Climate Change Risk For Financial Institutions: Predicting Corporate Greenhouse Gas Emissions

Ties Rothman
University of Twente
September 8, 2023

Abstract

In this research, we investigate the use of statistical analysis and machine learning methods to predict corporate greenhouse gas emissions. We trained and tested these models on corporations that disclose emission data, aiming to create models applicable to corporations that do not disclose this information. Due to a scarcity of Environmental, Social, and Governance (ESG)-related data, we used financial, geographical, and sector classification data as predictor variables. We applied a log transformation to both the predictor and output variables. Multiple imputation was employed to handle missing data, thereby enlarging our dataset while preserving underlying variable distributions. We evaluated the models in three rounds of testing: first on the imputed data, then on baseline data, and finally on baseline data after correcting for log transformation bias. In the log-transformed feature space, the models accurately predict corporate greenhouse gas emissions. However, in the original feature space, they fail to provide accurate predictions. Our findings suggest that the models struggle with the complexity of the data and do not generalize well. For more accurate predictions, additional ESG-related data, as well as information on production processes, materials, and other physical assets, are needed.

Faculty of Behavioral, Management, and Social Sciences
Industrial Engineering and Management - Financial Engineering and Management

University Supervisors: Prof. Dr. L. Spierdijk, Dr. B. Roorda
Company Supervisor: R. van der Sande

Contents

List of Figures

List of Tables

Nomenclature

CH_4	Methane
CO_2	Carbon dioxide
CO_2e	Carbon dioxide equivalents
N_2O	Nitrous oxide
NF_3	Nitrogen trifluoride
SF_6	Sulfur hexafluoride
AI	Artificial Intelligence
ANN	Artificial Neural Network
BICS	Business Industry Classification System
CSRD	Corporate Sustainability Reporting Directive
CV	Cross-Validation
ESG	Environmental, Social, and Governance
EU	European Union
FCS	Fully Conditional Specification
GHG	Greenhouse Gases
GICS	Global Industry Classification Standard
GWP	Global Warming Potential
HFCs	Hydrofluorocarbons
IPCC	Intergovernmental Panel on Climate Change
MAPE	Mean Absolute Percentage Error
MAR	Missing at Random
MICE	Multiple Imputation by Chained Equations
MI	Multiple Imputation

MSE	Mean Squared Error
NAICS	North American Industry Classification System
NFDR	Non-Financial Reporting Directive
PCAF	Partnership for Carbon Accounting Financials
PFCs	Perfluorocarbons
PMM	Predictive Mean Matching
RBF	Radial Basis Function
RMSE	Root Mean Squared Error
ROA	Return on Assets
ROC	Return on Capital
ROE	Return on Equity
RSS	Residual Sum of Squares
SME	Small- and Medium-sized Enterprises
SVM	Support Vector Machine
UN	United Nations
VIF	Variance Inflation Factor
XGBoost	Extreme Gradient Boosting

Chapter 1

Introduction

In the Chapter 1, we introduce the company where the research was conducted, discuss the context in which the research is performed, identify the core problem, define the research questions, discuss the research methodology, and give the research outline.

1.1 Company Introduction

Zanders was founded in 1994 and, in recent years, has grown into a leading consultancy firm specializing in treasury management, risk management, and corporate finance. The company has more than 250 consultants working for over 500 clients. The head office is based in Utrecht, the Netherlands. Other offices exist in Belgium, Germany, Switzerland, Sweden, the United Kingdom, the United States, and Japan. This research was carried out in the Financial Institutions department, which provides financial risk advice to banks, insurers, and asset managers. Within the Financial Institutions department, multiple employees work together in an Environmental, Social, and Governance (ESG) expert group which focuses, among others, on climate change risks. All banks are currently investigating, measuring, or modeling climate change risks in their portfolio. One of the components of climate change risks relates to measuring Greenhouse Gas (GHG) emissions. Reliable GHG emission data could ultimately be used in risk management practices and publication purposes. Unfortunately, not all corporations disclose GHG emission data. This research focuses on the implementation of statistical and machine learning methods to predict GHG emissions for corporations that do not disclose this information, based on corporations that do disclose this information.

1.2 Problem Context

On December 12th, 2015, 196 Parties adopted a legally binding, international UN treaty on climate change to limit global warming to well below 2, preferably 1.5 degrees Celsius, compared to pre-industrial levels. The formulated strategy to reach this goal is to reach global peaking of GHG emissions as soon as possible and, thereafter, undertake rapid reductions to achieve a carbon-neutral climate by 2050 (UNFCC, 2018). Financial institutions have an important role in the transition towards a carbon-neutral climate. First, because financial institutions are the instrument through which large investments can be facilitated to adhere to the Paris Agreement, and, second, because the effects of climate change result in financial risks that affect the financial system (González and Soledad, 2021). Climate change risks for financial institutions can be divided into physical risks and transition risks. Physical risks occur through gradual changes in ecosystems or sudden extreme weather phenomena, otherwise known as chronic and acute changes. Both categories lead to physical damage to assets, disruptions to supply chains, or expenditures to prevent these damages and disruptions (An et al., 2022). Transition risks occur through the transition towards a low-carbon economy that results in changing policies, regulations, technologies, and consumer preferences (González and Soledad, 2021). Given that we concern ourselves with the prediction of GHG emissions, this research focuses on transition risks.

The transition towards a low-carbon economy is not going as planned in the Paris Agreement. A German study stated that meeting the 1.5 degrees Celsius goal is already not plausible, and limiting the temperature rise to well below 2 degrees Celsius can only become plausible if efforts are rapidly increased (Engels et al., 2023). A more rapid transition poses risks for financial institutions that are exposed to corporations with business models that are unaligned with the transition towards a low-carbon economy, such as fossil fuel sectors and corporations with high GHG emissions (Nguyen, Diaz-Rainey, Kuruppuarachchi, et al., 2023). These exposures lead to increased levels of the traditional types of risk that financial institutions face. Credit risk is increased when loans are issued to corporations that fail to adapt, risking fines and loss of customers, resulting in increased probabilities of default. Market risk is increased when the transition towards a low-carbon economy leads to sudden repricing of assets, or to markets being dissolved over time. Other risks that may increase are, e.g., operational risk, liquidity risk, reputational risk, legal risk, and model risk. The increased level of risk that financial institutions face due to transition risk is likely to be significant. A transition risk stress test, using data from more than 80 Dutch financial institutions, showed that portfolio values can decline by up to 11%, and CET1-ratios can decline by up to 4.1% (Vermeulen et al., 2021).

1.3 Core Problem

The international trajectory towards a low-carbon economy has led financial institutions to evaluate their climate change exposures. Where physical risk is relatively easier to grasp due to its physical materialization, mapping transition risks is more of a challenge. A method often recognized in literature as a method to quantify transition risk is the evaluation of financial institutions' carbon footprints, also referred to as carbon footprinting (Yang, Li, and Pan, 2022). Investments and loans to corporations with large carbon footprints may lead to increased exposures to transition risks, as large carbon footprints may imply that these corporations are lagging in the transition towards a low-carbon economy. Carbon footprinting, thus, enables financial institutions to understand and monitor their climate change risks while simultaneously helping steer towards global climate change reduction goals (PCAF, 2019).

There is a need for reliable publicly available emission data, not only in order to incorporate corporations' GHG emission data into risk management practices, but also for reliable reporting of emissions. In 2014, an EU Directive stated that undertakings are required to prepare non-financial statements that should contain, among others, GHG emission data (EU, 2014). Problems as to the effectiveness of this directive were identified in a commission report which stated that there is significant evidence that many undertakings do not disclose material information on all major sustainability-related topics, including GHG emissions (EU, 2022a). In practice, we see that only a subset of corporations do disclose GHG emission data. On January 5th, 2023, the Corporate Sustainability Reporting Directive (CSRD), a new EU reporting directive, entered into force. With this directive, the set of corporations that are required to report on non-financial data is increased from 11,700 to approximately 50,000 corporations, as compared to the policies introduced in the Non-Financial Reporting Directive (NFDR) (EU, 2022b), the predecessor of the CSRD. Although the set of large corporations, as well as listed small- and medium-sized enterprises (SME), is significantly increased, there is still a large portion of SMEs that is not covered by the new directive. Thus, in the coming years, a large portion of SMEs will not yet be obliged to disclose their GHG emissions. This issue is the main motivation for this research. We will focus on creating models that predict GHG emission data for individual corporations. Subject to this, we may also question the reliability of GHG emission data that is disclosed, as was justified following the Volkswagen Diesel Scandal (Aurand et al., 2018). In such a scenario, a GHG emission prediction model could act as a benchmark to detect outliers in disclosures.

1.3.1 Previous Research

Relatively little research is done on predictions for individual corporations. Current research on predicting GHG emissions focuses mainly on predictions for sectors (Pandey and Agrawal, 2014), regions (Franco et al., 2022), and countries (I. Ulku and E. E. Ulku, 2022), for which a variety of regression and machine learning methods are used. In our search for previous research on predicting GHG emissions for individual corporations, we found 5 research articles and 1 white paper, which focused on GHG emission prediction through regression analysis, as well as machine learning methods.

First, there are two studies that focus solely on regression analysis. One of these studies generated an estimate of GHG emissions by regressing emissions for disclosing firms on a linear combination of several relevant variables used for a study on the effect of emission disclosure on equity value (Griffin, Lont, and Sun, 2017). The other study tested 5 hypotheses that focused on the effect of single variables on carbon footprint (Goldhammer, Busse, and Busch, 2017). Both report on the R^2, which measures a model's goodness of fit, but do not report on the models' accuracies. The first research on GHG emission prediction using machine learning follows up on the two previously discussed studies by implementing several machine learning methods and comparing these methods to the proposed regression methods. The best-performing model achieves an accuracy gain of 25% and 30% based on Mean Absolute Error (MAE) (Nguyen, Diaz-Rainey, and Kuruppuarachchi, 2021). Finally, there are three papers that are not linked to the previously mentioned papers. These three focus on creating Gradient Boosting Decision Trees (GBDT) which are compared to 'baseline' models, which are a variety of regression methods (Han et al., 2021) (Bloomberg, 2021) (Assael et al., 2023).

1.4 Research Problem

Financial institutions want to quantify their transition risk exposures by attaining carbon footprints for underlying corporations. The fact that a large subset of corporations does not disclose the information needed to map the GHG footprints results in the need for GHG emission prediction modeling. This research focuses on the prediction performance of traditional statistical analysis and the prediction performance of several machine learning methods. By doing so, we intend on mapping the emissions of SMEs that are not theoretically subject to new legislation. Hence, we have the following research question:

What model is best suited for the prediction of GHG emissions of corporations?

Following the research question, we state the following hypothesis:

Machine learning methods significantly outperform naive prediction and statistical analysis.

To approach the research question in a structured manner and either accept or reject our hypothesis, several sub-research questions are defined. The sub-research questions will help answer the research question and are listed below with a brief motivation:

1) What is the impact of climate change on financial institutions?
Prior to starting with prediction modeling of GHG emissions, it is important to understand what the climate change risks for financial institutions are, and why and how GHG emission predictions could be of value for financial institutions.

2) What are appropriate statistical and machine learning methods from literature for the application of GHG emission prediction?
We will assess and select appropriate statistical and machine learning methods from literature.

3) What variables are best suited as GHG emission predictors?
We will assess what variables can best be implemented in the prediction models in terms of prediction power. Here, the key is to select variables that are publicly accessible to financial institutions.

4) What is an appropriate way to compare traditional statistical analysis with machine learning methods?
As stated in the research question, we want to compare the prediction performance of traditional statistical analysis with the prediction performance of machine learning methods. A framework is proposed wherein the performance of the proposed methods will be assessed in terms of accuracy and applicability.

1.5 Problem Approach

This section gives an outline of the plan of approach for solving the research questions. Next to describing the methodology, we will discuss the outline of the thesis.

1.5.1 Methodology

The methodology for answering the research questions composes of qualitative research and quantitative research:

1. Qualitative Research
   The qualitative research consists of a literature study which enables us to create a theoretical framework necessary to understand the outcomes of this research. Furthermore, a literature study is conducted to create a framework wherein appropriate performance measures are defined for the comparison of the different prediction models. Finally, before doing quantitative research on predictors, we conduct qualitative research to find possible variables that theoretically could serve as important predictors.

2. Quantitative Research
   For the quantitative research, we need access to data from GHG emission-disclosing corporations. The first step is processing and visualizing this data. From there, we can create and tune the prediction models. The data will be extracted from Refinitiv Eikon, which has a large database of corporations with included emission data and other relevant financial data. The prediction models that we will focus on are regression models and machine learning models, specifically, tree-based models, Artificial Neural Networks (ANN), and Support Vector Machines (SVM).

1.5.2 Thesis Outline

We will give an overview of the outline of the thesis so that the reader has a clear overview of what to expect while reading the thesis. Chapter 2 provides the theoretical context in which the research is placed. First, we discuss climate change risk and its implications for financial institutions. Second, we evaluate and select important predictors based on previous research and literature. Chapter 3 discusses the methodology, specifically the prediction models that are used and the comparative framework through which the prediction models are evaluated. Chapter 4 describes the process of data selection and preparation, and provides a visualization of the used data. Chapter 5 describes the process of prediction modeling, which contains descriptions of model evaluations and improvement, the model results, and the feature importance. In Chapter 6, we formulate our final conclusions and reflect on these conclusions in the discussion.

Chapter 2

Theoretical Context

Chapter 2 describes, in further detail, the context in which we aim to create GHG prediction models, namely climate change risk. We discuss the global effects of climate change, the risks of climate change for financial institutions, and the increasing need for GHG accounting. Furthermore, we address previous research and literature to evaluate possible predictors.

2.1 Climate change risk

This section elaborates on the context of the research. First, we will describe the significance of climate change. Second, we will describe the way in which climate change risk impacts financial institutions, namely through risks and regulations. We want to emphasize that, considering the concept of double materiality, we are aware that climate change risk does not only have an impact on financial institutions, but that financial institutions also have an impact on climate change. Although the application of our research could be considered for both materialities, this section focuses on the former materiality.

2.1.1 Climate change

Climate change is one of the most significant global issues facing humanity today, with far-reaching impacts on the environment, economy, and society as a whole. Earth's climate is changing at an unprecedented rate, primarily due to human activities that have been releasing greenhouse gases into the atmosphere. These gases trap heat from the sun, leading to an increase in global temperatures and causing a wide range of environmental and social consequences. The Intergovernmental Panel on Climate Change (IPCC), a scientific body established by the United Nations, has been studying the causes and impacts of climate change for several decades. According to the IPCC's Fifth Assessment Report, published in 2014, "human influence on the climate system is clear, and recent anthropogenic emissions of greenhouse gases are the highest in history" (IPCC, 2014). In 2021, a contribution to this Sixth Assessment was published which focused on the latest physical science basis of climate change. Figure 2.1 shows the change in global surface temperature as observed and simulated using human and natural factors, and only natural factors (IPCC, 2021). The report states that the average global temperature has increased by approximately 1.2 degrees Celsius since 1880. Through the differentiation between human and natural factors, and only natural factors, it can be clearly observed what effect human activities have on global warming.

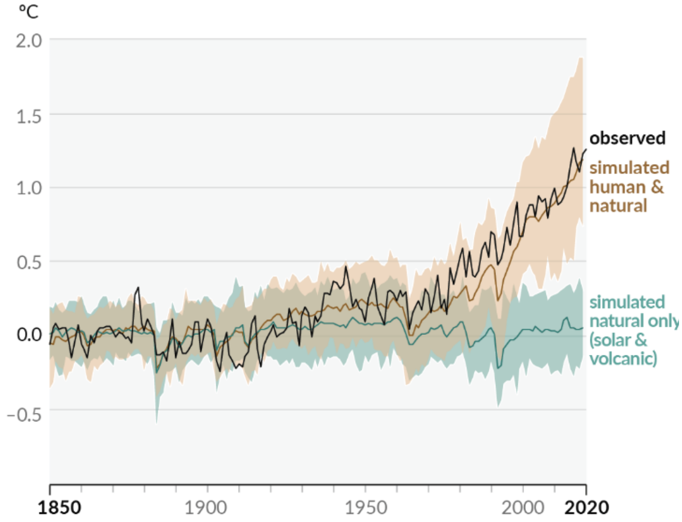


Figure 2.1: Change in average surface temperature (IPCC, 2021).

One of the most significant consequences of climate change is the melting of glaciers and polar ice caps, which leads to rising sea levels. Sea levels have risen by around 26 centimeters since 1880, and this rate is accelerating (IPCC, 2021). The IPCC predicts that sea levels could rise by up to 98 centimeters by the end of the 21st century if emissions continue to increase at current rates. This already affects several island groups, and could also have catastrophic consequences for coastal cities and low-lying regions, leading to increased flooding, erosion, and displacement of people and wildlife. Another major impact of climate change is changes in rainfall patterns, and the frequency and intensity of extreme weather events such as floods, droughts, heatwaves, and hurricanes. In 2022, monsoon rains caused severe floods in Pakistan, causing one-third of the country to be flooded. Such events have severe economic, social, and environmental impacts, causing crop failures, infrastructure damage, and loss of life. The IPCC predicts that extreme weather events will become more frequent and intense in the coming decades, especially in regions that are already vulnerable to these hazards. Climate change also affects ecosystems and biodiversity, leading to changes in species distributions, ecosystem productivity, and the spread of invasive species and diseases. The Partnership for Biodiversity Accounting Financials (PBAF) states that biodiversity is declining fast and that this decline is undermining nature's productivity, resilience, and adaptability, fueling risk and uncertainty for our economies and well-being (PBAF, 2022).

2.1.2 Risks for financial institutions

Section 1 briefly introduced the concept of climate change risk for financial institutions. As stated in Section 1, there are two main drivers of climate change risks for financial institutions: physical risks and transition risks. Although this research mainly focuses on transition risks, we will also treat the concept of physical risks for the sake of clarity.

- **Physical risks**
  Physical risks arise due to the physical materialization of climate change as discussed in section 2.1.1. The impacts of physical risks include the costs and losses that are incurred due to alterations in ecosystems (e.g. rising sea levels) and extreme weather phenomena (e.g. heavy precipitation that results in floods). The former is often referred to as chronic risks, whereas the latter is often referred to as acute risks. The physical risks are no longer risks that can be referred to as "risks of the future". Worldwide economic costs have exceeded the 30-year average more and more in the past 15 years. Furthermore, the number of extreme weather events has more than tripled since the 1980s (NGFS, 2019).

- **Transition risks**
  Physical risks, and even more the prospect of future physical risks, are the origin of transition risks. Physical risks have resulted in the adoption of laws, regulations, and directives that aim to decarbonize the economy as quickly as possible. Transition risks arise from this transition towards a low-carbon economy. As global efforts to address climate change intensify, there

is a growing recognition that the transition towards a low-carbon economy will have far-reaching impacts on economic activities, markets, and financial institutions. Transition risk drivers, such as changes in policies, technologies, and consumer preferences, lead to economic costs (e.g. stranded assets, unemployment), and financial impacts (e.g. portfolio losses, higher default probabilities). The economic costs and financial impacts directly result in macroeconomic impacts which can, again, fuel the transition risk drivers (Semieniuk et al., 2021).

Physical risks and transition risks both materialize through the traditional risks that financial institutions face, such as credit risk, market risk, operational risk, liquidity risk, reputational risk, legal risk, and model risk. Table 2.1 shows an overview of the possible implications of physical risks and transition risks for these traditional financial risks.

Table 2.1: Implications of physical risks and transition risks for traditional risks for financial institutions.

	Physical Risks	**Transition Risks**
Credit Risk	Increased expected loss due to physical damages to assets/supply chains.	Increased expected loss due to business models not being aligned to the transition towards a low-carbon economy.
Market Risk	Increased price volatility due to unavailable products/ processes caused by extreme weather.	Sudden repricing of assets.
Operational Risk	Extreme weather may affect business continuity.	Threat of greenwashing and fraud.
Liquidity Risk	Physical damages can affect the liquidity of assets (e.g. properties), or lead to increased withdrawals.	Liquidity can be affected by the shift in market sentiment towards low-carbon investments, causing increased withdrawals.
Reputational Risk		Increased risk if business models are not aligned with the transition towards a low-carbon economy.
Legal Risk		Increased risk of litigation if businesses do not adapt quickly to new regulations.
Model Risk	The incorporation of physical risks in e.g. credit risk models leads to risk due to its novelty.	The incorporation of transition risks in e.g. credit risk models leads to risk due to its novelty.

2.1.3 GHG accounting

As stated in Section 1.3, there is a need for reliable publicly available emission data in order to incorporate corporations' GHG emission data into risk management practices, as well as to comply to new regulations. The measurement of the amount of GHGs generated, avoided, or removed by a corporation is referred to as GHG accounting (PCAF, 2022). Loans and investments do not necessarily result in GHG emission generation. Projects that are focused on removing GHG from the atmosphere or projects that avoid GHG emissions are also considered in GHG accounting. The Kyoto Protocol defined what gases we should see as GHGs. These are the following: carbon dioxide (CO_2), methane (CH_4), nitrous oxide (N_2O), hydrofluorocarbons (HFCs), perfluorocarbons (PFCs), sulfur hexafluoride (SF_6), and nitrogen trifluoride (NF_3) (Brander, 2012). These gases have different properties and therefore differ in their impacts on global warming. In order to make comparisons between the gases possible, the Global Warming Potential (GWP), a GHG metric, is used to normalize GHGs as carbon dioxide equivalents (CO_2e) (Neubauer, 2021).

GHG accounting is one of the building blocks that allow financial institutions to quantify their transition risks and enables them to publish reliable GHG emission data. Comprehensive GHG accounting demands more data than one might initially anticipate. The complexity of GHG accounting arises from the fact that GHG emissions fall into two broad categories: direct emissions and indirect emissions. Direct emissions are generated by sources owned or controlled by the reporting corporation. Indirect emissions are generated by sources within the operating cycle owned or controlled by other corporations. These direct and indirect emissions are further categorized into three scopes: (PCAF, 2022)

- **Scope 1**
  Direct emissions generated by sources owned or controlled by the reporting corporation.
- **Scope 2**
  Indirect emissions generated by sources from which electricity, steam, heating, and cooling are purchased for use by the reporting corporation.
- **Scope 3**
  All other indirect emissions from activities from other corporations/entities, but part of the reporting corporation's operating cycle. These activities are divided into upstream and downstream activities.

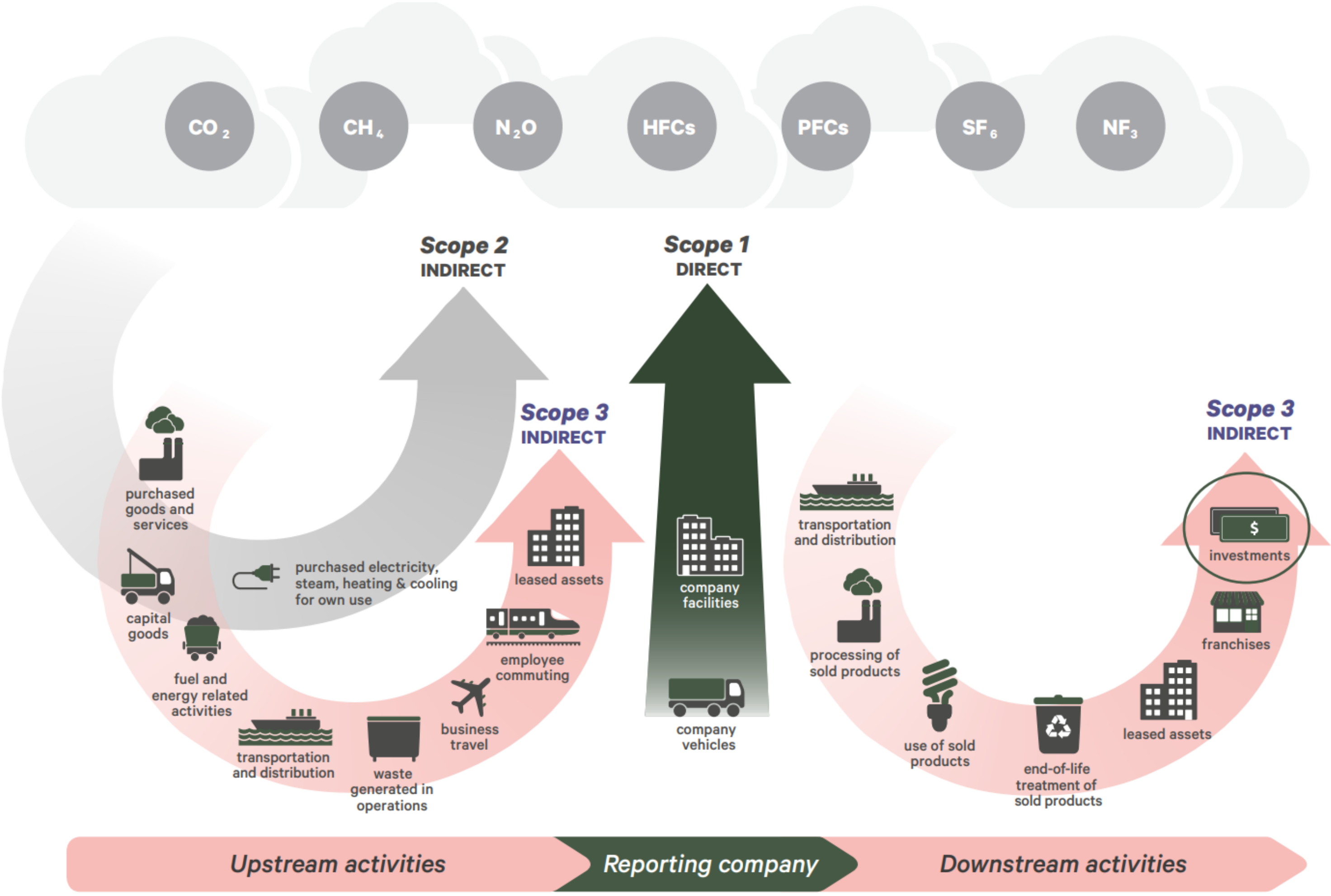


Figure 2.2: Schematic of scope 1, 2, and 3 emissions for reporting corporations (PCAF, 2022).

Figure 2.2 gives a clear overview of the direct and indirect emissions for a reporting corporation. It can be seen that the 'investments' category of scope 3 is circled. The 'investments' category, also known as scope 3 category 15 emissions, are called financed emissions. For financial institutions, financed emissions are the highest contributor to their overall emissions. The question then arises of how financial institutions should report their financed emissions. In general, there are three approaches to measuring financed emissions: the equity share approach, the financial control approach, and the operational control approach (PCAF, 2022).

- The equity share approach requires an organization to report emissions according to the share it has in the underlying organization. For example, holding an equity share of 15% means that 15% of all emissions (scopes 1, 2, and 3) should be reported as financed emissions in the respective scopes.
- The financial control approach requires an organization to report on 100% of emissions for all activities where it can directly influence financial and operational policies or benefit from the corporation's activities.
- The operational control approach requires the reporting of 100% of emissions from operations for which it can introduce and implement operational policies.

The Partnership for Carbon Accounting Financials (PCAF) requires financial institutions to report their financed emissions using the financial control approach or the operational control approach, as a loan or investment is not meant to result holding a controlling interest. As financial institutions do not have 100% control over underlying corporations, an attribution factor is used for the calculation of financed emissions. The attribution factor is calculated as follows (PCAF, 2022):

$$AttributionFactor = \frac{OutstandingAmount}{TotalEquity + Debt} \tag{2.1}$$

To calculate the financed emissions, the attribution is calculated. In order to comprehensively report all its financed emissions, financial institutions need corporations' GHG emission data for all three scopes. This again emphasizes the need for GHG emission data for scopes 1, 2, and 3.

2.2 GHG predictor evaluation

The goal of the research is to create prediction models that predict GHG emissions for corporations that do not disclose GHG emission data. We aim to do this by fitting prediction models on corporations that do disclose this information. To create a well-performing prediction model, it is important to find the right combination of predictor variables. In this section, we will explore what predictor variables were employed in previous research, what the rationale is behind the application of these variables, and how significant their contribution is to the overall prediction performance of the models. These include sector classifications, geographical information, financial information, energy consumption levels, and ESG scores. Furthermore, we assess corporation performance indicators from literature based on the sub-categories performance, efficiency, leverage, and liquidity.

2.2.1 Sector classification

The sector in which a corporation operates is likely to be a relevant GHG emission predictor. Logically, a corporation within the fossil fuel industry will have higher GHG emissions than a corporation within the renewable energy industry. Previous research on GHG emission prediction implements three different sector classification variables. These are the Global Industry Classification Standard (GICS), the North American Industry Classification System (NAICS), and the Business Industry Classification System (BICS). The BICS classifies a corporation by examining in which industry it is making the biggest fraction of its revenues. It is a rather detailed method as there are seven hierarchical levels. This sector classification method was found to be paramount for GHG emission prediction (Assael et al., 2023). The GICS and NAICS are less granular than the BICS with four and five hierarchical levels, respectively. Again, the industry classification helped achieve substantial accuracy gains (Nguyen, Diaz-Rainey, and Kuruppuarachchi, 2021).

2.2.2 Geographical classification

The country in which the corporation is headquartered is a relevant predictor. The country in which a corporation operates can give information on what regulatory environment a corporation operates. Consequently, it can be distinguished whether a country has relevant GHG laws and regulations. The presence or absence of such laws is seen as relevant information. This is represented by its relatively high level of predictor importance (Assael et al., 2023).

