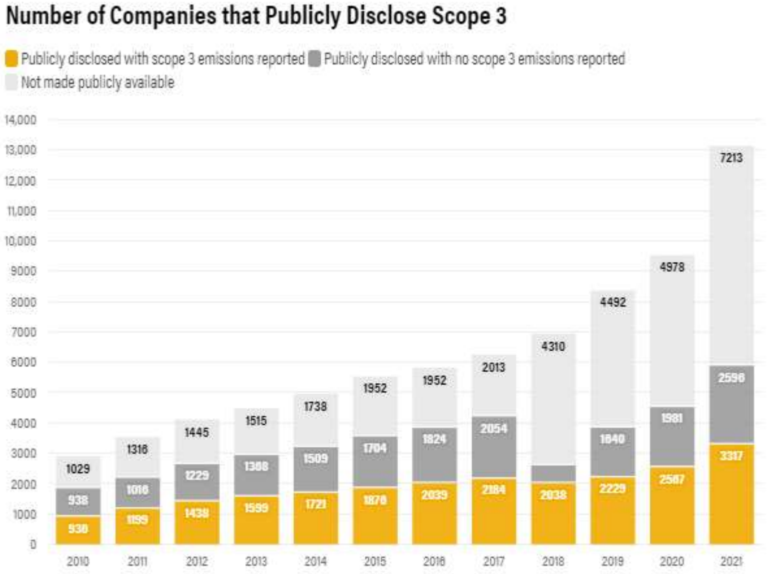


Figure 1: Number of Companies that publicly disclose Scope 3. Source: Data is from CDP. Research and analysis of the data were conducted by Concordia University.

In another research carried out by World Resource Institute (WRI), companies are considered to report scope 3 emissions if they report emissions for one or more of the fifteen scope 3 categories identified in the GHG Protocol Scope 3 Accounting and Reporting Standard. On average, these companies reported emissions for 5-6 scope 3 categories in recent years (WRI).

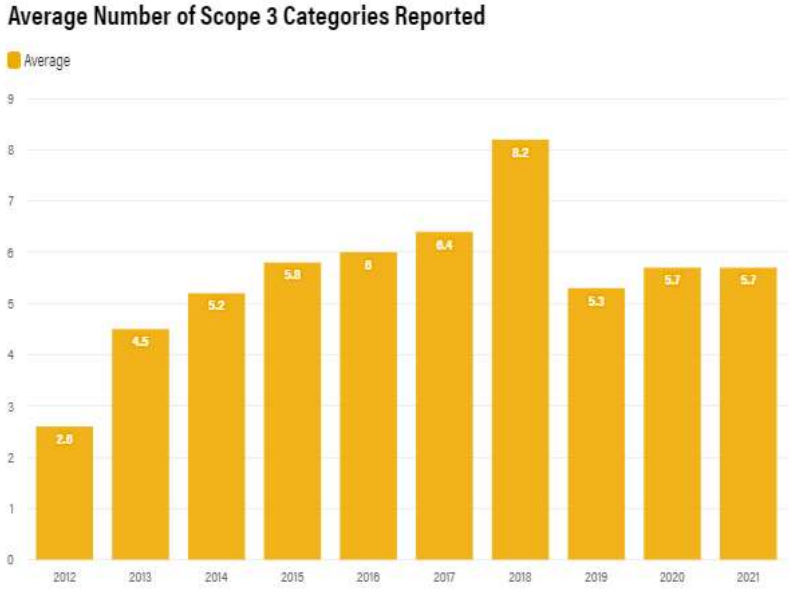


Figure 2: Average Number of Scope 3 Categories Reported. Source: Data is from CDP. Research and analysis of the data were conducted by Concordia University.

The number of companies reporting scope 3 emissions also varies across regions. According to data from CDP, in 2021, 71% percent of European companies and 80% of Australian companies disclosed emissions to CDP. The world average reporting rate is lower because of the high number of non-disclosing companies in China and Brazil. Specifically, companies in China accounted for 14% of disclosing companies and had a scope 3 emissions reporting rate of 27%, and companies in Brazil accounted for 6% of disclosing companies and had a scope 3 reporting rate of 37%.

In most industry sectors, two-thirds of companies or more reported scope 3 emissions, with the highest percentage (84%) of companies reporting scope 3 emissions in the power generation industry. In contrast, the industries with lower scope 3 reporting rates consist of those with supply chains that account for approximately 50% of the global GHG emissions, including freight, food, fashion, as well as electronics and automotive. Requiring these companies to report scope 3 emissions would ensure that companies with carbon-intensive value chains provide more complete information about their exposure to climate-related financial risks.

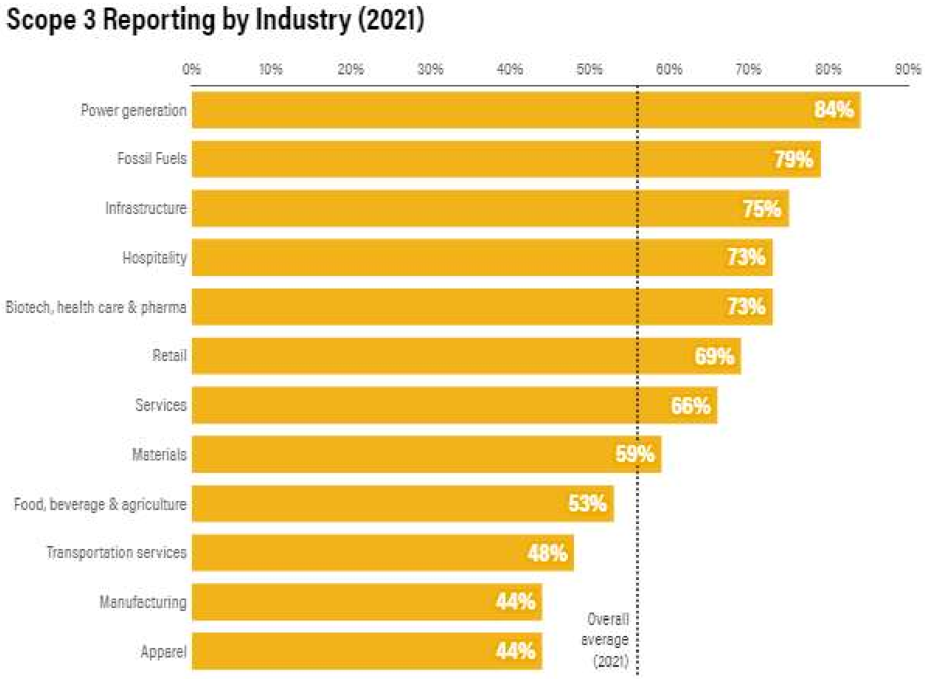


Figure 3: Scope 3 Reported by Industry (2021). ***Source:*** *Data is from CDP. Research and analysis of the data were conducted by Concordia University.*

2.2. Sustainability of machine learning

Machine learning is a rapidly growing field that has the potential to revolutionize the way we live, work, and interact with the world around us. AI community often aims at obtaining “state-of-the-art” results, which typically report accuracy measures but omit any mention of economic, environmental, or social cost (Schwartz et al., 2019). As machine learning models become more complex, requiring larger amounts of data, computing power, energy consumption; their environmental impact and sustainability have become a growing concern. A study by Strubell et al., exhibits that the process of training a single, deep learning, natural language processing (NLP) model (GPU) can lead to approx. 600,000 lb of carbon dioxide emissions (Strubell et al., 2019). This issue has ushered in a novel research direction with the advent of new terms such as "Sustainable AI", "Green AI (Green ML)". Green AI refers to AI research that yields novel results without increasing computational cost, and ideally reducing it (Schwartz et al., 2019). In another study, the sustainability of AI is proposed as a branch of sustainable AI that refers to how to measure carbon footprints, computational power for training algorithms, etc when developing and using AI/ML models (Van Wynsberghe, 2021). Although there is no official definition of sustainability of AI, we consider the term as a broad concept that encompasses social, economic, and environmental aspects. The social perspective includes the impact on individuals and communities, the economic aspect includes the cost and benefits of using the model, and the environmental aspect includes the energy consumption and carbon footprint of the model.

2.2.1 Environmental Sustainability

The environmental sustainability of a machine learning model refers to the extent to which the model's development, deployment, and ongoing use have a minimal negative impact on the environment. There is no standard set of evaluation metrics for assessing the environmental sustainability of machine learning models, as it depends on the specific context and application of the model. Some studies have proposed several factors to consider when assessing the environmental impact of machine learning. These factors could be energy consumption(García-Martín et al., 2019; Getzner et al., 2023), and carbon footprint (CodeCarbon; ML CO2 Impact; Gitzel et al., 2023; Lacoste et al., 2019).

Energy consumption

There are several methods for measuring the energy consumption of machine learning models, including both hardware-based and software-based measurements. Hardware-based measurements involve measuring the energy consumption of the hardware used to run the machine learning model. This can be done using specialized hardware sensors, such as power meters or current sensors, that are attached to the hardware running the model. These sensors can provide accurate measurements of the energy consumption of the hardware, but may not account for other factors, such as the efficiency of the hardware or the energy consumed during data transfer. Software-based measurements, on the other hand, involve measuring the energy consumption of the machine learning model through software profiling tools. These tools monitor the energy consumption of the software running on the hardware and can provide estimates of the energy consumption of the machine learning model. Software-based measurements are often less accurate than hardware-based measurements but can provide a more complete picture of the energy consumption of the model, accounting for factors such as the efficiency of the software and data transfer.

Carbon footprints

Kasper proposed a simple formula for computing carbon footprint, considering two factors: the number of electricity units consumed during some computational procedure (E) and the amount of CO_2e emitted from producing one unit of electricity (C).

$$\text{Carbon footprint} = E \text{ x } C$$

E is quantified as kilowatt-hours (kWh) and C as kg of CO_2e emitted per kilowatt-hour of electricity (CO2/kWh), sometimes referred to as the carbon intensity of electricity. Given this equation, there are several software tools estimating the carbon footprint of some computational procedure by measuring or estimating E and C.

Some web-based tools use key metrics such as training time, energy mix, and hardware information to estimate the electricity consumption € of the training procedure, then carbon emissions. Examples in this category are ML Emissions ML CO2 Impact (Lacoste et al., 2019) and Green Algorithms Tool1 (Lannelongue et al., 2021).

Other tools integrate directly with the ML code to measure carbon footprints through electricity consumption (E) of the GPU, CPU and RAM on which the code is executed. The Python package energyusage2 contains code that can be called for a Python function and passes a given set of parameters. It computes an estimate of energy use as well as carbon emissions based on the energy mix of the region where the code is assumed to be run. CPU power usage is computed using the RAPL

(Running Average Power Limit) interfaces found on Intel processors. Vendor data is used to make an estimate for computations run on the GPU. Another tool using ML code is Codecarbon, evaluating energy use and resulting emissions of code. For the evaluation, two function calls are needed – one that starts the analysis and one that stops it. The advantage of such ML code tools is that this way of coding is probably easier to add to existing scripts.

2.2.2 Economic sustainability

The economic sustainability of machine learning models can be evaluated by analyzing the cost and benefits of using the model. Several studies have proposed methods for estimating the cost of running machine learning models, which includes the cost of computing resources, data storage, and personnel. The benefits of using machine learning models can be measured in terms of the value generated by the model, such as increased efficiency, reduced costs, or improved performance.

The computational and environmental cost of producing a result in machine learning increase linearly with three factors: the cost of executing the model on a single example (E), the size of training dataset (D), the number of hyperparameter experiments (H) (Schwartz et al., 2019).

$$\text{Cost}(R) = E \times D \times H$$

The cost of executing a machine learning model on a single example depends on the complexity of the model itself. More complex models, such as deep neural networks, require a lot of computational resources and can be more time-consuming to execute. Additionally, the amount of data being processed can also impact the computational cost, as larger datasets require more processing power. Relying on more data to improve performance is notoriously expensive because of the diminishing return of adding more data (Sun et al., 2017). Another factor in the equation is the number of hyperparameter experiments, which controls how many times the model is trained during model development. Some models have poured large amounts of computation into tuning hyperparameters or searching over neural architectures such as model descriptions of neural networks (Melis et al., 2017; Zoph & Le, 2016)

Open-source tool Training Cost Calculator (TCC) developed by Aipaca team can estimate the time need to complete the training process for neural networks and the cloud computing costs for various machine learning tasks on different cloud instances by using feature of models, software environments and computing hardware.

2.2.3 Social sustainability

The social sustainability of AI and machine learning models is an emerging area of research. It involves assessing the impact of machine learning models on individuals and communities, including issues such as privacy, bias, and fairness. One critical aspect of social sustainability in machine learning models is privacy. As these models often rely on vast amounts of data, there is a pressing need to ensure that individuals' personal information is handled with care and respect. Müller points out in his research that many AI technologies amplify the issues linked to control of access to data such as face recognition (Müller, 2020). Facial recognition technology has raised concerns regarding privacy infringement, as it can potentially be used for surveillance and tracking purposes without individuals' consent or knowledge. Some researchers propose methods to preserve privacy such as anonymisation, access control, and encryption (Dwork, 2006; Stahl & Wright, 2018). Another vital aspect is the potential for bias in machine learning models. Recent research shows that machine learning algorithms can introduce gender and race discrimination (Buolamwini, J. &amp; Gebru, T., 2018). This raises important ethical and moral considerations regarding the deployment and use of AI systems. To address the issue of bias in machine learning models, researchers and practitioners have been actively working on developing techniques and methodologies to mitigate these biases. One approach involves ensuring that the training datasets used to develop machine learning models are diverse, representative, and inclusive. By incorporating data from various demographic groups and taking into account different perspectives, it becomes possible to reduce biases that may arise due to skewed or limited data samples. In conclusion, the social sustainability of AI and machine learning models is a multifaceted and evolving field of research. Addressing privacy concerns, mitigating bias, and navigating ethical challenges are essential steps towards building AI systems that respect individual rights, promote fairness, and contribute to the well-being of society as a whole.

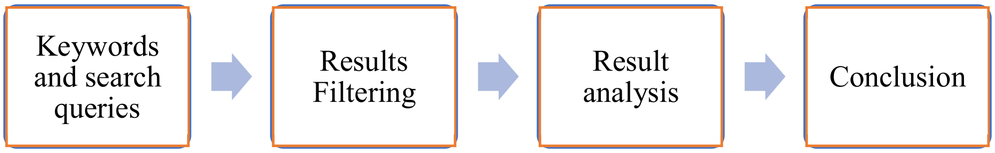
Chapter 3: Literature review

The purpose of this section is to provide a comprehensive overview of current research on the topic, identify research gaps and support the development of research questions in our thesis. With rising global warming as one of the most serious issues in the 21st century, recording carbon emissions has become increasingly important, resulting in a rising number of studies in the field. Research about a Systematic Literature Review in accounting carbon by Marlowe, J., & Clarke, A. identified main research themes, which consist of (i) increasing carbon emissions seen across the globe, (ii) carbon accounting faces a critical amount of measurement uncertainty, (iii) there is a lack of ability to compare reporting results, (iv) there is a need for the implementation of policy and integrated procedures surrounding carbon accounting and (v) there is immense opportunity for the accounting profession to act to make a difference within the field of carbon accounting. The second and third themes have ushered in new methods to tackle the incompleteness and a lack of transparency in carbon accounting, one of which would be building models to evaluate carbon emissions, Moreover, of all carbon estimating issues, carbon disclosure is the most extensively studied. Due to external and internal factors, not all companies or cities disclose their carbon emissions, leading to inadequate data to compare across companies in an industry or to create a benchmark for setting sustainable development goals. Current carbon accounting practices do not appear to be in step with the latest technological developments; therefore, researchers may explore the potential of technology for carbon emissions recording (He, R., Luo, L., Shamsuddin, A., & Tang, Q., 2022).

3.1. Comparson among publications

In order to investigate the current research on carbon emission prediction, we conducted analysis of published literature. We selected appropriate keywords and performed searches on three journal articles databases (Web of science, Google Scholar, ScienceDirect, GreenFILE), then filtered the results to only include relevant ones for analysis. Finally, we drew conclusions based on the results of our analysis.

The process for literature review is below.



3.1.1 Keywords and search queries

The data was obtained through the following filtering query with a focus on the English language:

- “carbon emission*” OR “carbon footprint*”

AND

- “predict*” OR “forecast*”

The first group of keywords relates to carbon dioxide. "Carbon emissions" and "carbon footprints" are not exactly equivalent terms, however, they are related concepts that are often used interchangeably in discussions about environmental impact and sustainability. Both terms relate to the release of greenhouse gases, particularly carbon dioxide (CO2), and are concerned with measuring the impact of human activities on the environment and identifying ways to reduce greenhouse gas emissions. Considering research on the two terms can help us to understand existing methods to measure the amount of CO2 associated with human activities. The second group of keywords contains “predict”, its lemma, and synonyms because our thesis focuses on using machine learning algorithms to estimate the scope 3 carbon emissions of non-disclosing companies rather than capturing emissions.

Moreover, in order to narrow down search results to more relevant and related work, we applied these criteria:

- Exclude publication before the year of 2000
- Publication language: English
- Exclude publications that are too short (less than 2500 words)
- Exclude publications with No concise abstract, summary, or conclusion.

3.1.2 Journal databases

We searched publications in journal databases as below.

Databases	Introduction	Website
Web of Science	A journal articles and citations database covering the leading academic journals in science, social science and humanities. Includes the databases: Arts & Humanities Citation Index, Social Sciences Citation Index. Science Citation Index Expanded.	www.webofscience.com
Google Scholar	Google Scholar is a freely accessible web search engine that indexes the full text or metadata of scholarly literature across an array of publishing formats and disciplines	scholar.google.com
GreenFile	References to journal articles and other publications about environmental concerns and global warming. Links to some fulltext, open access-titles.	web-s-ebscohost-com.ezproxy.library.bi.no
ScienceDirect	Articles within economics, psychology, medicine, science and technology.	www.sciencedirect.com

Table 2: Journal database for comparisons among research

3.1.3 Analysis result

After removing duplications and irrelevant research in journal databases, we end up with 114 publications, covering a wide range of research disciplines, including accounting, business, economics, science, and engineering.

We adopted a citation network to identify influential papers within the research field, as well as track the spread of ideas over time. The graph below shows the relationships between surveyed papers based on how they reference and cite each other. In the network, nodes represent individual papers, and links illustrate citations. Nodes are colored based on year of publication and are sized according to how many citations they have. This means that the larger a node is, the more affecting the corresponding paper is.

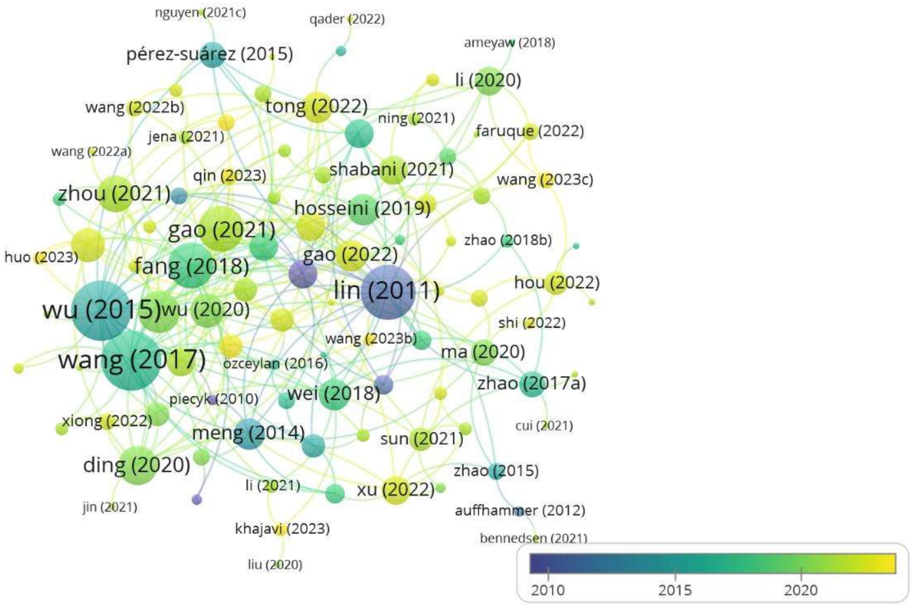


Figure 4: Citation network of research relating to Carbon Emissions Prediction

From the visualization, it is clear that the field of carbon emissions prediction has been gaining momentum over time, with frequent appearance of light-colored nodes representing more recent publications. This trend is not surprising, given the growing concern about climate change and the need to reduce greenhouse gas emissions. Carbon emissions prediction is a crucial area of research, enabling policymakers, businesses, and other stakeholders to anticipate future emissions trends and identify opportunities for reducing emissions. The increasing number of publications in this area also reflects the growing sophistication of modeling techniques used in carbon emissions prediction. As one of the largest nodes, Lin et all., 2011 apply the traditional grey model GM(1,1) with one variable and first order differential to forecast carbon emissions in Taiwan (Lin et al., 2011). In 2015, Wu et all., continue developing and use a novel multi-variable grey model to examine the relationship between energy consumption, urban population, economic growth and CO2 emissions in the BRICS countries (Wu et al., 2015). Wang et all., 2017 introduce a non-linear grey multivariable model, which includes the power exponential term of the relevant variables as exogenous variables (Wang & Ye, 2017). This model has been further improved by other researchers with grey relational analysis (Huang et al., 2019), multivariate grey prediction model using neural networks (Chiu et al., 2020), novel grey rolling prediction model (Zhou et al., 2021), Fractional Grey Prediction Model (Hu et al., 2021). Most of the nodes in the network have connections with three biggest nodes and adopt different versions of grey models. As such, we can draw a quick conclusion that grey modeling is the most commonly used method in carbon emissions prediction and still draw attention from researchers until now. Another branch of research on carbon emissions is applying intelligent algorithms to yield more accurate prediction results. Among surveyed research, Wei et all., (2018) combine random forest and extreme learning machine together for carbon dioxide emission forecasting (Wei et al., 2018) . Citing Wei's paper, Li et all., 2021 use a set of open access data and machine learning methods to estimate and predict city-level CO2 emissions across China (Li & Sun, 2021). Research employing machine learning and hybrid models have thrived recently, as denoted by yellow nodes. In the next section, we will come into details about methodology that researchers often used to predict carbon emissions.

a. Carbon emissions prediction methods

Based on our survey, we divide methods that have been used to predict carbon emissions into 4 categories: conventional statistical models, machine learning, and hybrid models.

Conventional Statistical models

A statistical model conducts modeling and calculation by identifying and quantifying the relationships between variables that influence carbon emissions. Common statistical models used in predicting carbon emissions include regression analysis, time series analysis, and factor analysis.

Kone and Buke (2010) employed trend analysis and regression analyses to forecast energy-related CO2 emissions. Firstly, scholars identify trends in CO2 emissions of the top-25 countries and the world total in the period 1971-2007. These data were regressed against the year using a least squares technique. Statistically significant trends were found in eleven countries namely, India, South Korea, Islamic Republic of Iran, Mexico, Australia, Indonesia, Saudi Arabia, Brazil, South Africa, Taiwan, Turkey and the world total with R2 larger than 0.94. The results obtained from the analyses showed that the models for those countries can be used for CO2 emission projections into the future planning (Köne & Büke, 2010).

Hosseini et all., 2019 forecast Iran's carbon emissions in 2030 under two scenarios: business as usual (BAU) and the Sixth Development Plan (SDP) with multiple linear regression (MLR) and multiple polynomial regression (MPR) analysis. The results show that R2 for both models had been found to be high 0.97; however, RSS for the MPR model was much lower in comparison.

Beside trend analysis, grey prediction models have been increasingly used in recent years for carbon emissions due to their ability to handle incomplete and uncertain data. The advantages of the GM are; the next unknown data can be produced by a few past data and the GM can use a first order differential equation to characterize the unknown system behavior (Huang and Huang, 1997). The simplest version of the grey prediction model is GM(1,1) which uses first-order differential equations to estimate the trend of the system and make predictions. It assumes that the system's future behavior is only affected by its past behavior and the error between the predicted and actual values should be minimized. In our survey, there are many studies adopting GM(1,1) to predict carbon emissions in countries such as Taiwan (Lin et al., 2011), China (Meng et al., 2014) and Turkey (Hamzacebi & Karakurt, 2015).

The main advantage of this approach is its simplicity, and projections are based on past data behaviors. However, the classical statistical model is not very complex and lacks flexibility to handle complex and dynamic systems. Relationships between all factors and CO2 emissions are regarded as linear in econometric methods which is inconsistent with the actual situations, requiring more robust forecasting methods.

Machine learning models

Recently, machine learning (ML) techniques have emerged as a promising approach for predicting carbon emissions. Qader et al., (2022) applied multiple methods, consisting of neural network time series nonlinear autoregressive, Gaussian Process Regression, and Holt's methods for forecasting CO2 emission to forecast the CO2 emission of Bahrain. The research concludes that the neural network model has outperformed other proposed models with RMSE of merely 0.206, compared with 1.0171 and 1.4096 of Gaussian Process Regression Rational Quadratic (GPR-RQ) Model and Holt's method, respectively (Qader et al., 2022).

Q Nguyen et al., use machine learning approach to predict corporate carbon emissions. The scholars applied a two-step framework that uses a meta-learner (Elastic Net) to aggregate predictions from six machine learning models (OLS, Elastic Net, Neural Network, K-Nearest Neighbours, Random Forest, Extreme Gradient Boosting) and prove that this approach outperforms existing conventional statistical models with a decrease in MAE by 25-35% (Q. Nguyen et al., 2021). However, the research uses carbon emissions data as estimated results from a third-party rather than data reported from companies. As such, results from the research appear to be sensitive to errors of original estimated models, as well as depend on the quality of estimated dataset.

Hybrid models

Nevertheless, a single intelligent algorithm may have limitations, many researchers use the hybrid model (two or more models combined) to forecast carbon emissions. In one study, W Sun and J. Sun (2017) proposes a novel hybrid model that combined principal component analysis (PCA) with regularized extreme learning machine (RELM) to make CO2 emissions prediction based on the data from 1978 to 2014 in China. Several conclusions can be obtained as follows: (a) the PCA process is conducive to improving the operation speed and forecasting accuracy; (b) the high prediction precision of RELM model is attributed to the introduction of the regularization part which enhances the global optimization and generalization ability with little time cost. (c) RELM combined with PCA outperforms other models with the lowest MAPE, MaxAPE, MdAPE, and RMSE, indicating that PC-RELM model is a promising technique for CO2 emission prediction.

J Zhou et al., propose a novel GM model and BPNN hybrid forecasting methodology to predict CO2 emissions, using China’s sample data from 1980 to 2015 (Zhou et al., 2021). While the GM model is used to address the problem of small sample datasets, BPNN is applied to handle the non-linear and non-stationary data of CO2 emissions. The forecasting performance of the single GM model and BPNN is evaluated to demonstrate the higher prediction precision of the proposed GMMN. Results from research indicate that the hybrid model produces the lowest MAE, MAPE, and RMSE, and the highest R2, thereby better capacity for forecasting CO2 emissions and capturing the non-linear and non-stationary characteristics of CO2 emissions.

b. Predictor variables

In many studies, a theoretical framework called STIRPAT is used to analyze the factors that contribute to environmental impact. STIRPAT stands for "Stochastic Impacts by Regression on Population, Affluence, and Technology.". In a classical version of the model, environmental impact is the product of population (P), affluence (A), and technology (T). The STIRPAT model suggests that as people become more affluent, they tend to consume more resources and generate more waste, which leads to increased environmental impact. At the same time, technological advancements can increase the efficiency of resource use, but can also contribute to environmental degradation if not properly managed. The model can also take the logarithm on both sides of the equation as follows:

$\ln I_i = a + b \times \ln P_i + c \times \ln A_i + d \times \ln T_i + e$

Applying this framework, Liu, Z et al., (2020) add 12 variables into regression model: permanent resident population (PRP), GDP, urban road mileage (RM, km), passenger capacity of public transportation (PCPT), industrial structure (IS, %), household consumption level (HCL), total electricity consumption (TEC), urban and rural household electricity consumption (HEC), per capita disposable income (PCDI, yuan), Engel coefficient (E), population of employees at the end of the year (PE), and the proportion of employees for the three main industry categories: primary industry, secondary industry, and tertiary industry (EI). Results of regression indicate that EC presents a low correlation value, whereas the other factors have high correlation values. The confidence level is higher than 99%, proving that the growth of carbon emissions is related to these factors (Liu et al., 2020). Using the same framework, in research of Zhao, Huiru, Guo Huang, and Ning Yan. (2018), economic structure, energy structure, urbanization rate and energy intensity are taken into consideration as the driving factors of CO2 emissions in China (Zhao et al., 2018). It is worth noticing that some models employ direct predictors, such as energy consumption, electricity consumption, or historical emission patterns. Such information is often not available at the firm level in a universal context.

3.2. Summary and research gaps

Predicting carbon emissions is a complex and challenging task that requires the use of sophisticated methods and the consideration of multiple factors. While conventional statistical models have been widely used in predicting carbon emissions, machine learning and hybrid models offer promising alternatives that can account for nonlinear relationships and interactions between variables. However, improving the accuracy and reliability of carbon emissions prediction requires addressing the challenges of data availability and model complexity, particularly at the company-level where these issues are more pronounced.

Based on our survey, the majority of research on the journal databases was conducted at the macro level such as for an industry or a country, taking social and economic factors into consideration when building models to predict carbon emissions. These models also employ direct predictors, such as energy consumption, electricity consumption, or historical emission patterns, which are often not available at the firm level in a universal context. The literature on methods of estimating corporate carbon emissions remains fairly new. Out of 114 papers in our survey, only 2 researchers predicted carbon emissions at the corporate level.

Another common feature of research on carbon emissions predictions is that scholars identified optimal solutions by looking at accuracy of the models such as R2, Mean Absolute Percentage Error (MAPE), etc. While accuracy metrics are certainly valuable in evaluating the effectiveness of models, they do not account for the computational complexity required to generate those predictions. Intelligent algorithms, such as machine learning and artificial neural networks, have become increasingly popular in carbon emissions prediction research due to their ability to analyze complex data sets and identify hidden patterns. However, these algorithms can be computationally expensive to run, requiring significant amounts of time and processing power to generate predictions. As carbon emissions prediction research continues to grow in importance, it is essential to consider both accuracy and computational complexity when evaluating the effectiveness of these models. This will help researchers to identify not only the most accurate predictions but also those that can be generated quickly and efficiently, allowing for more timely and effective interventions to mitigate climate change.

From the analysis of previous studies, we identify research gaps in predicting emissions at the corporate level. Hence, we place our focus on doing research to fill the gap. The GHG Protocol (WRI and WBCSD, 2020) divides carbon emissions into three categories: Scope 1, Scope 2, and Scope 3, forming a basis for mandatory reporting in many countries. While Scope 1 and 2 emissions are parts of mandatory greenhouse gas emission reporting in some countries, firms have the discretion on disclosing their Scope 3 emissions. For many businesses, especially in the energy sector, Scope 3 emissions account for more than 70 percent of their carbon footprint (Hertwich & Wood, 2018). However, the analyses of firm-level emissions are usually limited to Scope 1 and Scope 2 emissions (Boermans & Galema, 2019; Global Carbon Project, 2022; Griffin et al., 2017). For example, Han et al., 2021 estimated the emissions of non-disclosure companies by modeling the Scope 1 and Scope 2 emissions separately, using Gamma GLM, Gradient Boosting Decision Trees for Amortized Inference, Recalibration using Normalizing Flows, and Patterned Dropout for regularization (Han et al., 2021). A two-step framework that applies a Meta-Elastic Net learner to combine predictions from multiple base learners was introduced in the research of Q. Nguyen et al., 2021 to predict corporate carbon emissions for risk analyses (Q. Nguyen et al., 2021). In terms of Scope 3 emissions, some data providers estimate the numbers by employing process-based life cycle assessment (Carbon4Finance), and multi-variable regression models using metrics at the firm level (CDP, 2020). Unfortunately, there is no proof of the prediction ability of these estimating models. Therefore, we would like to develop regression machine learning models to leverage Scope 3 emissions of disclosure firms to predict unreported numbers of other firms, considering a set of available variables in financial statements or GHG emission reports.

Chapter 4: Methodology

4.1. Prediction model

To have an overview of how well each model performs and a relative comparison across models, we implement both distance-based models and tree-based models for predicting the target variables.

a. Distance-based models

Ordinary Least Squares Regression (OLS)

In this paper, the OLS regression model is used to explain the correlation or relationship between the target variable and the predictive ones. In addition, it also performs as a baseline model to evaluate the linear approximation of the target variable predicted by input features. Specifically, a linear regression fits a linear model between target variable and input features to minimize the sum of the squares in the difference between the observed and the predicted values of the dependent variable configured as a straight line. Our target variables include total scope 3 emissions and 16 scope 3 emissions categories. We use all input features as dependent variables for the OLS regression baseline model.

The OLS baseline model has the following form:

$$y_i = \beta_1 \, x_{i1} + \beta_2 \, x_{i2} + \cdots + \beta_p \, x_{ip} + \varepsilon_i,$$

Where y_i is the target variables, x_i is predictor.

K-nearest neighbor (KNN)

KNN regressor is a non-parametric algorithm that calculates predictions based on a measure of similarity defined by the minimal distance between samples (Pedregosa et al., 2011). In an intuitive manner, KNN approximates the association between independent variables and the continuous target by averaging the observations in the same neighborhood. The distance between samples in KNN is commonly calculated by Euclidean. In Python model library, the default distance metric is Minkowski and could be calculated as follows:

$$d_{a,b} = \left(\sum_{i=1}^{k} (|x_{ai} - y_{bi}|)^q \right)^{1/q}$$

Where q = 1 for the Manhattan distance and q = 2 for the Euclidean distance.

b. Tree – based models

Ensemble learning: The process of combining the output of multiple individual models is called ensemble learning. The idea behind this learning method is to let groups of weak learners come together to form a strong learner.

Bagging is a technique to decrease the variance in the prediction by generating additional data for training from a dataset using combinations with repetitions to produce multi-sets of the original data.

Boosting is an iterative technique that adjusts the weight of an observation based on the last classification. If an observation is classified incorrectly, the weight of this observation will be increased for the next classification.

Random Forest Regressor

Random Forest Regressor is a supervised machine learning algorithm and bagging technique that uses an ensemble learning method for regression. According to scikit-learn documentation, a random forest is a meta estimator that fits a specific number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control overfitting. The trees in random forest run in parallel, meaning there is no interaction between these trees while building the trees. Compared to the Decision Trees algorithm, the Random Forest Regressor reduces the variance and overfitting by introducing randomization into the construction and building an ensemble taking the average of prediction of developed trees.

Extra Tree Regression

Extra Tree Regression is almost similar to Random Forest Regression, in which a number of random decision trees (also known as extra-trees) are fitted on various sub-samples of the dataset and then the final decision is obtained by taking into account the arithmetic mean of every tree. Both two methods have the same developing tree procedure and the same random selection process of subsets of features. However, there are two main differences between the two algorithms. Firstly, Random Forest subsamples the input data with replacement, which is known as bootstrap replicas, while Extra Tree uses the whole original samples. Specifically, the extra trees algorithm creates many decision trees, but the sampling for each tree is random without replacement. Thus, this creates a dataset for each tree with unique samples. A specific number of features are also selected randomly for each tree from the total set of original features. The second difference is the selection of cut points to split nodes. While Random Forest chooses the optimal split, Extra Trees chooses it randomly. In detail, instead of calculating a locally optimal value using Gini or Entropy to split data, the Extra Trees algorithm randomly selects a split value. This makes the trees uncorrelated and diversified. Nevertheless, once the split points are chosen, the two models select the best one between all the subsets of features. Therefore, Extra Trees adds randomization but still has optimization. Extra Trees is expected to reduce both bias and variance in data evaluations.

Adaptive Boosting Regressor (AdaBoost)

AdaBoost is an ensemble learning method. According to scikit-learn definition, an AdaBoost regressor is a meta-estimator that begins by fitting a regressor on the original dataset and then fits additional copies of the regressor on the same dataset but where the weights of instances are adjusted according to the error of the current prediction. As such, subsequent regressors focus more on difficult cases. Specifically, after the first model is created, the second model learns from the mistaken classification of the first model. The data points which are mistakenly predicted by the first model are given higher weights. The correctly predicted points can also have their weights decrease. The second model is built by paying more attention to the data points with higher weight. AdaBoost could be more susceptible to outliers as it assigns high weights to misclassified samples.

Gradient Boosting Regressor

Gradient Boosting Regressor is also an ensemble learning method, but unlike AdaBoost, Gradient Boosting Regressor learns from the residual errors, which are the difference between the actual values and the predicted values, from the previous model to minimize it. In other words, this algorithm builds models sequentially and these subsequent models try to reduce the loss function such as mean squared error or cross-entropy of the previous model using gradient descent. In each iteration, the algorithm computes the gradient of the loss function with respect to the predictions of the current ensemble and then trains a new weak model to minimize this gradient. The predictions of the new model are then added to the ensemble, and the process is repeated until a stopping criterion is fulfilled. Gradient Boosting Regressor is less sensitive to outliers because it updates the weights based on the gradients.

4.2. Performance metrics

To assess the models, we report two distinct predictive accuracy metrics including R-squared and Root Mean Squared Error (RMSE), and three energy consumption metrics including execution time, power consumption, and equivalent carbon emissions. For the selection of the best model, we create a new metric called *energy efficiency* by taking the power consumption divided by RMSE. In addition, to increase the robustness of our selected machine learning model, we apply the k-fold cross-validation technique to train the model.

4.2.1 Machine learning predictive evaluation metrics

- R-squared

R-squared (R^2 or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by independent variables. In other words, the R^2 regression score evaluates the goodness of fit of the model where the best possible score is 1.0 (Pedregosa et al., 2011). The higher the R^2, the more variation in reported target values is explained by the input features.

Given that the variance is dataset-independent, R^2 is not comparable between different datasets or across different targets. To get R^2 from the machine learning model, Python scikit-learn library is implemented and calculated by the following formula:

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

Where the i-th sample model prediction is $\hat{yi}$; yi is the corresponding reported value and

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

R-squared evaluates the scatter of the data points around the fitted regression line and explains the extent to which the variance of one variable explains the variance of the second variable. Therefore, R-squared allows us to know which input features explain a certain percentage of observed variation. However, R-squared has limitations. Firstly, R^2 keeps increasing when more predictors are added to the model no matter how important the features could be. As a consequence, a model with more input features may appear to have a better fit not because it actually is but because it has more explanatory variables. Secondly, if a model has too many predictive variables and higher-order polynomials, it starts to model the random noise in the data, leading to overfitting and misleading high R^2 values and weakening the ability of the model to make predictions.

- Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) measures the average difference between a statistical model's predicted values and the actual values. Mathematically, it is the standard deviation of the residuals, in which the residuals represent the distance between the regression line and the data points. In other words, RMSE quantifies how dispersed these residuals are, releasing how tightly observed data clusters around the predicted values. The lower the RMSE, the better the model and its prediction. A higher RMSE indicates that there is a large deviation from the residuals to the ground truth. RMSE is calculated by taking the root square of Mean Squared Error retrieved by Python scikit-learn library package. The formula for RMSE calculation is as follows:

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

Where n is the number of observations, $\hat{y}_i$ is the predicted value for the i^{th} observation, y_i is the actual value for the i^{th} observation.

In terms of advantages, RMSE is computationally simple and easily differentiable. It serves as a heuristic for training models and is easy to understand. Compared to Mean Squared Error (MSE), RMSE penalizes less errors due to the square root. However, like MSE, RMSE is dependent on the scale of the data, leading to an increase in magnitude if the scale of the error increases. Moreover, RMSE is highly sensitive to outliers, therefore, the outliers must be replaced or removed for it to function properly. Lastly, RMSE increases when the size of the test dataset increases, causing an issue when calculating the results on different test samples.