2.2.3 Corporation size

In general, the assumption is made that when two corporations produce the same system, using the same processes under the same circumstances, ceteris paribus, the larger corporation will produce higher GHG emissions (Goldhammer, Busse, and Busch, 2017). Turnover and revenue are both often referred to as indicators of corporation size. As GHG are emitted in the process of creating revenue, this seems like a logical predictor to consider. Next, corporation size can be expressed through the number of employees. High numbers of employees could indicate that the production processes are labor-intensive, or it could indicate that employees are divided over several production facilities, indicating low production centrality. Both imply higher GHG emissions (Goldhammer, Busse, and Busch, 2017) (Nguyen, Diaz-Rainey, and Kuruppuarachchi, 2021). Other financial information is not necessarily classified under the numerator of ‘corporation size’. However, we will classify these as indicators of the size of corporations as we identify these as representatives of production processes and capital structure. These include capital expenditure, Property, Plan & Equipment (PPE), assets, intangibles, and leverage (Nguyen, Diaz-Rainey, and Kuruppuarachchi, 2021) (Goldhammer, Busse, and Busch, 2017) (Assael et al., 2023) (Han et al., 2021) (Griffin, Lont, and Sun, 2017).

2.2.4 Energy consumption

Energy consumption of production processes is naturally directly related to GHG emissions. On the one hand, we have energy that is used in production processes through the use of corporations’ facilities and vehicles, and on the other hand, we have energy that is purchased for own use. These differences are represented through the differences in scope 1 and 2 emissions. Energy consumption is shown to be one of the best predictors in the prediction models in previous research (Assael et al., 2023). However, no distinction was made between renewable and non-renewable energy sources. We believe that this distinction could be a relevant factor in the prediction models.

2.2.5 ESG

Refinitiv Eikon provides ESG scores that are designed to measure corporations' relative ESG performance. The ESG scores are built out of the three main pillars of ESG, namely Environmental, Social, and Governance. These pillars consist of several categories, for Environmental, the categories Emission, Innovation, and Resource Use, for Social, the categories Community, Human rights, Product responsibility, and Workforce, and for Governance, CSR strategy, Management, and Shareholders. As these ESG scores provide relevant corporate information with regards to ESG, and emissions do no necessarily need to be included for the calculation of the ESG score, the ESG scores could be a relevant predictor in the prediction models.

2.2.6 Performance

The financial performance of a corporation is related to its ability to manage environmental impact. Financial performance can be affected by market sentiment towards GHG emissions, exposure to transition risks, and cost savings due to energy efficiency measures. We have three indicators of corporations' financial performance: Return on Capital (ROC), Return on Equity (ROE), and Return on Assets (ROA). ROC measures performance by calculating the return a corporation earns over its own capital, the ROE measures performance by calculating the return over outstanding equity and the ROA measures performance by calculating the return over its total assets, which is equal to the sum of its total liabilities and its total equity (Brealy, Myers, and Allen, 2017).

2.2.7 Efficiency

Financial efficiency indicates the efficiency with which a corporation uses its assets. The asset turnover ratio measures a corporation's ability to generate revenue from its assets, by dividing its net sales over its average total assets. The inventory turnover ratio indicates how quickly a company is selling its inventory and replacing it with new inventory. A high inventory turnover ratio generally indicates that a company is efficiently managing its inventory and is selling its products quickly (Brealy, Myers, and Allen, 2017).

2.2.8 Leverage

Financial leverage can be an indicator of a corporation's GHG emissions, because companies that are heavily reliant on debt financing may be more vulnerable to transition risks. Financial leverage is measured by the debt-to-equity ratio, which measures the amount of debt a corporation has relative to its equity. Next to the debt-to-equity ratio, we consult the interest coverage ratio, which measures the extent to which interest obligations are covered by earnings, and the cash coverage ratio, which measures the extent to which interest obligations are covered by operating cash flow (Brealy, Myers, and Allen, 2017).

2.2.9 Liquidity

Financial liquidity can be an indicator of the availability of funds for investing in the transition towards a low-carbon economy and the exposure to transition risks. The liquidity measures that we use are the current ratio and the quick ratio. The current ratio measures a corporation's ability to pay its short-term liabilities with its short-term assets. The quick ratio also measures a corporation's ability to pay its short-term liabilities with its short-term assets, but excludes assets that cannot be liquidized quickly, such as inventories (Brealy, Myers, and Allen, 2017).

2.2.10 Conclusion

We evaluated predictor variables from previous research on GHG emission prediction, as well as other literature, and divided these variables into 9 categories. We consider all predictor variables discussed in this section as potential significant predictor variables for our models. We assess the availability of these variables in Section 4.

Chapter 3

Methodology

In this Chapter, we give background information on the proposed prediction methods and describe the comparative framework in which we focus on several accuracy measures.

3.1 Prediction methods

In Chapter 1, we described that we test multiple models to examine which of those models is best suited for the prediction of corporate GHG emissions. The models that we consider can be categorized into regression analysis and machine learning methods. Machine learning is the field within artificial intelligence (AI) that is involved with making machines learn from examples by recognizing patterns in data. By recognizing patterns in data, machine learning models extract information that is relevant for future data. This section gives a brief introduction to the concept of regression analysis and machine learning by introducing linear regression, supervised learning, and unsupervised learning. Next, we introduce several specific machine learning methods that we use to tackle the core problem of this research.

3.1.1 Regression

Regression analysis focuses on deriving a relationship between a dependent variable and one or more independent variables. The most known and basic form of regression analysis uses linear regression. In linear regression, it is assumed that the relationship between the dependent and independent variables is approximately linear. For example, if we would want to test the relationship of sales with marketing expenditure using linear regression, we get a representation of the linear relationship between these two variables through a linear equation. The parameters that are predicted are the coefficients of the independent variable and the intercept of the linear model, where, in our example, sales is the dependent variable and marketing expenditure is the independent variable (James et al., 2021). Linear regression can also be used for prediction purposes. In this case, linear regression predicts output for the dependent variable based on the linear equation of the independent variables.

3.1.2 Supervised Learning

Supervised learning is one of the most often used machine learning methods. When we refer to a machine learning method as a 'supervised' method, we indicate that we have training input for which we know the desired output (Müller and Guido, 2016). An example of supervised learning is spam detection. Using a training set that consists of emails that are labeled as 'spam' and emails that are not labeled as 'spam', we can train a model that detects future spam mail. Given that we know the input/output pairs available in our training set, we can verify the performance of the supervised learning model. There are two main categories of supervised learning methods: classification and regression.

3.1.2.1 Classification & Regression

Classification models are algorithms that have as a goal to predict a discrete class label, which is one from a sample of predetermined class labels (Domingos, 2012). Here, the model uses a training set consisting of input/output pairs to train the model so it can predict the class label for data not seen during the training of the model. The class can be a binary output, e.g. when predicting whether patients are dead or alive after a certain period of time, or the class can have more than two output possibilities, e.g. when predicting the color of flowers based on measurements of petal and sepal lengths.

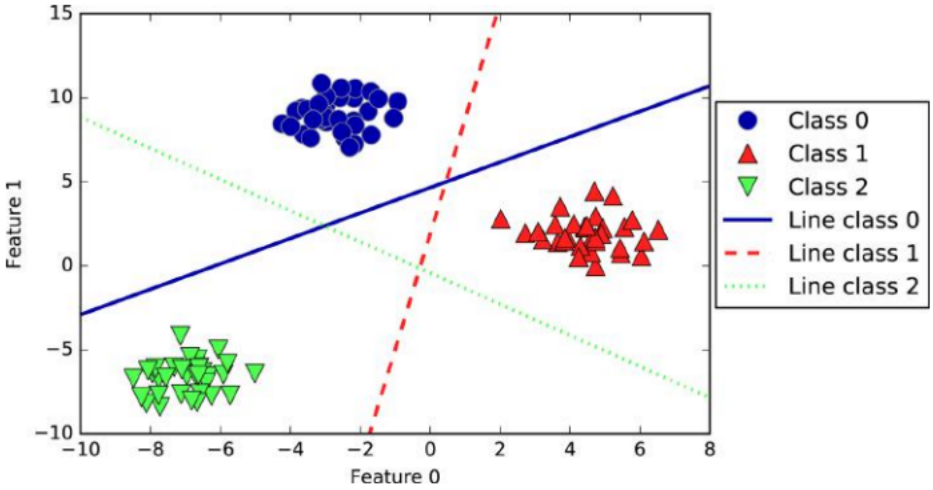


Figure 3.1: An unambigious example of the output of a classification model that uses three one-versus-the-others classification models (Müller and Guido, 2016).

Regression models are algorithms that focus on predicting a continuous output value. The prediction of an individual's income based on several variables is an example of a regression problem, as the income has a continuous value that can be any given number (within a reasonable range) (Müller and Guido, 2016). If we compare classification and regression models, it is clear that the prediction of GHG emissions is a regression problem. Note that there is a difference between regression in the context of supervised learning and in regression analysis (multiple linear regression). In regression analysis, regression focuses on deriving a causal relationship that can be used to predict output values, whereas, in supervised learning, regression focuses on purely predicting a continuous value. This difference is highlighted in Section 3.1.5.

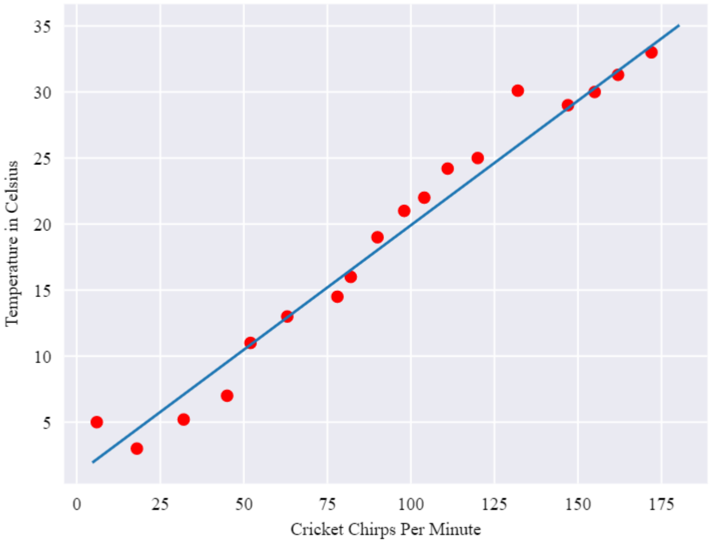


Figure 3.2: An example of the in-sample relationship of a regression model on cricket chirps per minute and the temperature in Celsius (Google, 2022).

3.1.3 Unsupervised learning

Unsupervised learning is the field of machine learning where we have no prior knowledge of the output. The measurement method is given, but information on what to expect is not (Müller and Guido, 2016). Unsupervised learning can be implemented to uncover hidden patterns in the data, but not for classification and regression problems as supervised learning can. There are two main categories of unsupervised learning methods: dimensionality reduction and clustering.

3.1.3.1 Dimensionality reduction & clustering

Dimensionality reduction is literally the reduction of the dimensions of the data. In other words, with dimensionality reduction, we decrease the number of input variables available in the training set. The result is that we have fewer input parameters which leads to less complex models. Furthermore, dimensionality reduction can be used to better visualize data. Clustering aims to divide data into groups with similar characteristics. The goal of clustering is that we end up with clusters that have high similarities within the cluster, but low similarities between the clusters. The similarities of the clusters are defined using the input variables.

3.1.4 Machine Learning methods

In the previous section, we discussed the main categories of machine learning. Looking at our research problem, we identify that we need to apply supervised learning methods, specifically supervised regression models. We have known input/output pairs, where the input is the set of predictor variables identified in Section 2.2, and the output is the set of corporate GHG emissions. As it is almost impossible to predict in advance what model suits a specific problem best, we will implement several supervised regression algorithms (Sterkenburg and Grünwald, 2021). In the following sections, we will introduce the methods that we will implement on our problem. Note that the following sections are not meant to be fully explanatory, but merely a short introduction to the concept of the methods. We will give more specific explanations of the applied methods later in this research.

3.1.4.1 Decision Trees

Decision trees are tree-structured machine learning algorithms that are assembled using a hierarchy of if/else questions (Müller and Guido, 2016). All if/else questions are linked to one of the predictor variables, and are represented by nodes. The nodes are connected by branches that represent the different answer options for the if/else questions. Take for example the problem of determining a person's salary using decision trees. The question at the first node could be "What industry sector does person X work in?" from which several industry branches can be picked. The goal is to create an accurate final prediction using as few if/else questions as possible. In itself, a single decision tree may lead to mediocre results. However, there are several methods to combine machine learning models into a more powerful model, often referred to as ensemble methods (James et al., 2021). This research will focus on two specific decision tree ensemble methods, namely bagging and boosting, and for these methods consider two specific applications.

Bagging
Bagging, short for bootstrap aggregating, is an ensemble method that builds multiple decision trees using random bootstrap samples of the data and aggregates the predictions of each decision tree into a single final prediction. Every individual decision tree is optimized for its bootstrap sample. By combining the results of a large number of decision trees, the variance of the decision trees is significantly decreased. Random forest is a popular bagging method that also builds trees by using a different bootstrap sample of the data for each tree. The difference with other bagging of decision tree methods is that each node within the trees is split by choosing the best predictor in a random subset of predictors (Liaw and Wiener, 2002). For a regression problem, each decision tree will generate a continuous output. The final prediction is typically the mean of the outputs of the individual decision trees. The main advantage of random forest is that the model, and its feature importance, can be easily visualized and understood by non-experts. Next to this, random forest works well for a mix of discrete and continuous variables as no pre-processing of the data is necessary (Müller and Guido, 2016).

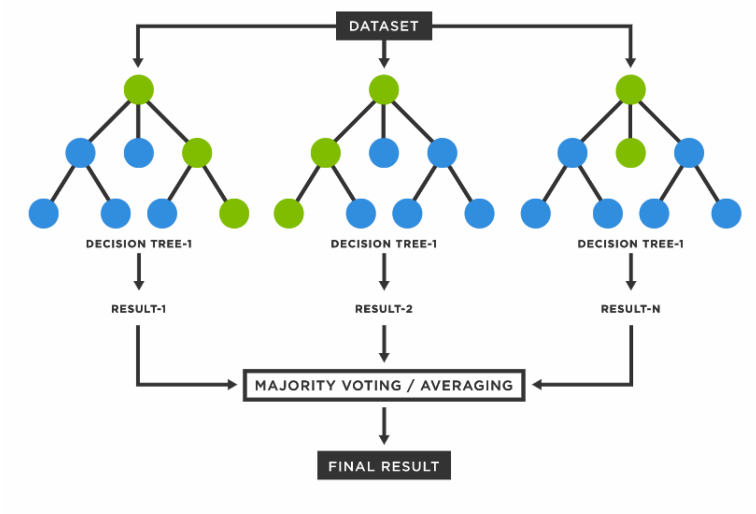


Figure 3.3: Simplified representation of a random forest model (TIBC, 2023).

Boosting
Boosting is an ensemble method that does not combine multiple decision trees, but creates a sequence of decision trees where each tree uses information from previous trees (Müller and Guido, 2016). Boosting is achieved by progressively adjusting the weights of the training samples based on the prediction errors of the preceding models. The sequence of decision trees is created one at a time, where each new tree is fine-tuned to correct the mistakes made by the prior tree. Extreme Gradient Boosting (XGBoost) is a tree boosting method that focuses on minimizing a specified error function by combining weak base learning models into a stronger learner in an iterative fashion. The XGBoost's biggest advantages over other boosting methods are its scalability and its speed (Laurensia, Young, and Suryadibrata, 2020).

3.1.4.2 Support Vector Machine

The goal of support vector machines is to find the perfect boundary of a hyperplane that separates different classes or gives a predicted output. SVMs can thus be used for classification and regression purposes. The simplest classification SVM is the maximal margin classifier which only works for linearly separable data. Although this means that this version can not be used often with real-world data, it is relevant for understanding the main concept of SVMs (Cristianini and Shawe-Taylor, 2014). The maximal margin classifier tries to find a decision boundary that maximizes the distance between the decision boundary and the closest of the data points. Figure 3.4 shows the maximal margin classifier for a two-dimensional space. The bold 'Os' and 'Xs' are the closest of the data points and are referred to as support vectors.

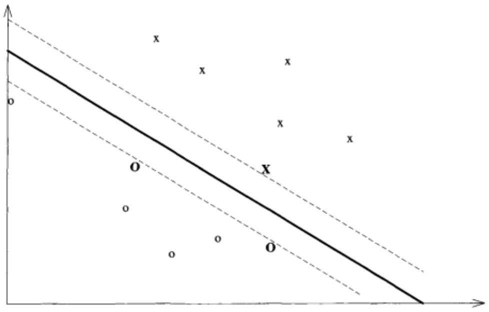


Figure 3.4: A maximal margin hyperplane with highlighted support vectors (Cristianini and Shawe-Taylor, 2014).

In SVMs for regression, also referred to as Support Vector Regression (SVR), the goal is to find the hyperplane that fits the training data whilst minimizing an ϵ-insensitive error function. The

ϵ-insensitive error function ignores errors that are within the bandwidth of the error margin and measures the cost of errors for errors outside of this margin, given by ξ. Again, the goal is to maximize the distance between the decision boundary and the support vectors.

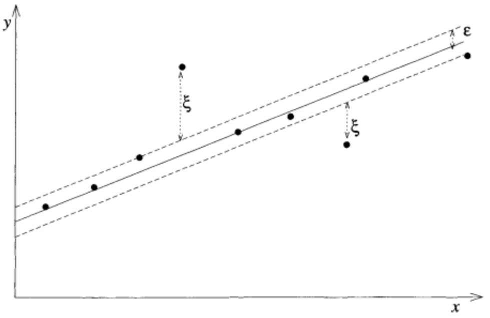
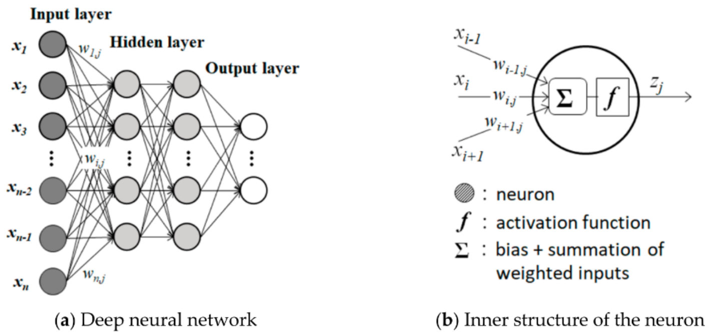


Figure 3.5: An ϵ-insensitive error function for a regression problem (Cristianini and Shawe-Taylor, 2014).

Figures 2.5 and 2.6 depict SVM examples for linear problems in low-dimensional input spaces. In practice, SVM applications have higher dimensional feature spaces. This is made possible by applying the kernel trick on the input space. The kernel trick allows the input to be mapped into a higher-dimensional space which in turn allows the SVM to capture non-linear relationships between the input and the target output (Bishop, 2006).

3.1.4.3 Neural Networks

Artificial Neural networks (ANNs) are based on biological neural networks and consist of several layers of nodes. Figure 3.6 shows the structure of an ANN and the structure of a single node. In the ANN, the first layer of nodes represents the input variables, the last layer of nodes represents the output variables, and the layers between the input and output layers are called hidden layers. The nodes are interconnected by links that represent the weights of each of the nodes (Bishop, 2006). Each node has an activation function that processes the weighted inputs of the previous layer and an added bias into an output value. The output value is then used as input for the nodes in the next layer. Finally, the last layer of nodes uses an activation function to create a set of output values. The output values are compared to the target values using an error function. The weights can be adjusted to minimize the error function using backpropagation, which uses gradient descent (Bishop, 2006).



(a) Deep neural network

(b) Inner structure of the neuron

Figure 3.6: Structure of an ANN and the structure of a single node (Kimura et al., 2019).

3.1.5 Causality versus prediction

We discussed the concept of regression analysis and the machine learning methods that we implement in this research. However, we want to make a final note on a key difference between the two distinct approaches used in regression analysis and machine learning.

Regression analysis is primarily concerned with understanding the causal relationship between a dependent variable and one or more independent variables. The goal of regression analysis is to identify the impact of independent variables on the dependent variable and to estimate the magnitude of this impact. In other words, regression analysis seeks to determine how much the dependent variable changes as a result of changes in the independent variables. On the other hand, machine learning is primarily focused on making predictions based on patterns and relationships in the data. Machine learning algorithms use mathematical models to learn from data and make predictions about new observations. Unlike regression analysis, machine learning is not necessarily concerned with understanding the causal relationship between variables. Instead, its sole focus is to attain accurate results in its predictions. Although, in general, machine learning methods are superior to regression analysis in terms of prediction accuracy, we consider both regression analysis and machine learning to provide a more comprehensive understanding of the relationship between variables and possibly improve the accuracy of the results.

3.2 Comparative framework

In the previous section, we discussed several machine learning methods that we employ for the prediction of corporate GHG emissions. These methods are employed to output predictions, $\hat{Y}_i$, based on data of GHG-disclosing corporations. Here, i refers to the corporation. We want to compare model accuracies, based on the difference between the observed values and model output, $\hat{Y}_i - Y_i$. This section outlines the metrics to evaluate these model accuracies. Furthermore, we discuss the difference between in-sample performance and out-of-sample performance, and the bias-variance trade off.

3.2.1 In-sample & out-of-sample performance

The goal of this research is to predict corporate GHG emissions for corporations that do not disclose this information. We approach this problem by taking emission data of corporations that do disclose GHG emissions, fitting a model on this data, and measuring whether the predictions made by the fitted model come close to the actual observed values. Given that we have good prediction accuracies, we can implement the fitted model for corporations that do not disclose GHG emission information. The assessment of the prediction accuracies can be done by assessing the in-sample prediction performance and the out-of-sample prediction performance. The in-sample prediction performance refers to a model's ability to accurately predict the model output for data it has seen during the training stage. The out-of-sample prediction performance refers to a model's ability to accurately predict the model output for data unseen during the training stage. A fitted model will always incorporate some of the sample-specific noise into the model fit. The in-sample performance is therefore always higher than the out-of-sample performance, as it complies with this fitted noise. Referring to the research goal, we want to predict GHG emissions for corporations unseen during the training stage of the models, as we do not have information on the output for these corporations. Therefore, for our research, the out-of-sample prediction performance is of more relevance than the in-sample prediction performance. In Section 5.3, we reflect on both in-sample and out-of-sample performance. However, our main conclusions are based on the out-of-sample performance.

3.2.2 Prediction accuracy

First, we measure the performance of the prediction models as compared to a no-skill predictor. This comparison allows us to inspect how our prediction models perform when compared to an untrained predictor. For this comparison, we compute Theil's inequality coefficient, which is obtained as follows: (Bikker et al., 2008)

$$U = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i^2} + \sqrt{\frac{1}{n}\sum_{i=1}^{n}Y_i^2}} \tag{3.1}$$

Theil's U is a scale-independent measure of forecast quality, where U is equal to zero for a perfect prediction ($\hat{Y}_i = Y_i$), and equal to one for no-skill prediction ($\hat{Y}_i = 0$). Consequently, we are aiming for a value close to zero. A value close to one would indicate that our models perform only slightly better than a no-skill predictor, in which case you can wonder whether our prediction models are a useful contribution. Another scale-independent error measure is the Mean Absolute Percentage Error (MAPE), for which we aim for a value close to zero, similar to Theil's U. The MAPE is obtained as follows:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{Y}_i - Y_i}{Y_i}\right| \tag{3.2}$$

Next to relative error measures, we include absolute error measures. Previous research on the topic of GHG prediction uses a variety of methods to measure the performance of the prediction models. (Griffin, Lont, and Sun, 2017) and (Goldhammer, Busse, and Busch, 2017) both reported only on the explanatory power of the used prediction variables using R-squared. These research papers did not report on the prediction errors of their respective regression models. The previously discussed research papers that used machine learning methods either used the Mean Squared Error (MSE) (Nguyen, Diaz-Rainey, and Kuruppuarachchi, 2021) or the Root Mean Squared Error (RMSE) (Assael et al., 2023) (Han et al., 2021) (Bloomberg, 2021) to compare model accuracy. The MSE and RMSE are calculated using formulas (3.3) and (3.4), respectively.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{3.3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{3.4}$$

The MSE takes the average squared distance between the predicted value, $\hat{Y}_i$, and the target value, Y_i. The RMSE takes the square root of the MSE. The advantage of the RMSE over the MSE is that the RMSE is expressed in the same ‘unit’ as the target value. In the context of GHG emission prediction, this means that the RMSE is expressed in tonnes, which makes the error rate of the model more easily understandable. For both RMSE and MSE, a lower value implies higher prediction accuracy.

Lastly, we evaluate the prediction performances through a goodness-of-fit measure, R-squared. The R-squared measures the proportion of variance in the dependent variable that can be explained by the independent variables. The R-squared is obtained as follows:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y_i})^2}{\sum(y_i - \bar{y_i})^2} \tag{3.5}$$

In Equation 3.5, the numerator is defined as the sum of squared errors, and the denominator is defined as the total sum of squares. Here, the sum of squared errors is the sum of the squared residuals produced by the prediction model, and the total sum of squares is the sum of the squared distance between the output and the mean of the training output. The ratio of these two metrics represents the proportion of the total variation in the dependent variables that is not explained by the prediction model. Subtracting this from 1 gives you the proportion of total variation that is explained by the prediction model.

3.2.3 Bias-Variance Tradeoff

Next to assessing the performance of the prediction models based on error measures, it is essential to discuss the bias-variance tradeoff of the models. It is well-known that we need to understand the bias-variance tradeoff to optimize our models' prediction accuracy. Ultimately, we want to minimize both bias and variance. However, the two are inversely related. Therefore, we ideally need to approach an equilibrium between the two to get the best results. Before we elaborate on this equilibrium, let us first explain the concepts of bias and variance.

Bias is the difference between the predicted values and the actual values. Models with high bias tend to perform poorly on both training and testing data. This indicates that a model with high bias underfits the training data, meaning it oversimplifies the relationship between dependent and independent variables. On the contrary, variance is the variability of the prediction of a single data point. Models with high variance tend to put too much weight on the training data. This leads to overfitting of the model to the training data, meaning the model does not generalize well, causing it to perform poorly on unseen data.

As stated earlier, we can improve model performance by finding an equilibrium for the bias-variance tradeoff. Figure 3.7 visualizes the bias-variance tradeoff. Here, the prediction error is made up of three aspects: the bias, the variance, and the irreducible error. Obviously, we are unable to reduce the irreducible error. The bias and variance can be reduced by finding a balance in model complexity. We can control the complexity of the model by tuning regularization parameters that control how 'aggressive' the learning process for a specific model is. Furthermore, we can control the number of predictor variables that are used in building the models and combine models using ensemble models, as we do with the XGBoost and the Random Forest. As bias and variance are important factors in the evaluation of predictor models, we will estimate and assess the bias and variance of our models to make coherent conclusions.

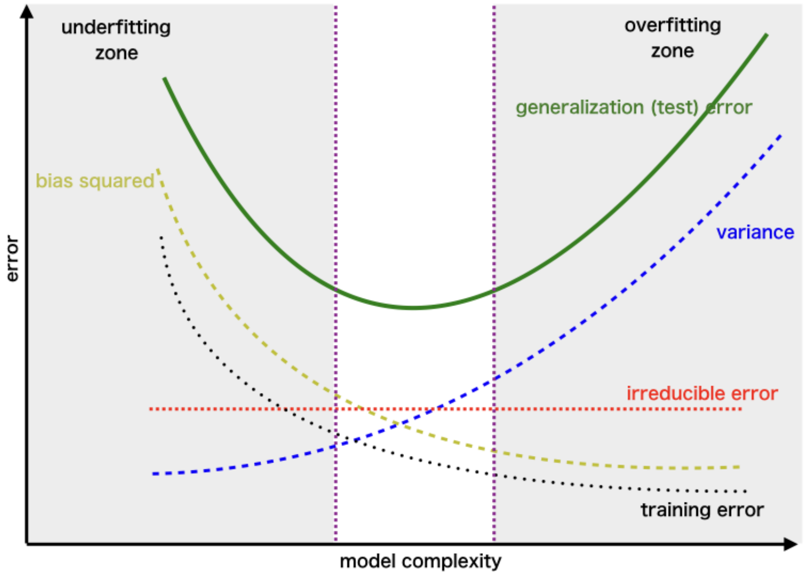


Figure 3.7: Visualization of the bias-variance tradeoff (Singh, 2018).

3.2.4 Conclusion

We introduced five performance metrics that we use for the assessment of our prediction models. We include two scale-independent error measures, where Theil's U measures the performance of the prediction models as compared to a no-skill predictor, and the MAPE measures relative prediction accuracy. Furthermore, we include two absolute error measures, the MSE and the RMSE. Lastly, we evaluate the prediction performance through a goodness-of-fit measure. Next to these five performance metrics, we introduced the distinction between in-sample and out-of-sample performance, and discussed the bias-variance tradeoff through which we can assess model generalization. With both accuracy and generalization measures, we expect to be able to make meaningful conclusions on the performance of our models.

Chapter 4

Data Selection & Preparation

In this Chapter, we describe our data source and describe how we extracted the data from this data source. Furthermore, we describe the data preparation process for both emission data and the predictor variables. Finally, we visualize our data to enable us to better understand the data.

4.1 Data description

As stated in Chapter 1, we extract data from Refinitiv Eikon, a financial analysis desktop. We manually extracted data on four emission categories: 'Total' emissions (scope 1 + scope 2), scope 1 emissions, scope 2 emissions, and scope 3 emissions. This data is gathered by Refinitiv Eikon through a team of content research analysts that are trained to collect reported ESG data, as well as a broad range of other data points. Next to emission data, we extracted data on all variables described in Section 2.2, ordered by the main industries as described by Refinitiv, namely Energy, Basic Materials, Industrials, Consumer Cyclical, Consumer Non-Cyclical, Financials, Healthcare, Technology, Utilities, Real Estate, Academic & Educational Services.