- K-fold cross-validation

K-fold cross-validation is a resampling technique used to evaluate the performance of machine learning models as an estimate of generalizability or expected model performance on unseen data (Pedregosa et al., 2011). The procedure has a single parameter called k referring to the number of groups that a given dataset is split into. A k-fold cross-validation process consists of the following steps. Firstly, a value of k is selected. Common values are k = 3, k = 5, and k = 10. However, the most popular value used in applied machine learning to evaluate models is k = 10 because it is proved to provide a good trade-off between the low bias and low computation cost in an estimate of model performance. Once k is chosen, the given dataset is split into k non-overlapping folds. K-1 folds are used as a training set while the remaining set is used as a validation or test set. The model is trained on the training dataset and validated on the test dataset. The validation score of each iteration is saved. The training and validation steps are repeated until all iterations are completed. Finally, validation scores of all iterations are averaged giving the final score. In other words, the performance ability of models reported by k-fold cross validation is the average of the scores computed in the k-fold loop.

- Feature importance

In the field of machine learning, feature importance scores are used to assess the relative importance of each input feature in a dataset during the construction of a predictive model. In other words, it measures the contribution of variables toward predicting the target, helping identify which features have the most significant impact on the model's prediction. Feature importance can be calculated using various techniques, such as the coefficient magnitude in linear regression models or the Gini importance for Extra Trees algorithm. Feature importance is useful in machine learning task as it enables practitioners to discern the primary contribution among the features in a dataset that substantially influences the final prediction, while also identifying the relatively less significant features. Specifically, feature importance improves the feature selection process, diminishes dimensionality, and minimizes noise in the data. In addition, it enhances model interpretability and model performance by reducing training time and mitigating overfitting issues.

4.2.2 Energy consumption

To measure the execution time, power consumption, and equivalent carbon emissions of models, we use a Python package called *eco2Ai.* The eco2Ai is an open-source library capable of tracking equivalent carbon emissions while training or inferring machine learning models for energy consumption of CPU, GPU, RAM devices (Budennyy et al., 2022). Thanks to this capability, the eco2Ai library can monitor the power consumption and carbon footprint of training model in real time and supports to demonstrate and implement various memory and power optimization algorithms. In Python, eco2Ai could be imported from the available library.

Specifically, in terms of energy consumption, it is measured in *kilowatt-hours* (kWh). We focus on CPU, GPU, and RAM energy evaluation because of their direct and most significant impact on the machine learning process. Regarding equivalent carbon emissions, the total equivalent emissions value of *carbon footprint* (CF) generated during machine learning model learning is defined by multiplication of the total power consumption from CPU, GPU, and RAM by the *emission intensity coefficient* γ (kg/kWh) and the PUE coefficient (Budennyy et al., 2022).

$$CF = \gamma PUE(E_{\mathrm{CPU}} + E_{\mathrm{GPU}} + E_{\mathrm{RAM}}).$$

Where *emission intensity coefficient* γ is defined by the regional energy consumption, and PUE is the power usage effectiveness of data center when training is done on cloud. PUE is an optional parameter with a default value of 1 and it is defined manually in the eco2AI library (Budennyy et al., 2022)

Within the scope of this paper, we measure energy consumption metrics on the following device specifications:

- CPU: 11th Gen Intel(R) Core (TM) i5-1135G7 @ 2.40GHz   2.42 GHz
- RAM: 16.0 GB
- System type: 64-bit operating system, x64-based processor
- Operating system: Windows 10 Enterprise

Equivalent carbon emissions are measured based on the carbon intensity of Oslo, Norway, 2023. The energy measurement could vary across devices and locations because of the difference in carbon intensity and CPU specifications.

Energy efficiency

This metric is created by taking the power consumption divided by RMSE of the model. It explains how much energy is consumed to reduce one unit of RMSE. In other words, the efficient prediction metric tells us about the trade-off between the predictive accuracy of the model and the total power consumed to train it. The formula to calculate the energy efficiency is as follows:

$$Energy\ efficiency\ = \frac{Power\ consumption}{RMSE}$$

Chapter 5: Data

5.1. Data collection

5.1.1 Data sources and data license

Data sources

For the purpose of predicting unreported scope 3 corporate emissions, we utilize publicly available data that are accessible to a wide range of users, including researchers, analysts, and policymakers. First, we attain a list of companies joining carbon disclosure projects in the period of 2015-2022 from CDP website. Companies that do not satisfy our predefined conditions are eliminated from the list, including:

- Companies are not given a ESG score by CDP as they do not report sufficient emissions information in CDP's questionnaires.
- Companies are not listed on the stock exchange as it is difficult for us to collect their financial data.

Then we search company names on Bloomberg to get companies' information and financial data. However, in many data retrieval requests, Bloomberg provides incomplete or unavailable data. We supplement this data by employing an alternative source, Refinitiv Eikon. To ensure cross-reference, we use company tickers obtained from Bloomberg to filter corporate information in Refinitiv Eikon. Features we used in this thesis can be divided into four groups as shown in the table below:

Data categories	**Sources**	**Description**
Target	Bloomberg, Refinitiv Eikon	Scope 3_total, 16 scope 3 categories
Country information	International Monetary Fund	Country, GDP, Population, Carbon tax, Carbon Intensity
Financial data	Bloomberg, Refinitiv Eikon	Total Assets, Capital Expenditures, Operating Expenses, SGA expense, Cost of Goods & Industrials Sold, Inventories, Revenue,

		Property Plant & Equipment Net, Inventories, Revenue, Asset Turnover, Inventory Turnover, Number of employees
Industry classification	Refinitiv Eikon	The Refinitiv Business Classifications (TRBC) is the global, comprehensive, industry classification system owned and operated by Refinitiv. It divided companies into 13 economic sectors: Energy, Basic Materials, Industrials, Consumer Cyclicals, Consumer Non-Cyclicals, Financials, Healthcare, Technology, Utilities, Real Estate, Associations & Organizations, Government Activity, Academic & Educational Services.
ESG metrics	Bloomberg, Refinitiv Eikon	CRS/Sustainability committee, ESG Disclosure score, ESG News Sentiment ES Positive, Emission scope 1, Emission scope 2

Table 3: Data sources

Data dictionary can be found on appendix of this thesis.

Data license

For data from Bloomberg and Refinitiv Eikon, we get access to the platforms and collect data by using BI student accounts. Our data collection from the International Monetary Fund (IMF) website is permitted and aligns with their terms and conditions of usage. Further details regarding data access, copyright, and usage of the IMF are found on their website: https://www.imf.org/external/terms.htm.

5.1.2 Feature selection explanation

To provide accurate predictions about unreported scope 3 corporate emissions, the selection of predictors plays a crucial role. We carefully choose a set of predictors based on extensive research and industry knowledge. In this section, we explain the reasons why we use proposed features.

a. Country information

Information relating to the country where companies operate can have a significant impact on corporate Scope 3. Firstly, countries have varying regulations and policies related to emissions, energy efficiency, and other environmental aspects, which can directly influence the emissions generated by a company's operations and its supply chain. Companies operating in countries with stricter emission standards may need to invest in cleaner technologies, adopt renewable energy sources, or implement emission reduction measures, thereby affecting their Scope 3 emissions. Among environmental regulations, ***carbon tax*** is a fiscal policy tool implemented by governments to encourage businesses and individuals to reduce their carbon emissions by making them financially responsible for the pollution they generate. Secondly, economic factors of a country such as ***GDP*** provide insights into the scale of economic activity, energy consumption, transportation infrastructure. Generally, countries with higher GDP tend to have more industrial activities and a larger commercial sector, which can lead to increased emissions. Companies operating within these countries may have larger production volumes, higher energy consumption, and more extensive supply chains, resulting in higher Scope 3 emissions. In addition, infrastructure and transportation systems available in a country can affect the efficiency and emissions associated with transportation within the supply chain. Countries with high GDP appear to have well-developed transportation networks, including efficient logistics, public transportation options, and alternative fuel infrastructure, which can help companies to reduce emissions from transportation activities. Thirdly, ***Population size*** is closely linked to consumption patterns. Larger populations typically have higher levels of consumption, causing an increased demand for goods and services. Companies located in countries with larger populations may experience greater Scope 3 emissions due to the scale of production and consumption.

b. Financial information

Financial metrics provide insights into various aspects of a company's operations, expenditures, and revenue generation, which have certain effects on corporate Scope 3 emissions. ***The number of employees*** *directly* impacts the emissions related to employee commuting and business travel. If a company has a large workforce, it means more employees having business trips and traveling to/from the workplace. Commuting emissions can be significant, especially if employees rely on personal vehicles or use high-emission modes of transportation. Business travel emissions, including air travel, can be substantial and contribute to Scope 3 emissions. Moreover, the number of employees can *indirectly* impact

Scope 3 emissions through supply chain management and procurement decisions. Companies with a larger workforce often require more resources and goods, which can increase emissions upstream in the supply chain. The production, transportation, and disposal of products and materials needed to support a larger workforce can contribute to Scope 3 emissions.

Capital expenditures relating to expansion or increased production capacity can have a direct impact on Scope 3 emissions. Scaling up operations typically leads to higher energy consumption, transportation requirements, and related emissions. When predicting Scope 3 emissions, the anticipated increase in production volume resulting from capital expenditures must be factored in to ensure accurate estimations. ***Property, Plant & Equipment*** are often utilized in manufacturing and production processes. For example, machinery and equipment used in industrial operations may consume energy and emit greenhouse gases during operation. The emissions resulting from these processes would be classified as Scope 1 emissions. However, the production activities enabled by the PP&E assets may also generate indirect emissions along the value chain, which fall under Scope 3 emissions. These can include emissions from the extraction and processing of raw materials used in the production process, as well as transportation emissions from the delivery of goods to customers. PP&E assets also impact Scope 3 emissions through their eventual disposal or decommissioning. When PP&E reaches the end of its useful life, the disposal process can generate emissions, particularly if not handled in an environmentally sustainable manner.

COGS, operating expenses and SGA expenses often include logistics cost for shipping products, delivering goods to customers, or distributing materials. The emissions generated by transportation activities, including the burning of fossil fuels by trucks, ships, or planes, contribute to a company's Scope 3 emissions. SG&A expenses can include costs related to employee salaries, benefits, travel, or payments to suppliers for goods and services. These expenses give insights into transportation-related or operation emissions.

Revenue represents the income generated by a company from its core operations. It is often correlated with the volume of goods sold, services provided, or the scale of business activities. Higher revenue generally indicates larger business operations, which can result in increased Scope 3 emissions. For example, if a company sells products that have a high carbon footprint or relies on energy-intensive processes, its revenue growth may be in line with a corresponding increase in Scope 3 emissions. In some cases, the revenue generated by a company can serve as a proxy for the usage or consumption of its products or services by customers. Depending on the nature of the business, the use of its products or services can have significant environmental implications. For instance, companies in the energy, transportation, or consumer goods sectors may generate Scope 3 emissions through customer usages, such as emissions from fuel combustion or energy consumption. ***Total Assets*** variable provides an indication of the company's scale of operations. Larger companies with extensive assets and operations tend to have a broader supply chain and customer base, which can result in higher Scope 3 emissions. ***Inventories*** reflect the flow of goods through a company's supply chain. The emissions from storing, and distributing inventories can contribute to Scope 3 emissions. This includes emissions from transportation vehicles, such as trucks or ships, as well as emissions from warehousing and logistics operations.

c. Industry classification

The industry classification of a company can significantly affect its Scope 3 emissions due to the specific characteristics, operations, and value chains of compaies in the same industries. Industry classification relates to corporate scope 3 emission in the following aspects:

Emissions intensity: Different industries have varying levels of emissions intensity based on their production processes, energy consumption, and use of raw materials. For example, industries such as manufacturing, mining, and power generation tend to have higher emissions intensity compared to sectors like information technology or financial services. The industry classification can serve as an indicator of the potential emission sources and intensity, providing a basis for estimating Scope 3 emissions.

Supply chain complexity: Industries differ in terms of the complexity and length of their supply chains. Industries with complex supply chains, such as retail, consumer goods, or automotive, tend to have a higher potential for Scope 3 emissions due to the involvement of multiple suppliers, transportation networks, and distribution channels.

Regulatory and market drivers: Industries are subject to specific regulations and market dynamics that can influence their emissions profile. For instance, industries with high carbon emissions may face stricter regulatory requirements or carbon pricing mechanisms that incentivize emission reductions.

d. ESG metrics

ESG metrics such as CSR/Sustainability committees, ESG Disclosure scores, ESG News Sentiment, Emission Scopes 1 and 2 are relevant factors that can indirectly relate to Scope 3 emissions.

A CSR (Corporate Social Responsibility) or Sustainability committee within a company is responsible for overseeing and driving sustainability initiatives. While the committee's primary focus may be on internal operations (Scope 1 and Scope 2 emissions), it can also influence and promote sustainable practices throughout the value chain, including suppliers and customers. By engaging with suppliers, encouraging sustainable sourcing practices, and collaborating with customers to reduce emissions associated with product use or disposal, the committee can indirectly contribute to Scope 3 emission reduction efforts.

ESG Disclosure score reflects a company's level of transparency and reporting on environmental and social factors. A high ESG Disclosure score indicates that a company is committed to comprehensive reporting and addressing sustainability impacts across its value chain. It suggests that the company is more likely to consider and manage Scope 3 emissions, even if they are not explicitly reported.

ESG News Sentiment refers to the overall tone and perception of a company's environmental, social, and governance practices in news and media. Positive ESG News Sentiment indicates that a company is receiving favorable attention for its sustainability efforts. This positive sentiment can lead to increased stakeholder expectations and pressure for the company to address its Scope 3 emissions and demonstrate a commitment to sustainable practices.

Emission Scope 1 and Emission Scope 2: While Emission Scope 1 and Emission Scope 2 specifically pertain to a company's direct emissions from its own operations), they can indirectly influence Scope 3 emissions. By actively managing and reducing their direct emissions (Scopes 1 and 2), companies can set an example and inspire similar actions among their suppliers and customers. Additionally, emission reduction efforts in Scopes 1 and 2 may involve energy efficiency measures, renewable energy adoption, or process improvements that can indirectly lead to reduced emissions in the value chain and Scope 3 emissions.

5.2. Data overview

5.2.1 Statistics summary

We study 2015-2022 as it had the most complete data across the datasets. We used company ticker and reporting year to match data points across two data providers and ended up with 8581 firm-year observations.

Variables	Non-Null	%Non-Null	Mean	Std	Min	Median	Max
GDP	8581	100%	238633.9	693978.6	234.558	4233.538	2957423
Population	8581	100%	6332.131	17749.94	5.184	126.221	65647
Scope 3 Purchased Goods and Industrials Emissions	3747	44%	4600.971	20770.45	0	772.8	896595
Scope 3 Capital Goods Emissions	2049	24%	590.2923	2972.178	0	102.8	45162
Scope 3 Fuel & Energy Related Activities Emissions	3325	39%	1263.231	5953.685	0	40.7	138000
Scope 3 Upstream Transportation and Distribution Emissions	2804	33%	558.8304	2871.645	0	74.65	45176
Scope 3 Waste Generated in Operations Emissions	3346	39%	165.541	2346.672	0	3.2	45182
Scope 3 Employee Computing Emissions	2911	34%	276.2088	3117.262	0	13.9	45174
Scope 3 Business Travel Emissions	5314	62%	150.5443	2329.362	0	4.6	45191
Scope 3 Upstream Leased Assets Emissions	1017	12%	81.57248	1428.777	0	1.2	45079
Scope 3 Downstream Transprttn and Distrbtn Emissions	1873	22%	577.4451	2458.215	0	62.4	45171
Scope 3 Processing of Sold Products Emissions	691	8%	9574.147	52993.85	0	25.173	556464
Scope 3 Use of Sold Products Emissions	2204	26%	33012.66	95488.91	0	1460.2	1162800
Scope 3 EOL Treatment of Sold Products Emissions	1704	20%	717.0033	3444.255	0	28.65	54601.3
Scope 3 Downstream Leased Assets Emissions	871	10%	214.8873	1172.045	0	1.7	16100
Scope 3 Emissions from Franchises	647	8%	471.0227	1340.412	0	0	8644.7
Scope 3 Emissions from Investments	1026	12%	2322.975	18965.02	0	16.1	461600
Scope 3 Emissions Other	1073	13%	2079.56	14894.14	-6	15.9	250370
Scope 3_total	8555	100%	12993.56	60360.61	0	93.6	1339550
ESG News Sentiment ES Positive	4851	57%	-0.05525	0.303998	-0.76	0	5.39
ESG Disclosure Score	7577	88%	52.61717	10.73424	12.9363	52.84	85.71

Number of Employees	8113	95%	44707.51	100269.8	3	16400	2300000
Total Assets	8548	100%	90324.87	285944.9	17.06335	12390.9	3743567
Capital Expenditures	8488	99%	-1109.45	3067.945	-71311	-266.55	0
Operating Expenses	8540	100%	5920.179	13423.48	0	1674.35	220101
SGA expense	7720	90%	3214.814	8644.09	0	828.9	156337.8
Cost of Goods & Industrials Sold	5927	69%	13835.34	36094.74	0	3811.1	627771.1
Inventories	7311	85%	1746.393	4744.557	0	378	89461.7
Revenue	8573	100%	19531.3	45218.58	-89294	6128	819169.4
Property Plant & Equipment Net	8504	99%	8874.852	28695.15	0	1958.95	699406
Asset Turnover	8537	99%	0.701684	0.60195	-0.2833	0.61	5.49
Inventory Turnover	6525	76%	22.58875	161.3744	-17.17	5.07	5314.8
Emission scope 1	7291	85%	41265.58	1198872	0	75.72	45170000
Emission scope 2	7288	85%	2927.177	79316.21	0	100	4460977

Table 4:Statistics Summary

Statistical summary of data shows that 16 scope 3 emission categories have a low level of availability. Scope 3 Business Travel Emissions was reported the most, with 62% companies because of its relative ease of calculation. Scope 3 Purchased Goods and Industrials Emissions is the second most reported type (44%). However, figures in this category vary in a wide data range, with a standard deviation of 20,770. The reason would be differences in the boundaries set for reporting and the methodologies employed. Different companies, based on the size and complexity of their supply chains, make different choices regarding the extent of their reporting within this category. Some companies opt for a comprehensive approach and calculate emissions for the entire value chain of purchased goods and services. This calculation often involves conducting a life cycle analysis (LCA) of the complete value chain to account for emissions at each stage. On the other hand, some companies choose to focus their reporting on emissions directly associated with their immediate suppliers, limiting the scope of their calculations to the emissions generated by those suppliers. This approach is narrower and excludes emissions that occur further upstream in the value chain. As a result of these different reporting approaches, companies with similar profiles may report inconsistent emissions figures for purchased goods and services. In contrast, only 8% of companies in the data set disclose Scope 3 Emissions from Franchises and Scope 3 Processing of Sold Products Emissions due to lack of control. Collecting data on Scope 3 emissions from franchises can be complex and challenging. Since franchises operate as separate entities and companies have limited direct control over the day-to-day operations and emissions of individual franchise partners, gathering consistent and accurate emissions data from a decentralized network can be difficult.

For predictors, the proportion of non-missing values is considerably high, at an average of 85%. However, these features record high variations in data. For example, Property Plant & Equipment Net variable has a mean of 8874.85 and a standard deviation of 28695.15. There could be 2 possible reasons: (1) differences in firms' size and (2) data errors. Companies vary in terms of their size, operational capacity, and market presence. These differences can directly impact the data generated by these firms. Larger firms typically have more extensive operations, higher revenues, and a larger customer base compared to smaller firms. As a result, the data produced by larger firms might exhibit higher values or greater fluctuations compared to smaller firms. These variations in data can be attributed to the differences in the scale and complexity of business operations across firms of varying sizes. Another reason for high variations in the data could be data errors or inconsistencies. Data collection and reporting processes are prone to errors, such as incorrect data entry, data processing issues, or inconsistencies in the reporting standards across different firms. These errors can introduce noise or distortions in the data, leading to higher variations. In addition, we found some weird patterns in financial variables such as Revenue and Asset Turnover and Inventory Turnover. Some companies recorded negative numbers, as can be seen in the "min value" in the summary table. For instance, min value of Revenues is -89,294. We double-check companies' annual reports to find the causes of the issue. These corporates are all in Financials sector and the negative numbers are explained by unrealized losses on financial investments. The net loss is primarily driven by a significant increase in interest rates, reducing the value of the fixed-maturity securities. As a result, we conclude that negative numbers are reasonable. Nevertheless, we decided to remove observations with negative numbers in these financial variables as they account for a small proportion and can skew overall data distribution.

5.2.2. Data issues

After exploring the dataset, we have identified several issues that potentially impact the accuracy and reliability of any analysis or insights derived from the data. It is essential to address these problems to ensure the integrity of our findings.

Data Availability: One of the main challenges with scope 3 emission data is the limited availability of comprehensive and standardized data. Not all companies disclose their scope 3 emissions, and even when they do, the level of detail and transparency can vary significantly. To address these data availability challenges in Scope 3 emissions, there is a growing need for collaboration and standardization in reporting frameworks and regulations. However, solving this issue is beyond our ability. We analyze data based on what we collected and the topic could be further researched in the future when more data is publicly available. Additionally, for variables other than scope 3 emissions, data unavailability may arise due to limitations in our access to data provider platforms. To overcome this, we utilize substitute data from another provider, Eikon. Furthermore, we employ imputation techniques to fill missing data.

Zero values: Another worth-noting pattern in the dataset is the appearance of zero values in Scope 3 total emissions and its categories. The presence of these zero values introduces ambiguity as they can be interpreted in different ways and there is no explanation from Bloomberg. In some cases, a zero value in Scope 3 total emissions may indicate that no emissions were generated by reporting companies. However, we are leaning toward the interpretation that zero values can be misleading or unavailable. It is possible that emissions data for specific categories may not have been captured and calculated by companies. Thus, we exclude zero values in emissions features.

Inconsistent data: The issue is caused by both companies and data providers. In some scope 3 category variables, we spot differences between 2 data providers (Bloomberg and Refinitiv). This inconsistency was also confirmed by previous research. Busch et al. (2022) investigated the consistency of emissions data among third-party data providers (including Bloomberg, CDP, ISS, MSCI, Sustainalytics, Thomson Reuters Refinitiv, and Trucost) in the period of 2005- 2016 and concluded that the divergence in reported Scope 3 is much more significant than those observed in scope 1 and scope 2 emissions. When it comes to companies themselves, capturing accurate carbon emissions data is a challenging task. Companies employ different approaches to record their indirect emissions. We address this issue by grouping countries into industrial groups and creating models to predict corporate scope 3 emission in each industry.

5.2.3. Descriptive analysis

Total scope 3 carbon emissions

For trend analysis, we omit Year 2022 as the number of companies reporting scope 3 emissions at the time we were collecting data is relatively small.

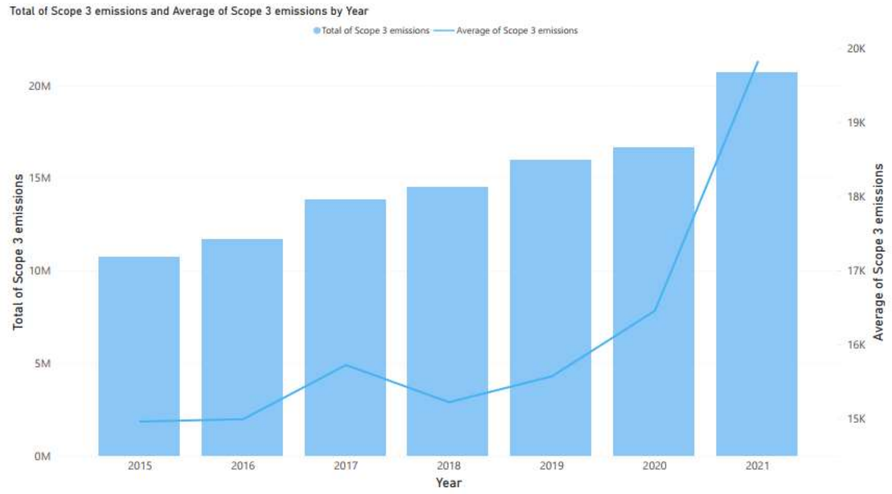


Figure 5: Scope 3 Carbon Emissions in the period of 2015-2021

The line graph shows that on average, corporate scope 3 carbon emissions increase significantly from 15,000 thousand tons to 20,000 tons throughout the period, especially after 2020. The total carbon emissions have also grown gradually. Despite our expectation of a decrease in carbon emissions in 2019-2020 due to the global lockdown and economic downturn caused by the Covid-19 pandemic, the figure paints a different picture. One of the reasons could be that outliers in the dataset contribute to the higher average carbon emissions. Outliers can represent specific companies with exceptionally high emissions, skewing the overall average. Another possibility is many carbon-intensive industries, such as manufacturing, energy production, and transportation, were deemed essential and continued their operations during the pandemic. These sectors contribute significantly to global carbon emissions, and their activities were not curtailed to the same extent as other industries. Additionally, disruptions in supply chains may have caused companies to rely on alternative, less environmentally friendly sources or methods, prioritizing reliability, and resilience over sustainability. To mitigate risks and ensure consistent supplies, organizations might have chosen to collaborate with suppliers that were not aligned with their carbon reduction goals.

Scope 3 carbon emissions by Sectors

We plot scope 3 carbon emissions by Year and Sectors for further analysis to explore how they contribute to the total number.

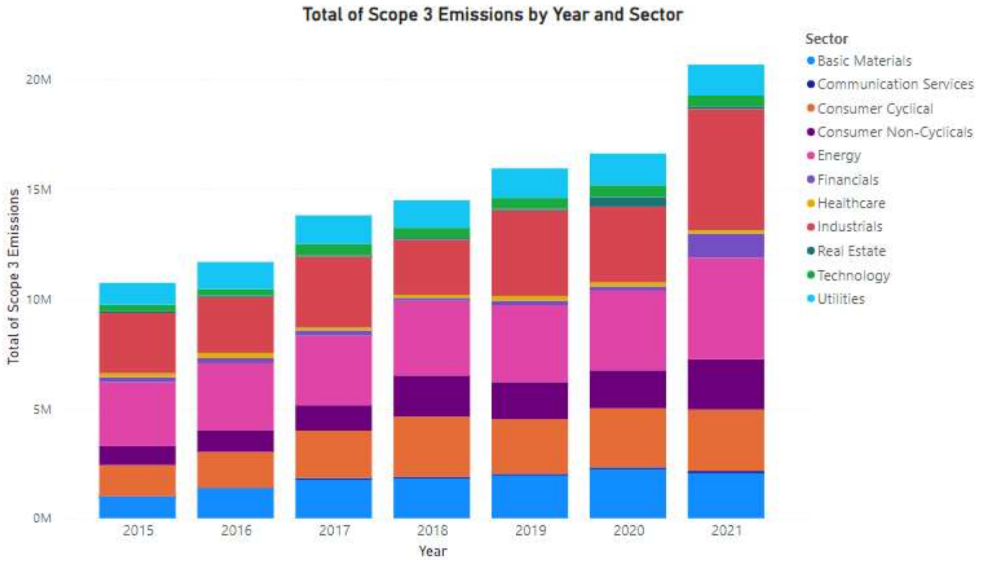
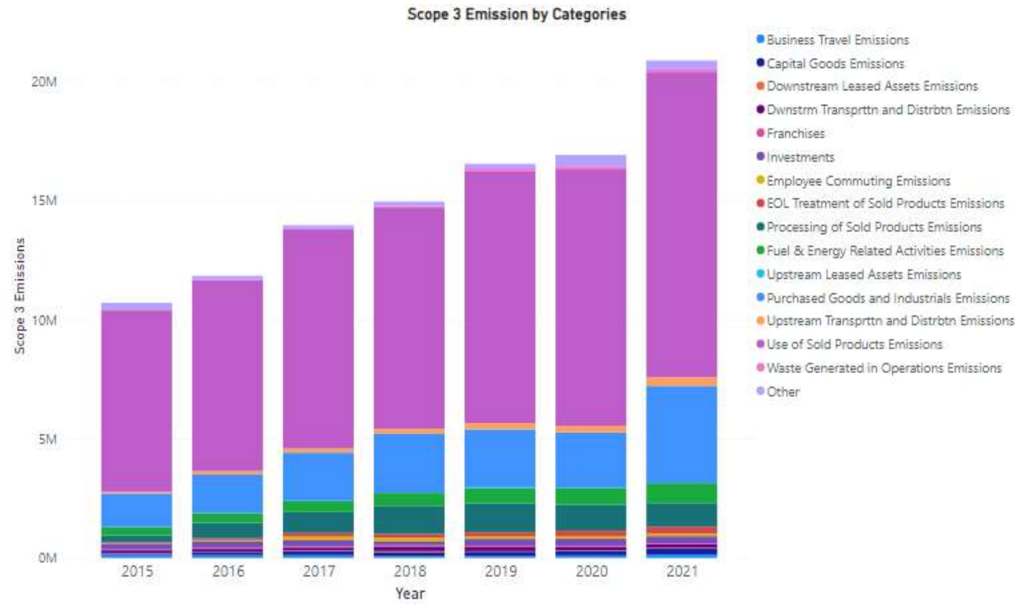


Figure 6: Total Scope 3 Emissions by Sectors

In general, firms in different economic groups experience a rise in scope 3 carbon emissions. Among sectors, Energy and Industry sectors contribute carbon emissions to the total the most, accounting for about 50% of the total. The products and services offered by Energy and Industrials companies often have significant emissions associated with their use. For instance, energy producers provide fuels for transportation and electricity generation, which contribute to emissions when consumed by end-users. Similarly, industrial manufacturers generate products such as cement, steel, and chemicals, which are used in various sectors and can result in emissions during their lifecycle. Additionally, the energy-intensive nature of these sectors can indirectly drive emissions up in connection with electricity consumption, extraction and production of raw materials. The consumer cyclical sector, responsible for 15% of carbon emissions, ranks as the second-largest contributor. This sector comprises products such as cars and household appliances that tend to consume significant amounts of energy during their usage. The carbon emissions linked to the utilization and upkeep of these products contribute to the Scope 3 emissions of the manufacturing companies. Additionally, consumer cyclical corporates generate waste during their operations, including packaging waste and the disposal of products at the end of their lifecycle, thereby producing indirect emissions.

Scope 3 carbon emissions by Year and Categories



16 categories experience the same increasing trend with total scope 3 carbon emissions in the period of 2015-2021. A significant portion of their Scope 3 emissions comes from the Use of Sold Products and Purchased Good & Industrials. This means that the emissions are generated when customers use the products that these firms sell, such as burning fossil fuels in vehicles or using energy-intensive products. Emissions from these 2 categories make up 75% of total scope 3 emissions.

Scope 3 categories by Sectors

A deeper look at each sector reveals variation in data of 16 emissions categories

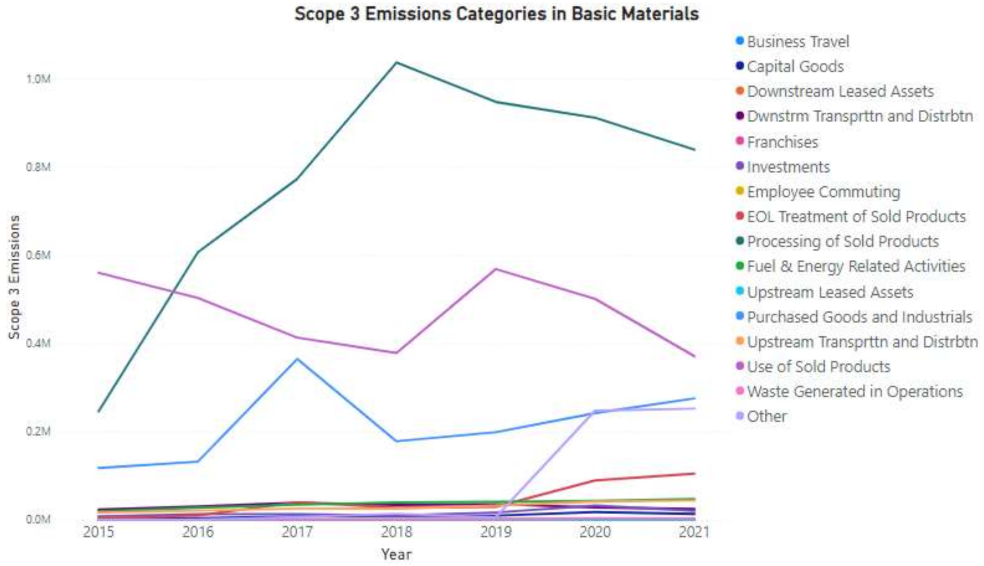


Figure 7: Scope 3 Emission Categories in Basic Materials group

For corporations operating in the Basic Materials sector, Scope 3 emissions typically arise from two primary sources: Processing of sold products and Use of sold products. Companies in this sector are involved in extracting, refining, and processing raw materials such as minerals, metals, chemicals, and other materials. These operations can result in significant emissions. The total emissions of researched firms in this sector from Processing of Sold Product go up to 1M thousands of tones in 2018 before decreasing gradually. Regarding Use of Sold Products, it includes emissions resulting from the combustion or utilization of products made from the company's materials. For example, if a corporation produces coal, the emissions are from burning that coal by end-users (e.g., power plants or households). The Use of Sold Products emissions go down to around 0.4M thousand tones in 2018, then slightly increase in 2019. On the other hand, companies within the Financials industry contribute to emissions through investment portfolios. Therefore, Scope 3 emissions for these corporations are primarily generated from the Investment category (purple line).

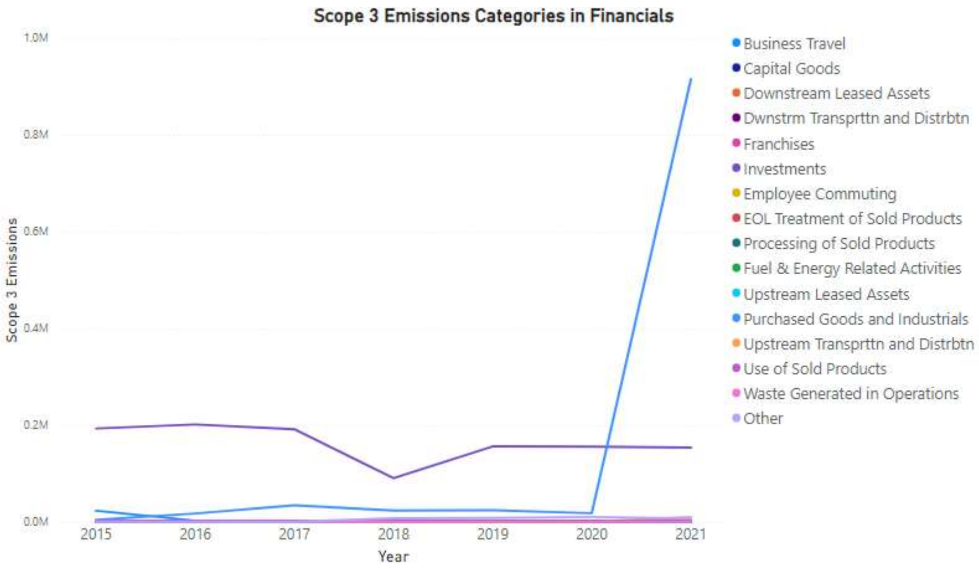
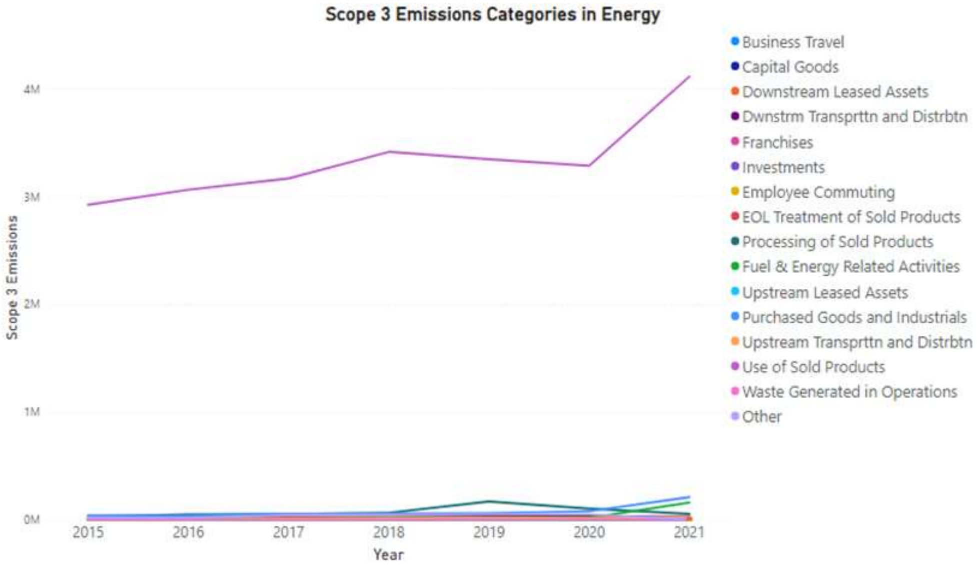


Figure 8: Scope 3 Emission Categories in Financial sector

The Energy and Industrials sectors, which encompass oil and gas companies as well as manufacturers of fuel-based cars, exhibit a significant share of carbon emissions stemming from the Use of their sold products.