4.2 Data preparation

4.2.1 Emission data

We extracted corporate emission data for the last 20 years for all emission categories. This implies that we have a mixture of time-series data and cross-sectional data, i.e. panel data. Note that not all corporations in our data frame have published emissions for the last 20 years, indicating that we have an unbalanced data panel. We convert the panel data into a cross-sectional data frame, meaning that corporations now have multiple data rows, one for each year that they have published their emissions. So, we now have a cross-sectional data frame consisting of emission data for all emission categories over a differing number of years.

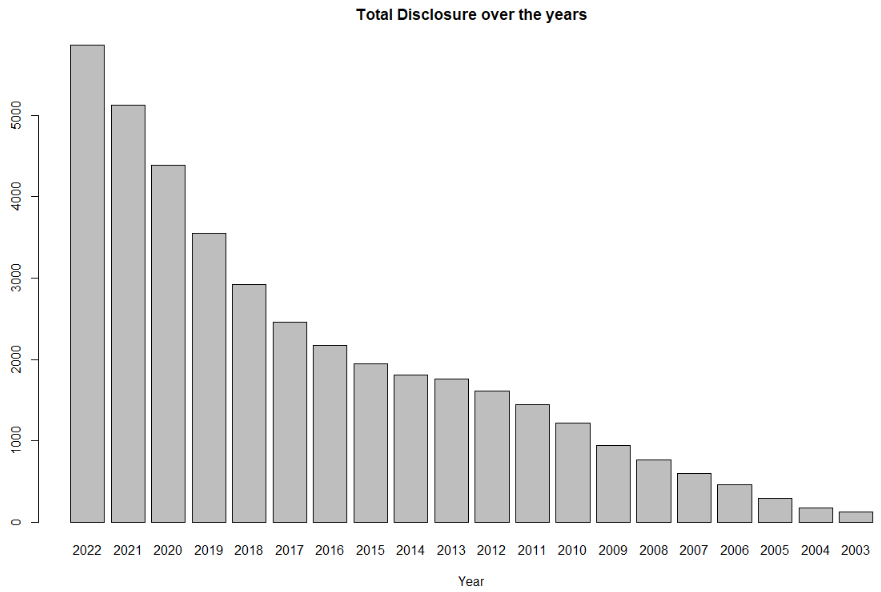


Figure 4.1: Total Disclosures in the previous 20 years.

Figure 4.1 depicts the number of 'Total' emission disclosures over the past 20 years. Disclosures of the other emission categories show similar counts, as can be seen in Appendix A.1. The figure shows that the number of emission disclosures decreases as we go back in time. This can be explained by the increasing importance of GHG disclosures. For the sake of relevance of the data and decrease of data instability, we exclude emission data from the years before 2018. Furthermore, we divide our data frame into four separate data frames, one for each emission category. In each of these data frames, information on the other emission categories is deleted. Table 4.1 shows the number of data rows for each of the four data frames.

Data frame	Year	Number of disclosures
Total	2022	5863
	2021	5126
	2020	4387
	2019	3555
	2018	2923
	Total	**21854**
Scope 1	2022	5329
	2021	4605
	2020	3904
	2019	3067
	2018	2466
	Scope 1	**19371**
Scope 2	2022	5334
	2021	4613
	2020	3891
	2019	3052
	2018	2451
	Scope 2	**19341**
Scope 3	2022	3291
	2021	2889
	2020	2338
	2019	1791
	2018	1404
	Scope 3	**11713**

Table 4.1: Disclosure counts for all emission categories (2018-2022).

4.2.2 Predictor data

In section 2.2, we discussed possible predictors found in previous research and literature. Subsequently, we assessed which of these predictors were available in Refinitiv Eikon. With regards to the three 'sector classification' predictors, Refinitiv Eikon only provides the GICS classifier. In total, we selected 29 possible predictors. Although Refinitiv Eikon has a large range of other possible variables, we choose to focus on the 29 predictors that are discussed in this section. Appendix A.2 gives a more detailed description of what these predictors denote. Now that we have four data frames, one for each emission data frame together with data on the 29 possible predictors, we inspect the usability of these predictors.

Predictor Category	Predictor variable	Percentage of missing values
Year	Fiscal Year	0.0%
Sector Classification	GICS Sub-Industry Name	0.79%
	GICS Industry Name	0.79%
	GICS Industry Group Name	0.79%
	GICS Sector Name	0.79%
Geographical Classification	Country	0.0%
	Region	0.0%
Corporation Size	Revenue	10.36%
	Employees	14.22%
	Capital Expenditure	25.45%
	Net PPE	2.71%
	Net Intangibles	12.44%
	Operating Expenses	0.77%
	Total Assets	0.56%
Energy Consumption	Energy Purchased	19.31%
	Energy Produced	75.67%
	Renewable Energy Purchased	72.12%
	Renewable Energy Produced	78.28%
ESG	ESG Score	0.01%
Financial Performance	ROE	14.18%
	ROC	40.87%
	ROA	23.28%
Financial Efficiency	Asset Turnover	9.95%
	Inventory Turnover	24.94%
Financial Leverage	DE Ratio	23.28 %
	Interest Coverage Ratio	25.41%
	Cash Flow Coverage Ratio	3.82%
Financial Liquidity	Current Ratio	12.83%
	Quick Ratio	12.83%

Table 4.2: The percentage of missing values of predictor variables in the 'Total' emission data frame.

Table 4.2 shows that most of the predictor variables have missing values. A high number of missing values can introduce an unwanted bias into our prediction models. Therefore, we disregard variables that have a proportion of missing values higher than 30%. This results in the deletion of the variables 'Energy Produced', 'Renewable Energy Purchased', 'Renewable Energy Produced', and 'ROC'. Next to this, we delete the rows that have missing data on variables with a percentage of missing values lower than 5%. These include the GICS sector classification variables, 'Operating Expenses', 'Total Liabilities', 'Total Assets', and 'Cash Flow Coverage Ratio'. So, we delete rows with missing values for one or more of these variables. Lastly, we evaluate the availability of the predictor 'ESG Score' in the subset of corporations that do not disclose emission data. We observe that 'ESG score' is missing for most of the data in this subset. Furthermore, one can argue about the reliability of ESG data. Therefore, we exclude 'ESG Score' as a predictor variable. We are now left with four data frames that still contain missing values. Next to this, we need to address the possible collinearity, and outlier values in our data frames. In the following sections, we will first describe the procedure of handling outliers, discuss the procedure of multiple imputation for missing values, and discuss the issue of collinearity.

4.2.2.1 Outliers

Outliers are data points for which the values lie at an abnormal distance from other observed values. Outliers can arise due to errors in observations or measurements, or due to extreme observations. We assume that the outliers in our data frames are due to extreme observations and we will treat them as such.

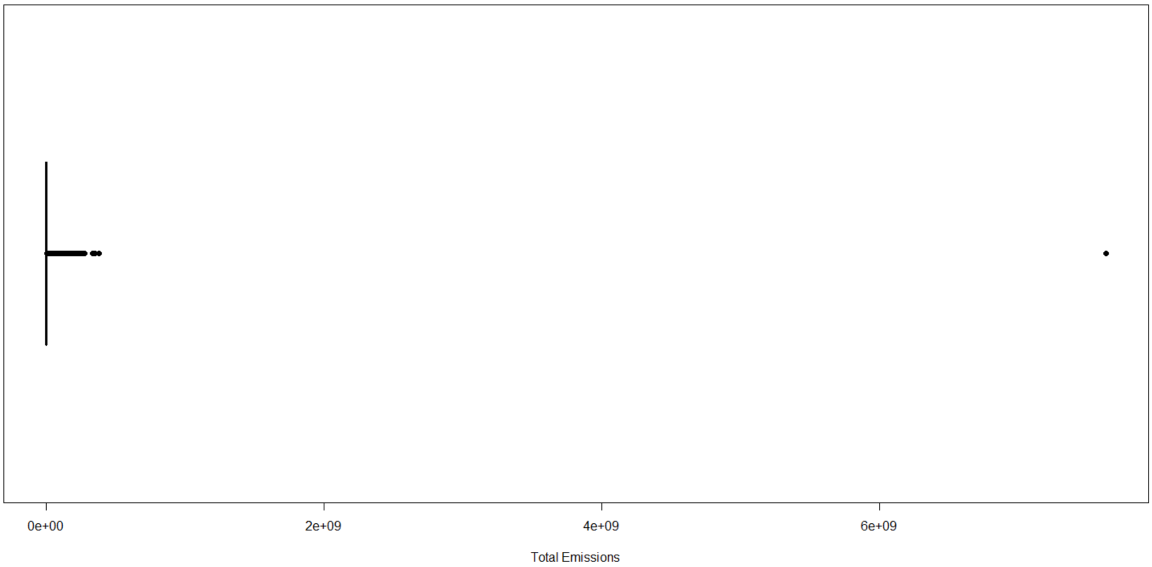


Figure 4.2: Total Emission data with extreme observation.

Figure 4.2 shows an example of the influence of an extreme observation in the ‘Total’ emission variable. Extreme observations in large sample data are common and therefore not necessarily an issue. One can argue that the existence of outliers in data is noise that can create a bias. However, on the other hand, we can argue that extreme observations are relevant data points from which a model can learn. For the former argument, we would need to do an extensive analysis of the extreme observations for all variables. As this is not in the scope of this research, we will consider all extreme observations but one as relevant data points. The outlier in Figure 4.4, which is almost 20 times larger than the closest value is deleted from the data set. After the deletion of this specific outlier, the data for the predictor variables remain skewed due to outliers. To ensure that smaller values are not overwhelmed by their abnormally large counterparts, we use log transformation to transform the data of continuous variables. In other words, variables x will be transformed in $log(x)$. Log transformation de-emphasizes outliers and enables the visualization of the distribution of the variables to be clearer (Metcalf and Casey, 2016).

Figure 4.3 shows a boxplot of the ‘Total’ emission variable. The distribution is more clearly depicted than the distribution prior to the log transformation. Furthermore, we see that the data is more closely distributed, indicating that the outliers are de-emphasized.
Preliminary test results show that, even after log-transformation, the models are unable to predict outlier values. This implies that the inclusion of outliers in the training and testing of the prediction models leads to highly inaccurate results. The prediction models will make sense only for ‘central’ data points. Therefore, we eliminate outlier values based on the interquartile ranges.

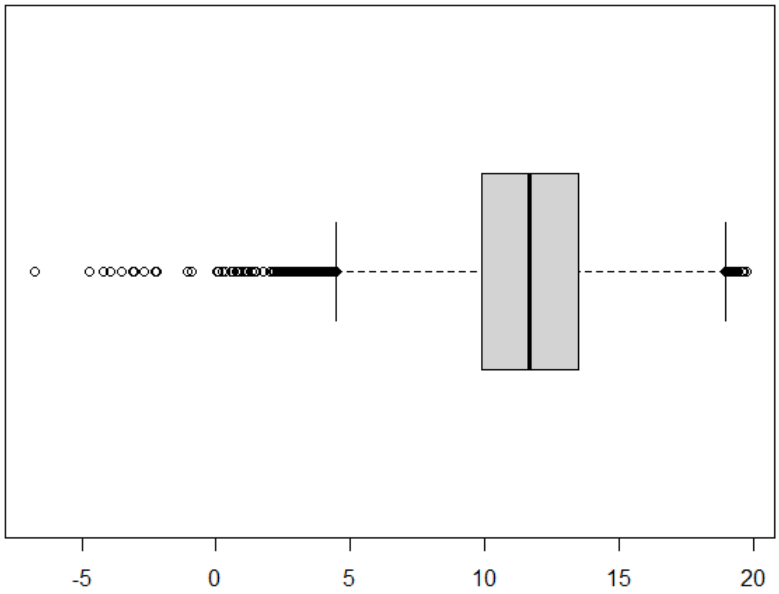


Figure 4.3: Total Emission data without extreme observation.

4.2.2.2 Data imputation

Table 4.2 shows that a large portion of predictor variables has missing values in their data. The deletion of all rows with missing values can have an impact on the validity of our prediction methods, as the deletion of rows will create biased results. On the other hand, if we choose to impute all rows, the validity of our prediction methods will be biased due to data rows with a large portion of imputed values. We want to increase the validity of our prediction methods by deciding on the trade off between the two former issues. We inspect the pattern of missing data to see how the missing values are dispersed over the data set. We choose to use multiple imputation for data rows that have at most 3 missing values. Multiple Imputation (MI) is a statistical technique that has gained popularity in recent years as a way to handle missing data in research studies. The principle behind MI is that the imputed values are not known with certainty, and thus multiple imputed data frames are created, each with different plausible values for the missing data. These data frames are then analyzed separately, and the results are combined to produce a final result that accounts for the uncertainty introduced by the missing data.

There exist several MI methodologies for which the usability depends on the subjected data frame. Unfortunately, there is no universal method that works well for all data frames unconditionally. Therefore, there are several choices to be made. First, we make the assumption that the missing values in our data frames are Missing at Random (MAR). Data is MAR if the probability of data missing is equal within subsets defined by observed data. Second, we choose the form of the imputation model. Our data contains multivariate missing data, meaning that missing data is not limited to one variable, but occurs everywhere in the data set. One method of dealing with multivariate missing data is through the use of Fully Conditional Specification (FCS). FCS imputes variables one by one conditional on the other variables in the data set. We will implement one specific application of FCS: the MICE algorithm (Buuren and Groothuis-Oudshoorn, 2011). MICE, or Multiple Imputation by Chained Equations, imputed variables one by one conditional on other variables using a pre-specified univariate imputation model. This leads us to the third choice to make, the choice of the univariate imputation model. MICE offers several built-in univariate imputation techniques. The most commonly used technique for continuous variables is predictive mean matching (PMM). PMM takes place in several steps. First, a linear regression model is fitted on the target variable, that is the variable to be imputed, which returns regression coefficients for the predictor variables. Second, Bayesian regression coefficients for the predictor variables are defined. All observed target values are then predicted by the linear regression model, and the missing target values are predicted by the Bayesian regression model. Finally, the Bayesian prediction for one target value is subtracted from all predictions of observed values. The five smallest differences are pooled together from which a value is randomly drawn. This is repeated for all missing values within the target variable (Heymans and Eekhout, 2019). The main advantages of predictive mean matching are that imputations are restricted to the observed values, that it is fairly robust to transformations of the target variable, and that it can preserve non-linear relations even if the structural part of the imputation model is wrong (Buuren and Groothuis-Oudshoorn, 2011).

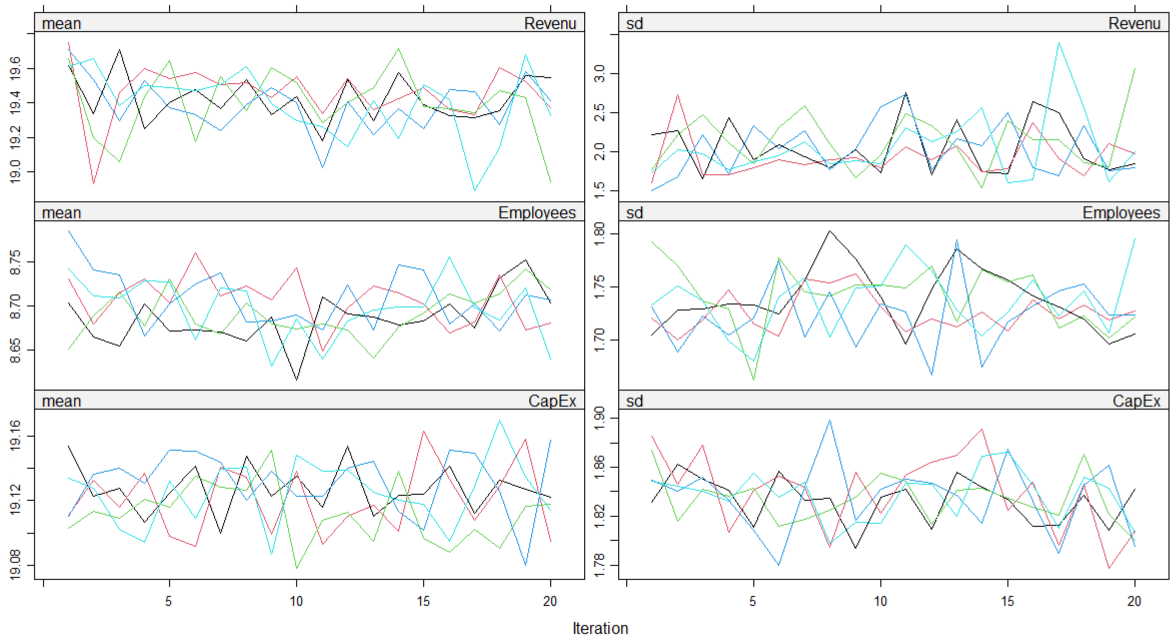


Figure 4.4: Convergence of MI of 'Revenue', 'Employees', and 'Capital Expenditure'.

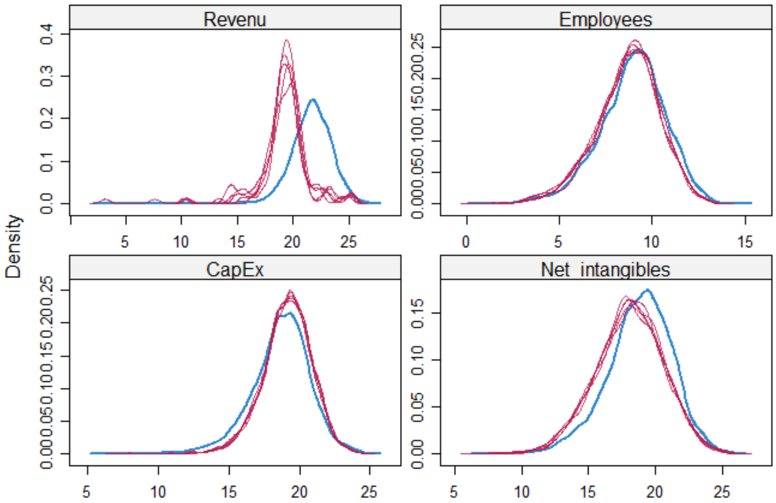


Figure 4.5: Density of observed data (blue) against the density of imputed data (red) for the variables 'Revenue', 'Employees', 'Capital Expenditure', and 'Net Intangibles'.

Figure 4.4 shows us the convergence of the imputation of the variables ‘Revenue’, ‘Employees’, and ‘Capital Expenditure’. The convergence of the imputations of the remaining variables in the ‘Total’ emission category are found in Appendix A.3. In every iteration, five imputation streams are projected. The results are combined into a final data frame that does not contain missing values. Figure 4.3 shows the density plots of the variables ‘Revenue’, ‘Employees’, ‘Capital Expenditure’, and ‘Net Intangibles’. We observe that the density of the distribution of the imputed values is similar for these four variables, with only ‘Revenue’ observably deviating. This indicates that the underlying distribution of the variables is preserved after the MI process. Density plots are created for all imputed variables of the data frame of the ‘Total’ emission category and are found in Appendix A.3. We repeated the MI process for the emission categories ‘Scope 1’, ‘Scope 2’, and ‘Scope 3’.[1] We now have four data frames without missing values. It is possible to even further analyze the imputed values, however, this is out of the scope of this research.

4.2.2.3 Collinearity

Collinearity occurs when 2 or more predictor variables are correlated with each other, which can cause problems in prediction methods that assume the independence of predictors. Collinearity is especially a problem for regression methods that investigate the causal relationship of predictors with the dependent variable. In other words, it can be difficult to assess the interaction of 2 predictors with the dependent variable separately, when the 2 tend to increase or decrease together (James et al., 2021). The prediction model would then overfit the training data, causing poor results on out-of-sample data. We describe the procedure for testing for collinearity using the data frame of the emission category ‘Total’. The procedure is done for all four data frames.

Before we test for collinearity, we create a correlation matrix for all numerical predictor variables. First, we inspect whether there are predictors that are approximately perfectly correlated. Appendix A.4 depicts the plotted correlation matrix. We find that ‘Quick Ratio’ has an almost perfect correlation with ‘Current Ratio’, and ‘Revenue’ has an almost perfect correlation with ‘Operating Expenses’. For these highly correlated pairs, we use a correlation higher than 0.8 as a cut-off (Mason and Perreault, 1991). Thus, we drop the variables ‘Current Ratio’ and ‘Revenue’. This causes no problems as to the loss of explanatory power, as these predictors add no independent information to the prediction methods. The deletion of highly correlated variables, however, does not directly indicate an absence of collinearity. It is possible that collinearity exists between more than 2 variables without high correlation. This is called multicollinearity. We measure multicollinearity by calculating the Variance Inflation Factor (VIF) for all predictors. Formula 4.1 shows the VIF calculation: (James et al., 2021)

$$VIF = \frac{1}{1 - R^2_{X_j|X_{-j}}} \tag{4.1}$$

Here, $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all other predictors. A VIF value that exceeds 5 or 10 is an indication of a problematic amount of collinearity, whereas a value of 1 indicates an absence of multicollinearity (James et al., 2021). Table 4.3 gives an indication of multicollinearity when fitting a linear regression model on the ‘Total’ data frame.

[1] For figures on convergence and densities on the other data frames, please consult the author.

Predictor variable	**VIF**
Fiscal Year	1.08
Country	3.32
GICS Sector Name	7.09
Revenue	6.01
Employees	3.59
Capital Expenditure	4.45
Net PPE	5.43
Net Intangibles	2.55
Total Assets	6.02
Energy Purchased	2.59
Asset Turnover	2.06
Inventory Turnover	1.02
ROE	1.07
ROA	1.24
DE Ratio	1.01
Interest Coverage Ratio	1.00
Cash Flow Coverage Ratio	1.01
Quick Ratio	1.17

Table 4.3: VIF values for all predictor variables for the 'Total' data frame.

Table 4.3 shows the presence of considerable multicollinearity in the 'Total' data frame. As stated earlier, this is especially a problem for regression methods. Thus, we want to select a regression method that can handle multicollinearity. Regression methods that eliminate multicollinearity can be categorized into two methods: variable selection and modified estimates. Variable selection methods select a subset of predictor variables. Often used variable selection models are the forward and backward selection models. The former starts with zero predictors and iteratively adds a predictor, whereas the latter starts with all predictors and iteratively drops predictors. These methods are known to be interpretable but do not indicate why predictor variables are included or dropped from the final model (Chan et al., 2022). Modified estimates, also known as shrinkage methods, fit the predictor variable coefficients towards zero, which significantly reduces the variance and consequently overfitting of the model (James et al., 2021). There are two well-known shrinkage methods, namely ridge regression and lasso. Ridge regression works well for multicollinearity, but the coefficients are not reduced to zero, which makes interpretability difficult (Chan et al., 2022). The lasso regression, however, does shrink the coefficients to zero which leads to variable selection. Variable selection through the use of Lasso regression eliminates the problem of multicollinearity for the regression method. Because of this characteristic, we choose to implement Lasso regression.

4.3 Data Visualization

In the previous sections, we selected, processed, and finalized our data frames. In this section, we will examine the data frames through data visualization. This allows for a better understanding of the data, before going into the prediction modeling process.

4.3.1 Demographics

The demographics give us information on the population of corporations in our data set. The corporations are headquartered in 79 different countries divided over 5 regions: Africa, Oceania, Americas, Asia, and Europe. Figure 4.6 shows the regional distributions of the disclosures of all emission categories. The regions Europe and Asia are represented the most in our data set. Asia and Europe account for the largest portion of disclosures. The Americas come in third place and Africa and Oceania are least represented in the data set. Figure 4.7 gives an overview of the top 10 countries with the most disclosing corporations in our data set, for all emission categories. We see that the United States of America has by far the most disclosures, followed by the United Kingdom and Japan. Finally, Figure 4.8 shows the distribution of disclosing corporations in the highest segment of the GICS classification segments, the sector name. Industrials is the leading sector within our data set when it comes to GHG emission disclosures, followed by Materials and Consumer Discretionary. Remarkable to notice is that although there is a relatively small number of disclosing corporations from the Energy sector in our data set, it still sums up to be one of the most emitting sectors.

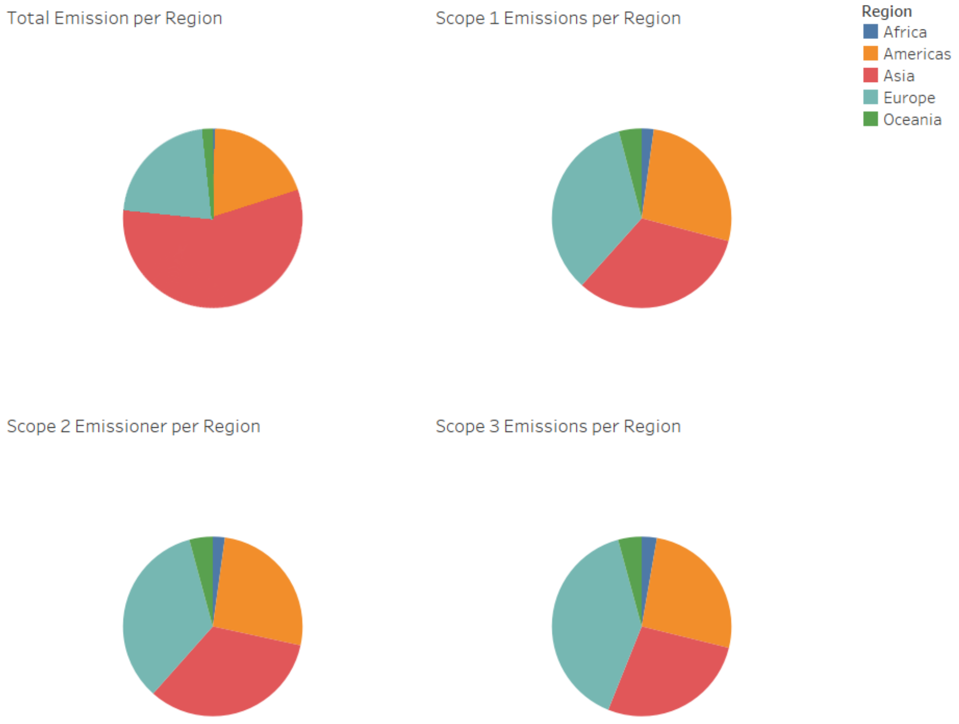


Figure 4.6: Regional distribution of emission disclosures.

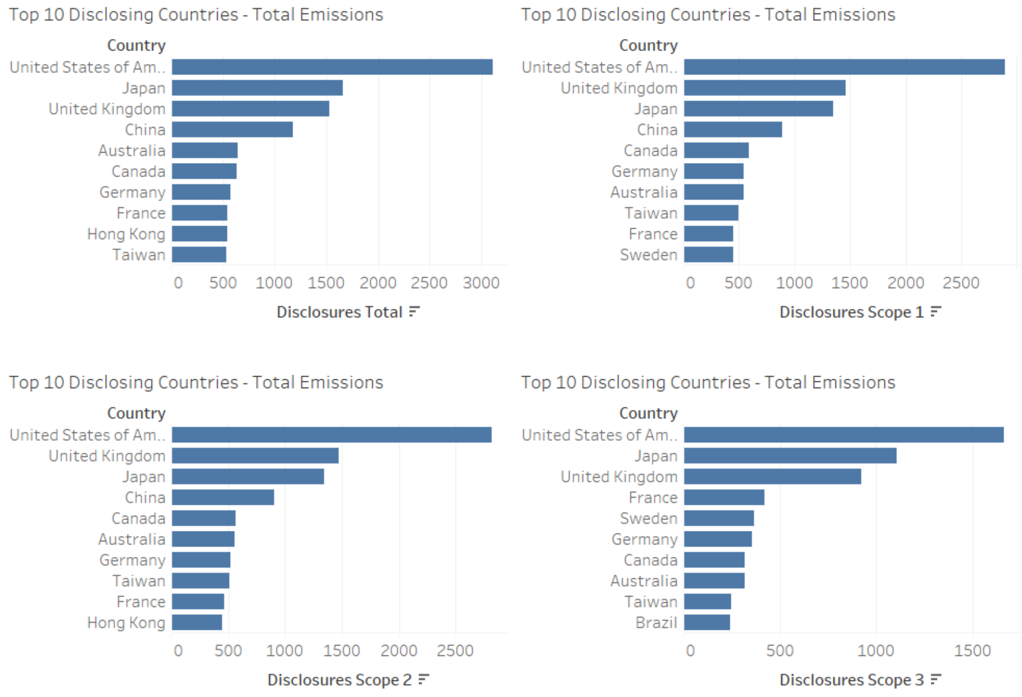


Figure 4.7: Top 10 disclosing countries for all emission categories.

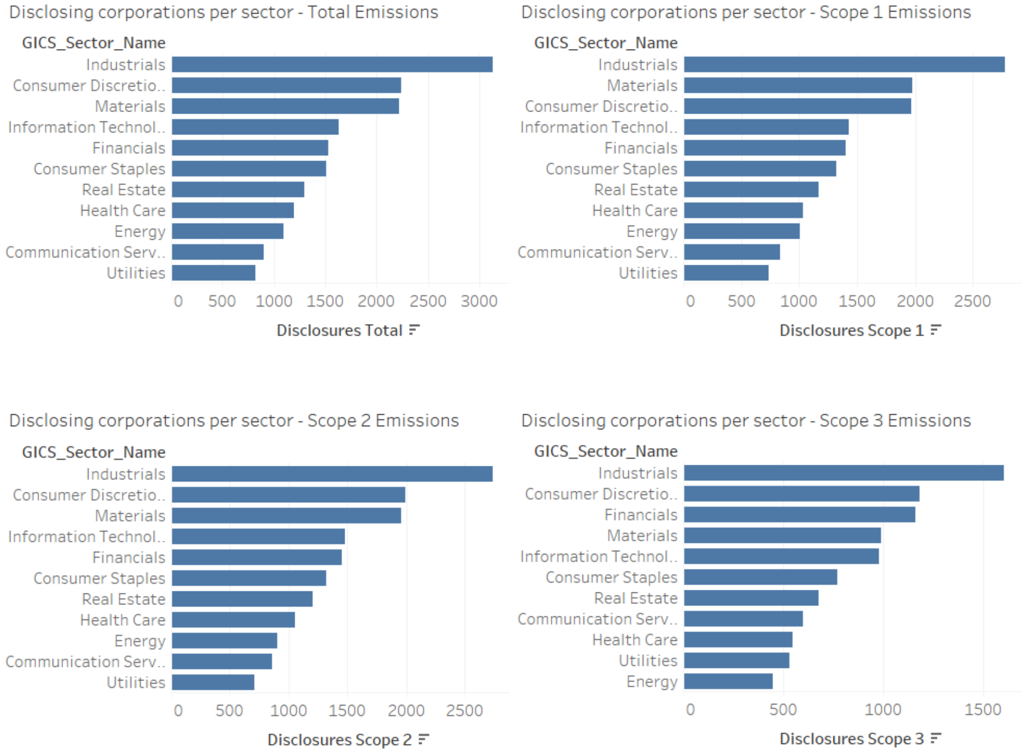


Figure 4.8: Disclosing corporations per sector name for all emission categories.