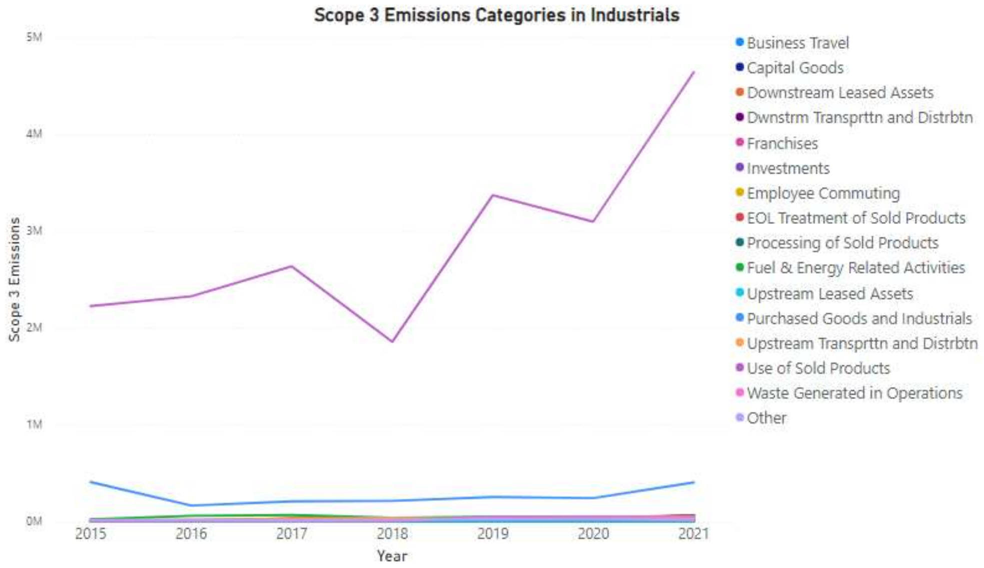


Figure 9: Scope 3 Emission Categories in Energy and Industrials Sectors

5.3. Data processing

5.3.1 Outliers

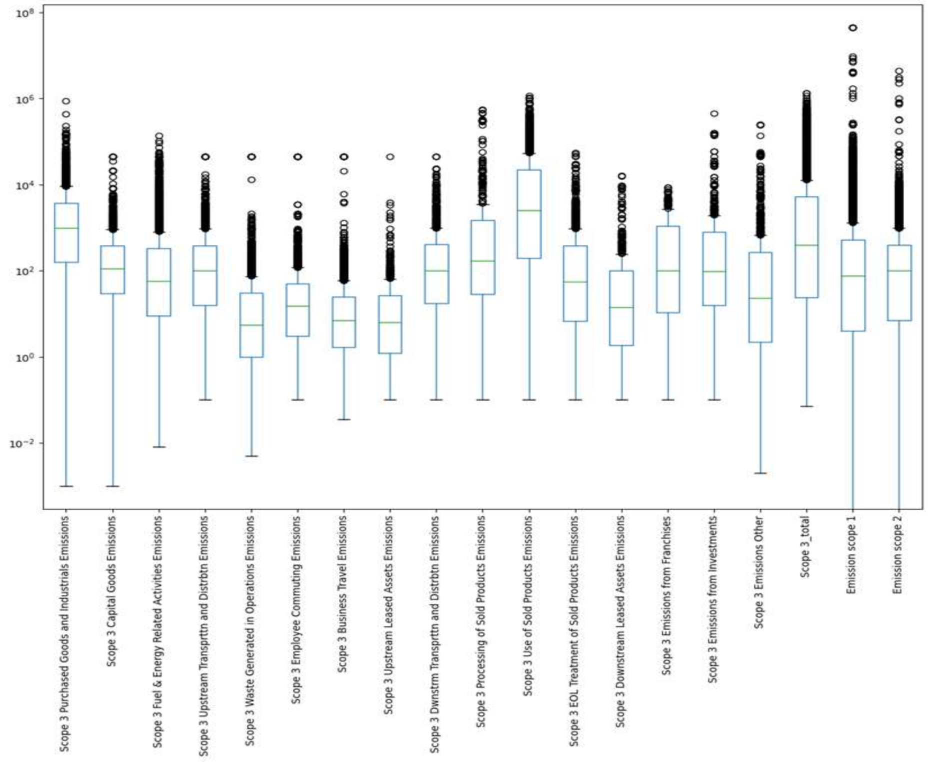


Figure 10: Boxplot of Carbon Emissions variables

The dataset contains substantial presence of outlier values. We have two justifications for these extreme outliers in the data set. On the one hand, outlier could be incorrect records from either reporting companies or data providers

(Bloomberg, Refinitiv). These errors might occur during the recording, estimating, or transforming, causing anomalous data points to be included in the dataset. For data providers, inconsistencies in the measurement units used for different variables within the dataset can contribute to outliers. Specifically, Emission variables such as scope 1 and scope 2 are recorded in million metric tons, but scope 3 emissions are in thousands of metric tons, a simple mistake in unit conversion can result in a data point that is 1,000 times larger than its actual value. Such a discrepancy can significantly skew the dataset, introducing extreme outliers. For companies, recording scope 3 emissions is a challenging task as scope 3 is indirect greenhouse gas emissions that occur throughout a company's value chain, involving a broader range of stakeholders and factors outside the company's immediate control. Companies rely on complex assumptions to calculate these emissions, which introduces a degree of uncertainty. Different methodologies and calculation approaches can lead to variations in reported scope 3 emissions, resulting in outliers in the data set. In some companies in our data set, there is a significant and abnormal increase in emissions figures throughout the periods of 2015-2022.

On the other hand, extreme outliers in the data set can also be genuine representations of actual data. This justification aligns with the assumption that large companies or industry giants, due to their business size and global operations, might have significantly higher emissions and financial figures compared to other entities. Additionally, it is widely acknowledged that a substantial portion of total greenhouse gas emissions is contributed by major players in various sectors. Changes in a company's emissions profile over time can be also other reason for outliers within the dataset. For example, if a company undertakes significant expansions, acquisitions, or changes in its business operations, it may experience a substantial increase or decrease in its data. These changes can lead to outliers, especially when comparing data over different time periods.

As a result, we do not remove outliers from the data set because they can often contain valuable information and insights. Instead, we apply Winsorization technique to upper tails of distribution to address right skewness in almost all of variables. Values in the dataset that exceed the 99th percentile are classified as outliers and are substituted with the values at the 99th percentile. This ensures that the extreme values are retained but reduced to a more moderate level, mitigating their impact on statistical analysis and modeling.

5.3.2 Encoding

Categorical data should be encoded into a numerical format because most machine learning algorithms are designed to work with numerical data. By encoding categorical variables (Sector, Sustainability Committee, Carbon tax), we convert qualitative information into a quantitative form that algorithms can understand. Furthermore, encoding categorical data can help reduce the dimensionality of the dataset by transforming a categorical variable with multiple categories into binary variables. Ordinary encoding and one-hot encoding are two different techniques used in data preprocessing. Ordinary encoding is used when there is an inherent order or hierarchy among the categories, while one-hot encoding is employed when there is no such order and each category should be treated independently. For this reason, we apply one-hot encoding to 3 categorical variables in our dataset. After encoding, we have these following binary variables:

Variables	**Types**	**Numbers of 0**	**Numbers of 1**
Carbon tax	Binary	2766	3981
CSR/Sustainability Committee	Binary	3820	2927
Sector_Communication Services	Binary	6437	310
Sector_Consumer Cyclical	Binary	5949	798
Sector_Consumer Non-Cyclicals	Binary	6069	657
Sector_Energy	Binary	6474	273
Sector_Financials	Binary	5700	1047
Sector_Healthcare	Binary	6389	358
Sector_Industrials	Binary	5349	1398
Sector_Real Estate	Binary	6433	314
Sector_Technology	Binary	6168	579
Sector_Utilities	Binary	6278	469

5.3.3 Imputation

Handling missing data is an important step in data preprocessing. As missing data appears in almost all of the explanatory variables, we will not opt for removing option. Instead, we perform 3 data imputation techniques and compare how well imputation works for data analysis and modeling. First and foremost, we identify missingness types of variables to choose effective imputation strategies. The type of missingness refers to the underlying patterns or reasons why the data is missing. In general, there are three common types of missing data: Missing Completely at Random (MCAR), Missing at Random (MAR), Missing Not at Random (MNAR).

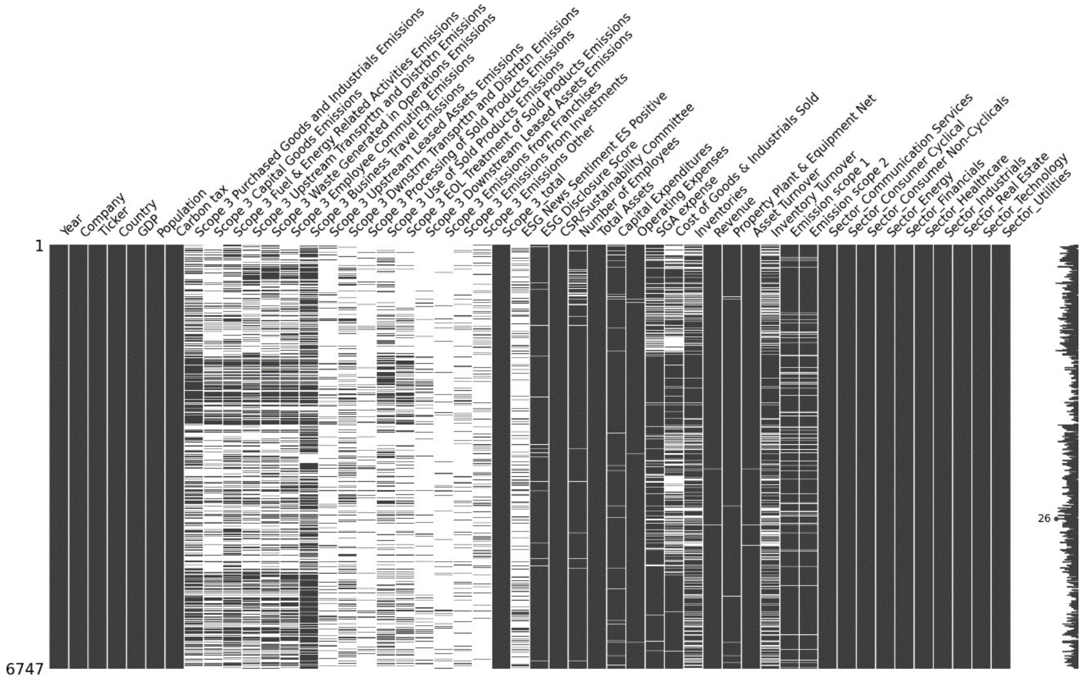


Figure 11: Missing value matrix

The matrix represents how null values are scattered across the dataset. White segments or lines illustrate missing values. As can be seen in the matrix, object, and categorical columns have no missing values, but numerical columns have. There is no systematic missingness in the dataset, except for two pairs of columns: Emission scope 1 - Emissions scope 2 and Inventories - Inventory turnover. For Emission Scope 1 and Emission Scope 2, the reason would be that companies not reporting Scope 1 have a tendency not to disclose Scope 2 and vice versa. However, we do not spot any relationship between the missing data in these two columns and those of other variables. It is clear to conclude that these 2 columns are MCAR. The same patterns are found in Inventories and Inventory Turnover variables. These two columns have the same null-value distribution because Inventories is one of the factors to calculate Inventory Turnover. All in all, missing values are totally random in the dataset. Based on the above analysis, we end up with 3 methods of implementation with increasing levels of complexity.

a. Simple imputation

The technique involves replacing missing values with a single value, typically based on summary statistics or fixed values. We choose median imputation since it is less sensitive to outliers compared to mean imputation. However, replacing missing value with the median of non-null values in the whole dataset might introduce bias and distort the underlying data distribution. As a result, the missing value is covered with the median of available values grouped by sectors and years. This is because we assume that companies in the same economic group share similarities and each year has specific effect on variables.

b. KNN imputation

Imputation using k-Nearest Neighbors (k-NN) is a technique used to fill in missing values in a dataset by using the values of its neighboring data points. k-NN imputation is a popular method because it takes into account the similarity between data points to estimate the missing values. The imputation method works as follows:

- Determine the value of k: Decide the value of k, which represents the number of nearest neighbors that will be used to impute a missing value. Because our dataset is not so large, we choose k = 3 as a starting point.
- Calculate the distance: Calculate the distance between the missing value and all other data points in the dataset. Euclidean distance is used for numerical variables.
- Select k-nearest neighbors: We select 3 data points with the shortest distances to the missing value. These data points will be considered the nearest neighbors.
- Impute the missing value: Once we have identified the nearest neighbors, take the average (for numerical data) or mode (for categorical data) of the values of the missing feature from these neighbors. This average or mode value will be used to impute the missing value.
- Repeat the process: Repeat steps for all missing values in the dataset.

c. **Iterative imputation**

Iterative imputation is an advanced imputation technique used to fill in missing values in a dataset. The key idea behind iterative imputation is that missing values are imputed using information from other variables in the dataset. By iteratively updating the imputed values based on the estimated models, the imputations become more refined with each iteration. The mechanism behind the techniques include the following steps:

- Split the dataset: Split the dataset into two parts: a complete cases dataset and an incomplete dataset. The complete cases dataset contains observations with no missing values, while the incomplete dataset contains observations with missing values for the variables of interest.
- Initialize imputed values: Start by filling in the missing values with initial estimates. This can be done using the median imputation method
- Build imputation model: For each variable with missing values, a model using the 3 other nearest variables in the dataset as predictors. Nearness between features is measured using the absolute correlation coefficient between each feature pair (after initial imputation). The model we used is Random Forest Regressor.
- Estimate missing values: Use the built imputation model to estimate the missing values for each variable. The model takes the available data as input and predicts the missing values.
- Update the dataset: Replace the missing values in the incomplete dataset with the estimated values obtained in the previous step. This creates a new dataset that combines the imputed values with the complete cases from the original dataset.
- Repeat steps 3 to 5 or a predefined number of iterations. In each iteration, the imputation model is rebuilt using the updated dataset, and missing values are estimated based on the model.
- Final imputed values: After completing the iteration, the imputed values obtained in the final iteration represent the best estimates for the missing values based on the available data. These imputed values replace the initial estimates from step 3.

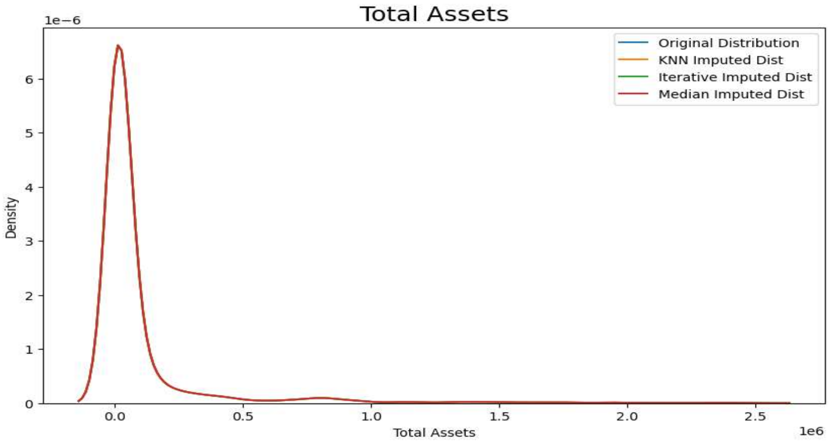
d. **Compare imputation techniques**

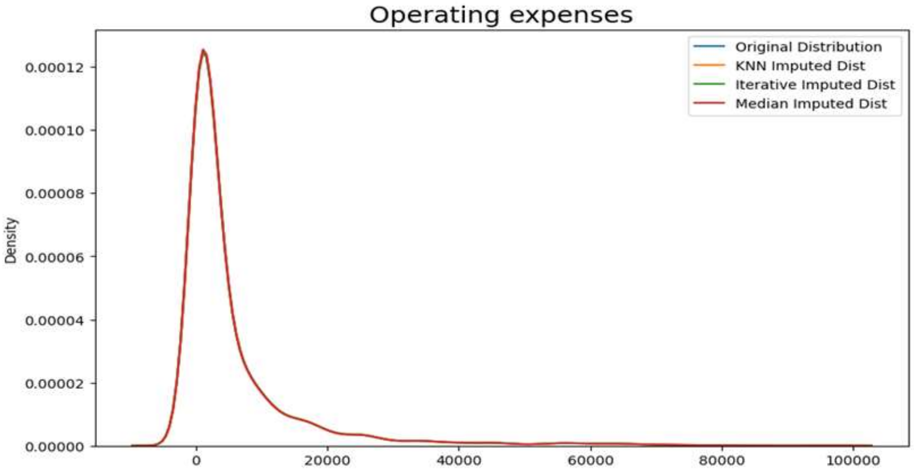
We compare the 3 imputation techniques based on 3 aspects: assessing data distribution, analyzing downstream effects, and evaluating computational efficiency.

Assess distributional properties

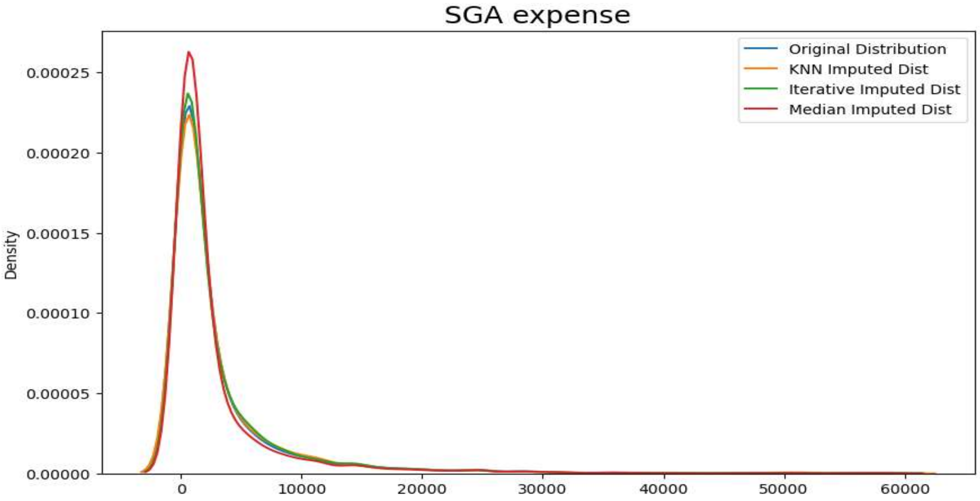
For each type of imputation technique, we visualize Kernel Density Estimation plot, which estimates the probability density function (PDF) of a continuous variable in our dataset. The purpose is to examine the distributional properties of variables with imputed data and check if they are preserved or distorted, compared to the original distribution.

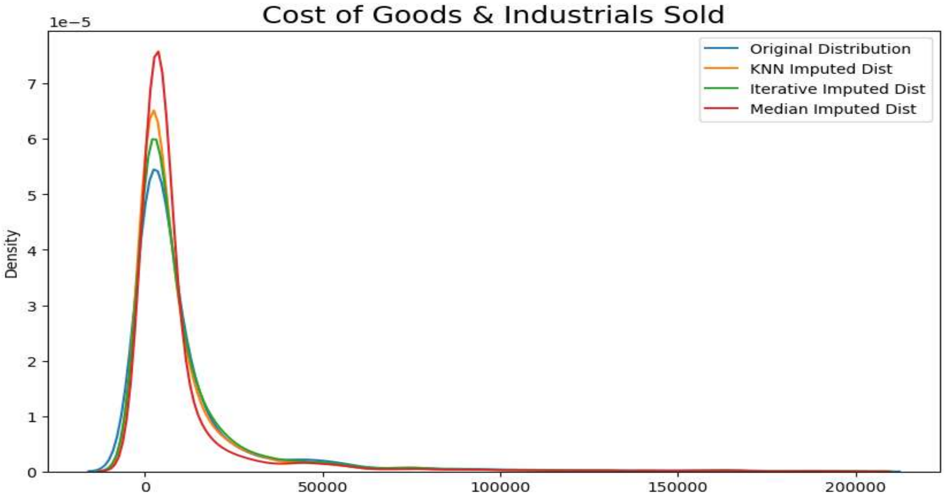
The overlapping distributions of imputed data sets indicate that there is no significant difference between the imputation techniques for the variables Total Asset, Capital Expenditure, Operating Expense, Revenue, Property Plant & Equipment, and Asset Turnover. The reason for the similarity in distributions could be attributed to the small proportion of missing data in these columns. When the proportion of missing data is small, the choice of imputation techniques is likely less influential on the overall distribution of the data.



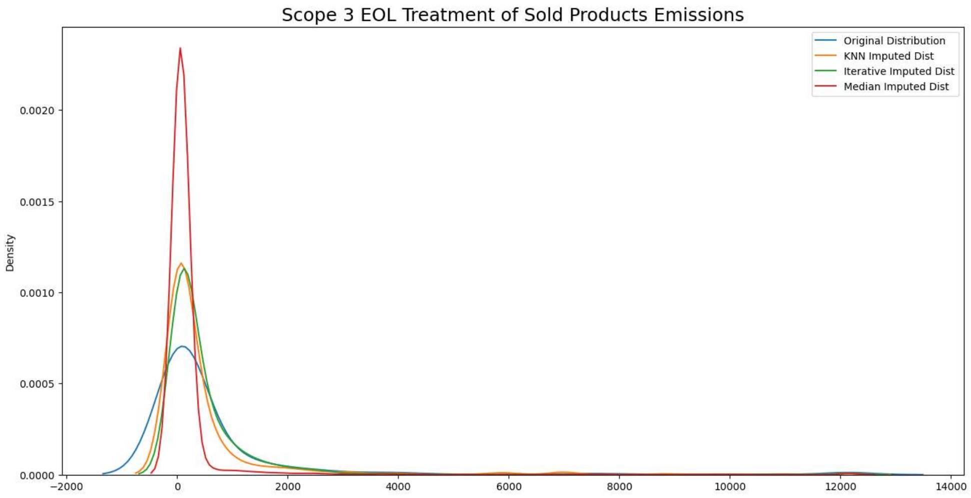


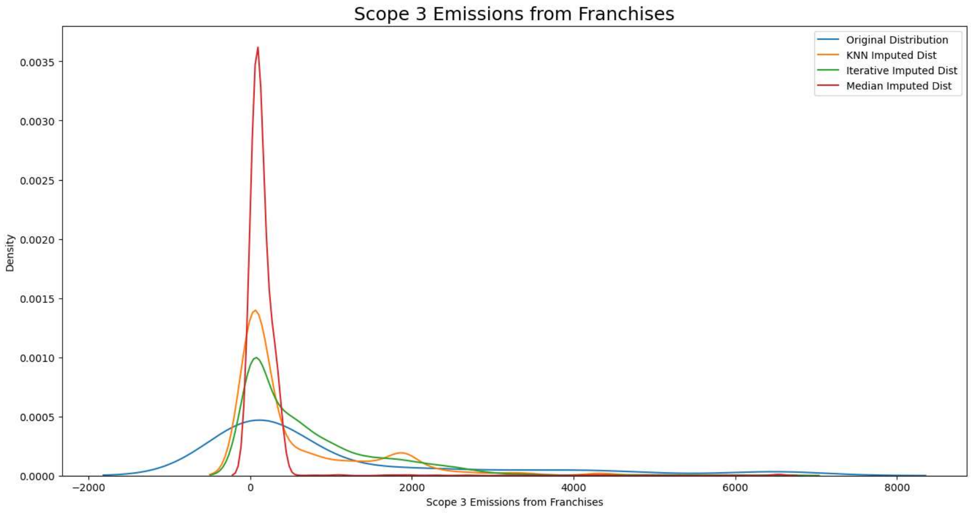
For some predictors such as SGA expense and COGS, the three techniques do not fully fit the original distributions, with Iterative and KNN imputations showing superiority over Median one. A similar property is found in Inventories, Inventory Turnover variables.





The three imputation techniques work poorly for 16 scope 3 categories variables due to a high proportion of missing data. Imputation methods rely on patterns and relationships within the observed data to impute missing values. If a substantial portion of the data is missing, these patterns and relationships may be insufficient. Furthermore, variables with a majority of missing data introduce uncertainty into the imputation process. The imputed values are essentially educated guesses based on the available information, and with limited data, the imputation becomes more speculative. This increased uncertainty can result in imputed values that deviate further from the true values, leading to poorer imputation. The following plots show some examples of this case.





Evaluate computational efficiency

We use eco2Ai package in Python to capture execution time, power consumption and equivalent CO2 emissions of imputation process.

Computational efficiency	Simple Imputaion	KNN Imputation	Iterative Imputation
Execution time (seconds)	8.78	26.13	2768.36
Power consumption (kWh)	0.000002	0.000009	0.003847
CO2 emissions (kg)	5.006117e-08	2.813225e-07	1.191930e-04

Table 5: Computational metrices between imputation techniques

Simple Imputation: Median imputation is a relatively simple method. It typically has a low execution time 8.6s and does not require intensive computations.

The computation resources primarily depend on the dataset size and the number of variables with missing values. With the size of 6734 observations, simple imputation consumes 5.006117e-08 kWh.

KNN Imputation: KNN imputation is a more computationally intensive method compared to median imputation. Execution time of KNN imputation is three time as that of simple imputation (26.13 seconds). Execution time and power consumption of KNN imputation depends on factors such as the dataset size, the number of variables, the chosen value of k (number of nearest neighbors), and the distance metric used. Generally, KNN imputation can be more time-consuming, especially for larger datasets and higher values of k.

Iterative Imputation: Execution time and Power consumption of iterative imputation are highest (2768 seconds and 0.003847 kWh) due to the more complex computations involved in the iterative modeling and imputation steps. The power consumption depends on the complexity of the imputation model, the number of iterations, and the dataset size. It is much higher than median imputation but can vary depending on the specific implementation.

Analyze downstream effects

We consider the impact of the imputed values on scope 3 emissions prediction by comparig the performance of models using datasets generated by three imputation techniques to evaluate how well each technique supports our analysis objectives. We will report more details in Chapter 6: Results

5.3.4 Scaling

As our variables have various ranges, scaling is crucial data transformation to ensure that all variables or features in a dataset have comparable magnitudes and units, which can be beneficial for distance-based algorithms such as OLS regression or k-nearest neighbors. We perform scaling on the training set and then apply consistently to the test sets using the parameters obtained from the training set to avoid leakage. This ensures that the scaling is done based on the training data's characteristics and maintains consistency across different datasets. There are two common scaling techniques: Standardization (or Z-score normalization) and Min-Max scaling (or normalization). Standardization scales the data to have a mean of 0 and a standard deviation of 1. Standardization preserves the shape of the distribution and is less affected by outliers. It is useful when the data does follow a specific distribution or when the algorithm benefits from centered variables.

$$x_{stand} = \frac{x - mean(x)}{standard\ deviation(x)}$$

In contrast, Min-Max scaling transforms the data to a predefined range, often between 0 and 1. Min-Max scaling brings all the data within the range, preserving the relative relationships between data points. It is suitable when the algorithm requires all variables to have the same scale or when there are specific requirements for the input range.

$$x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

As a test for appropriate scaling, a normal quantile-quantile plot is generated to check if a variable follows a normal distribution to apply appropriate scaling. If the points on the Q-Q plot form a 45-degree straight line, it indicates that the predictive feature is normally distributed. On the other hand, if the points deviate significantly from the straight line, it suggests that the feature does not follow a normal distribution. However, it is worth noting that for tree-based algorithms, model performance is not heavily impacted by the choice of scaling techniques After creating Q-Q plot, we found that only ESG Disclosure Score variable displays normal distribution. Therefore, we apply Standardization to this variable. Other features are treated with Min-Max scaling.

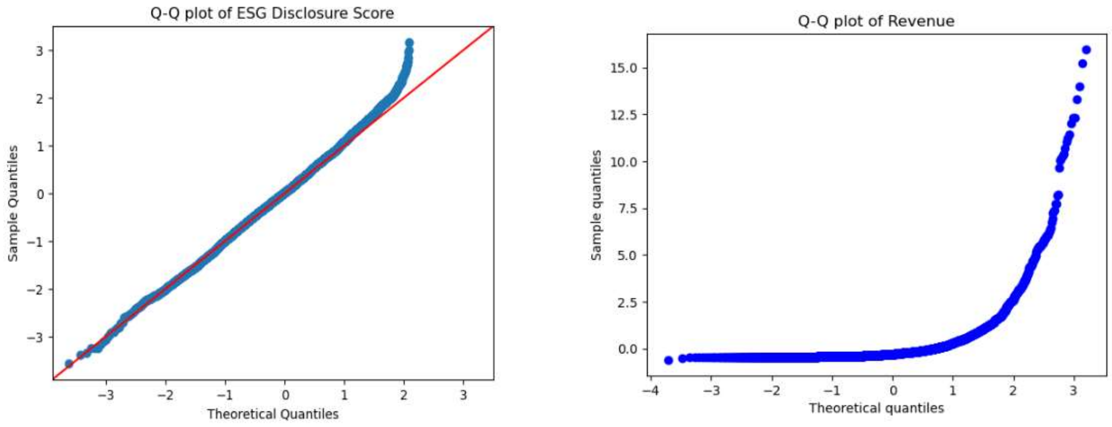


Figure 12: Q-Q plot of predictors

Chapter 6: Result

6.1. Predict total scope 3 emissions

a. Train_test_split technique

Based on three imputation techniques explained in the *Data processing* part, we have three datasets accordingly used for training machine learning models. The target variable is total scope 3 emissions. All models are trained on 80% of the full sample dataset and the prediction performance metrics are achieved by testing on the holdout test set comprising of 20% of the total sample not seen by the model previously. We used *train_test_split* syntax from scikit-learn to conduct this data split. The results are then compared to selecting the best model having the best prediction metrics in alignment with the most efficient energy consumption.

Dataset 1*: Simple imputation*

	R^2	**RMSE**	**Execution time (s)**	**Power consump-tion (kWh)**	**Equivalent carbon emissions (kg)**	**Energy efficiency**
KNN Regressor	0.0334	43773.82	5.0946	4.350e-08	1.347e-09	9.94e-13
Linear regression	0.3791	35085.96	4.6845	4.574e-08	1.417e-09	1.30e-12
Adaptive Boosting Regressor	0.4029	34405.62	4.9767	5.439e-08	1.685e-09	1.58e-12
Gradient Boosting Regressor	0.6803	25176.02	7.5968	1.113e-06	3.451e-08	4.42e-11
Random Forest Regressor	0.7990	19959.75	21.1364	1.118e-05	3.466e-07	5.60e-10
Extra Trees Regressor	0.8743	15784.63	6.9669	2.527e-07	7.830e-09	1.60e-11

Table 6. Performance results of models running on simple imputation dataset

Table 1 shows us the results of multiple models running on the dataset imputed by the simple imputation technique. We observed an increase in R-squared and a decrease in RMSE from KNN Regressor (0.0334; 43773.82) to Extra Trees Regressor (0.8743; 15784.63). With this considerable improvement in predictive accuracy across models, Extra Trees Regressor could be concluded as the best model for this imputed dataset. In addition, Extra Trees Regressor performs well in terms of generating less equivalent carbon emissions (7.830e-09 kg) and consuming less power energy (2.527e-07 kWh) compared to other well-performing models such as Gradient Boosting Regressor (3.451e-08 kg; 1.113e-06 kWh) and Random Forest Regressor (3.466e-07 kg; 1.118e-05 kWh).

Dataset 2: *KNN imputation*

	R^2	**RMSE**	**Execution time (s)**	**Power consumption (kWh)**	**Equivalent carbon emissions (kg)**	**Energy efficiency**
KNN Regressor	0.0091	44321.22	4.408650	6.937e-07	2.149e-08	1.57e-11
Linear regression	0.3616	35574.95	4.527451	7.567e-07	2.344e-08	2.13e-11
Adaptive Boosting	0.2786	37817.73	5.784304	1.081e-06	3.352e-08	2.86e-11
Gradient Boosting	0.6655	25748.25	8.097750	3.674e-06	1.138e-07	1.43e-10
Random Forest	0.7838	20703.05	20.113382	2.301e-05	7.131e-07	1.11e-09
Extra Trees	0.8751	15734.78	8.415064	3.798e-06	1.176e-07	2.41e-10

Table 7. Performance results of models running on KNN imputation dataset.

Table 2 shows us the results of models running on the dataset imputed by the KNN imputation technique. For this imputed dataset, we observe that OLS Regression model has a higher R^2 and a lower RMSE than those of AdaBoost, whereas this is not the case in the simple imputation dataset. For the other models, KNN still has the lowest R^2 (0.0091) and highest RMSE (44321.22) while Extra

Trees Regressor has the highest R^2 (0.8751) and lowest RMSE (15734.78). In terms of energy consumption and equivalent carbon emissions generation, Extra Trees Regressor also appear to be an energy-efficient model with 3.798e-06 kWh of power consumption and 1.176e-07 kg of equivalent CO_2, resulting in an energy efficiency of 2.41e-10. This means that a model consumes 2.41e-10 kWh of energy power to reduce one unit of error in RMSE.

Dataset 3*: Iterative imputation*

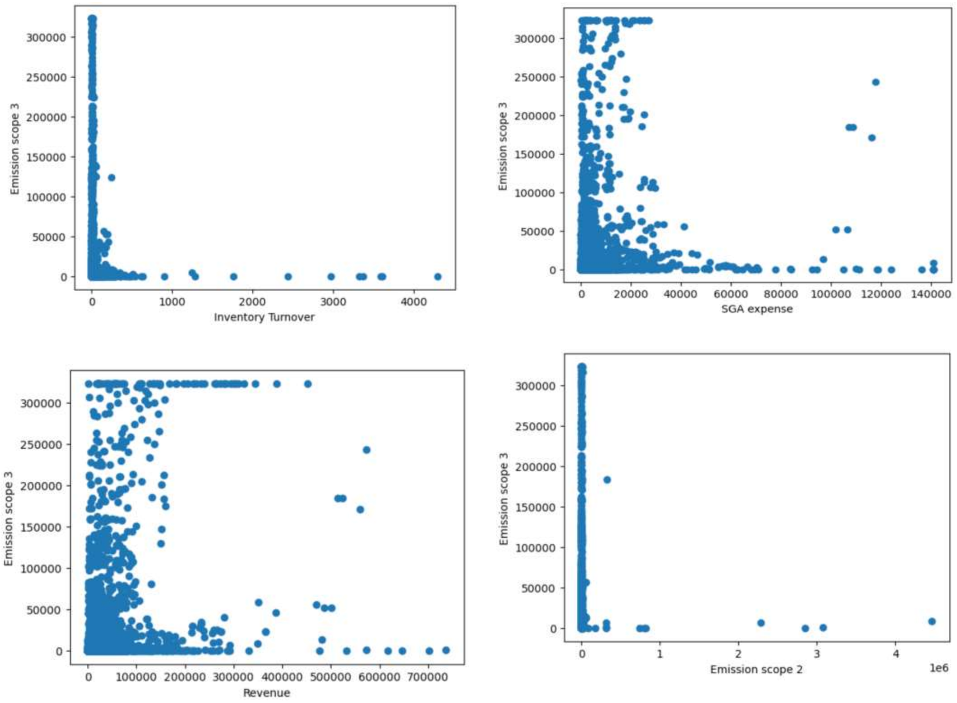
	R^2	RMSE	Execution time (s)	Power consumption (kWh)	Equivalent carbon emissions (kg)	Energy efficiency
KNN Regressor	0.0369	43695.96	5.502040	9.090e-07	2.816e-08	2.08e-11
Linear regression	0.3698	35345.32	5.144282	9.292e-07	2.879e-08	2.63e-11
Adaptive Boosting	0.2523	38501.00	7.604740	1.626e-06	5.038e-08	4.22e-11
Gradient Boosting	0.6613	25909.95	8.115260	3.498e-06	1.084e-07	1.35e-10
Random Forest	0.7590	21856.96	20.416519	2.427e-05	7.522e-07	1.11e-09
Extra Trees	0.8581	16771.68	11.620457	6.402e-06	1.983e-07	3.82e-10

Table 8. Performance results of models running on iterative imputation dataset.

Table 3 reports the performance metrics of models running on the dataset imputed by iterative imputation technique. Like the KNN imputation dataset, AdaBoost running on the iterative imputation dataset has a less effective performance than OLS Regression, and Extra Trees Regressor still has the best performance with an R^2 of 0.8581 and the lowest RMSE of 16771.68 while KNN Regressor has the lowest R^2 of 0.0369 and the highest RMSE of 43695.96.

Taking all models developed from the three datasets into consideration, we observe that Extra Tree Regressor running on the simple imputation dataset appears to be the most effective in predicting the target and the most efficient in respect of energy consumption.

Specifically, within each dataset, Extra Tree Regressor always performs better than the other models in terms of accurately predicting the target. It has the highest R-squared and the lowest RMSE while distance-based models such as OLS and KNN Regression perform much less effectively. There are several main reasons explaining for the poor performance of these two models on the dataset used in this paper. With respect to the OLS regression, firstly, it is a parametric approach, meaning that it makes assumptions about the data for the purposes of analysis. One of the essential assumptions is the linearity of residuals. In other words, there needs to be a linear relationship between the dependent variable and the independent variables. Whereas not all the relationships between the target and the predictors in our dataset are linear, which is demonstrated by the following scatter plots. Secondly, OLS regression assumes that the error terms should follow a normal distribution with a mean equal or close to zero. If the error terms are not normally distributed, meaning that there are some abnormalities affecting the predictive ability of the model. For our dataset, the distribution of data points is still right-skewed after the data normalization step. For these reasons, linear regression model fails to predict the target accurately.



Regarding KNN regression model, it is a non-parametric algorithm because no assumptions about the training data are required. Therefore, this algorithm could prevent the issue of non-linear relationships between target and predictors in OLS regression. However, KNN might be very sensitive to the scale of data because it is built based on computing distance among observations. For features having a high range of values, the calculated distance can be high and might result in poor predictions. For our dataset, even though we treat extreme outliers, the difference between the top and the bottom values is still substantial. As a consequence, KNN regression is not suitable for this dataset.

In some previous research, AdaBoost is concluded as the best model reaching the highest R-squared and the lowest RMSLE (George & Gladys, 2022). Contrary to our expectations, the AdaBoost model's performance on our dataset is not good across three imputation datasets. One of distinguishing advantages of AdaBoost regressor is that it is less prone to overfitting because it runs each model in a sequence and has a weight associated with them. However, AdaBoost needs quality data for training since it is highly sensitive to outliers and noisy data. As a result, with a large range of values in our dataset, AdaBoost suffers the same issue as KNN regression.

For Gradient Boosting regressor and Random Forest regressor, they have quite good R-squared scores with approximately 66% and 76%, and lower RMSE values of around 25,000 and 20,000 respectively across three datasets. As with other tree-based models, both Gradient Boosting and Random Forest handle non-linearity well. In other words, when the relationship between input features and target variable is not perfectly linear, both models are not substantially affected. The reason is that there are no attempts to fit a linear trend to the data during trees developing process. Furthermore, both Gradient Boosting and Random Forest are also not heavily affected by outliers. Therefore, these two models solve the issues of outliers and non-linearity of the OLS regression and KNN. However, there are two main differences between the random forest regressor and gradient boosting trees. Firstly, trees in the random forest are constructed independently, therefore, many trees in the forest could be trained in parallel, whereas Gradient Boosting builds one tree at a time. Secondly, the principle of making output decisions is different between the two models. Random Forest constructs independent trees, therefore, it could determine the outputs of each tree in any order. After, based on the average aggregation across all developed trees, the outcome is made. On the other hand, the gradient boosting trees run in a fixed order, meaning the tree developing sequence could not change, leading the trees to admit only sequential evaluations. Theoretically, gradient boosting could give better performance compared to Random Forest because the newly developed tree fixes the error of the previous tree in the sequence while the Random Forest takes the average of all independent trees in the forest. However, the performance of gradient boosting deteriorates if the data are noisy. Specifically, the boosted trees may overfit the data and start modelling the noise. The dataset used in the paper is not complete with a pretty high percentage of missing data in predictors and outliers in target variable. In spite of being imputed, the data could still contain noise, leading the poorer performance of gradient boosting than random forest.

Similar to Random Forest, Extra Trees regressor is also an ensemble machine learning method training various decision trees. The results from developed trees are then aggregated to output a prediction. Nevertheless, if random forest uses bagging method to select different variations of the training set to ensure the sufficient difference across decision trees, Extra Tree regressor uses the unique training subsets to train decision trees. As this approach, Extra Trees regressor randomly selects the value to split features and generate child nodes. Using the whole data set allows Extra Trees to decrease the bias of the model. However, the randomization of feature value at which to split nodes increases the variance and bias. According to the *Extra Tree regressor introduction* paper of Pierre, Damien, and Louis (2006), a bias-variance analysis was carried out for six different tree-based models. The paper concluded that Extra Trees has higher bias and lower variance than Random Forest. Nonetheless, the study also reported that the higher bias in Extra Trees compared to the Random Forest is due to the inclusion of irrelevant features in the models. Therefore, when irrelevant variables are eliminated, Extra Trees gets a similar bias score as that of Random Forest. With the dataset used in this paper, Extra Trees regressor not only has a better predictive performance, but also outperforms Random Forest in terms of energy consumption. It executes in shorter time, consumes less power energy, and generates less equivalent carbon emissions. The reason why Extra Trees algorithm saves more time and energy is because the tree developing procedure is the same, and it randomly selects the split point instead of calculating the optimal one.

When three imputation datasets are compared, there is no significant difference in the predictive performance of models, whereas the computational cost increases gradually from the simple imputation to KNN imputation and then iterative technique. Therefore, with the goal of obtaining a model consuming as little energy as possible, the simple imputation technique is selected.