4.3.2 Distribution of Predictor Variables

In this section, the distribution of four numerical predictor variables, for which we observed a visible relation after log transformation, is plotted against the 'Total' emission category. Before log-transformation, we observe a distribution that is similar to the distributions of other combinations of emission categories and predictor variables, as seen in Figure 4.9. The distribution of data points lies close to the axes with several outlier values. From these scatter plots, we can conclude that there is no clear relationship to be found between 'Total' emissions and the predictor variables 'Total Assets', 'Net PPE', 'Employees', and 'Energy Purchased', in the original feature space.

We discussed earlier that we decided to log-transform the emission and predictor variables such that the outliers are de-emphasized and the visualization of the distribution of the variables is more clear. Figure 4.10 shows the scatter plots of 'Total' emission with the predictor variables 'Total Assets', 'Net PPE', 'Employees', and 'Energy Purchased'. We now observe a sample distribution that seems to have a small, but observable relationship. Appendix A.5 shows the distributions of the remaining predictor variables. Besides the variables 'Capital Expenditure' and 'Revenue', the remaining variables show no observable relationship with 'Total' emissions.

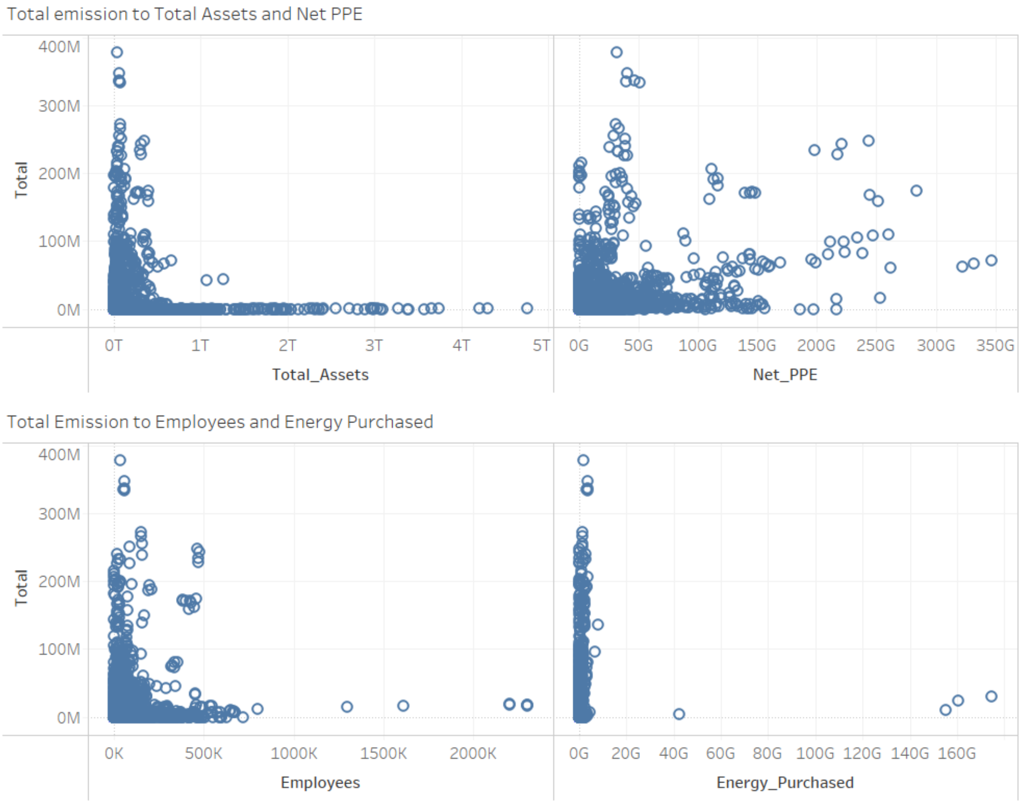


Figure 4.9: Scatter plots of Total Emission with the predictor variables 'Total Assets', 'Net PPE', 'Employees', and 'Energy Purchased' before log transformation.

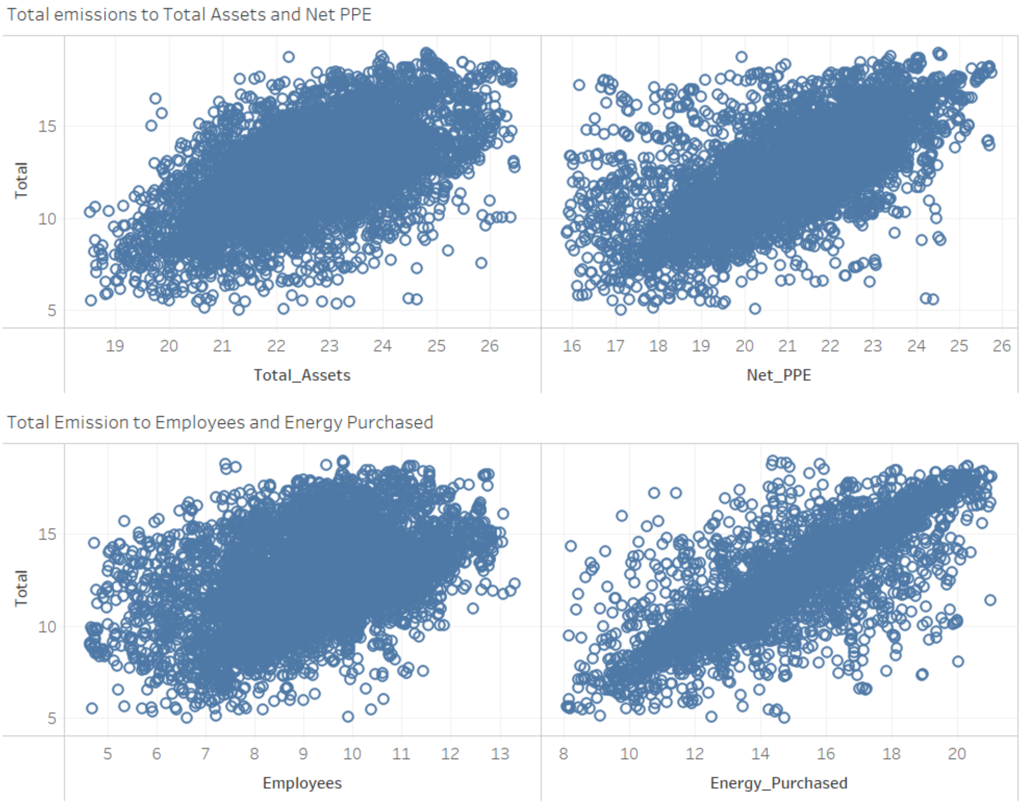


Figure 4.10: Scatter plot of Total Emission with the predictor variables ‘Total Assets’, ‘Net PPE’, ‘Employees’, and ‘Energy Purchased’ after log transformation.

Chapter 5

Prediction Modeling

In this chapter, we will discuss the process of prediction modeling. The goal of the prediction models is to attain accurate GHG emission predictions, based on the predictor variables described in chapters 2 and 4. We use the programming language ‘Python’ for the prediction modeling, specifically using the libraries ‘Skicit-learn’, ‘XGBoost’, and ‘Keras’. This chapter describes the process of cross-validating the results, the grid search process through which we tune hyperparameters, and for each model, the specific hyperparameters that we tune. Finally, we present the results of the grid search and give, for each model and scope, the specific hyperparameters that we use.

5.1 Model Evaluation and Improvement

5.1.1 Train-test split

To train and evaluate our prediction models, we split the data frames into training sets and test sets. The training set is the subset on which we fit the prediction models. Prediction models try to learn from the training data by recognizing relationships and patterns between the predictor variables and the output variable. The test set is used to assess the performance of the fitted prediction models on new, unseen data. The prediction models are used to predict the output variables based on the predictor variables, and these predictions are compared to the observed values in the test set. In Section 3.2.1, we discussed the difference between in-sample and out-of-sample performance. The in-sample performance is measured by assessing the prediction performance of the prediction models on the training set, whereas the out-of-sample performance is measured by assessing the prediction performance of the prediction models on the test set.

5.1.2 Group K-Fold Cross-Validation

Before we describe the prediction models and the tuning of hyperparameters, we introduce the method with which we will evaluate the models. Cross-validation is a method of evaluating the generalization of our models. We described the splitting of the data into a training set and a test set. Cross-validation is seen as an improvement over such a split, as it splits the data multiple times and trains a model for each of these splits. K-fold validation splits the data set into K folds. If we take $K = 5$, the data set is split into five folds, with which five models are trained. The first model is trained using the first four folds and tested on the fifth fold, the second model is trained using folds 2-5, and tested on the first fold, and so on.

Earlier, we described that our data frames consist of panel data, meaning we have multiple measurements over time for individual corporations. We want to prevent that the data of a single corporation is present in both the training and the test set. After all, if the model is tested on corporations it has previous information on due to its presence in the training set, the results of the generalization of the model are biased. To ensure that we have independence between the training and test set, we use group k-fold cross-validation. Figure 5.1 shows how, for each fold (CV iteration), the groups are kept intact by placing them in their entirety in either the training set or, the test set. In our application, the groups are labeled by ‘Company ID’. The ‘class’ section of

Figure 5.1 can be ignored for our specific application. Group K-fold cross-validation results in K measurements per performance metric. The final prediction performance results are calculated as the mean of the K measurements.

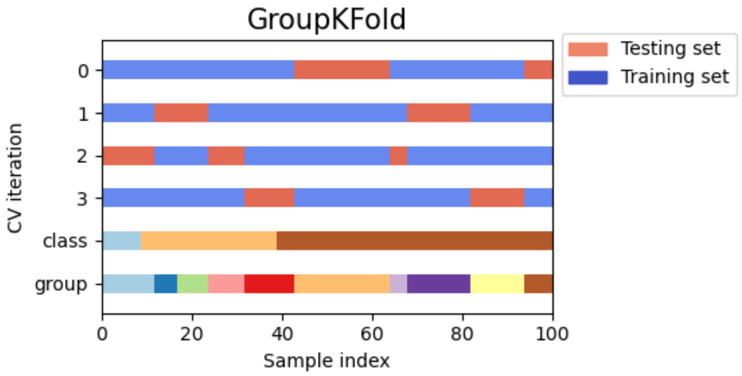


Figure 5.1: Group K-Fold Cross-Validation (Pedregosa et al., 2011).

5.1.3 Grid Search Cross-Validation

Grid search cross-validation is used for hyperparameter tuning, meaning that we use grid search cross-validation to find the combination of hyperparameters that gives us the best results for a specific model. The rationale behind applying grid search CV to our prediction models is that it is not possible for us to know beforehand what combinations of hyperparameters result in the best model performances. Next to this, we note that we cannot test for all possible hyperparameter combinations due to the high computational cost of this procedure. Therefore, we use a selection of hyperparameters for all models in our grid search.

We specify a parameter grid in which we give the parameters and their possible values (e.g. the learning rate of the XGBoost model). The grid search then applies all possible combinations of parameter values in the parameter grid by building models using these combinations. The grid search compares model performance by comparing the negative MAE and returns the best combination of hyperparameters. We use the MAE as we try to minimize the distance between observed values, and predicted values. The grid search CV's scoring function links higher values to better performance. As our scoring function is MAE, we use the negative MAE within the grid search so that the highest scoring value refers to the model with the smallest MAE.

5.1.4 Random Search Cross-Validation

Grid search cross-validation is a useful tool to inspect all possible combinations of hyperparameters for some models. However, for the models that are significantly more computationally intensive, we need an alternative method of hyperparameter tuning. In our case, the random forest and artificial neural network are significantly more computationally intensive. For these two methods, we implement the random search cross-validation. This model is similar to the grid search cross-validation in terms of selection criteria. The difference is that it does not test for all possible hyperparameter combinations, but makes a specified number of random combinations for which it tests the model performance. Although this method may overlook the optimal hyperparameter combination, it gives us the possibility to cross-validate several hyperparameter combinations while keeping the computational effort in check.

5.2 Model Specifications

In Section 5.2, we explain, shortly how the prediction models operate, followed by introducing the hyperparameters that we tune with the grid search CV and the results of the grid search.

5.2.1 Lasso

We use lasso as our model for regression analysis. Lasso regression is a regression model that incorporates feature selection by shrinking coefficients toward zero. The Lasso coefficients are selected such that the following quantity is minimized:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|$$

The left-hand side of the equation is known as the Residual Sum of Squares (RSS), used for estimating coefficients in linear regression. The right-hand side is known as the lasso penalty, also referred to as the L1 penalty. This penalty has the advantage of shrinking coefficients towards zero, and even equal to zero when parameter λ is large enough. The result is that the model performs feature selection, meaning only a subset of the possible predictor variables is considered in the final model. Note that we use a general λ instead of a β-specific λ. A β-specific λ could be introduced when predictor variables are in different scales, or prior information on predictor variables requires that a specific variable needs a higher or lower λ. Both do not apply to our research. The main advantages of Lasso are its simplicity and interpretability, its feature selection, the handling of collinearity (as explained in Chapter 4), and the dampening of model variance.

In panel data regression, it is common practice to include fixed or random effects in the regression model to account for endogeneity caused by unobserved heterogeneity, i.e. the correlation between the independent variable and the error term, caused by the unobserved dependency of other independent variables. This is particularly useful in the case of estimation, where you are interested in the causal relationships between dependent and independent variables. As stated before, we are in the first place not interested in causal relationships, but in prediction accuracy. While including fixed or random effects are very useful in estimation models where the goal is to understand causality, it is not necessary and might not be beneficial to include them in prediction models.

The hyperparameter that is most interesting for the model improvement, is the tuning parameter λ. This parameter determines the amount of shrinkage of the coefficients. When $\lambda = 0$, there is no shrinkage and the coefficients will be equal to the RSS estimates. As λ increases, the amount of shrinkage increases. The shrinkage is based on the relative predictor importance of the variables, variables with low importance have a higher shrinkage than those with high importance. Other tuning parameters are the number of iterations, the tolerance for optimization, i.e. at what level of improvement do we conclude that the model is no longer significantly improving and stop the model, whether to fit an intercept, and whether the updating of coefficients follows a sequence or happens randomly.

Hyperparameter Grid	**Total**	**Scope 1**	**Scope 2**	**Scope 3**
λ: [0.1, 0.01, 0.001, 0.001]	0.001	0.0001	0.0001	0.001
Intercept: [True, False]	True	False	False	True
Max iterations: [10000, 100000, 1000000, 10000000]	100000	10000	10000	1000000
Tolerance: [1e-1, 1e-2, 1e-3, 1e-4, 1e-5, 1e-6, 1e-7]	1e-2	1e-7	1e-4	1e-2
Selection: ['cyclic', 'random']	'random'	'random'	'random'	'random'

Table 5.1: Results of the hyperparameter grid search cross-validation for the Lasso-model.

Table 5.1 shows the results of the hyperparameter grid search cross-validation for the Lasso model. These combinations of hyperparameters performed best in terms of negative MAE. We observe that models for ‘Total’ and scope 3 emissions perform better with a higher λ, while models for scope 1 and scope 2 perform better with a lower λ. This implies that, for the latter models, a lower shrinkage penalty is applied, meaning that more predictor variables have non-zero coefficients. Furthermore, we observe that for 'Total' and scope 3, the Lasso model performs best with similar hyperparameter values, whereas the same goes for scope 1 and scope 2, with the exception of optimization tolerance.

5.2.2 Decision Trees

As discussed in Section 3.1.4.1, decision trees are tree-structured machine learning algorithms that can be combined into two decision tree ensemble methods, namely bagging and boosting. We consider two specific applications that we use in our research.

5.2.2.1 Extreme Gradient Boosting

The XGBoost is an ensemble method that uses a sequence of many weak learners, in this case, shallow decision trees, and combines them into a superior model. Here, each subsequent decision tree learns from the previous tree. This optimization is done using gradient descent, which is a commonly used optimization algorithm in the field of machine learning for finding the minimum of a loss function. XGBoost is known for its speed and superior performance over other methods in many machine learning competitions. Furthermore, the XGBoost provides feature importance scores for the models which makes it easier to interpret the outcomes of the models and the relative importance of the predictor variables.

The are two common strategies for the generalization of tree-based models, namely pre-pruning and post-pruning. Pre-pruning involves stopping the construction of the trees early, whilst post-pruning involves removing splits in the tree that have a low information gain after the model is built. We will focus only on pre-pruning. For the XGBoost, we apply the grid search on five hyperparameters. First, we have the learning rate. In gradient descent optimization, a learning rate is used as a step size by which the gradient is updated. In other words, the learning rate decides at what rate the gradient descent optimization algorithm converges to the optimal solution, that is, the minimum of a loss function. Thorough consideration of the choice of the learning rate is important, as a too-low value leads to slow convergence, whereas a too-high value leads to possible overshooting of the optimal solution. Other tuning parameters are the maximum depth of the trees, the minimum sum of instance weights needed in a child, i.e. a measure to control the minimum number of samples that must be present in a node during training, the minimum loss reduction, and the ratio of the training sample that is used for constructing a single tree.

Hyperparameter Grid	**Total**	**Scope 1**	**Scope 2**	**Scope 3**
Learning rate: [0.001, 0.01, 0.1, 0.15, 0.20, 0.30]	0.2	0.2	0.15	0.1
Max depth: [6, 8, 10, 12]	12	10	12	12
Min child weight: [1, 5, 7, 9, 11]	1	1	1	1
Gamma: [0.0, 0.001, 0.01, 0.1]	0.0	0.001	0.01	0.0
Training sample ratio: [0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]	1	0.8	1	0.9

Table 5.2: Results of the hyperparameter grid search cross-validation for the XGBoost.

Table 5.2 shows the results of the hyperparameter grid search cross-validation for the XGBoost. We observe that models for 'Total' and scope 1 emissions need a more 'aggressive' learning rate to attain better predictions that models for scope 2 and scope 3. The remaining hyperparameters do not differ significantly from each other.

5.2.2.2 Random Forest

Random Forest is an ensemble method that aggregates many separate decision trees into a superior model. Where a single decision tree has a high likelihood of overfitting the data, a collection of decision trees decreases this likelihood whilst preserving high prediction performance. A disadvantage of Random Forest is that it loses interpretability relative to single decision trees. Single decision trees can be easily visualized. This becomes inherently more difficult as the number of trees increases. The hyperparameters of the random forest that we focus on are the number of decision trees, the number of predictor variables considered when looking for a split with the highest information gain, the maximum depth of the tree, the minimum number of samples to split an internal node, i.e. the minimal number of observations that need to be present in a node for it to split in further child nodes, the minimum number of samples to split a leaf node, i.e. the minimal number of observations needed to form a leaf node (end node), and whether bootstrap samples are used to build the decision trees, instead of the entire data frame.

Hyperparameter Grid	**Total**	**Scope 1**	**Scope 2**	**Scope 3**
Number of trees: [500, 1000, 1500]	500	500	500	500
Maximum depth: [10, 20, 30 ,40, 50]	50	50	50	50
Maximum features: [‘sqrt’, ‘log2’, None]	‘sqrt’	‘sqrt’	‘sqrt’	‘sqrt’
Minimum sample split: [2, 3, 4, 5, 6]	1	1	1	1
Minimum sample leaf: [1, 2, 3, 4, 5]	2	2	2	2
Bootstrap: [True, False]	True	True	True	True

Table 5.3: Results of the hyperparameter random search cross-validation for the Random Forest.

5.2.3 Support Vector Regression

Support Vector Regression uses Kernel functions to project data in a higher-dimensional feature space to increase the prediction and computation power of linear models. In this higher-dimensional feature space, the SVR estimates a function that approximates the relationship between the dependent and independent variables. The goal of the model is to minimize the residuals, whilst allowing a bandwidth in which the model may deviate from the actual values, specified by the ϵ-insensitive error function. The advantage of SVR is that for small- and medium-sized data frames, such as our data frame. The disadvantage is that the use of higher-dimensional feature spaces leads to a decrease in the interpretability of the model.

The first hyperparameter that we tune is the type of kernel that is used in the algorithm. We test for two different kernel types, namely a linear transformation, and a transformation using a Radial Basis Function (RBF). For these specific kernel types, we have different hyperparameters to apply. First, we have gamma for the RBF, which accounts for the consideration of other data points when looking at a single data point. So, when we have a low gamma, we consider a large number of data points close to the considered data point. This would lead to possible over smoothing of the regression boundary, which can be described as underfitting. The opposite happens for a high gamma. Second, we have for both kernel types the regularization parameter. Similar to other methods, the regularization parameter determines the tradeoff between minimizing the training error and the complexity of the model.

Hyperparameter Grid	**Total**	**Scope 1**	**Scope 2**	**Scope 3**
Kernel: ['rbf', 'linear']	'rbf'	'rbf'	'rbf'	'rbf'
Gamma: (only for rbf) [0.001, 0.01, 0.1]	0.01	0.01	0.01	0.01
Regularization: [0.001, 0.01, 0.1, 1, 10]	10	10	10	10

Table 5.4: Results of the hyperparameter grid search cross-validation for the SVR.

5.2.4 Artificial Neural Network

ANNs consist of several layers of nodes. The input layer represents the predictor variables, where we have one node for each of the predictors, and an added bias. The input layer is followed by several hidden layers. The hidden layers take the summation of the weighted inputs and bias, which are the output of the nodes in the previous layer, and incorporate these in an activation function. Finally, we have one output node which takes the summation of the weighted output of the last layer and incorporates it in an activation function to produce a final output. The ANN learns by a process called backpropagation. The final output is compared to the actual value using a loss function, and the error is propagated back through the network, adjusting the weights to minimize the error function.

For the ANN, we do not only use the random grid search for determining the hyperparameters but also for determining the structure of the neural network. Naturally, the input and output layers both have a predetermined size. The structure of the hidden layers, however, is for us to determine. We use the grid search to determine the number of hidden layers to use, and the number of nodes present in these hidden layers. Next to the structure, we use grid search to determine the learning rate with which the weights are updated.

Hyperparameter	**Total**	**Scope 1**	**Scope 2**	**Scope 3**
Number of layers	13	12	12	12
Number of nodes per layer	[480, 288, 512, 288, 224, 512, 288, 32, 224, 160, 512, 224, 64]	[512, 480, 320, 32, 512, 320, 288, 192, 224, 256, 32, 32]	[512, 480, 320, 32, 512, 320, 288, 192, 224, 256, 32, 32]	[512, 480, 320, 32, 512, 320, 288, 192, 224, 256, 32, 32]
Activation function	ReLu	ReLu	ReLu	ReLu
Learning rate: [0.1, 0.01, 0.001, 0.0001]	0.001	0.01	0.01	0.01

Table 5.5: Results of the random search cross-validation for the ANN.

Table 5.5 shows the results of the random grid search for the ANN models. Note that the number of layers does not include the input and output layers. The random search comprised of a hyperparameter grid of 10 to 50 hidden layers, with a possible number of nodes between 32 and 512, looped over with a step size of 32. The learning rate grid was [0.1, 0.01, 0.001, 0.0001] and we tested solely for the 'ReLu' activation function.

5.3 Model Results

In the previous section, we discussed the hyperparameter selection for each model. In this section, we apply these hyperparameters and discuss the results of our prediction models. We performed several rounds of testing. First, we have our imputed data frames where we applied log transformation and outlier deletion. Second, we compare the results of the previous round of testing to a baseline data frame, namely the data frames where no imputations were performed. Third, we perform a bias correction for the bias introduced through the log transformation of the data.

5.3.1 Imputed Data

In the imputed data frames, we corrected the data for missing values. The goal of the imputation was to enlarge the available data while preserving the relations within the original data. We applied five prediction models and one naive predictor (mean prediction) on the imputed data frames, the results of which we discuss in this section. We present the results for each emission category separately, such that comparisons can be made between models' performances for each category. First, we look into the results of the prediction models without back-transformation to assess the performance of the models for log-transformed output. Note that we need to transform the output back to the original scale in order to make meaningful conclusions.

	Metric	**Mean**	**Lasso**	**XGB**	**RF**	**SVR**	**ANN**
Out-of-Sample	Theil's U	0.094	0.043	0.044	0.043	0.044	0.062
	MAPE	0.166	0.065	0.065	0.068	0.063	0.098
	MSE	5.736	1.235	1.330	1.259	1.316	2.772
	RMSE	2.373	1.078	1.113	1.090	1.108	1.570
	R-Squared	-0.056	0.764	0.747	0.763	0.751	0.516
In-Sample	Theil's U	0.095	0.043	0.003	0.014	0.036	0.061
	MAPE	0.166	0.063	0.004	0.020	0.043	0.096
	MSE	5.729	1.171	0.005	0.118	0.850	2.679
	RMSE	2.394	1.082	0.071	0.344	0.922	1.534
	R-Squared	0	0.800	0.999	0.979	0.852	0.532
Bias-Variance	Bias	-	1.238	1.211	1.284	1.310	5.819
	Variance	-	0.010	0.164	0.011	0.063	1.145

Table 5.6: Results for the emission category 'Total' for the imputed data frame, without back-transformation of the output.

Table 5.6 shows the results for the emission category 'Total' for the imputed data frame before the transformation of the output back to the original scale. We observe that, in general, the in-sample performance is better than the out-of-sample performance, as we expected. The naive estimator shows inferior prediction performance in terms of all accuracy measures. The MAPE has a promising value of 0.166. However, this is most probably caused by the log transformation that pushes the output into a relatively small output range. The R-squared takes a negative value. Referring to Equation 3.5, this can happen when the sum of squared residuals of the prediction model is higher than the sum of the squared distance between the output and the mean of the training output. Generally, a negative R-squared indicates that the prediction model does quite a poor job of fitting the data. Looking at our prediction models, we observe that the out-of-sample performance is similar for the Lasso, XGBoost, Random Forest, and SVR models. Theil's U is close to zero for all models, indicating that the models' predictions are significantly outperforming the no-skill predictor, and the predictions are close to the observed values. This is also reflected in the low scores for the error measures MAPE, MSE, and RMSE. Based on the MAPE, we can conclude that these models are able to accurately predict the emission output in the log-transformed feature space. For in-sample predictions, the XGBoost is clearly superior to other prediction models, with error terms close to zero for all accuracy measures. Finally, we observe that the bias-variance trade off is tilted towards model bias. Theoretically, this implies that all models are currently underfitting the data, and that model complexity needs to be increased for the errors to decrease.

As stated earlier, we need to transform the output back to the original scale in order to make comparisons that are meaningful within the context of this research. Therefore, we need to take the exponential of both the predicted values and the observed values. This implies that we take $e^{\hat{Y}}$, and e^{Y} as the input for our accuracy measures.

	Metric	**Mean**	**Lasso**	**XGB**	**RF**	**SVR**	**ANN**
Out-of-Sample	Theil's U	0.930	0.366	0.477	0.600	0.389	0.460
	MAPE	14.398	3.739	3.806	3.045	5.521	5.635
	MSE	1.18e14	5.44e13	7.17e13	7.67e13	6.3e13	6.26e13
	RMSE	8378767	5330841	6328310	8422895	5711859	7587840
	R-Squared	-0.147	0.308	0.073	0.303	0.138	0.425
In-Sample	Theil's U	0.971	0.367	0.049	0.252	0.338	0.458
	MAPE	14.367	2.536	0.067	0.256	3.683	5.653
	MSE	1.17e14	5.01e13	1.09e12	2.1e13	4e13	5.99e13
	RMSE	10838823	7079848	1039885	4579171	6333988	7728440
	R-Squared	-0.073	0.542	0.990	0.808	0.633	0.452
Bias-Variance	Bias	-	1.238	1.211	1.284	1.310	5.819
	Variance	-	0.010	0.164	0.011	0.063	1.145

Table 5.7: Results for the emission category 'Total' for the imputed data frame.

Table 5.7 shows the results of the prediction models for the emission category 'Total' for imputed data after transforming the output back to the original scale. Immediately, we observe that, in the original scale, the prediction models do not perform well for out-of-sample predictions. Although Theil's U suggests that all models perform significantly better than a no-skill predictor, the accuracy measures show large errors in predicting the 'Total' emissions. The Random Forest shows the lowest MAPE, meaning that it is able to make predictions closest to the observed values. However, the best prediction model is still not able to accurately approach the actual output values. The MAPE of the Random Forest suggests that the difference between its predicted values and the observed values is, on average, approximately 300%. The absolute error measures are substantial and also indicate a large deviation between the predicted values and the observed values. After transforming the output back to the original scale, the residuals exponentially increase, causing the error measures to significantly increase. Especially for outliers, the back-transformation leads to a significant increase in prediction error. Next to high error measures, the R-squared is consistently low for all prediction models, indicating that our independent variables are not adequately explaining the variability in our dependent variable. Furthermore, consistent with the results in Table 5.6, the bias-variance trade off is tilted towards high model bias, implying that the prediction models are underfitting the data. For most prediction models, the in-sample predictions also show poor performance in terms of accuracy. The tree-based models have a relatively low MAPE, suggesting that they perform well on data that was used during the training phase of the model. However, due to poor model generalization, this does not results in significantly higher results for out-of-sample predictions.

Overall, we observe that the prediction models perform best for the 'Total' emission category, followed by scope 2, scope 1, and scope 3 emissions, respectively. Scope 2 emission prediction accuracies are in the same order of magnitude as the prediction accuracies of 'Total' emissions. We observe a performance decline when looking into the results of scope 1 emissions, whereas the performance plummets when we look into the results of Scope 3 emissions. We assess the differences in prediction performance in later sections. The prediction results of scope 1, 2, and 3 emissions are found in Appendix B.1. In Appendix B.2, we visualized the increase in error after transforming the output back to the original scale by plotting the observed values against the predicted values, plotting a regression line for the observed and predicted values (red), and comparing the plotted regression line to the line $y = x$ (blue).

5.3.2 Baseline data

In the previous section, we showed the results of the prediction models on our imputed data frames. As stated earlier, the initial motivation to implement multiple imputation on our data frame, was to enlarge the data frames, while preserving the variable distributions. It is possible that the multiple imputation process added bias to the data. Predictive mean matching assigns values that are already in the data to impute missing values. This could have an impact on the dynamics with which independent variables interact with each other within the prediction models. Therefore, it is important that we consider unimputed (complete cases) data as a baseline with which we compare the results of the previous section. Consequently, the number of available data rows for the training and testing of the prediction models is cut in more than half the size of that of the imputed data, for all emission categories. This could have a negative impact on the prediction performance, as there is less data available to train the models.

	Metric	**Mean**	**Lasso**	**XGB**	**RF**	**SVR**	**ANN**
Out-of-Sample	Theil's U	0.808	0.232	0.399	0.487	0.239	0.336
	MAPE	9.243	4.178	3.185	1.972	5.188	4.094
	MSE	7.28e13	1.86e13	3.37e13	4.08e13	2.09e13	2.54e13
	RMSE	5842261	2656402	3868111	5756435	2841653	4628389
	R-Squared	-0.216	0.334	-0.073	0.446	0.546	0.596
In-Sample	Theil's U	0.951	0.253	0.008	0.191	0.178	0.352
	MAPE	9.208	2.656	0.010	0.174	2.718	4.494
	MSE	7.28e13	1.64e13	1.92e10	8.12e12	8.32e12	2.56e13
	RMSE	8530302	4047998	136617	2848105	2884342	5019654
	R-Squared	-0.091	0.754	0.999	0.878	0.875	0.616
Bias-Variance	Bias	-	0.589	1.350	0.822	1.163	5.718
	Variance	-	0.021	0.169	0.009	0.069	1.484

Table 5.8: Results for the emission category 'Total' for the baseline data frame.

Table 5.8 shows the results for the emission category 'Total' for the baseline data frame. We observe a decrease in Theil's U for all prediction models, indication that the overall prediction performance of the models has improved. The accuracy measures decreased for most prediction models. The Random Forest model shows the most significant accuracy improvement, where the MAPE decreased by 1.073, to 1.972. The overall increase in prediction performance indicates that the multiple imputation method imposed on the data has added a bias to the data frames. As predictive mean matching assigns values available in the data to missing data, it is possible that bias was added to the data, which causes inferior prediction performance compared to predictions done on the baseline data. Another possible explanation is that the cutoff proportion for the imputation of an allowed maximum percentage of missing values of 30% is too high. The Lasso model is the only model that performs better in the imputed data frame. This could be explained by the fact that the Lasso model performs better with an enlarged sample size.

Overall, we observe that the trend described above also applies on the remaining emission categories. Notable is the fact that the prediction models perform best on scope 2 emissions for the baseline data frames, followed by 'Total', scope 1, and scope 3 emissions, respectively. The prediction results of scope 1, 2, and 3 emissions are found in Appendix B.3.

5.3.3 Log transformation bias correction

During pre-processing of the data, we decided to apply log transformation to improve model performance. Without such a transformation, the prediction models are not able to accurately predict the emission data, due to the skewed nature of our data (Appendix B.4 shows the results of the Random Forest model applied on untransformed data). However, introducing a log transformation means that we will have to transform the data back to the original format in order to make meaningful comparisons and conclusions. When transforming the output back to the original format, we introduce a transformation bias to the predictions.

As a logarithmic function is a concave function, the log transformation of the dependent and independent variables can be seen as a concave transformation. Jensen's inequality states that we have the following for concave transformations (More, 2022):

$$E[f(X)] \leq f(E[X]) \tag{5.1}$$

In the context of our research, this looks as follows:

$$E[log(Emissions)] \leq log(E[Emissions]) \tag{5.2}$$

$$exp(E[log(Emissions)]) \leq E[Emissions] \tag{5.3}$$

Equations 5.2 and 5.3 show that a transformation bias is introduced when transforming the output of our prediction models back to the original format. Given this bias, and the poor observed performance in previous data frames, we introduce a bias correction. We assume that for all test folds, the residuals are not normally distributed. This gives us the following bias correction:

$$BC = e^{\epsilon} = \frac{\sum_{i=1}^{N} e^{\epsilon_i}}{N} \tag{5.4}$$

The corrected prediction is then calculated as follows:

$$E[Y] = BC * f^{-1}(E[f(Y)]) \tag{5.5}$$

The bias correction of the prediction models is applied on the baseline data frames. Theoretically, this should imply that both imputation and transformation bias are both corrected for.

	Metric	**Mean**	**Lasso**	**XGB**	**RF**	**SVR**	**ANN**
Out-of-Sample	Theil's U	0.557	0.263	0.404	0.349	0.278	0.895
	MAPE	72.774	6.686	7.785	3.242	7.070	114.291
	MSE	5.86e13	5.40e14	1.63e14	2.83e13	6.64e14	1.31e17
	RMSE	5256347	638853	7289789	4875121	6479286	2.61e08
	R-Squared	0	-4.989	-22.002	0.582	-6.436	-1628
In-Sample	Theil's U	0.715	0.299	0.008	0.179	0.206	0.897
	MAPE	78.562	6.702	0.010	0.182	3.582	230.359
	MSE	6.67e13	3.26e13	1.90e10	7.31e12	1.45e13	8.26e16
	RMSE	8167803	5710122	136023	2702696	3806028	2.44e08
	R-Squared	0	0.511	0.999	0.890	0.783	-1260
Bias-Variance	Bias	-	0.606	1.336	0.818	1.149	5.313
	Variance	-	0.010	0.173	0.009	0.066	1.990

Table 5.9: Results for the emission category 'Total' for the baseline data frame after bias correction.

Table 5.9 shows the results for the emission category 'Total' for the baseline data frame after bias correction. It becomes immediately clear that the bias correction did not have its desired result. Theoretically, the negative bias created by the log transformation should be reduced by incorporating a bias correction of a value larger than 1, consequently leading to better prediction performance. However, we observe that the prediction performance has decreased with respect to previous results. The MAPE, MSE, and RMSE have all increased, meaning that the overall prediction accuracies of the model have decreased. Also, most models show a negative out-of-sample R-squared. Again, this indicates that the model predictions are wildly different from the actual emission output. Although, theoretically, the log transformation leads to a negative bias in the prediction output, we observe that for approximately half of the predictions, the model overestimates the emissions. Applying a bias correction of a value larger than 1 will in those cases increase the prediction errors, thus explaining the decreased prediction performances.

Overall, the best model performance is attained in the 'Total' emission category, followed by scope 2, scope 1, and scope 3 emissions, respectively. The prediction results of scope 1, 2, and 3 emissions are found in Appendix B.5.

5.3.4 Follow-up testing

In the previous sections, we observed that the prediction models could not accurately predict corporate GHG emissions. We follow up on these results by performing additional tests using linear regression, scaled variable predictions, and other combinations of sector and geographical classifications.

Appendix B.8 shows the results of the linear regression model on the baseline data frame. We observe that, for scope 2 emissions, linear regression outperforms the other prediction models in terms of out-of-sample prediction accuracy. We do not observe improved performance for the other emission categories. Furthermore, we tested the prediction performance of the random forest model on input and output variables, both normalized for 'Revenue'. In other words, we assess whether the prediction performance increases when we predict corporate GHG emissions per dollar of revenue. Again, we see a slight increase in prediction accuracy, without coming significantly close to observed emission values.

Appendix B.9 through B.11 show the results for the application of other specified sectors and geographical classifications. In the previous sections, we used 'GICS Sector Name' as the sector classifier, and 'Country' as the geographical classifier. Appendix B.9 shows the results for sector classifiers 'GICS Sub-Industry Name', 'GICS Industry Name', and 'GICS Industry Group Name'. However, we observe no significant improvements in terms of prediction accuracy. Appendix B.10 shows the results for predicting within sector subsets. We divided the imputed data frame into 11 subsets, based on the variable 'GICS Sector Name'. Within these data frames, we added the sector classifier 'GICS Industry Name', such that we could differentiate sectors within the sector subsets. We observed the highest prediction accuracy for the 'Industrials' sector and the lowest prediction accuracy for the 'Energy' sector. Lastly, we tested for the geographical classifier 'Region', instead of 'Country'. Again, we observed no significant increase in prediction accuracy.

5.4 Feature Importance

In the previous section, we observed that our models are unable to accurately predict corporate GHG emissions. In order to make meaningful conclusions on these performances, we want to interpret the output by assessing feature importance. Feature importance tells us the relative 'importance' a predictor variable, or feature, has in the prediction model. We perform this assessment by implementing SHAP values (SHapley Additive exPlanations).

5.4.1 SHAP

SHAP is a method originating from cooperative game theory, where the output for a player depends on the relative contribution of that player. When applied to machine learning, we can assess the relative contribution of a single predictor variable on the prediction output. Using SHAP values allows us to increase the interpretability of our models, and provides useful information for further research. In this section, we investigate three explanatory SHAP plots on the prediction performance of the random forest model on the baseline data frame (our best-performing model)[1]. Using these plots, we will explain what happens within the model.

The random forest model applied on the baseline data frame is our best-performing model for the 'Total' emission category, with a MAPE of 1.972. This is the result of the use of group K-validation using 10 folds. We analyze the best-performing and the worst-performing fold to assess whether we observe differences in the models that may explain the variability in prediction accuracy among the 10 validation folds. Furthermore, the additive explanation may inform us about the cause of our poor prediction performances.

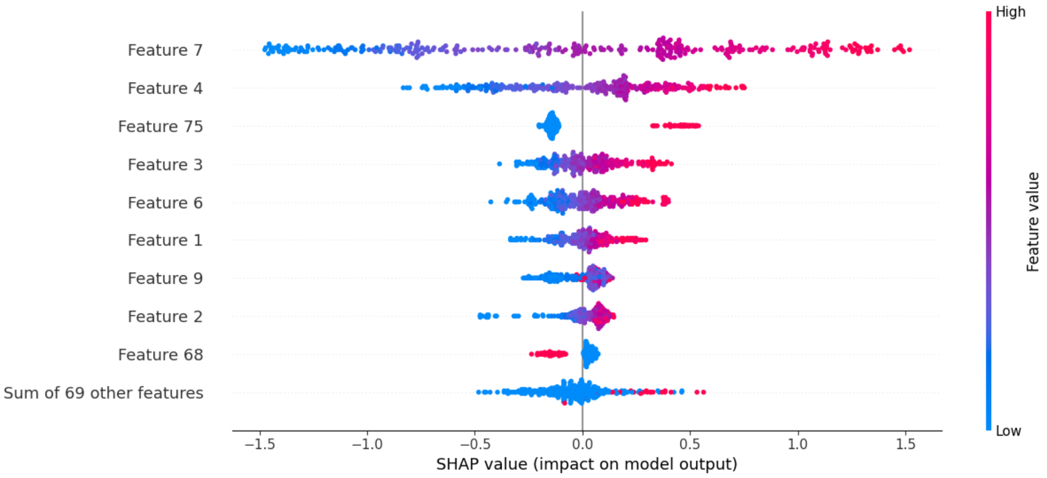


Figure 5.2: Beeswarm plot of the best-performing fold of the random forest model on the baseline data for the 'Total' emission category.

[1] For figures on SHAP values of other models, emission categories, or data frames, please consult the author.

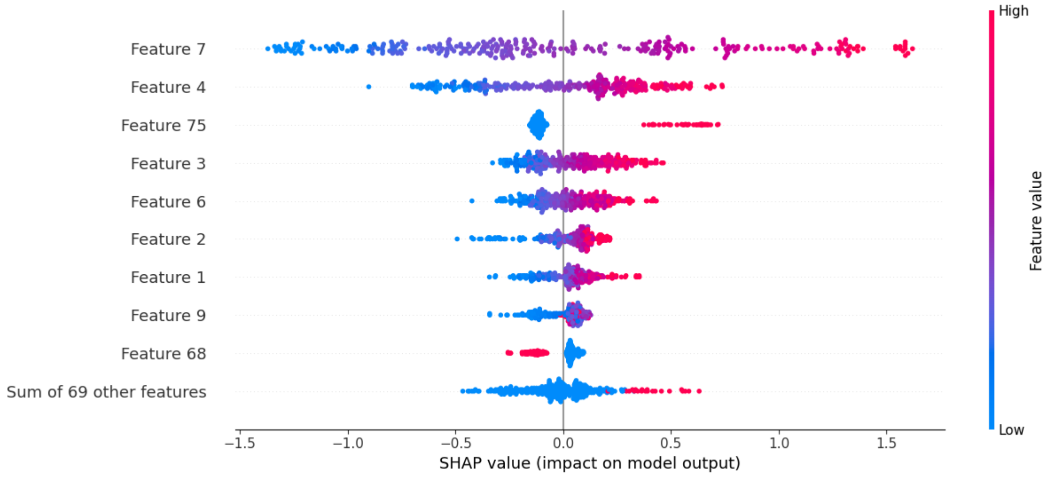


Figure 5.3: Beeswarm plot of the worst performing fold of the random forest model on the baseline data for the 'Total' emission category.

Figures 5.2 and 5.3 show beeswarm summary plots of the best-performing fold and the worst-performing fold, respectively, of the random forest model on the baseline data for the 'Total' emission category. An overview of all features and their corresponding variable name can be found in Appendix B.6. The beeswarm plots give summaries of how the most important input features affect the models' output. The x-axis displays the SHAP value, which is defined as the impact the feature has on model output. Each dot represents the impact of that feature on a single data row. The dots are stacked on top of each other to represent the density of SHAP values for a single feature, whereas the color of the dot represents the value of a feature.

The first thing that we observe is that there are no major differences between Figures 5.2 and 5.3. For both folds, we see that the most important feature is feature 7, 'Energy Purchased'. The higher the level of 'Energy Purchased', the higher the level of 'Total' emissions. This corresponds with what we would expect, as 'Energy Purchased' is classified under scope 2 emissions. After 'Energy Purchased', feature 4 is the most important feature, which is represented by 'Net PPE'. 'Net PPE' is short for Net Property, Plant & Equipment and represents the physical assets that a corporation has on its balance sheet. The SHAP value for 'Net PPE' implies that the higher the level of net property, plant, and equipment a corporation has on its balance sheet, the higher the level of Total emissions. This can be explained by a simple correlation of more machinery equals more emissions. However, 'net' implies that it is net of accumulated depreciation of physical assets. Thus, a lower level of 'Net PPE' could also indicate a high level of machinery that is amortized over the years, implying a higher level of outdated machinery. Remaining features 75, 3, 6, 1, 9, 2, and 68 are represented by 'GICS Sector Name: 'Information Technology', 'Capital Expenditure', 'Total Assets', 'Revenue', 'Inventory Turnover', 'Employees', and 'GICS Sector Name: Communication Services', respectively.

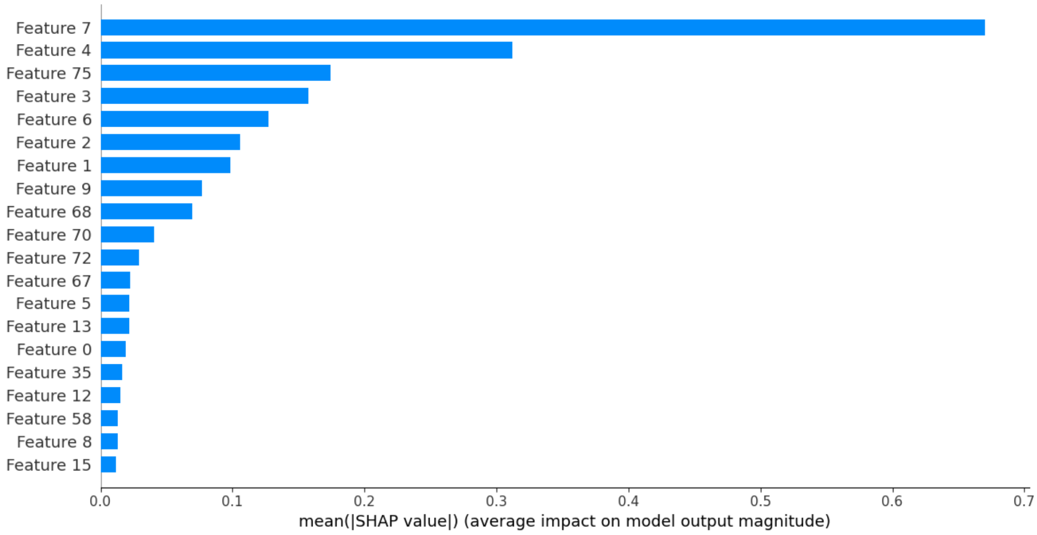


Figure 5.4: Mean absolute SHAP values of the best-performing fold of the random forest model on the baseline data for the “Total” emission category.

The importance that the predictor variables have relative to each other is denoted in Figures 5.2 and 5.3 through the span width of the SHAP values for each of the displayed variables. However, the relative importance is best portrayed by the mean absolute values of the SHAP values. Figure 5.4 shows the mean absolute SHAP values of the best-performing fold of the random forest model on the baseline data for the ‘Total’ emission category. When observing Figure 5.4 it becomes inherently clear that ‘Energy Purchased’, together with ‘Net PPE’, define the prediction performance. The fact that only a small portion of the predictor variables has such a large effect on the prediction output, combined with the fact that we have high prediction errors and high model bias, states that our models correctly capture the high-level relationships in the data, but do not capture remaining relationships that could explain the variability in corporate GHG emissions.

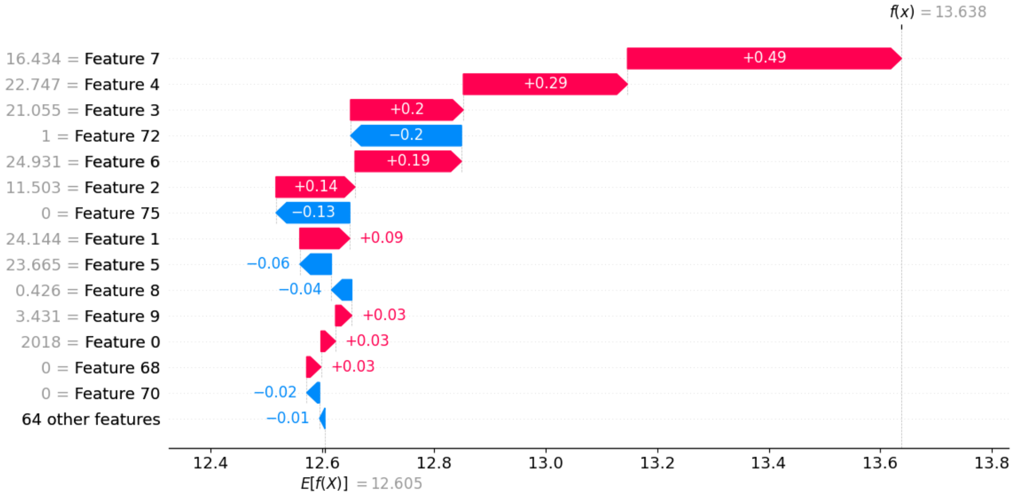


Figure 5.5: Waterfall plot for the first corporation in the best-performing fold of the random forest model on the baseline data for the ‘Total’ emission category.

We assess a single prediction within the previously mentioned random forest fold to understand how the SHAP values, i.e. the predictor variables, affect the model output. Figure 5.5 shows the waterfall plot for the first corporation in the best-performing fold of the random forest model on the baseline data for the ‘Total’ emission category. The bottom of the waterfall plot starts at the expected value of the model output, $E[f(x)]$, which is the average model output. From there, we see the effect that the 14 most important features have towards reaching the model output, $f(x)$, for this specific corporation. The three most important predictor variables for this prediction are ‘Energy Purchased’, ‘Net PPE’, and ‘Capital Expenditure’, which all have a positive (red) impact on the model output. Again, we observe that ‘Energy Purchased’ supplies a significant portion of the model output.

The SHAP values of our predictor variables give a clear insight into how our prediction models operate and increase the interpretability of our models. Here, we again need to note the difference between prediction and causality, referring to Section 3.1.5 on causality versus prediction. We observe high SHAP values for the predictor variables ‘Energy Purchased’ and ‘Net PPE’, meaning that these variables have a strong influence on the models’ prediction output. However, this does not necessarily imply that a direct causal relationship between these variables and the output is present in the real world. In other words, the SHAP values do not imply that unilaterally changing the value for one predictor variable, for example ‘Energy Purchased’, has a direct proportional impact on the actual observed emissions. What it does imply, is that given our data and models, the predictor variable ‘Energy Purchased’ is crucial in determining corporate GHG emissions. Given that our goal is purely prediction-oriented, it is useful that there is a strong positive relationship between ‘Energy Purchased’ and corporate GHG emissions. However, the important predictor variables do not provide us with enough information as to how each feature independently relates to corporate GHG emissions.

5.4.2 Monotonic Relationships

The previous section gave us insight into the importance of the predictor variables in our prediction models. We observed that the predictor variable ‘Energy Purchased’ is crucial in determining the prediction output and that a large portion of the predictor variables is trivial. The SHAP value analysis helps us understand the results of our prediction models. In turn, we can explain the SHAP values by assessing the presence or absence of monotonic relationships between the numeric predictor variables and corporate GHG emissions. The relationship between an independent variable and the dependent variable is monotonic when the value of the dependent variable consistently increases or consistently decreases as the value of the independent variable increases. The presence of a monotonic relationship between the dependent variable and an independent variable increases the probability that the feature will have high importance in the prediction model.

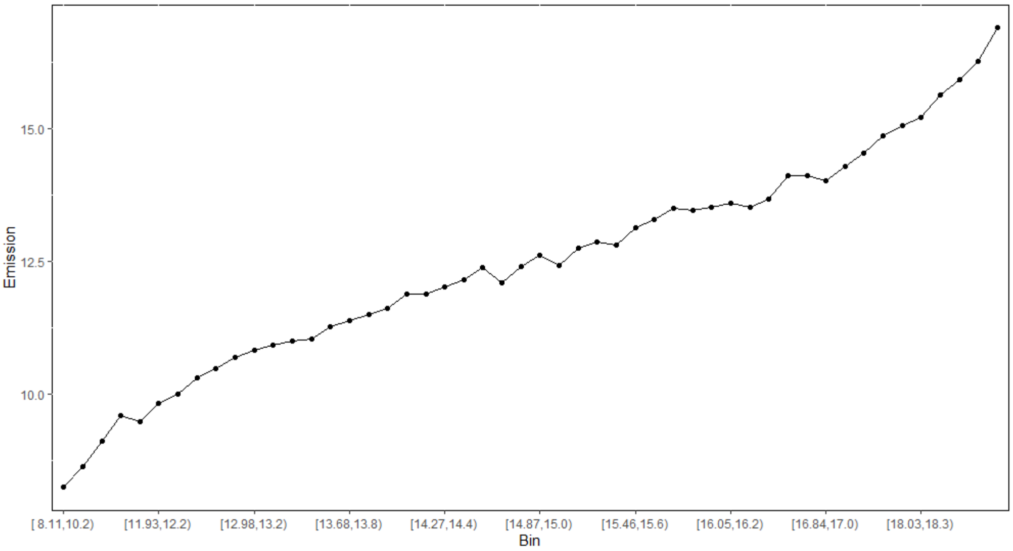


Figure 5.6: Monotonic relationship between ‘Energy Purchased’ and ‘Total’ emissions.

Figure 5.6 shows the relationship between ‘Energy Purchased’ and ‘Total’ emissions. We divided the values of ‘Energy Purchased’ into 50 quantiles, and, for each quantile, we calculated the mean ‘Total’ emissions. We observe an increasing monotonic relationship between the two variables. This observation is in line with what we observe in Figure 5.3; both figures imply that there is a constant positive correlation between ‘Energy Purchased’ and ‘Total’ emissions. While this monotonicity implies a consistently positive relationship, the existence of this relationship does not guarantee a high SHAP value. However, in the context of our research, we observe that more important features display a higher level of monotonicity, hence giving some explanation as to why certain features have higher importance than others. Take for example feature 11, ‘ROE’, which is not present in the top 20 most important features for the ‘Total’ emission category depicted in Figure 5.4. We observe a non-monotonic relationship, which could explain the low level of importance in the prediction model. Figures on the remaining relationships between numeric features and ‘Total’ emissions are found in Appendix B.7.

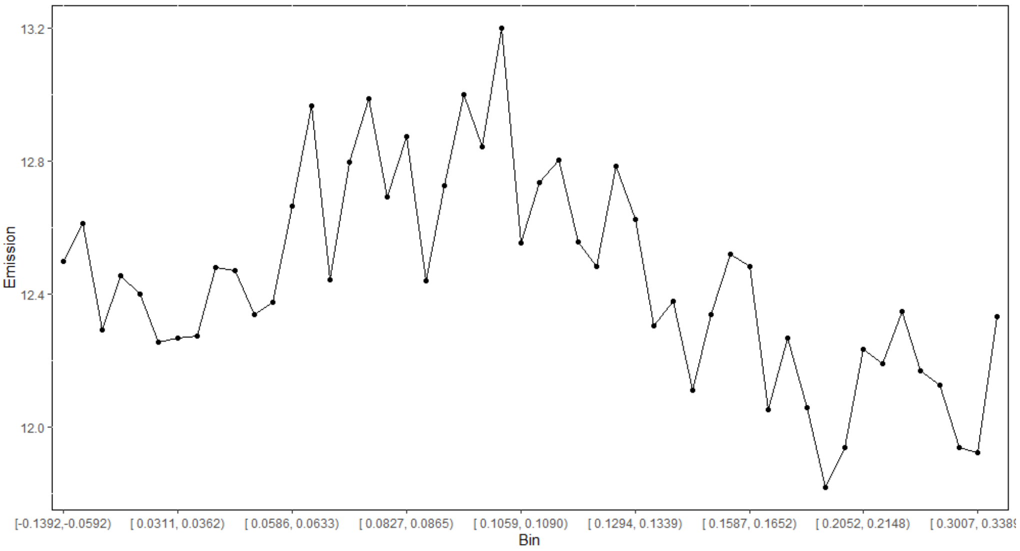


Figure 5.7: Monotonic relationship between ‘ROE’ and ‘Total’ emissions.

Chapter 6

Conclusions & Discussion

In Chapter 1, we identified the need for corporate GHG emission prediction modeling, due to the large portion of corporations that are currently not obliged to report their emissions. As a result, corporate GHG emission predictions could enable financial institutions to incorporate these emissions in their climate change risk framework. Through this research, we investigated whether we can acquire accurate corporate GHG emission predictions through the use of traditional statistical analysis, and the use of machine learning methods. In this chapter, we provide the conclusions of the research. First, we assess our research questions and discuss the results. Thereafter, we reflect on the results in the discussion.

6.1 Conclusions

The goal of the research was to implement statistical analysis and machine learning methods for the prediction of corporate GHG emissions for corporations that do not disclose this information, by using data of corporations that do disclose GHG emissions. Moreover, we focused on the distinction between the prediction performance of statistical analysis and machine learning methods. Therefore, we formulated the following research question:

What model is best suited for the prediction of GHG emissions of corporations?

Following the research question, we stated the following hypothesis:

Machine learning methods significantly outperform naive prediction and statistical analysis.

To answer the research questions, we first provided the theoretical context of the research by discussing the impact of climate change on financial institutions, and selecting possible predictor variables from literature. We identified that reliable GHG emission data could be used by financial institutions to quantify their exposure to climate change risks, specifically transition risks.
Next, we gave background information on the proposed prediction methods and described the comparative framework through which we compared the prediction performance of the models. We identified regression analysis, specifically Lasso regression (and later linear regression as a baseline), and four machine learning models, specifically Random Forest, XGBoost, Support Vector Regression, and Artificial Neural Network, as the prediction models to be implemented for the prediction of corporate GHG emissions. The introduced comparative framework consisted of two relative prediction accuracy measures, Theil's U and MAPE, two absolute prediction accuracy measures, MSE and RMSE, and a goodness-of-fit measure, R-squared. Furthermore, we introduced the concept of the bias-variance tradeoff, through which we can assess model fit.
Subsequently, we discussed data selection, preparation, and visualization, and assessed the results of our prediction models. We implemented group K-fold cross-validation for the evaluation of our models, we used grid search cross-validation and random search cross-validation to tune our hyperparameters, and we discussed the results of the prediction on imputed data, baseline data, and prediction with a log transformation bias correction. Hereafter, we assessed the feature importance of the predictor variables and discussed the results of some residual testing. With these findings, we can answer the research question.