Overfitting

Besides comparing R-squared and RMSE scores across models, we also evaluate the data overfitting to increase the robustness of our selected models. We carry out overfitting checks for the 3 best models (Gradient Boosting, Random Forest, and Extra Trees) on the most efficient imputation technique we have found (Median Imputation). Specifically, we plot a learning curve, which shows model learning performance over experience. As our best models are tree-based, we perform overfitting analysis by varying a key hyperparameter (max_depth) and evaluating the model performance (R_squared) on the train and test sets for each configuration. Shallow trees with low max_depth value cannot learn data patterns and have poor performance (high bias, low variance), whereas deep trees with high max_depth generally have good performance and do overfit (low bias, high variance). Our objective is to find an optimal point where the tree is shallow enough to generalize well to unseen data, having reasonable bias and variance.

We go through each level of tree depth from 1 to 20, where we train a model with a specific depth using the training dataset. Subsequently, we assess the performance of the model on both the train and test sets. We anticipate that as the tree depth increases, there will be an improvement in performance on both the training and test sets up to a certain extent. However, when the tree becomes excessively deep, it will start to overfit the training dataset, leading to a decline in performance on the independent test set.

The below graph shows the performance of Gradient Boosting Regressor on training and test sets. It is clear that the model is overfitting as it memorizes all points in training data but behaves poorly on test sets. After max_depth reaches 6, R-squared values obtained from the test set start decreasing.

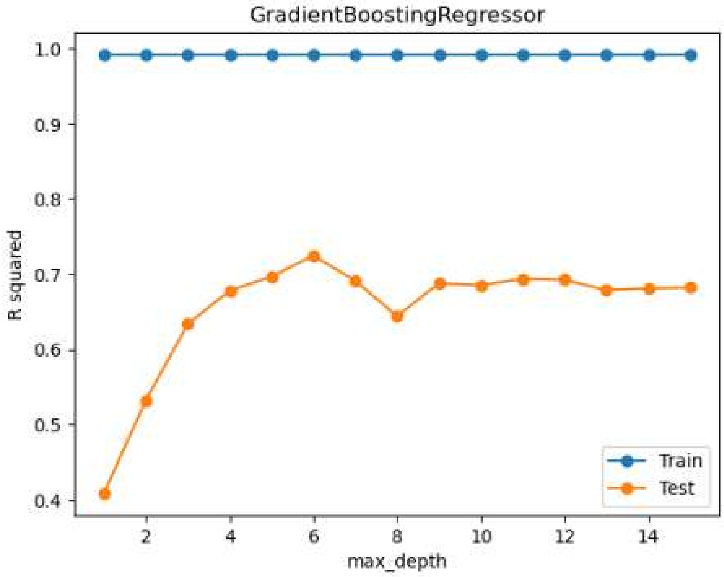


Figure 13:Test_Train split: model performance vs max_depth

For Random Forest, the performance of both train and test sets continues rising at max_depth=14 before reaching a plateau, where R squared values of training and test sets are 0.96 and 0.79, respectively. The gap between the R-squared values of the training and test sets is relatively small. This suggests that the model is not significantly overfitting the training data. Model performance on the test set indicates it has generalizability and is capable of making reasonably accurate predictions on unseen data. We choose optimal max_depth= 6 for Random Forest.

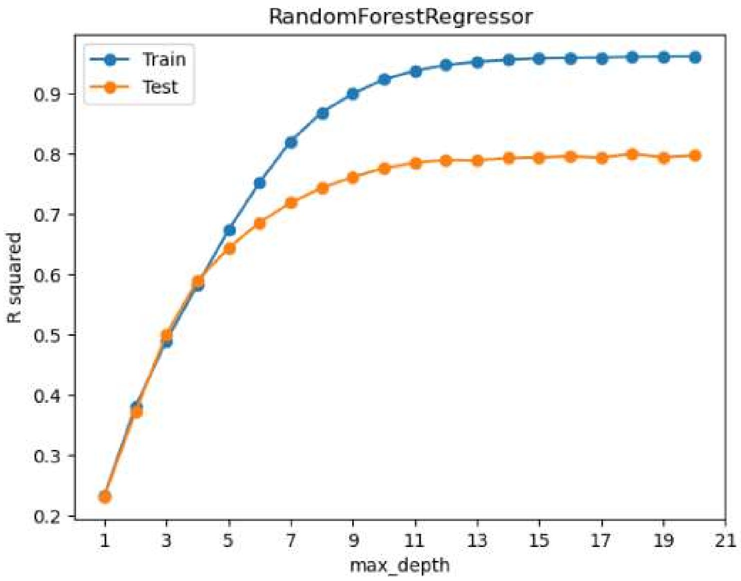


Figure 14:Test_Train split: Performance of Random Forest vs max_depth

At first, Extra Tree Regressor model seems to be slightly underfitting. A small value for max_depth restricts the depth of individual decision trees, resulting in a simple model that may not capture complex patterns effectively. As the depth of decision trees in the ensemble increases, model performance is improved. However, increasing the max_depth excessively can lead to overfitting. Hence, we set max_depth = 10 when running Extra Trees Regressor.

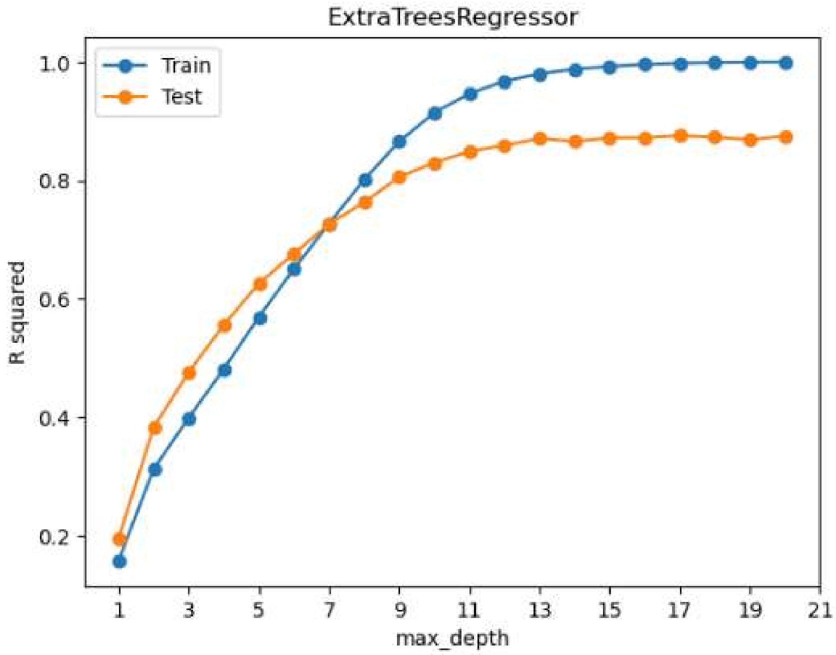


Figure 15:Test_Train split: Performance of Extra Trees Regressor vs max_depth

After finding optimal max_depth, we train the model again with new configurations and evaluate the performance metrics on the test set. The results are shown in table below

	R^2	**RMSE**	**Execution time (s)**	**Power consump-tion (kWh)**	**Equivalent carbon emissions (kg)**
Random Forest Regressor	0.72	21907	20.5036	1.002e-05	3.257e-07
Extra Trees Regressor	0.81	18085	6.2687	2.348e-07	7.52e-09

Table 9: Perfomance results of models with optimal max_depth

b. K-fold cross validation

While the train-test-split data technique is simple to examine the detailed result of the testing process, it might be not appropriate when the dataset is small. This is because when the dataset is split into training and test set, the data size in the training set is not enough for the model to learn an effective mapping of inputs to output. Similarly, the test set is also inadequate in data size to accurately assess the performance of the model. To address disadvantages of simple train-test-split data technique, k-fold cross validation technique is expected to improve the model performance in terms of overcoming overfitting and utilizing the available data. The training process is as follows: (1) The original dataset is split into training set (80%) and test set (20%), then (2) The training set is implemented k-fold cross validation with k = 10, in which the training set is split into *training folds* and *validation fold*. K value is a hyperparameter and selected by taking into account the trade-off between bias/variance and computational cost. In this paper, we selected k = 10 because this is a commonly used value and has been shown to provide a good balance between bias and variance in many cases. (3) After k-fold cross validation is carried out, the model having the highest average performance on validation sets across all folds is selected. (4) The selected model is evaluated on the test set (20%) for ensuring the effectiveness and the generalizability of the model. Based on the conclusions drawn from the train-test-split technique above, we only implement k-fold cross validation for three models: Extra Trees regressor, Random Forest, and Gradient Boosting, on three imputation datasets. The following tables present the average values of validation sets across all folds from 80% data implemented cross-validation.

Simple imputation

	R^2	**RMSE**	**Execution time (s)**	**Power consump-tion (kWh)**	**Equivalent carbon emissions (kg)**	**Energy efficiency**
Gradient Boosting	0.5939	30489.96	329.283718	0.000207	0.000006	6.7891E-09
Random Forest	0.7108	25416.86	870.785796	0.001242	0.000038	4.8865E-08
Extra Trees	0.7849	22046.41	334.615385	0.000315	0.000009	1.3807E-08

KNN imputation

	R^2	RMSE	Execution time (s)	Power consumption (kWh)	Equivalent carbon emissions (kg)	Energy efficiency
Gradient Boosting	0.5923	30388.86	345.65878	0.000241	0.000075	7.9305E-09
Random Forest	0.7058	25693.80	898.712844	0.001467	0.000045	5.7095E-08
Extra Trees	0.7763	22089.75	351.964647	0.000305	0.000095	1.4288E-08

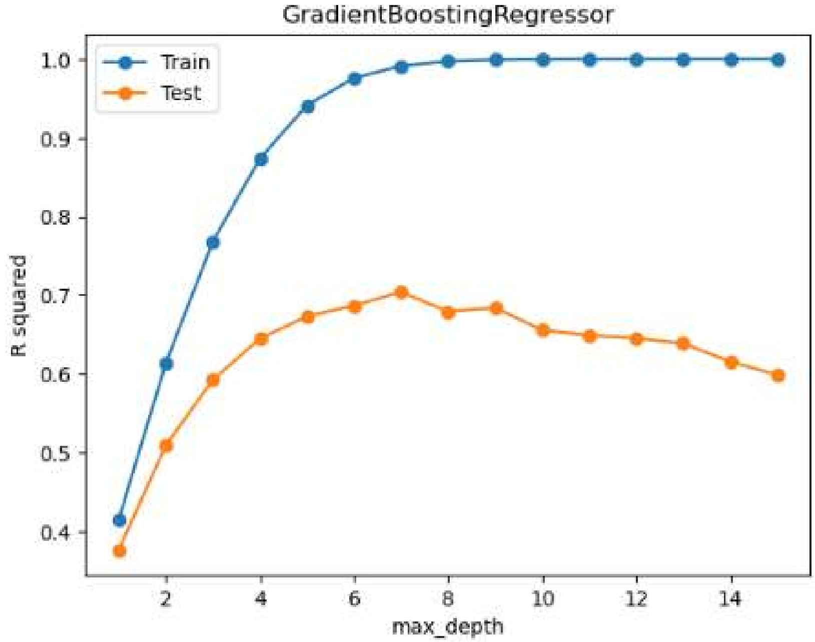
Iterative imputation

	R^2	RMSE	Execution time (s)	Power consumption (kWh)	Equivalent carbon emissions (kg)	Energy efficiency
Gradient Boosting	0.5856	30626.35	353.102129	0.000254	0.000008	8.2935E-09
Random Forest	0.7035	26000.62	882.755323	0.001328	0.000041	5.1076E-08
Extra Trees	0.7756	22374	361.642974	0.000395	0.000012	1.7654E-08

According to results presented in tables, when k-fold cross-validation is applied, Extra Trees regressor running on simple imputation dataset is still the model having the best performance in terms of prediction accuracy as well as energy efficiency. The model is then evaluated on the test set (20%).

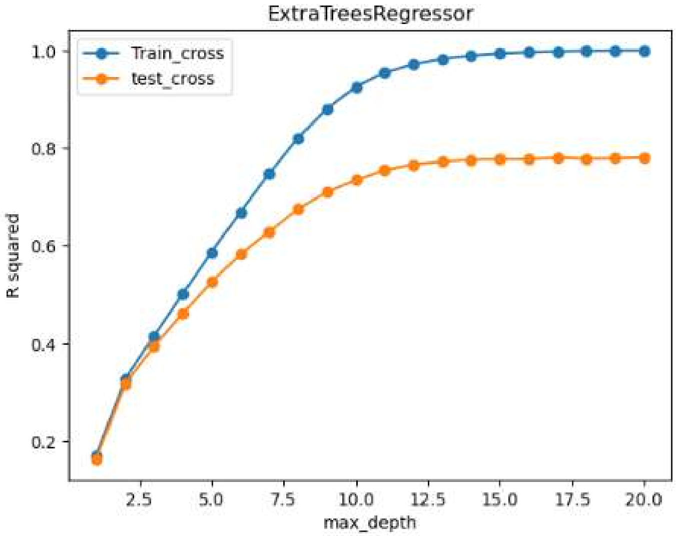
Overfitting

We compare the training performance with the validation performance (average performance across all folds) when max_depth values change from 1 to 20 to check whether there exists the overfitting issue. If the model performs significantly better on the training data compared to the validation data, it may be overfitting. However, there is no fixed threshold that definitively determines when the difference between the two performances is large enough to conclude overfitting.



Gradient Boosting is still overfitting when we apply K-fold cross validation. Average R squared value in training sets is reaching 1 but poorly performing on the validation sets. This indicates that the gradient boosting model is failing to generalize and is not the best fit for predicting scope 3 carbon emissions.

For Extra Trees Regressor, the model is slighly overfitting when value of max_depth increases. To avoid overfitting issue, we select max_depth =10 as optimal hyperparameter.



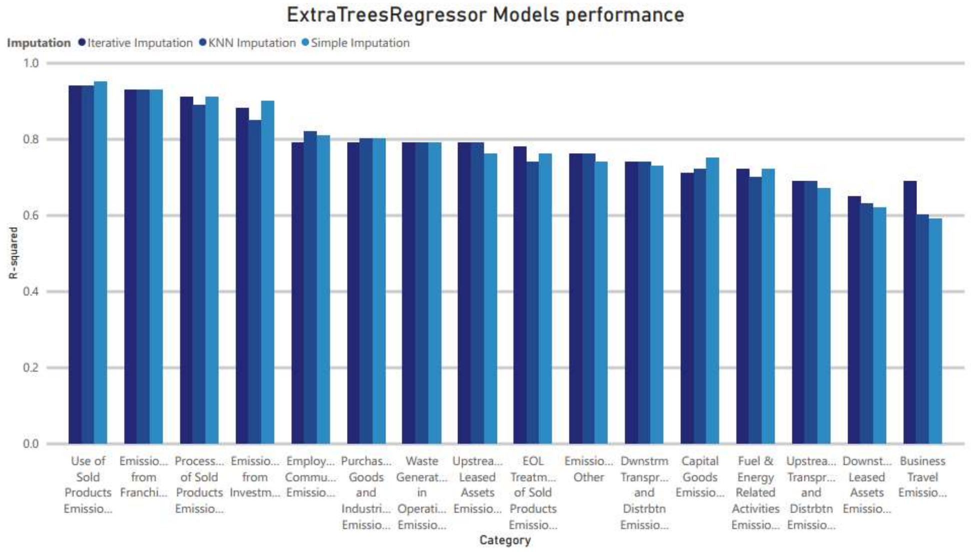
Compared with Train-Test Split technique, K-fold cross validation is able to reduce overfitting because the model is trained and evaluated on different subsets of the data. This helps to ensure that the model's performance is not overly influenced by the specific training and testing data split. Furthermore, K-fold cross validation allows for a better utilization of the available data since each data point is used for both training and testing, ensuring that the model is exposed to a larger portion of the dataset during training. However, K-fold cross validation does come with a computational cost. As the model is trained K times, the training process can be more time-consuming compared to a single Train-Test Split. The computational cost increases linearly with the number of folds used in the cross-validation process. In summary, there is trade-off between comprehensive evaluation of the model's performance and computational efficiency. The choice of splitting techniques depends on the dataset size and priority of companies. For smaller datasets or situations where computational resources are limited, a single train-test split may be more practical. On the other hand, if a more robust evaluation and better data utilization are desired, K-fold cross validation can be a suitable choice, despite its increased computational cost.

c. Feature importance

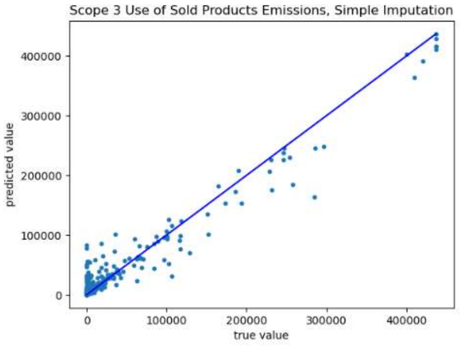
In alignment with the conclusion that Extra Trees regressor model outperforms other similar tree-based models, we proceed to examine the importance of various input features in predicting total scope 3 emissions. Extra Trees model calculated feature importance by Gini importance. In scikit-learn, Gini importance is employed to compute the node impurity and feature importance is basically a reduction in the impurity of a node. The result points out that the most significant features for predicting total scope 3 emissions include *Property Plant and Equipment net, Inventories, Cost of goods and industrial sold, Scope 1 & 2, Number of employees, ESG Disclosure score, Inventory turnover,* and *Total assets.* These important features align with our expectations. Take *Number of employees* as an example, if a company has more employees, the amount of carbon emissions associated with travelling and commuting will be higher. For manufacturing firms, *Cost of goods and industrial sold* is expected to take up a large proportion of total scope 3 emissions. In contrast, many encoded Sector features appear to have no influence on total scope 3 emissions, such as sector *Financials, Healthcare, Communication services, and Real Estate.* A possible explanation for this is because few companies operating in these sectors report total scope 3 emissions, leading to the lack of data affecting model performance.

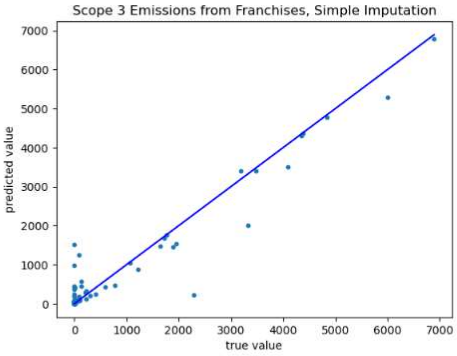
6.2. Predict scope 3 carbon emissions by categories

We apply the same data pipeline to predict 16 carbon emissions compositions. Dataset after imputation and scaling is split into 80% training and 20% test. The aim is to explore prediction power of models against breakdown emission categories. As we compare model performance across categories, R squared is the main metric we use.