6.1.1 General conclusions on results

In the log-transformed feature space, the prediction models are able to accurately predict corporate GHG emissions. The best prediction performance is acquired by the XGBoost, with a MAPE of 6.3% for 'Total' emissions. The log transformation dampens the effect of outliers and coerces the variables into a smaller value range. Furthermore, the log transformation causes the variables to have an approximately normal distribution. These factors induce the prediction models to utilize the added variable distributions, simplifying the process of recognizing patterns and finding relationships. Although the results in the log-transformed feature space are positive, the observed emissions and predicted emissions need to be transformed back to the original scale to make meaningful conclusions in the context of our research.

In the original feature space, the prediction models are unable to accurately predict corporate GHG emissions. Logically, when both observed emissions and -predicted emissions are exponentiated, the error also increases exponentially. The result is that the prediction models have high error scores, making them not pragmatically useful for their intended purpose. The best prediction results for 'Total' emissions, in terms of MAPE, are 304.5%, 197.2%, and 324.2%, for the imputed data frame, the baseline data frame, and the application of a bias correction, respectively. These results suggest that the imputation process added a bias to the data that negatively impacted the prediction performance. Besides, the bias correction did not have the desired effect and led to a decrease in prediction performance. Predictions for scope 2 emissions had results similar to 'Total' emissions. Scope 1 emissions performed significantly worse with twice as high values for the MAPE, as compared to 'Total' and scope 2 emissions. For scope 3 emissions, the prediction errors increased by a factor of 100 as compared to the errors of the other emission categories.

In general, the prediction models are not able to grasp the complexity of the data, and as a result, do not generalize well. We observe that some prediction models perform well for in-sample prediction, hinting at the fact that the models are overfitting on the training data. However, the observed bias-variance tradeoffs suggest that the prediction models are in the underfitting zone of the tradeoff. The cause is most probably the fact that the models are correctly capturing the most important and prominent patterns and relationships in the training data, causing high in-sample performance, but missing the variability inferred by remaining relationships, hence the high bias that suggests an oversimplification of the variable relationships.

The analysis of the SHAP values confirms the above statement, as we observe that the predictor variable 'Energy Purchased' has a significantly higher mean absolute SHAP value than the remaining predictor variables. The fact that 'Energy Purchased' supplies such a significant portion of the final prediction output suggests that the models capture the positive relationship between this variable and 'Total' emissions, whilst not recognizing usable relationships amongst the remaining predictor variables, indicating poor generalization performance. The superiority of 'Energy Purchased' in the prediction models is most logically explainable for 'Total' emissions and scope 2 emissions, as purchased energy falls directly under scope 2 emissions. However, we observe that 'Energy Purchased' is also the superior feature in the prediction models for scope 1 and scope 3 emissions. This indicates that the models for these emission categories overestimate the relationship between 'Energy Purchased' and these emission categories, due to a lack of other significant predictors, causing worse prediction results for scope 1 and scope 3 emissions, as compared to 'Total' and scope 2 emissions.

In the follow-up testing, we affirmed that, given our data frames, we did not overlook viable combinations of predictor variables that lead to enhanced prediction performances. We included tests for the remaining sector classifiers and the remaining geographical classifier. We observed that increasing the granularity of the sector classification and decreasing the granularity of the geographical classification, does not improve the prediction performances for emission categories 'Total', scope 1, and scope 3. We observe a slight increase in prediction performance for scope 2 emissions, however, this increase in performance is not significant. Hence, increasing the granularity of the sector classification, and consequently increasing model complexity, does not improve the models' performances. Finally, the prediction performances within sectors lead us to conclude that, even within sectors, there is a high level of variability and complexity not grasped by the prediction models.

6.1.2 Conclusion on the research question

The research found that the prediction models are unable to accurately predict corporate GHG emissions in the original feature space. The research question focuses on the difference in model performance, specifically distinguishing between the performance of naive prediction, statistical analysis, and machine learning methods. Both statistical analysis and machine learning methods significantly outperform a naive predictor, in our case the mean of the subjected data frame, for emission categories 'Total', scope 1, and scope 2. For scope 3, taking the mean of the subjected data frame gives higher prediction accuracies for some instances. Next to this, the values of Theil's U, for both statistical analyses and machine learning, indicate that our prediction models outperform a 'worthless' predictor.

The distinction between the performance of statistical analysis and machine learning methods is not unambiguous. In the baseline data frame, the best-performing machine learning method outperforms statistical analysis for emission categories 'Total', scope 1, and scope 3. For scope 2 emissions, both lasso regression and linear regression (Appendix B.8) outperform machine learning methods. In Chapter 1, we stated the hypothesis that machine learning methods significantly outperform naive prediction and statistical analysis. Given that machine learning methods do outperform naive prediction but not statistical analysis, we must reject the hypothesis.

6.2 Discussion

In the discussion, we reflect on the results and conclusions of the research by considering data limitations, discussing our assumptions and choice of actions, describing the theoretical contribution, and formulating recommendations for further research.

6.2.1 Data limitations

The foremost limitation of our research is data availability. The need to predict corporate GHG emissions identified earlier in this research indicates that the current availability of data on corporate GHG emissions is scarce. Hence, we have a relatively low number of data points with which we can train our prediction models, and, in general, machine learning methods benefit from a high number of data rows. In addition, we obtained our data exclusively from Refinitiv Eikon, which also limited the number of corporations from which we could extract data.

Next to data availability, we have the limitation of data quality and reliability. Currently, many corporations are in the early stages of tracking emission data, and, therefore, may lack the necessary processes and expertise to accurately measure their emissions. Although there are several corporate standards that provide requirements and guidance for corporations that disclose emissions, there is a lack of a single globally accepted standard that emission-disclosing corporations follow. This may lead to the fact that similar corporations in the same industry measure their emissions differently. The variability in emission measurement processes between similar corporations makes it difficult to detect generalized patterns and relationships in the data, making prediction modeling inherently complicated.

Finally, we note the lack of available ESG data. Besides a scarcity of emission data, there is also a scarcity of other ESG-related data, such as energy consumption, waste, and water usage. This scarcity lead us to examine if we were able to accurately predict corporate GHG emissions without the usage of ESG-related predictor variables, but instead, use corporate financials, sector classifiers, and geographical classifiers as predictor variables. The results of our research imply that without ESG data, it is very difficult to accurately predict emissions, because, one, the use of corporate financials, sector classifiers, and geographical classifiers leads to poor prediction accuracies, and two, because we see that the most important predictors in our models are information on energy consumption ('Energy Purchased'), and information on corporations' physical assets ('Net PPE'). Thus, we believe that the disclosure of information on ESG data, physical assets, production processes, and materials is critical for improving the prediction performance of our models.

6.2.2 Assumptions

We made several assumptions in the processes of data preparation and model improvement that may have significantly affected the outcomes of our research. First, we disregarded disclosed emission data prior to the year 2018. We did this for the sake of relevance as a large portion of the recovered emission data was derived from the last 5 fiscal years. This led to the deletion of a large number of data rows that were not taken into account for the training of the prediction models. It is possible that the inclusion of these years would have increased the prediction performances of the models due to a higher number of available data points during the training stage.

During the handling of outliers in the data, we deleted a large number of outliers. We deleted data outside of the interquartile range before getting useful results. Before outlier deletion, the MAPE could be as high as 1500%. Although the outlier deletion decreased prediction errors, it introduced a bias in the prediction models toward data within the selected data range. The models are not able to generalize well for corporations outside of this selected range, causing the prediction errors for these corporations to be high. Furthermore, we log-transformed the input and output variables to coerce the variables into a smaller variable range, and to decrease skewness. There are several other possible techniques to transform the data. We did not test the prediction performance on alternative variable transformation methods, and, therefore are unable to say whether the usage of these alternatives could lead to an increase in prediction performance.

Due to the large number of missing values, we chose to correct the data for these missing values. The goal of the data imputation was to enlarge the available data while preserving the relations within the original data. In the results, we identified that, through the data imputation, we introduced a bias that negatively affected prediction performance. We made several choices during the imputation stage that could have affected this outcome. First, we chose to apply the imputation method ‘predictive mean matching’. The main argument for this choice is the fact that it is the most commonly used imputation technique for continuous variables due to the fact that imputations are restricted to observed values, and it can preserve non-linear relations. Although we assessed and tested several other imputation methods, there are also methods we did not consider. Second, we decided on a maximum percentage of missing values a variable could have to be considered for the imputation process, namely 30%. This could have affected the outcome of our research in two ways: one, this cutoff led us to disregard variables that could have been significant predictors, and two, the cutoff of 30% could have led to a too high number of data points being imputed. Finally, we decided that a data row could have a maximum of three imputed data points for it to be considered for the training and testing of our prediction models. Again, this cutoff could have led to similar consequences as discussed for the maximum percentage of missing values.

The selection and tuning of hyperparameters is an important part of setting up prediction models and can improve the performance of these models. We used two methods for hyperparameter tuning: grid search cross-validation and random search cross-validation. For these two methods, we specified the hyperparameters that were to be tuned, and the hyperparameter grids that consisted of the values to be tested for. It has to be noted that we did not do this for all possible hyperparameters, and all hyperparameter values. Therefore it is possible, and even very likely, that we did not implement the ‘optimal’ combination of hyperparameters for all prediction models. Hence, it is also very likely that we did not attain the highest possible prediction accuracies for these models. However, we are confident about the fact that finding such an ‘optimal’ combination of hyperparameter values will not lead to dramatically better outcomes.

6.2.3 Contribution

This research represents a theoretical contribution to academic literature in the application of statistical analysis and machine learning methods for predicting corporate GHG emissions. Although the research resulted in prediction accuracies below anticipated levels, the value of this research lies in the examination and interpretation of the challenges and limitations currently associated with predicting corporate GHG emissions. Specifically, the research reveals that, with current data limitations, it is extremely challenging to detect patterns and relationships that aid the prediction models due to the inherent variability that is present in emission data, even within sectors. Despite poor prediction accuracies, the research demonstrated the potential of statistical analysis and machine learning methods in the analysis of corporate GHG emissions. We hope that this research serves as a stepping stone towards future research, through careful consideration of our conclusions and our recommendations for further research, discussed in the next section.

6.2.4 Recommendations

In this section, we elaborate on our main recommendations for further research. These recommendations focus on data and an alternative prediction approach.

6.2.4.1 Data

As stated earlier, we encountered several significant limitations regarding the data used in this research. Future research should focus on identifying and selecting corporations that have similar processes for measuring their GHG emissions. By doing so, prediction models will have a higher probability of finding useful patterns and relationships in the data. Furthermore, the volume of available data points should be increased by combining several data providers, so that the prediction models have a higher number of corporations to train on. Again, this would result in a higher probability of finding useful patterns and relationships in the data. Finally, future research should complement the set of available predictor variables with variables on ESG-related data, and data on production processes, materials, and other physical assets. Naturally, under the assumption that in the coming years, more data will be made available for these data categories.

6.2.4.2 Classification models

Our research focused on applying regression methods to predict a continuous output: GHG emissions. Future research might focus on adjusting the corporate GHG emission prediction problem from a regression problem, into a classification problem. Both earlier research and our research resulted in poor prediction accuracies when predicting continuous emission outputs. An alternative methodology would be to divide corporate emission data into several classes based on the magnitude of the emissions. Consequently, prediction models can be trained to predict the emission class of a corporation, instead of its actual emissions. Such a method has not yet been examined in academic literature, as to our knowledge, and, would therefore be an interesting idea for future research.

References

An, Qiguang, Lin Zheng, Qingzhao Li, and Chengwei Lin (Aug. 2022). “Impact of transition risks of climate change on Chinese financial market stability”. In: *Frontiers in Environmental Science* 10. ISSN: 2296-665X. DOI: 10.3389/fenvs.2022.991775.

Assael, Jérémi, Thibaut Heurtebize, Laurent Carlier, and François Soupé (Feb. 2023). “Greenhouse Gases Emissions: Estimating Corporate Non-Reported Emissions Using Interpretable Machine Learning”. In: *Sustainability* 15 (4), p. 3391.

Aurand, Timothy W, Wayne Finley, Vijaykumar Krishnan, Ursula Y Sullivan, Jackson Abresch, Jordyn Bowen, Michael Rackauskas, Rage Thomas, and Jakob Willkomm (2018). “The VW Diesel Scandal: A Case of Corporate Commissioned Greenwashing.” In: *Journal of Organizational Psychology* 18.1.

Bikker, Jacob A., Laura Spierdijk, Roy P. M. M. Hoevenaars, and Pieter Jelle Van der Sluis (Jan. 2008). “Forecasting market impact costs and identifying expensive trades”. In: *Journal of Forecasting* 27 (1), pp. 21–39. ISSN: 02776693. DOI: 10.1002/for.1052.

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning.* Springer.

Bloomberg (Feb. 2021). *Distributional Greenhouse Gas Emissions Estimates: Data Challenges and Modeling Solutions.* Bloomberg Content Data Solutions.

Brander, Matthew (Aug. 2012). *Greenhouse Gases, CO2, CO2e, and Carbon: What Do All These Terms Mean?* Ecometrica.

Brealy, Richard A., Stewart C. Myers, and Franklin Allen (2017). *Principles of Corporate Finance.*

Buuren, Stef van and Karin Groothuis-Oudshoorn (Dec. 2011). “mice: Multivariate Imputation by Chained Equations in R”. In: *Journal of Statistical Software* 45 (3).

Chan, Jireh Yi-Le, Steven Mun Hong Leow, Khean Thye Bea, Wai Khuen Cheng, Seuk Wai Phoong, Zeng-Wei Hong, and Yen-Lin Chen (Apr. 2022). “Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review”. In: *Mathematics* 10 (8), p. 1283. ISSN: 2227-7390. DOI: 10.3390/math10081283.

Cristianini, Nello and Shawe-Taylor (2014). *An Introduction to Support Vector Machines and other kernel-based learning methods.* 16th ed. Cambridge University Press.

Domingos, Pedro (Oct. 2012). “A few useful things to know about machine learning”. In: *Communications of the ACM* 55 (10), pp. 78–87. ISSN: 0001-0782. DOI: 10.1145/2347736.2347755.

Engels, Anita, Jochem Marortzke, Eduardo Gonçalves Gresse, Andrés López-Rivera, Anna Pagnone, and Jan Wilkens (2023). *Hamburg Climate Futures Outlook 2023.* Cluster of Excellence Climate, Climatic Change, and Society.

EU (2014). *DIRECTIVE 2014/95/EU OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL - of 22 October 2014 - amending Directive 2013/34/EU as regards disclosure of non-financial and diversity information by certain large undertakings and groups -*. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32014L0095.

– (2022a). */43/EC and Directive 2013/34/EU, as regards corporate sustainability reporting (Text with EEA relevance).* URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022L2464.

– (Dec. 2022b). “Corporate Sustainability Reporting Directive”. In: *Official Journal of the European Union* 65.

Franco, Camilo, Giulia Melica, Aldo Treville, Marta Giulia Baldi, Enrico Pisoni, Paolo Bertoldi, and Christian Thiel (2022). “Prediction of greenhouse gas emissions for cities and local municipalities monitoring their advances to mitigate and adapt to climate change”. In: *Sustainable Cities and Society* 86, p. 104114. ISSN: 2210-6707. DOI: https://doi.org/10.1016/j.scs.2022.104114. URL: https://www.sciencedirect.com/science/article/pii/S2210670722004279.

Goldhammer, Bernhard, Christian Busse, and Timo Busch (Oct. 2017). “Estimating Corporate Carbon Footprints with Externally Available Data”. In: *Journal of Industrial Ecology* 21 (5), pp. 1165–1179.

González, Clara I. and Núñez Soledad (Oct. 2021). *Markets, Financial Institutions and Central Banks in the Face of Climate Change: Challenges and Opportunities.* Banco de Espana.

Google (2022). *Descending into ML: Linear Regression.* URL: `https://developers.google.com/machine-learning/crash-course/descending-into-ml/linear-regression`.

Griffin, Paul A., David H. Lont, and Estelle Y. Sun (Mar. 2017). “The Relevance to Investors of Greenhouse Gas Emission Disclosures”. In: *Contemporary Accounting Research* 34 (2), pp. 1265–1297.

Han, You, Achintya Gopal, Liwen Ouyang, and Aaron Key (Sept. 2021). “Estimation of Corporate Greenhouse Gas Emissions via Machine Learning”. In.

Heymans, Martijn W. and Iris Eekhout (2019). *Applied Missing Data Analysis With SPSS and (R)Studio.*

IPCC (2014). *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II, III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* IPCC.

– (2021). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change.* IPCC.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2021). *An Introduction to Statistical Learning with Applications in R Second Edition.*

Kimura, Nobuaki, Ikuo Yoshinga, Kenji Sekijima, Issaku Azechi, and Daichi Baba (Dec. 2019). “Convolutional Neural Network Coupled with a Transfer-Learning Approach for Time-Series Flood Predictions”. In: *Water* 12 (1).

Laurensia, Yunika, Julio Christian Young, and Alethea Suryadibrata (2020). “Early Detection of Diabetic Retinopathy Cases using Pre-trained EfficientNet and XGBoost”. In: *Int. J. Advance Soft Compu. Appl* 12.3.

Liaw, Andy and Matthew Wiener (2002). “Classification and regression by randomForest”. In: *R news* 2 (3), pp. 18–22.

Mason, Charlotte H. and William D. Perreault (Aug. 1991). “Collinearity, Power, and Interpretation of Multiple Regression Analysis”. In: *Journal of Marketing Research* 28 (3), pp. 268–280. ISSN: 0022-2437. DOI: `10.1177/002224379102800302`.

Metcalf, Leigh and William Casey (2016). “Introduction to data analysis”. In: *Cybersecurity and Applied Mathematics.*

More, Sushant (Aug. 2022). “Identifying and Overcoming Transformation Bias in Forecasting Models”. In: URL: `http://arxiv.org/abs/2208.12264`.

Müller, Andreas C and Sarah Guido (2016). *Introduction to Machine Learning with Python.* O’Reilly Media, Inc.

Neubauer, Scott C (2021). “Global Warming Potential Is Not an Ecosystem Property”. In: *Ecosystems* 24 (8), pp. 2079–2089. ISSN: 1435-0629. DOI: `10.1007/s10021-021-00631-x`. URL: `https://doi.org/10.1007/s10021-021-00631-x`.

NGFS (Apr. 2019). *A call for action: Climate change as the source of financial risk.* Banque de France. URL: `https://www.ngfs.net/sites/default/files/medias/documents/ngfs_first_comprehensive_report_-_17042019_0.pdf`.

Nguyen, Quyen, Ivan Diaz-Rainey, and Duminda Kuruppuarachchi (2021). “Predicting corporate carbon footprints for climate finance risk analyses: A machine learning approach”. In: *Energy Economics* 95, p. 105129. ISSN: 0140-9883. DOI: `https://doi.org/10.1016/j.eneco.2021.105129`. URL: `https://www.sciencedirect.com/science/article/pii/S0140988321000347`.

Nguyen, Quyen, Ivan Diaz-Rainey, Duminda Kuruppuarachchi, Matthew McCarten, and Eric K M Tan (2023). “Climate transition risk in U.S. loan portfolios: Are all banks the same?” In: *International Review of Financial Analysis* 85, p. 102401. ISSN: 1057-5219. DOI: `https://doi.org/10.1016/j.irfa.2022.102401`. URL: `https://www.sciencedirect.com/science/article/pii/S1057521922003519`.

Pandey, Divya and Madhoolika Agrawal (2014). “Carbon Footprint Estimation in the Agriculture Sector”. In: *Assessment of Carbon Footprint in Different Industrial Sectors, Volume 1*, pp. 25–47. DOI: `10.1007/978-981-4560-41-2_2`. URL: `https://doi.org/10.1007/978-981-4560-41-2_2`.

PBAF (June 2022). *Taking biodiversity into account.* PBAF.

PCAF (Dec. 2019). *Accounting for and steering carbon: harmonised approach for the financial sector.* PCAF Netherlands.

– (2022). *The global GHG accounting and reporting standard part A: Financed emissions.* PCAF.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.85, pp. 2825–2830. URL: http://jmlr.org/papers/v12/pedregosa11a.html.

Semieniuk, Gregor, Emanuele Campiglio, Jean-Francois Mercure, Ulrich Volz, and Neil R. Edwards (Jan. 2021). "Low-carbon transition risks for finance". In: *WIREs Climate Change* 12 (1). ISSN: 1757-7780. DOI: 10.1002/wcc.678.

Singh, Seema (2018). *Understanding the Bias-Variance Tradeoff.* URL: https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229.

Sterkenburg, Tom F and Peter D Grünwald (2021). "The no-free-lunch theorems of supervised learning". In: *Synthese* 199 (3), pp. 9979–10015. ISSN: 1573-0964. DOI: 10.1007/s11229-021-03233-1. URL: https://doi.org/10.1007/s11229-021-03233-1.

TIBC (2023). *What is a Random forest?* URL: https://www.tibco.com/reference-center/what-is-a-random-forest.

Ulku, Ilayda and Eyup Emre Ulku (2022). "Forecasting Greenhouse Gas Emissions Based on Different Machine Learning Algorithms". In: ed. by Cengiz Kahraman, A Cagri Tolga, Sezi Cevik Onar, Selcuk Cebi, Basar Oztaysi, and Irem Ucal Sari. Springer International Publishing, pp. 109–116. ISBN: 978-3-031-09176-6.

UNFCC (Nov. 2018). "The Paris Agreement". In: United Nations.

Vermeulen, Robert, Edo Schets, Melanie Lohuis, Barbara Kölbl, David-Jan Jansen, and Willem Heeringa (2021). "The heat is on: A framework for measuring financial stress under disruptive energy transition scenarios". In: *Ecological Economics* 190, p. 107205. ISSN: 0921-8009. DOI: https://doi.org/10.1016/j.ecolecon.2021.107205. URL: https://www.sciencedirect.com/science/article/pii/S0921800921002640.

Yang, Sitong, Shouwei Li, and Zhilei Pan (July 2022). "Climate transition risk of financial institutions: measurement and response". In: *Applied Economics Letters.* doi: 10.1080/13504851.2022.2097630, pp. 1–11. ISSN: 1350-4851. DOI: 10.1080/13504851.2022.2097630. URL: https://doi.org/10.1080/13504851.2022.2097630.

Appendix A

Data Preparation & Visualization

A.1 GHG emission disclosures over the past 20 years

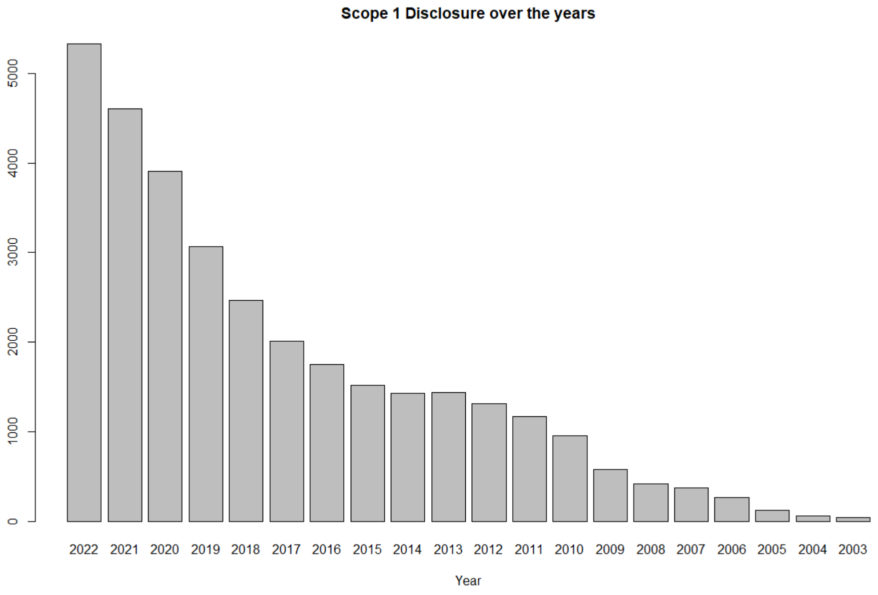


Figure A.1: Scope 1 disclosure in the previous 20 years.

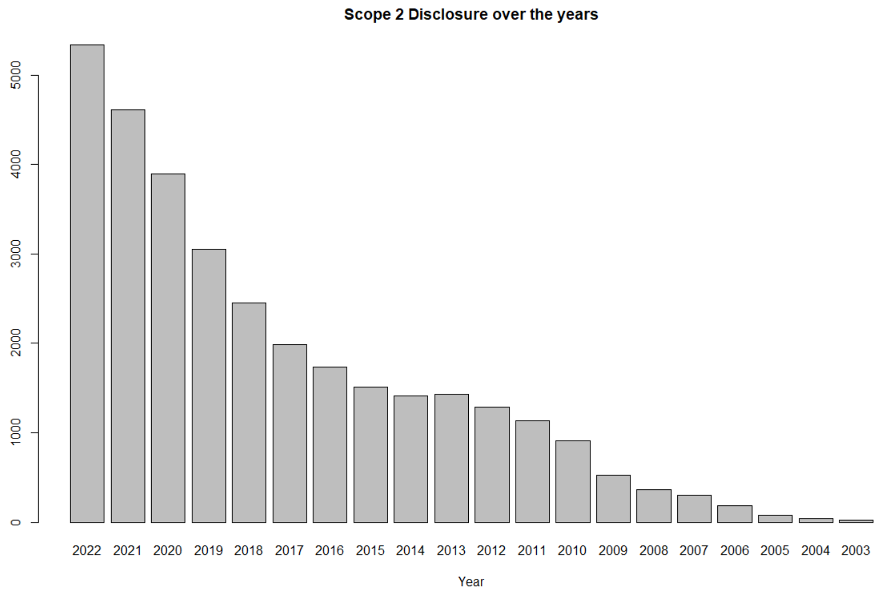


Figure A.2: Scope 2 disclosures in the previous 20 years.

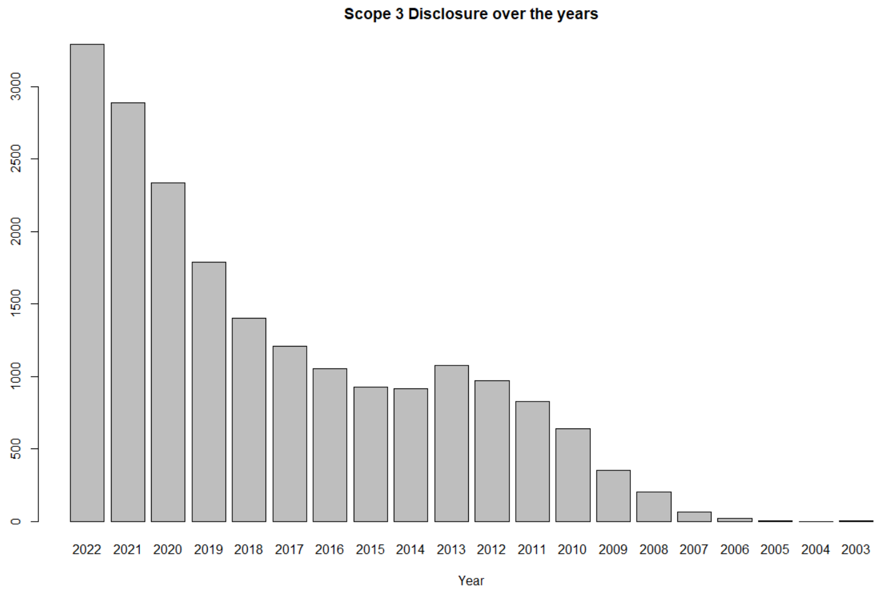


Figure A.3: Scope 3 disclosures in the previous 20 years.

A.2 Predictor variables descriptions

Predictor variable	Description
Fiscal Year	Reported fiscal year.
GICS Sub-Industry Name	Global Industry Classification Standard (GICS) Sub-Industry description.
GICS Industry Name	Global Industry Classification Standard (GICS) Industry description.
GICS Industry Group Name	Global Industry Classification Standard (GICS) Industry Group description.
GICS Sector Name	Global Industry Classification Standard (GICS) Sector description.
Country	Country of headquarters.
Region	Region of headquarters.
Revenue	Represents gross sales and other operating revenue less discounts, returns, and allowances.
Employees	The number of full-time employees.
Capital Expenditure	Expenditure for acquiring or maintaining fixed assets.
Net PPE	Represents the total gross value of fixed assets net accumulated depreciation expenses.
Net Intangibles	Net intangibles under GAAP.
Operating Expenses	Total cost of operations.
Total Assets	Represents the total assets reported by a company.
Energy Purchased	Direct energy purchased.
Energy Produced	Direct energy produced.
Renewable Energy Purchased	Total energy purchased from primary renewable energy sources.
Renewable Energy Produced	Total energy purchased from primary renewable energy sources.
ESG Score	Refinitiv ESG score based on reported information in the environmental, social, and corporate governance pillars.
ROE	Net income to the average of fiscal year's common equity.
ROC	Net income to the average of fiscal year's capital.
ROA	Net income to the average of fiscal year's assets.
Asset Turnover	The amount of revenue generated for each unit of assets.
Inventory Turnover	Total cost of revenue to the average total inventory.
DE Ratio	Total debt to total equity.
Interest Coverage Ratio	Net earnings to total interest expense.
Cash Flow Coverage Ratio	Net cash flow from operating activities to total debt.
Current Ratio	Total current assets to total current liabilities.
Quick Ratio	Total current assets less inventory to total current liabilities.

Table A.1: Overview of all considered predictor variables.

A.3 Convergence and density plots for remaining predictor variables in ‘Total’ data frame

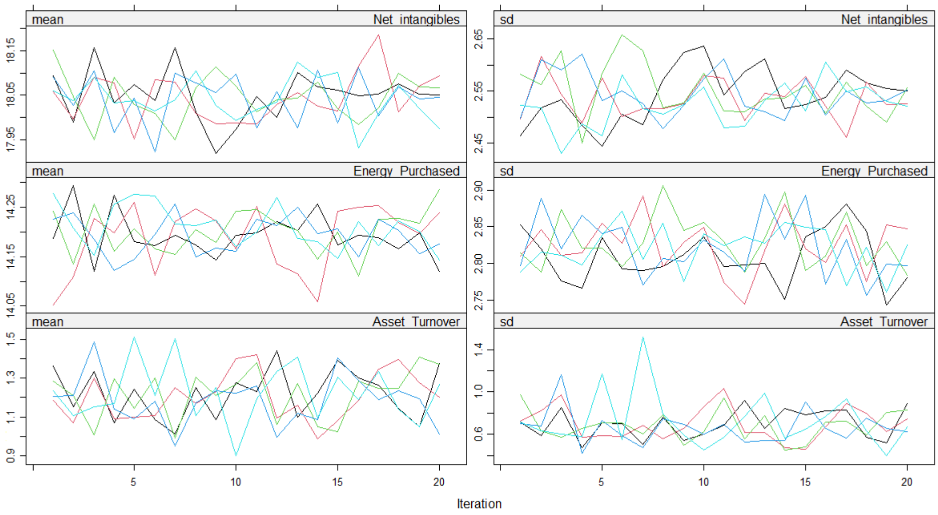


Figure A.4: Convergence of MI of ‘Net Intangibles’, ‘Energy Purchased’, and ‘Asset Turnover’.

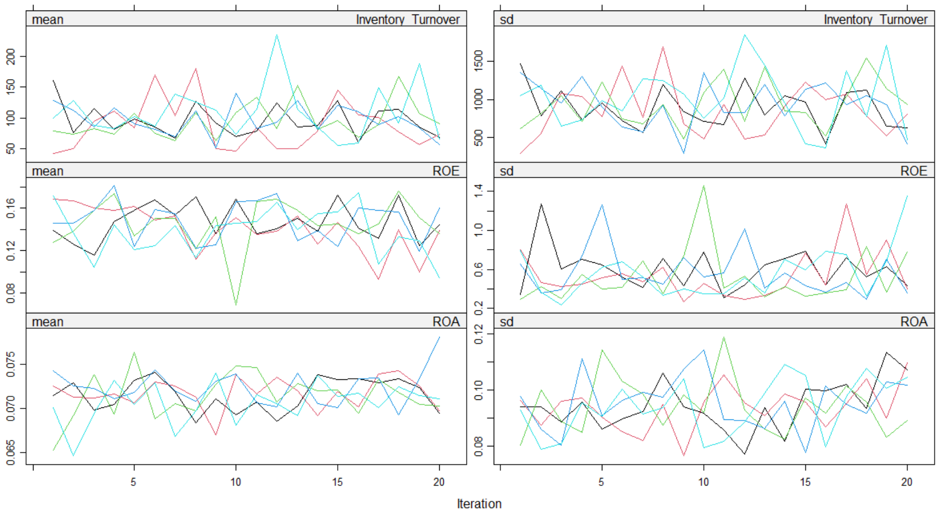


Figure A.5: Convergence of MI of ‘Inventory Turnover’, ‘ROE’, and ‘ROA’.

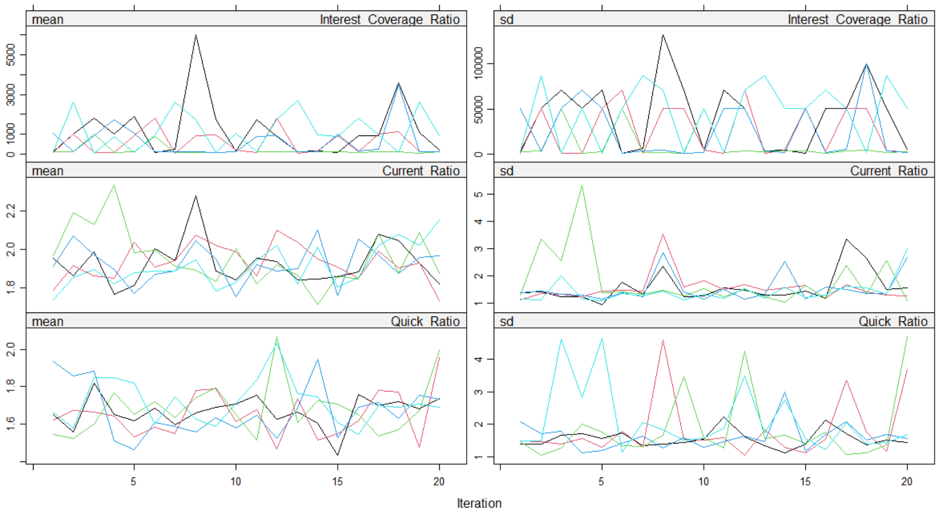


Figure A.6: Convergence of MI of 'Interest Coverage Ratio', 'Current Ratio', and 'Quick Ratio'.

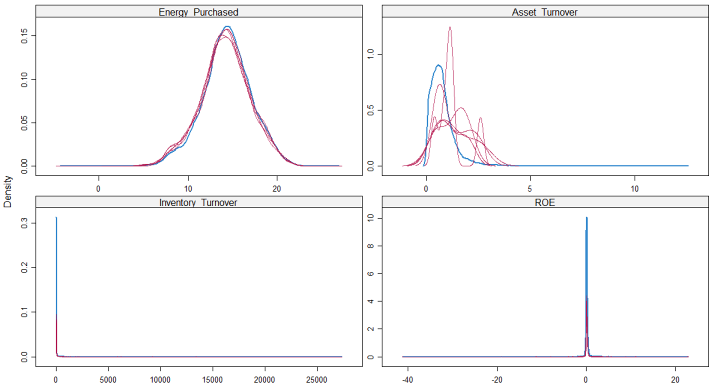


Figure A.7: Density of observed data (blue) against the density of imputed data (red) for the variables 'Energy Purchased', 'Asset Turnover', 'Inventory Turnover', and 'ROE'.

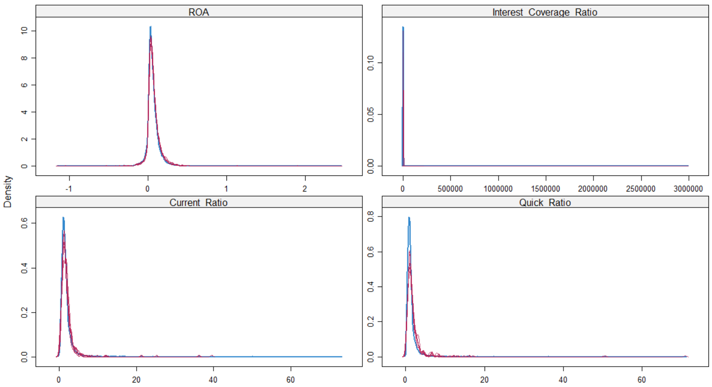


Figure A.8: Density of observed data (blue) against the density of imputed data (red) for the variables 'ROA', 'Interest Coverage Ratio', 'Current Ratio', and 'Quick Ratio'.

A.4 Correlation matrix of ‘Total’ data frame

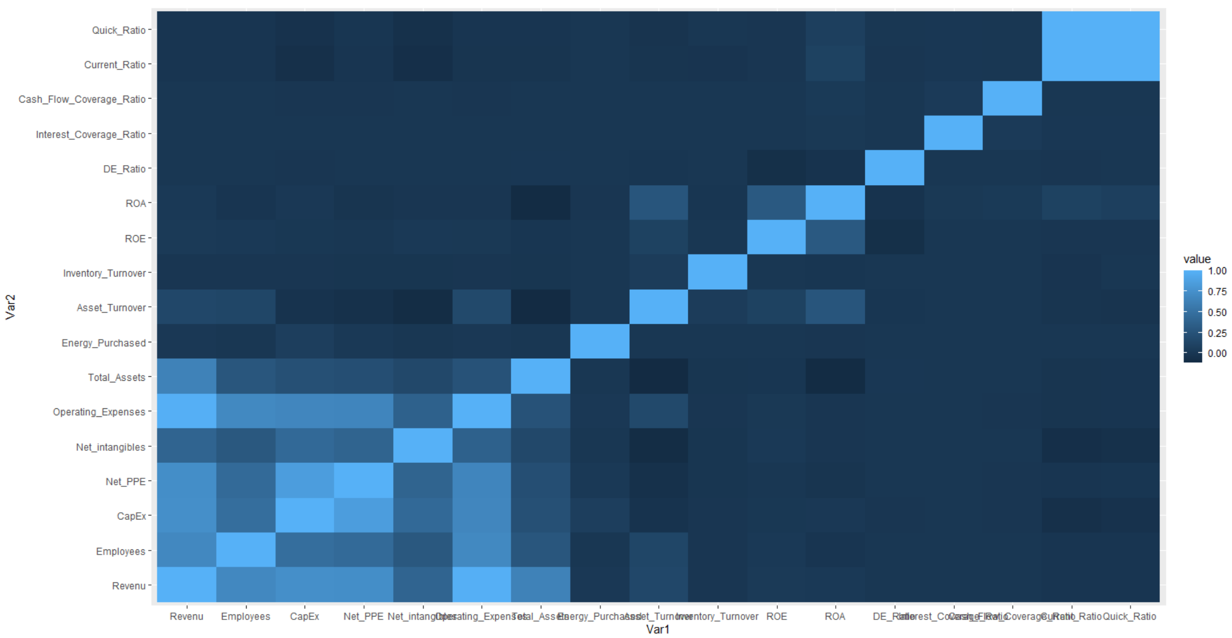


Figure A.9: Correlation matrix for continuous predictor variables in the ‘Total’ data frame.

A.5 Predictor variable distributions after log-transformation

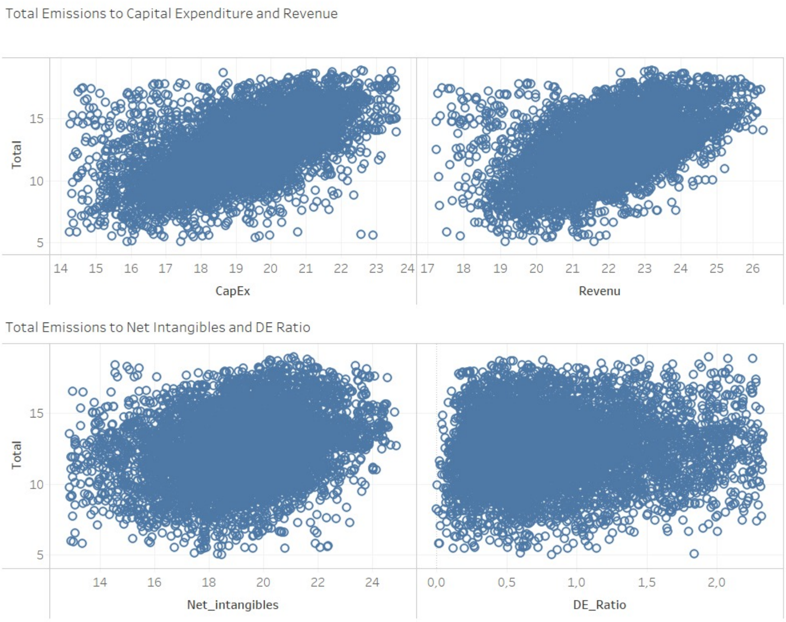


Figure A.10: Distribution of Total Emission to the predictor variables ‘Capital Expenditure’, ‘Operating Expenses’, ‘Net Intangibles’, and ‘DE Ratio’.

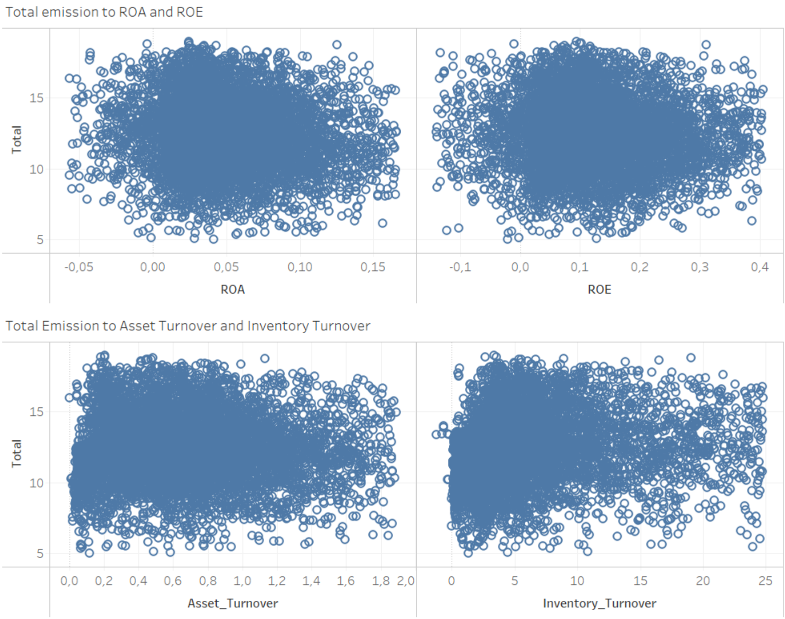


Figure A.11: Distribution of Total Emission to the predictor variables 'ROA', 'ROE', 'Asset Turnover', and 'Inventory Turnover' after log transformation.

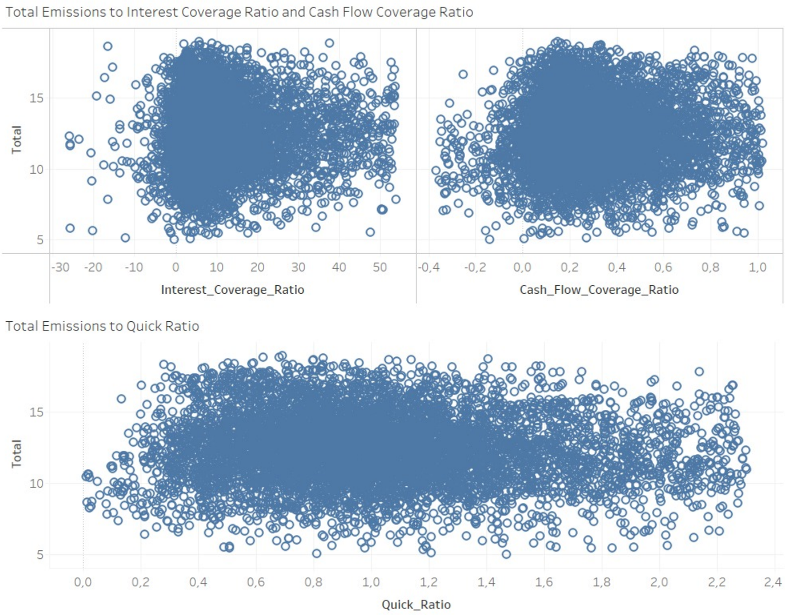


Figure A.12: Distribution of Total Emission to the predictor variables 'Interest Coverage Ratio', 'Cash Flow Coverage Ratio', and 'Quick Ratio'.

Appendix B

Model Prediction Results

B.1 Results Scope 1, 2, and 3 for the imputed data frames

	Metric	**Mean**	**Lasso**	**XGB**	**RF**	**SVR**	**ANN**
Out-of-Sample	Theil's U	0.963	0.397	0.443	0.638	0.435	0.478
	MAPE	51.969	5.291	6.751	6.664	13.882	10.909
	MSE	1.1e14	5.62e13	6.81e13	7.84e13	7.36e13	6.42e13
	RMSE	7820773	5355780	5743010	8527126	5963146	7745319
	R-Squared	-0.147	-0.322	-0.931	0.274	-3.712	0.382
In-Sample	Theil's U	0.991	0.371	0.103	0.304	0.387	0.482
	MAPE	51.766	3.239	0.137	0.356	8.032	11.092
	MSE	1.1e14	5301e13	4.24e12	2.69e13	4.84e13	6.14e13
	RMSE	10494089	7077533	2052743	5181364	6954864	7828517
	R-Squared	-0.066	0.515	0.959	0.740	0.532	0.407
Bias-Variance	Bias	-	1.737	1.934	2.109	2.237	9.590
	Variance	-	0.064	0.237	0.020	0.078	3.138

Table B.1: Results for the emission category 'Scope 1' for the imputed data frame.

	Metric	**Mean**	**Lasso**	**XGB**	**RF**	**SVR**	**ANN**
Out-of-Sample	Theil's U	0.839	0.383	0.413	0.589	0.398	0.115
	MAPE	12.072	6.751	3.575	3.235	4.212	5.646
	MSE	2.56e12	1.19e12	1.72e12	1.86e12	1.60e12	1.64e12
	RMSE	1175976	809727	870822	1259721	878687	1184076
	R-Squared	-0.217	0.245	0.305	0.281	-0.310	0.346
In-Sample	Theil's U	0.927	0.365	0.072	0.262	0.322	0.529
	MAPE	12.061	3.828	0.093	0.300	2.718	5.065
	MSE	2.56e12	1.02e12	5.21e10	4.90e11	8.00e11	1.53e12
	RMSE	1599620	1012318	227429	703018	894993	1232203
	R-Squared	-0.085	0.565	0.978	0.790	0.660	0.352
Bias-Variance	Bias	-	1.354	1.372	1.494	1.579	4.893
	Variance	-	0.043	0.147	0.012	0.066	0.914

Table B.2: Results for the emission category 'Scope 2' for the imputed data frame.

	Metric	**Mean**	**Lasso**	**XGB**	**RF**	**SVR**	**ANN**
Out-of-Sample	Theil's U	0.964	0.605	0.592	0.678	0.589	0.576
	MAPE	288.720	205.790	358.838	144.191	295.519	548.424
	MSE	8.58e15	8.15e15	6.34e15	6.82e15	4.7e15	6.59e15
	RMSE	51135562	48065804	41498637	69538946	4700000	69620926
	R-Squared	-0.204	-0.640	-0.218	0.212	-1.091	0.157
In-Sample	Theil's U	0.996	0.559	0.132	0.366	0.476	0.663
	MAPE	287.444	147.013	0.181	0.878	198.694	268.961
	MSE	8.62e15	7.40e15	5.10e14	2.78e15	5.10e15	7.45e15
	RMSE	92846795	86031963	22530500	52618284	7200000	85942058
	R-Squared	-0.041	0.106	0.938	0.662	0.379	0.099
Bias-Variance	Bias	-	6.093	6.489	6.459	6.546	13.521
	Variance	-	0.238	0.525	0.062	0.087	3.221

Table B.3: Results for the emission category 'Scope 3' for the imputed data frame.

B.2 Visualization increase error after back-transformation for 'Total' data frame.

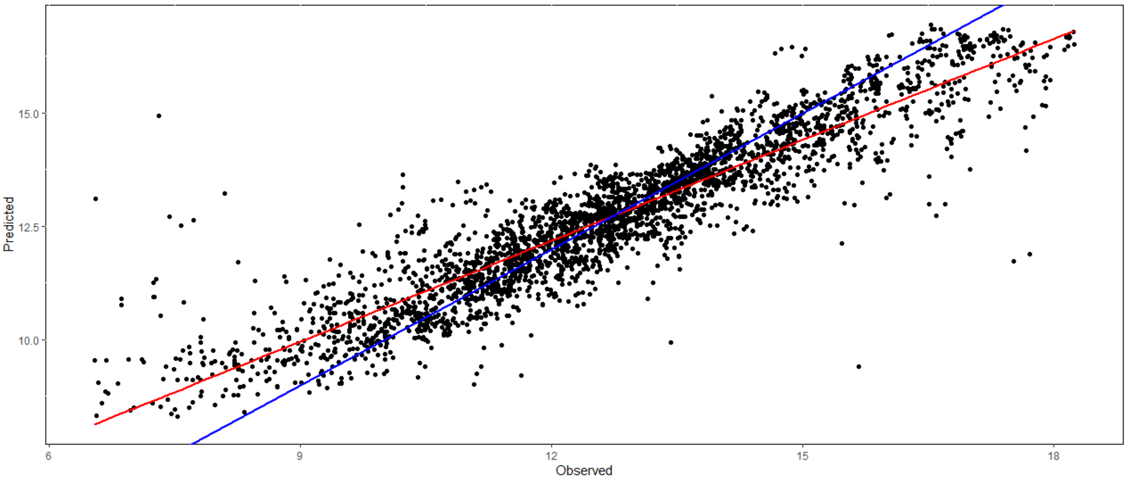


Figure B.1: Visualization error for log-transformed feature space.

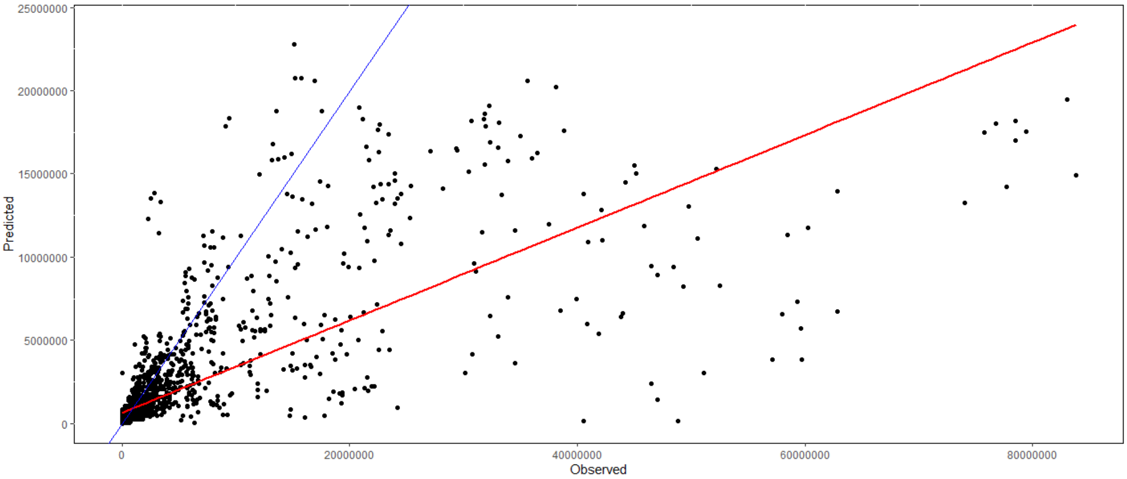


Figure B.2: Visualization error for back-transformed feature space.

B.3 Results Scope 1, 2, and 3 for the baseline data frames

	Metric	Mean	Lasso	XGB	RF	SVR	ANN
Out-of-Sample	Theil's U	0.892	0.318	0.335	0.542	0.304	0.362
	MAPE	24.020	6.701	7.808	3.551	11.534	6.824
	MSE	6.57e13	2.22e13	2.28e13	3.93e13	1.84e13	2.37e13
	RMSE	5462330	2886113	2938878	5867271	2727630	4673147
	R-Squared	-0.221	-0.983	-0.479	0.389	-0.0086	0.571
In-Sample	Theil's U	0.984	0.247	0.017	0.201	0.191	0.344
	MAPE	23.975	3.017	0.021	0.249	5.067	8.285
	MSE	6.59e13	1.54e13	8.18e10	7.87e12	8.58e12	2.42e13
	RMSE	8115129	3920373	282343	2803682	2928553	4836543
	R-Squared	-0.081	0.748	0.999	0.870	0.859	0.600
Bias-Variance	Bias	-	1.089	1.154	1.407	1.876	8.004
	Variance	-	0.063	0.157	0.017	0.091	2.407

Table B.4: Results for the emission category 'Scope 1' for the baseline data frame.

	Metric	Mean	Lasso	XGB	RF	SVR	ANN
Out-of-Sample	Theil's U	0.721	0.318	0.332	0.460	0.309	0.100
	MAPE	6.390	1.483	1.799	1.713	1.733	1.710
	MSE	1.47e12	7.03e11	7.31e11	8.51e11	6.11e11	8.00e11
	RMSE	925605	606066	609010	845969	571227	824731
	R-Squared	-0.285	0.200	0.220	0.364	0.266	0.353
In-Sample	Theil's U	0.853	0.323	0.021	0.174	0.206	0.425
	MAPE	6.367	1.022	0.031	0.210	0.667	1.489
	MSE	1.47e12	5.08e11	2.77e09	1.53e11	2.24e11	6.88e11
	RMSE	1213412	712725	52526	390484	473738	826870
	R-Squared	-0.137	0.608	0.998	0.882	0.827	0.465
Bias-Variance	Bias	-	0.815	0.898	5.277	1.295	3.677
	Variance	-	0.051	0.009	0.061	0.062	1.108

Table B.5: Results for the emission category 'Scope 2' for the baseline data frame.

	Metric	Mean	Lasso	XGB	RF	SVR	ANN
Out-of-Sample	Theil's U	0.853	0.575	0.580	0.711	0.519	0.830
	MAPE	155.804	731.259	309.931	351.199	680.532	717.120
	MSE	1.1e16	2.28e16	9.50e15	9.90e15	9.36e15	3.25e16
	RMSE	51303902	63524053	45379845	83358382	44169424	1.80e08
	R-Squared	-0.402	-70.478	-4.751	0.191	-3.314	-0.057
In-Sample	Theil's U	0.991	0.570	0.030	0.419	0.564	0.515
	MAPE	154.706	390.760	0.038	0.757	479.769	298.545
	MSE	1.13e16	8.53e15	4.03e13	4.41e15	7.75e15	9.81e15
	RMSE	1.06e08	92315140	6262845	66197641	87984471	99051408
	R-Squared	-0.057	0.203	0.996	0.589	0.276	-0.162
Bias-Variance	Bias	-	5.501	5.660	5.277	6.345	7.041
	Variance	-	0.395	0.497	0.061	0.450	0.789

Table B.6: Results for the emission category 'Scope 3' for the baseline data frame.

B.4 Results of Random Forest applied on untransformed data

	Metrix	Total	Scope 1	Scope 2	Scope 3
Out-of-Sample	Theil's U	0.411	0.438	0.443	0.498
	MAPE	29.435	246.968	17.310	3335.981
	MSE	5.51e13	5.85e13	1.52e12	5.83e15
	RMSE	7130185	7332103	1136120	65602838
	R-Squared	0.492	0.446	0.400	0.240
In-Sample	Theil's U	0.098	0.114	0.108	0.125
	MAPE	9.233	83.726	5.268	967.263
	MSE	4.01e12	4.96e12	1.06e11	4.48e14
	RMSE	2001985	2223463	325009	21127342
	R-Squared	0.963	0.952	0.955	0.946
Bias-Variance	Bias	5.61e13	5.68e13	1.54e12	5.9e15
	Variance	4.59e11	5.86e11	1.24e10	5.55e13

Table B.7: Results of the Random Forest model applied on untransformed data.

B.5 Results Scope 1, 2, and 3 for the baseline data frames after bias correction

	Metric	Mean	Lasso	XGB	RF	SVR	ANN
Out-of-Sample	Theil's U	0.604	0.347	0.356	0.410	0.349	0.987
	MAPE	556.545	10.121	32.446	7.855	16.968	1817
	MSE	5.44e13	2.83e14	6.18e15	3.49e13	5.01e14	6.44e18
	RMSE	4961275	5624184	13154211	5687538	6465324	1.87e09
	R-Squared	0	-13.223	-837.606	0.274	-24.814	-104469
In-Sample	Theil's U	0.745	0.362	0.017	0.181	0.224	0.989
	MAPE	546.441	10.766	0.021	0.264	8.383	2214.182
	MSE	6.09e13	5.3e13	8.08e10	6.66e12	1.61e13	3.11e18
	RMSE	7803574	7278129	280691	2578585	4008383	1.65e09
	R-Squared	0	0.129	0.999	0.890	0.736	-51608
Bias-Variance	Bias	-	1.168	1.178	1.403	1.524	8.436
	Variance	-	0.025	0.156	0.017	0.097	2.647

Table B.8: Results for the emission category 'Scope 1' for the baseline data frame after bias correction.

	Metric	Mean	Lasso	XGB	RF	SVR	ANN
Out-of-Sample	Theil's U	0.492	0.313	0.329	0.371	0.309	0.823
	MAPE	20.521	3.250	4.613	3.431	3.552	36.267
	MSE	1.12e12	2.29e13	5.35e13	1.00e12	4.56e13	2.05e14
	RMSE	812097	1154904	1512612	888188	1360371	12214124
	R-Squared	0	-13.979	-31.847	0.269	-27.232	-145.77
In-Sample	Theil's U	0.617	0.337	0.021	0.161	0.200	0.823
	MAPE	24.771	2.214	0.031	0.216	1.294	33.185
	MSE	1.29e12	8.39e11	2.71e09	1.34e11	2.70e11	1.34e14
	RMSE	1137764	915832	51928	365428	519094	11190444
	R-Squared	0	0.352	0.998	0.896	0.792	-102.758
Bias-Variance	Bias	-	0.891	0.898	1.033	1.894	3.55349
	Variance	-	0.023	0.095	0.010	0.106	1.265

Table B.9: Results for the emission category 'Scope 2' for the baseline data frame after bias correction.

	Metric	Mean	Lasso	XGB	RF	SVR	ANN
Out-of-Sample	Theil's U	0.528	0.313	0.616	0.371	0.309	0.978
	MAPE	7128	3.250	19277	3.431	3.552	236981
	MSE	8.87e15	2.29e13	1.45e17	1.00e12	4.56e13	8.01e20
	RMSE	44771485	1154904	1.64e08	888188	1360371	1.89e10
	R-Squared	0	-13.979	-4.751	0.269	-27.232	-140008
In-Sample	Theil's U	0.784	0.337	0.029	0.161	0.200	0.981
	MAPE	4932	2.214	0.038	0.216	1.294	525292
	MSE	1.07e16	8.39e11	3.80e13	1.34e11	2.70e11	6.22e20
	RMSE	1.03e08	915832	6081991	365428	519094	2.06e10
	R-Squared	0	0.352	0.996	0.896	0.792	-59510
Bias-Variance	Bias	-	0.891	5.674	5.302	6.548	11.673
	Variance	-	0.023	0.512	0.061	0.083	4.989

Table B.10: Results for the emission category 'Scope 3' for the baseline data frame after bias correction.