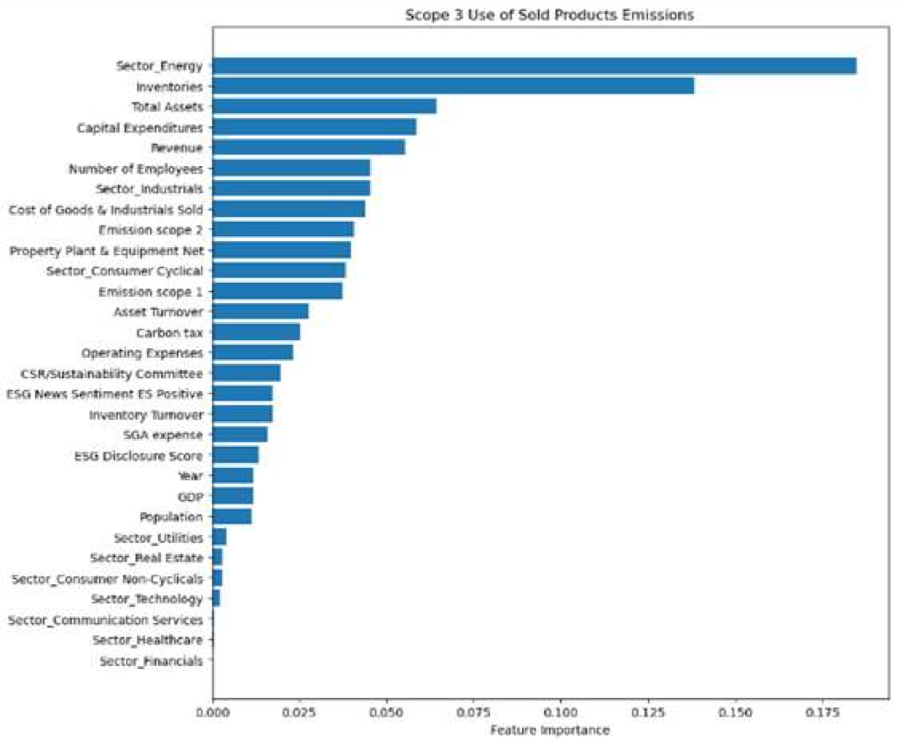


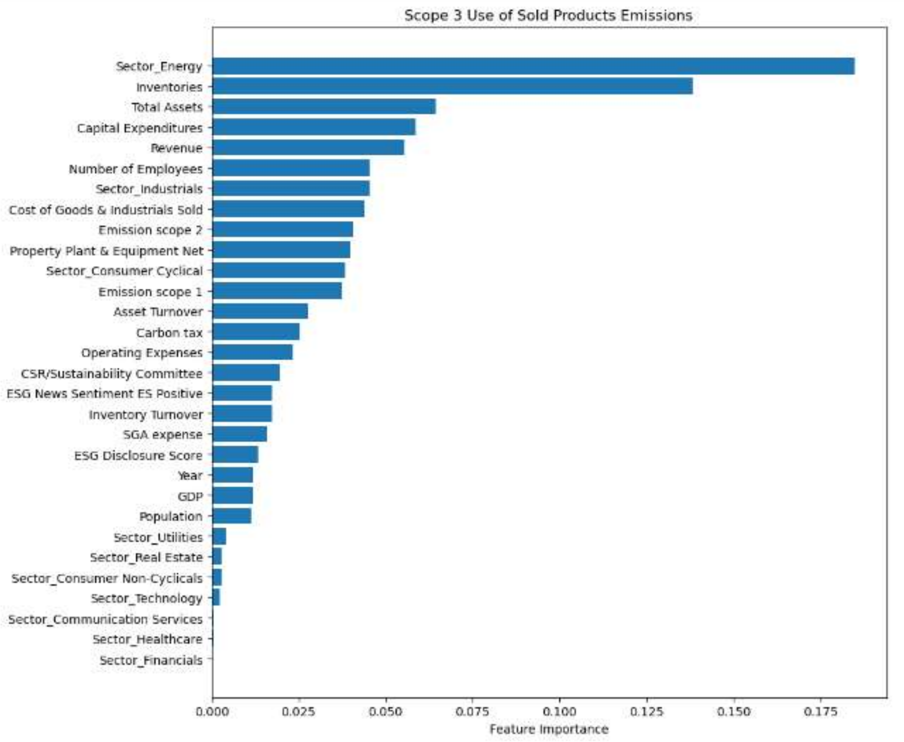
The performance of machine learning model on test set indicates that ExtraTreesRegressor is likewise the best model to predict scope 3 emissions components. A comparison of model performance across different imputed data sets is presented in the chart above. Remarkably, the R-squared values remain almost the same across the three imputation techniques, hence, the Simple imputation technique is preferred due to its ease of implementation. However, it is worth noting that the R-squared values exhibit variations across different categories. Some categories have higher R-squared values, surpassing even the total scope 3 prediction. Notably, the Use of Sold Product Emissions and Franchise Emissions categories demonstrate the highest R-squared values, approximately 0.9. This signifies that the model is highly effective in explaining the variations in emissions data. To gain further insights into the predictive power of the model, scatter plots are visualized to compare the predicted values with the observed values on the test set. Ideally, all data points should closely align with a diagonal regression line, where the predicted values match the true values. Upon examining the plots, we observe that the ExtraTreesRegressor model makes reasonably accurate predictions, even for extremely high observed values in the Use of Sold Product Emissions and Franchise Emissions categories. However, the residuals, which represent the differences between predicted and observed values, display heteroscedasticity. This means that the variance of the errors is not constant across different levels of the dependent variable. Specifically, the heteroscedasticity is pronounced, around value of zero.


Scope 3 Use of Sold Products Emissions, Simple Imputation

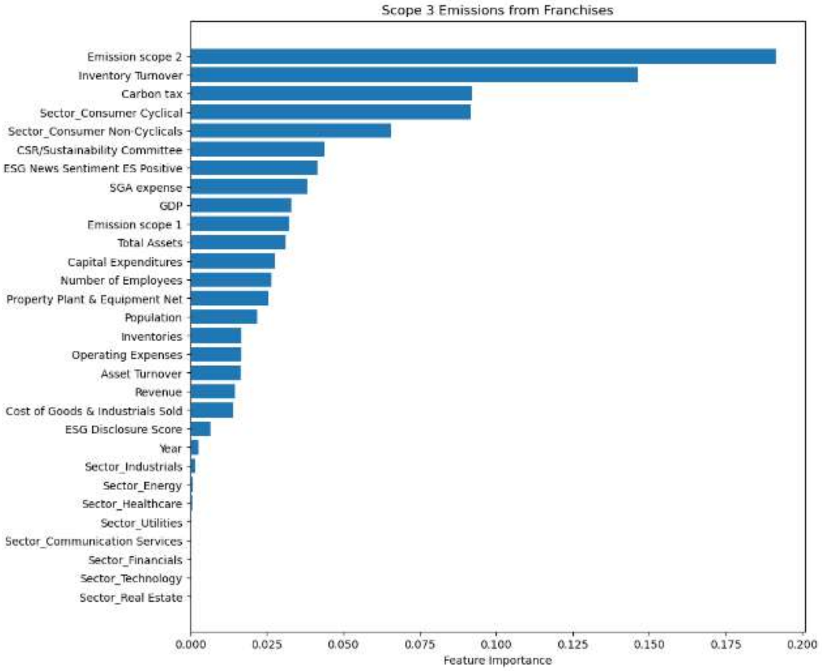

Scope 3 Emissions from Franchises, Simple Imputation

Here, we also explore feature importance to understand the underlying factors that are driving the model's predictions. For Use of Sold Product Emissions, predicted emission is highly associated with the size of the business, reflected by variables such as Inventory, Total Assets, Revenue, and Number of employees. Businesses with larger scale of operation (higher inventory levels, larger total assets, greater revenues, more employees) tend to have higher emissions resulting from product usage. Moreover, the Sector_energy variable has been identified as the most significant contributor to the performance of the model. Companies operating in this sector are likely to record higher emissions resulting from the use of their sold products. encompasses a wide range of industries, including oil and gas, electricity generation, renewable energy, and more. These industries are responsible for producing and distributing energy resources that power our modern society. Thus, emissions from sold product usages are larger in energy sector.

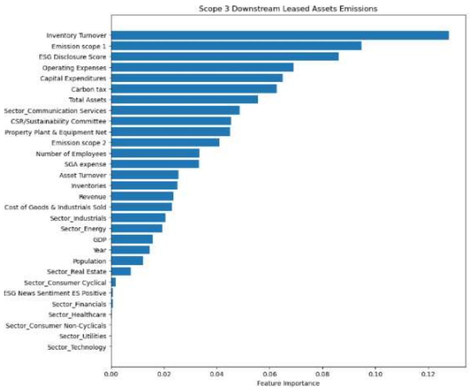


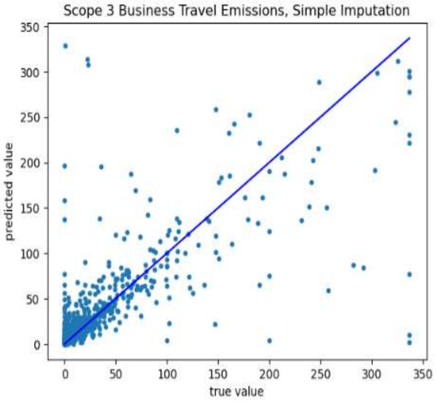


For Franchise Emissions, various factors related to the operating environment (Carbon tax, GDP) and the sustainability practices (CSR Committee, ESG News Sentiment) has a high impact. This is reasonable as Franchises operating in countries with higher GDP levels often experience increased demand for their products or services, resulting in greater production and energy consumption. On the other hand, sustainability variables such as the presence of a Corporate Social Responsibility (CSR) committee and Environmental, Social, and Governance (ESG) news sentiment also play a significant role in predicting franchise emissions. Having a dedicated CSR committee within a franchise demonstrates the company's commitment to ethical and sustainable practices. These committees oversee sustainability initiatives, assess environmental impacts, and implement strategies to reduce emissions.



In contrast, for scope 3 business travel emissions, the model is able to explain only 60% of variation in emissions data. This indicates that there are other factors influencing business travel emissions that are not adequately captured by the current model. One possible explanation for the limited explanatory power of the model in the case of scope 3 business travel emissions could be the influence of individual choices and behaviors. Factors such as employee travel preferences, booking decisions, and transportation modes used may vary widely and are difficult to quantify accurately using the available data.



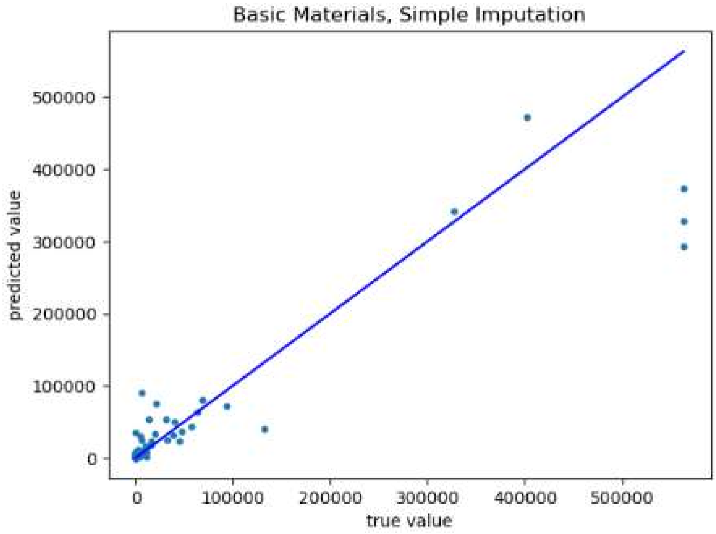


6.3. Predicting scope 3 carbon emissions by sectors

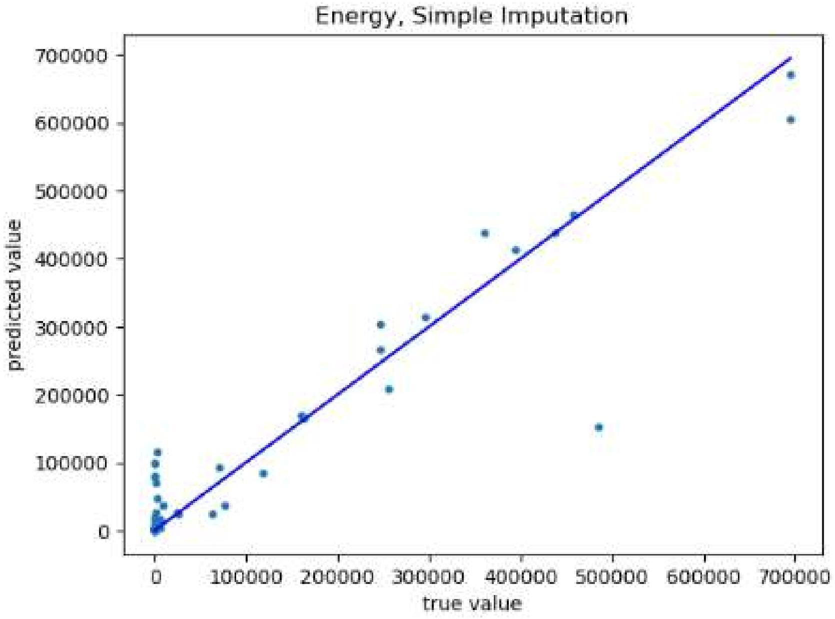
In this section, we would like to test whether the chosen models perform better when considering companies in the same sector. Companies within the same sector often share similarities in terms of their operations, processes, and emission sources, which may be crucial in accurately predicting scope 3 emissions. Moreover, by considering companies within the same sector, we are likely to have a more homogeneous dataset in terms of emission sources, measurement methods, and reporting standards. This homogeneity can reduce data variations and inconsistencies, making it easier to identify relevant features and train models effectively. We filter the original dataset based on sector and apply the same pipeline to predict scope 3 emissions for each sector. In total, we have 10 sectors: Utilities, Basic Materials, Consumer Cyclical, Financials, Industrials, Real Estate, Energy, Communication Services, Consumer Non-Cyclicals, Technology, Healthcare.

Models perform differently across sectors data set. They perform well with data in sector Utilities, Basic Materials, Consumer Cyclical, Energy, Consumer Non-Cyclicals with R squared ranging from 0.78 to 0.92, while poorly in Industrials, Real Estate, Technology and Healthcare

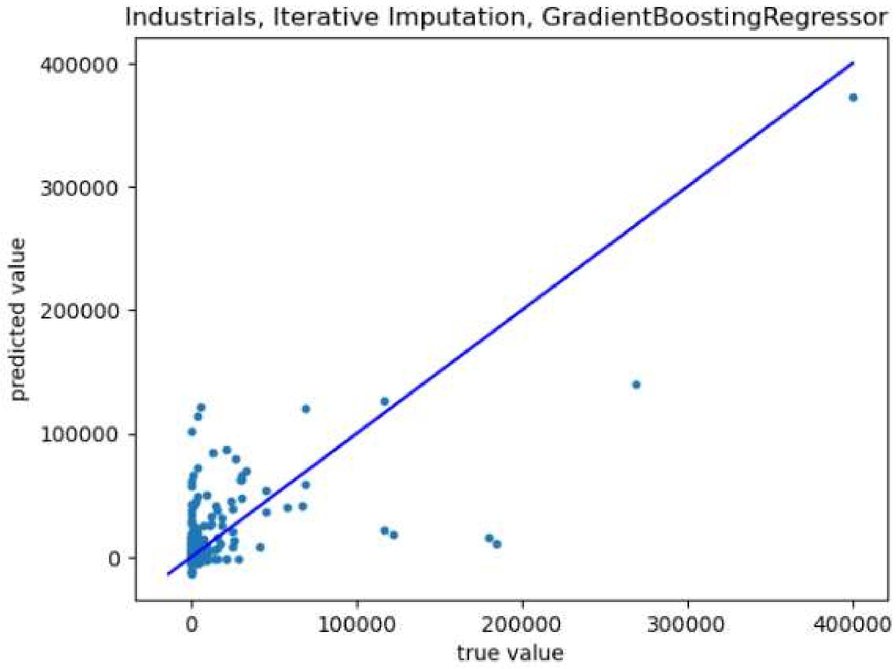
For Basic Materials sector, Extra Regressors and Simple Imputation yield best predictions ($R^2 = 0.92$). However, the models provide goood estimations only with small emissions values. There could be several reasons why the models struggle to provide accurate estimations for larger emissions values within the Basic Materials sector. Firsty, the relationship between the predictor variables and emissions becomes non-linear as the emission values increase. Extra Regressors may not capture these non-linear dynamics effectively, leading to poorer performance for larger emissions values. Secondly, outlier observations deviate significantly from the general emissions patterns. As the model is not robust to outliers, their performance deteriorates when dealing with larger emissions values.



Energy Sector emissions are also well-predicted by Extra Regressors and Simple Imputation with R square 0.91. In contrast with Basic Materials, the model performs quite good with high value. With companies reporting small emissions, model cannot provide reasonable estimations.



In the Healthcare and Industrials sectors, the model has a low performance. Data scarcity poses a significant hurdle for emission prediction models in these sectors. The size of data is insufficient enough for the model to learn and capture patterns accurately. Thus, we suggest that more data needs to be collected to train the model for emission prediction in the Healthcare sector.



Chapter 7: Discussion

The paper shows that there is no decline in the volume of carbon emissions generated during the pandemic covid-19. On the contrary, the trend demonstrates a continuous increase from 2015 to 2022. Based on this evidence, it is possible to conclude that companies are persistently generating escalating carbon emissions for their production and operation purposes no matter what globally unexpected events suddenly happen. Accordingly, this conclusion answers our first research question of whether scope 3 emissions are affected by covid -19. Given the current reporting scope 3 emissions status and the increasing urgency of reducing global carbon emissions, this paper proposes a machine-learning solution for calculating scope 3 emissions. By leveraging a broadly available dataset, the machine learning approach appears to bring a better result in estimating scope 3 emissions compared to previously used calculation methods such as conventional linear regression model. In pursuit of an energy-efficient and easily implementable model, the study demonstrates that the Extra Trees regressor model produces the best performance in terms of predictive accuracy as well as energy efficiency. Specifically, the selected model has a good R-squared score, the lowest error calculated by the difference between the actual values and the predicted values of data points, and consumes as little energy as possible. Input features selected for the predictive model present their importance and contribution to predicting the target variables. The robustness of the model's outcome is further ensured through hyperparameters tunning and overfitting treatment. Besides having total scope 3 emissions as the predictive target, we also predict scope 3 emissions by 16 different categories and 10 mainly reporting sectors. The paper points out that the scope 3 emissions volume as well as model performance varies across categories and sectors.

Calculating scope 3 emissions has consistently presented a challenge for many companies, especially small and medium–size ones. The scarcity of data, limited control over suppliers' and customers' decisions, and the inherent difficulty in calculation are all contributing factors. As a result, many firms lack the necessary resources and capabilities to accurately measure their scope 3 emissions (George & Gladys, 2022). Therefore, we strongly believe that implementing machine learning for estimating scope 3 emissions is a potential approach that promises to enable more short-resource companies to calculate and report these value chain carbon emissions efficiently. Once companies obtain better control over their value chain emissions, they could utilize the resource, hence identifying opportunities to reduce carbon emissions volume. Moreover, whereas there are numerous effective machine learning algorithms supporting us to accomplish this objective, we had better take the trade-off between the efficiency and the sustainability of the model into consideration. For this reason, we could achieve a double target of implementing a sustainable approach to solve a sustainable problem.

However, there are several existing limitations in this paper that could be solved and improved by future research. Firstly, the number of observations is quite small due to the limit on data access. After removing all missing data from the target variable, which is Scope 3 total emissions, our dataset remains 6,734 company-year observations. We collected data from multiple sources including mainly CDP, Bloomberg, and Refinitiv Eikon. Even though BI Norwegian Business School supports us in offering Bloomberg accounts to collect data, there is a monthly limit on the number of requests that we could send to collect data. Besides, as specifically mentioned in the Data Collection part, we must manually collect data from these sources in alignment with the time limit, leading the data collected to concentrate on North American, Scandinavian, European, and some Asian countries. For this reason, the data might not reflect the overall status of scope 3 emissions reporting all around the world. Secondly, more predictive features could be added to improve the prediction metrics of models. The primary objective of implementing machine learning algorithms in this paper is to utilize available scope 3 emissions reports of disclosing companies to train the model and predict the result for non-disclosing companies. Therefore, we prioritize using broadly available accounting data and the mandatory scope 1 and scope 2 emissions. This approach enables companies to estimate scope 3 emissions more easily and save company's resources. However, many important features could have been missed, such as carbon intensity, product choices, or the monopoly of the supply chain, which could have impacts on scope 3 emissions. Concerning sustainable machine learning, we succeeded in measuring and comparing the energy consumption and carbon emissions equivalent of different missing data imputation techniques as well as machine learning algorithms. The results show that the Extra Trees Regressor algorithm developed with the simple imputation technique gives the best performance in terms of prediction metrics and energy consumption. However, with this outcome, we are being limited to Python's package without having cross-comparison to other