B.6 Features with corresponding variable name

Feature	Predictor Variable	**Feature**	Predictor Variable
Feature 0	Fiscal Year	**Feature 39**	Country: Italy
Feature 1	Revenue	**Feature 40**	Country: Japan
Feature 2	Employees	**Feature 41**	Country: Jersey
Feature 3	Capital Expenditure	**Feature 42**	Country: South Korea
Feature 4	Net PPE	**Feature 43**	Country: Luxemburg
Feature 5	Net Intangibles	**Feature 44**	Country: Malaysia
Feature 6	Total Assets	**Feature 45**	Country: Mexico
Feature 7	Energy Purchased	**Feature 46**	Country: Netherlands
Feature 8	Asset Turnover	**Feature 47**	Country: New Zealand
Feature 9	Inventory Turnover	**Feature 48**	Country: Norway
Feature 10	ROE	**Feature 49**	Country: Peru
Feature 11	ROA	**Feature 50**	Country: Philippines
Feature 12	DE Ratio	**Feature 51**	Country: Poland
Feature 13	Interest Coverage Ratio	**Feature 52**	Country: Portugal
Feature 14	Cash Flow Coverage Ratio	**Feature 53**	Country: Russia
Feature 15	Quick Ratio	**Feature 54**	Country: Saudi Arabia
Feature 16	Country: Australia	**Feature 55**	Country: Singapore
Feature 17	Country: Austria	**Feature 56**	Country: South Africa
Feature 18	Country: Belgium	**Feature 57**	Country: Spain
Feature 19	Country: Bermuda	**Feature 58**	Country: Sweden
Feature 20	Country: Brazil	**Feature 59**	Country: Switzerland
Feature 21	Country: Canada	**Feature 60**	Country: Taiwan
Feature 22	Country: Chile	**Feature 61**	Country: Thailand
Feature 23	Country: China	**Feature 62**	Country: Turkey
Feature 24	Country: Colombia	**Feature 63**	Country: Ukraine
Feature 25	Country: Cyprus	**Feature 64**	Country: UAE
Feature 26	Country: Denmark	**Feature 65**	Country: UK
Feature 27	Country: Faroe Islands	**Feature 66**	Country: USA
Feature 28	Country: Finland	**Feature 67**	GICS: Communication Services
Feature 29	Country: France	**Feature 68**	GICS: Consumer Discretionary
Feature 30	Country: Germany	**Feature 69**	GICS: Consumer Staples
Feature 31	Country: Greece	**Feature 70**	GICS: Energy
Feature 32	Country: Hong Kong	**Feature 71**	GICS: Financials
Feature 33	Country: Hungary	**Feature 72**	GICS: Health Care
Feature 34	Country: Iceland	**Feature 73**	GICS: Industrials
Feature 35	Country: India	**Feature 74**	GICS: Information Technology
Feature 36	Country: Indonesia	**Feature 75**	GICS: Materials
Feature 37	Country: Ireland	**Feature 76**	GICS: Real Estate
Feature 38	Country: Israel	**Feature 77**	GICS: Utilities

B.7 Relationships between features and 'Total' emissions

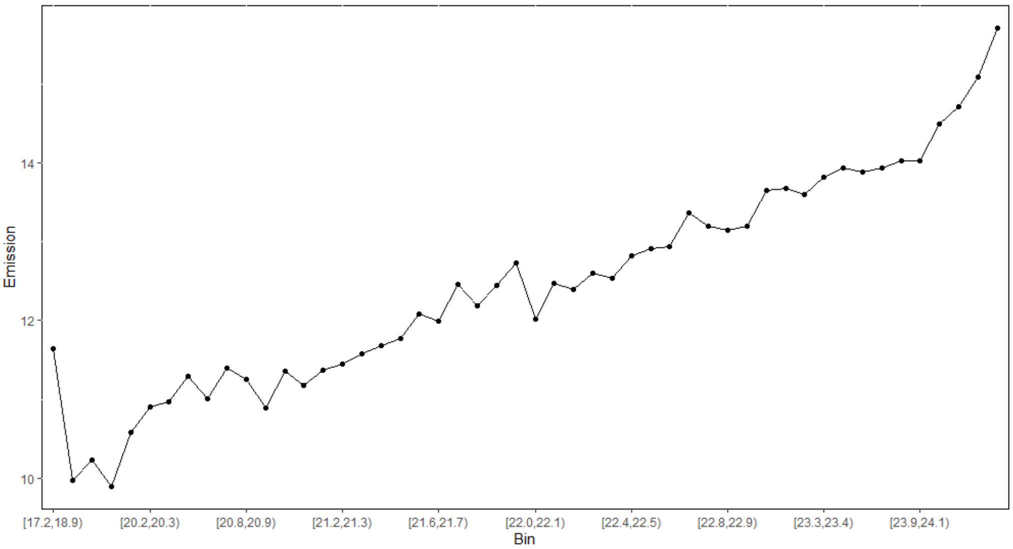


Figure B.3: Relationship between 'Revenue' and 'Total' emissions.

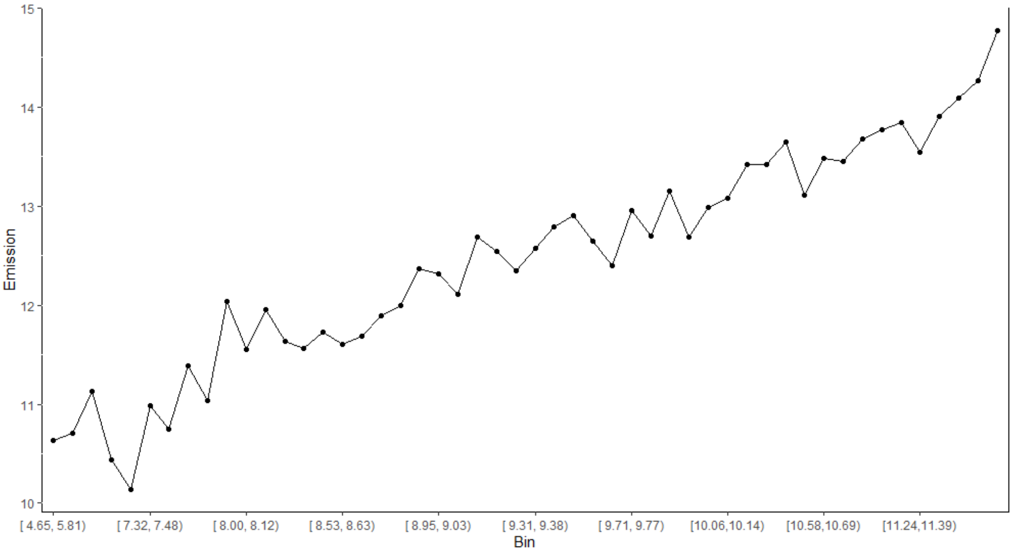


Figure B.4: Relationship between 'Employees' and 'Total' emissions.

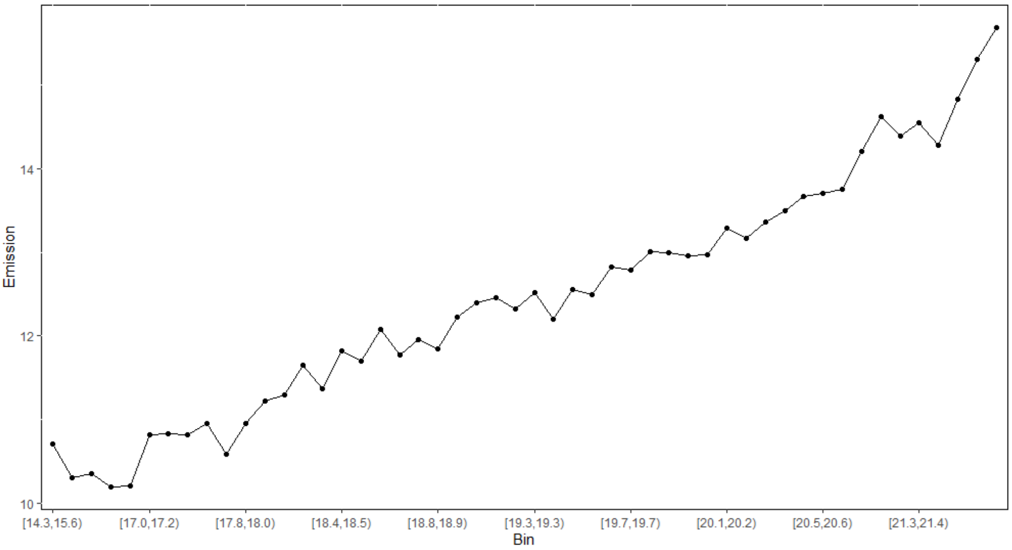


Figure B.5: Relationship between ‘Capital Expenditure’ and ‘Total’ emissions.

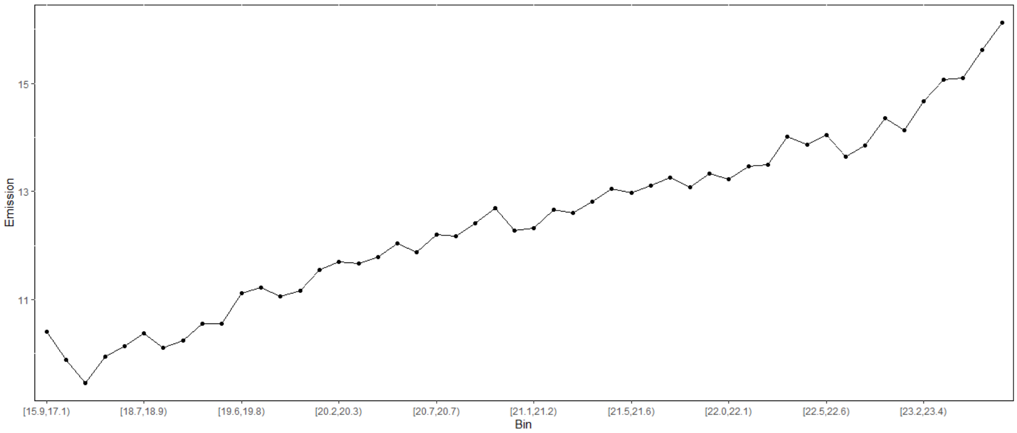


Figure B.6: Relationship between ‘Net PPE’ and ‘Total’ emissions.

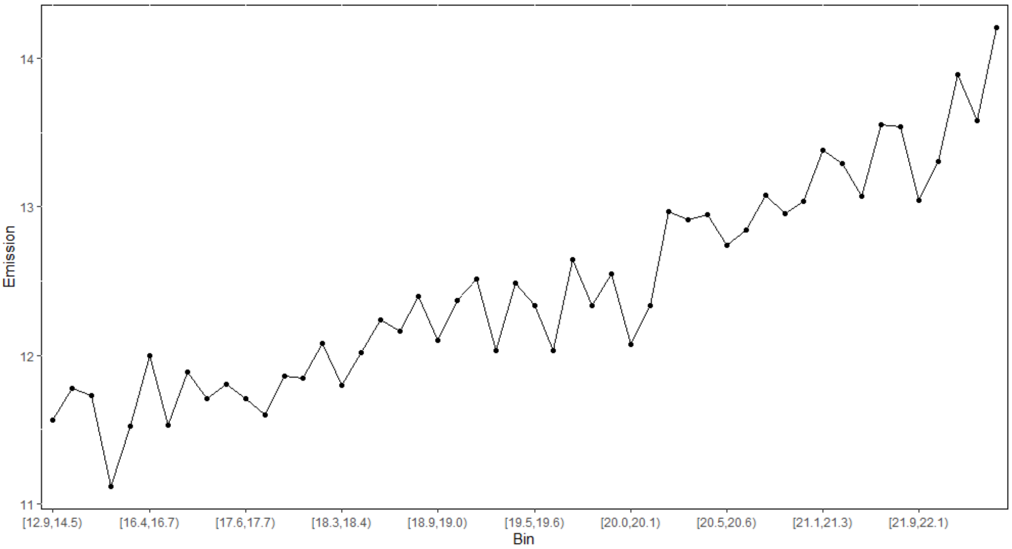


Figure B.7: Relationship between ‘Net Intangibles’ and ‘Total’ emissions.

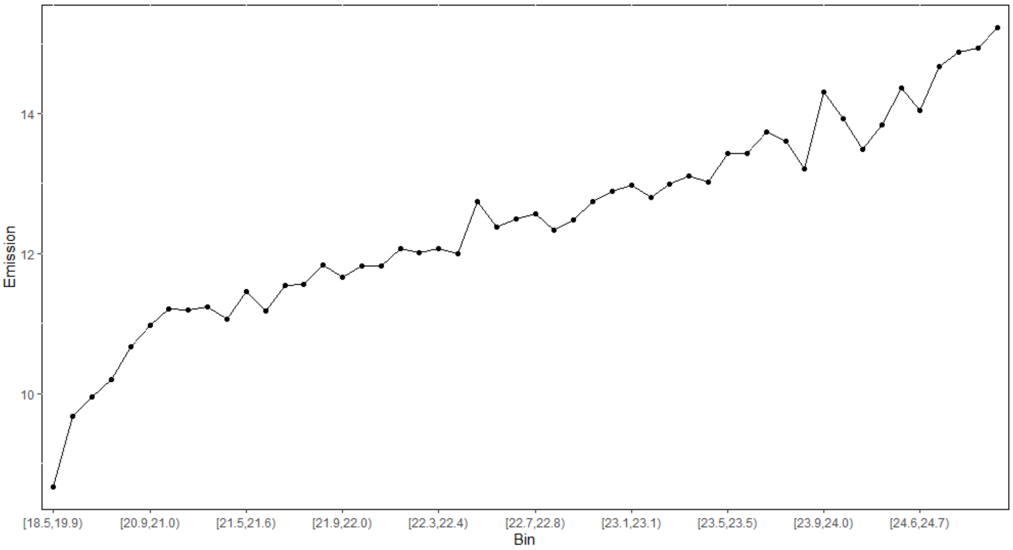


Figure B.8: Relationship between 'Total Assets' and 'Total' emissions.

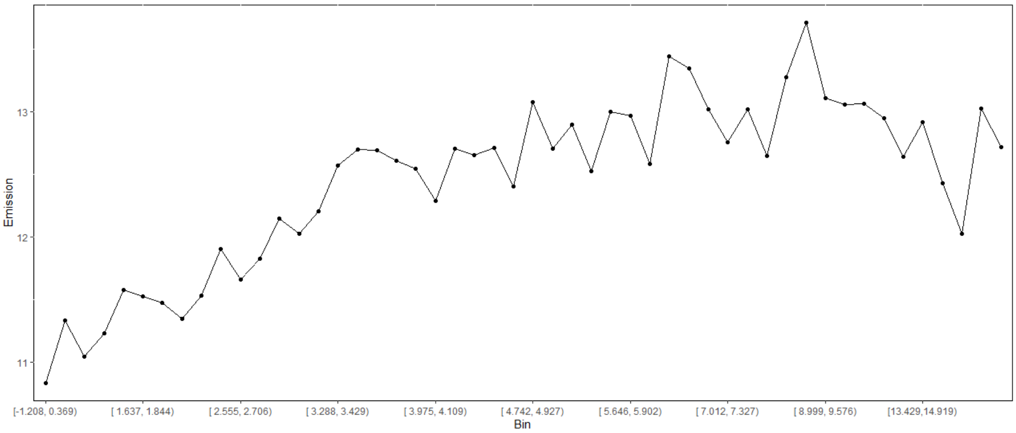


Figure B.9: Relationship between 'Inventory Turnover' and 'Total' emissions.

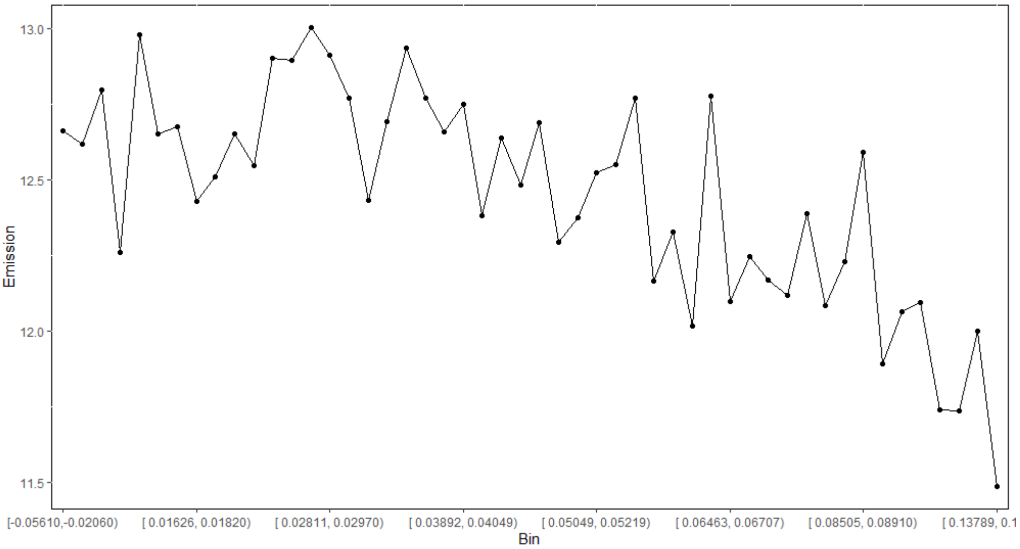


Figure B.10: Relationship between 'ROA' and 'Total' emissions.

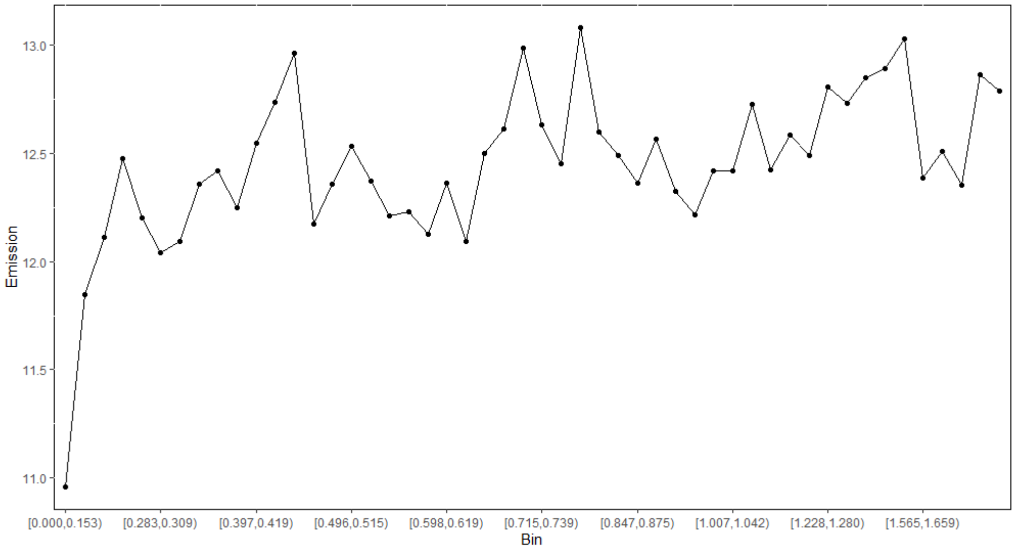


Figure B.11: Relationship between ‘DE Ratio’ and ‘Total’ emissions.

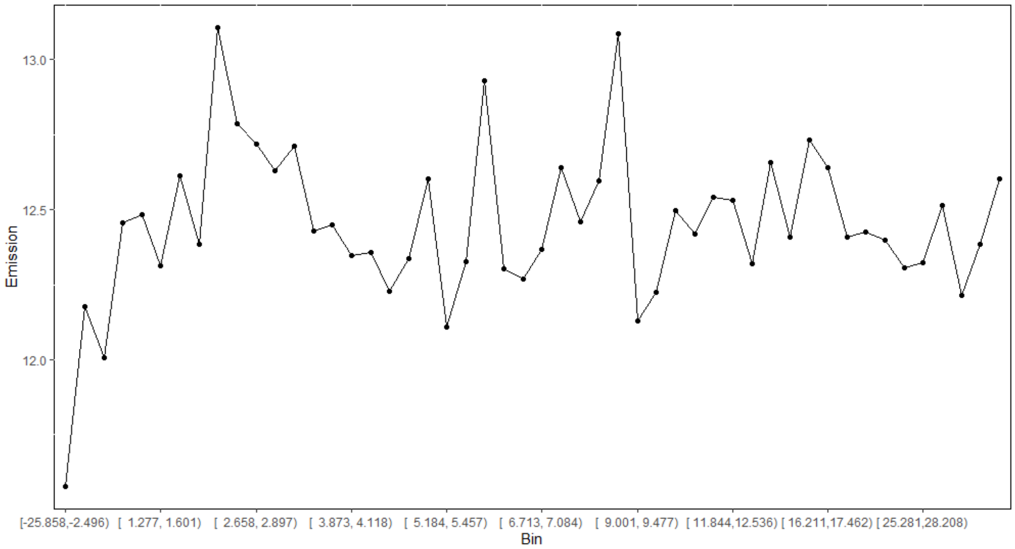


Figure B.12: Relationship between ‘Interest Coverage Ratio’ and ‘Total’ emissions.

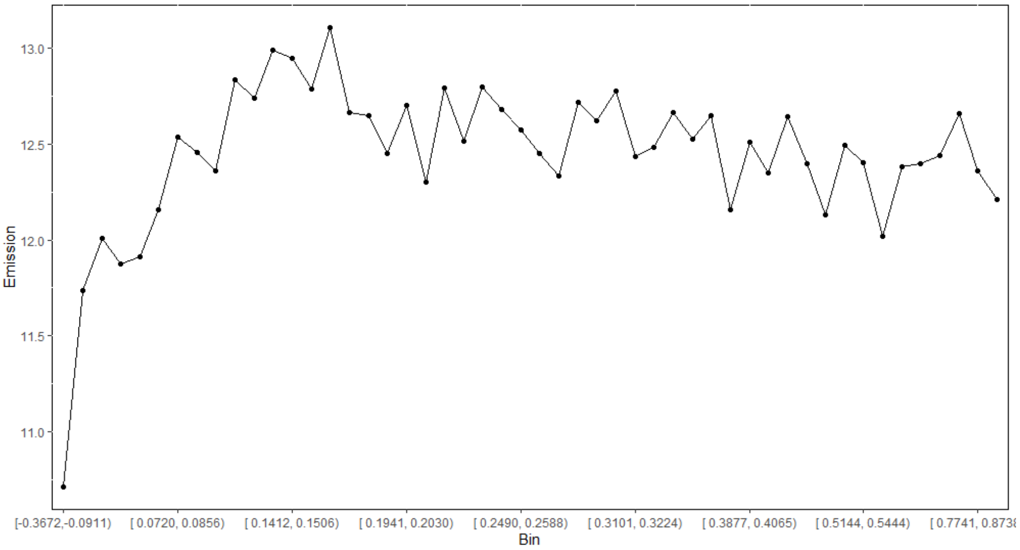


Figure B.13: Relationship between ‘Cash Flow Coverage Ratio’ and ‘Total’ emissions.

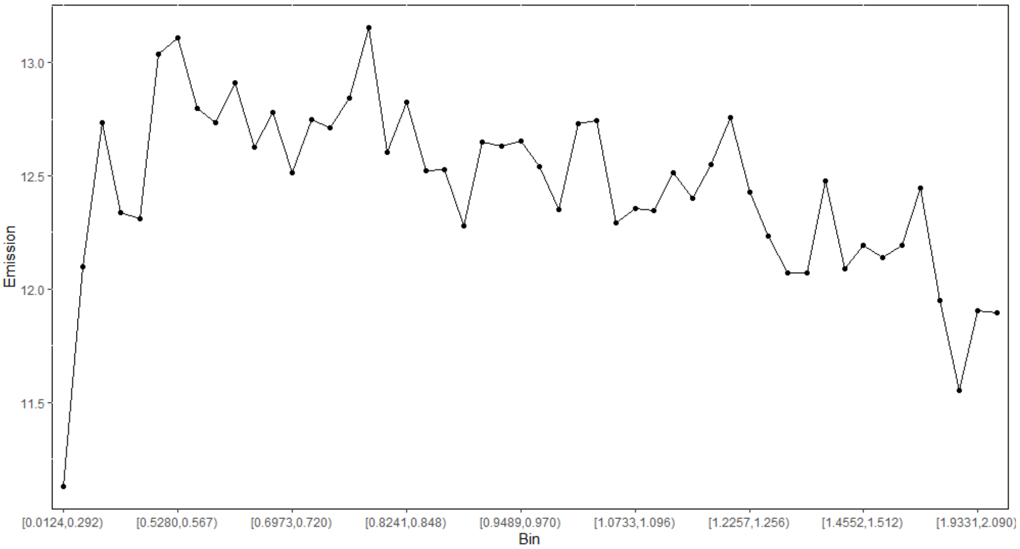


Figure B.14: Relationship between ‘Quick Ratio’ and ‘Total’ emissions.

B.8 Linear Regression

	Metric	Total	Scope 1	Scope 2	Scope 3
Out-of-Sample	Theil's U	0.259	0.349	0.329	0.585
	MAPE	4.440	6.933	1.639	1102.02
	MSE	2.60e13	2.58e13	7.5e11	9.99e15
	RMSE	3088550	3185268	616457	46535016
	R-Squared	0.273	0.275	0.014	-7.037
In-Sample	Theil's U	0.328	0.339	0.365	0.648
	MAPE	3.655	5.365	1.321	813.741
	MSE	2.37e13	2.36e13	6.27e11	9.45e15
	RMSE	4869142	4858369	791627	97192243
	R-Squared	0.645	0.612	0.516	0.116456
Bias-Variance	Bias	3.81e15	1.175	0.879	5.694
	Variance	1.52e16	0.027	0.022	0.178

Table B.11: Linear regression applied on the baseline data frame.

B.9 Remaining GICS classifications

	Metric	Total	Scope 1	Scope 2	Scope 3
Out-of-Sample	Theil's U	0.522	0.581	0.479	0.716
	MAPE	2.253	4.447	1.631	423.074
	MSE	4.27e13	4.23e13	9.01e11	9.96e15
	RMSE	5922909	6059366	866122.4	83590814
	R-Squared	0.405	0.354	0.342	0.186
In-Sample	Theil's U	0.199	0.212	0.188	0.403
	MAPE	0.181	0.264	0.209	0.745
	MSE	8.71e12	8.50e12	1.74e11	4.14e15
	RMSE	2946716	2914483	416694	64147827
	R-Squared	0.869	0.860	0.865	0.614
Bias-Variance	Bias	0.876	1.569	1.053	5.375
	Variance	0.010	0.018	0.011	0.061

Table B.12: Results of Random Forest model with GICS classifier sub-industry, for 'Total' data frame

	Metric	Total	Scope 1	Scope 2	Scope 3
Out-of-Sample	Theil's U	0.501	0.570	0.469	0.715
	MAPE	2.207	4.465	1.678	355.430
	MSE	4.17e13	4.16e13	8.81e11	9.93e15
	RMSE	5815001	6003152	865400	83554950
	R-Squared	0.431	0.367	0.355	0.187
In-Sample	Theil's U	0.189	0.205	0.182	0.415
	MAPE	0.175	0.252	0.208	0.751
	MSE	7.98e12	8.15e12	1.64e11	4.35e15
	RMSE	2821990	2852241	405094	65726154
	R-Squared	0.880	0.866	0.873	0.595
Bias-Variance	Bias	0.828	1.467	1.039	5.386
	Variance	0.010	0.017	0.010	0.060

Table B.13: Results of Random Forest model with GICS classifier industry name, for 'Total' data frame.

	Metric	Total	Scope 1	Scope 2	Scope 3
Out-of-Sample	Theil's U	0.489	0.544	0.458	0.718
	MAPE	1.927	3.833	1.719	306.435
	MSE	4.11e13	3.97e13	8.45e11	9.94e15
	RMSE	5779046	5887475	842594	83694483
	R-Squared	0.441	0.387	0.368	0.183
In-Sample	Theil's U	0.191	0.202	0.174	0.417
	MAPE	0.173	0.244	0.206	0.757
	MSE	8.18e12	7.94e12	1.52e11	4.40e15
	RMSE	2856986	2816607	389022	66085231
	R-Squared	0.877	0.869	0.883	0.591
Bias-Variance	Bias	0.820	1.377	1.023	5.342
	Variance	0.010	0.017	0.010	0.064

Table B.14: Results of Random Forest model with GICS classifier industry group name, for 'Total' data frame.

B.10 GICS classification subsets

	Metric	Communications	Consumer Discretionary	Consumer Staples	Energy	Financials
Out-of-Sample	Theil's U	0.409	0.396	0.415	0.273	0.464
	MAPE	2.526	1.765	2.460	20.502	2.115
	MSE	2.89e11	1.89e13	1.95e12	6.97e13	1.75e13
	RMSE	412650	1815118	1146991	7175951	1401123
	R-Squared	0.338	0.519	0.416	0.636	-1.475
In-Sample	Theil's U	0.256	0.691	0.210	0.077	0.388
	MAPE	0.229	0.224	0.208	0.352	0.313
	MSE	1.08e11	1.62e13	4.97e11	5.79e12	5.49e12
	RMSE	321649	3832815	702471	2404458	2224052
	R-Squared	0.735	0.241	0.826	0.969	0.666
Bias-Variance	Bias	1.202	1.205	1.051	1.671	3.373
	Variance	0.024	0.016	0.017	0.028	0.065

Table B.15: Results of Random Forest model within 'GICS Sector Name' classified subsets for 'Total' emissions (1/2).

	Metric	Healthcare	Industrials	IT	Materials	Real Estate	Utilities
Out-of-Sample	Theil's U	0.324	0.574	0.441	0.425	0.574	0.626
	MAPE	1.129	0.969	1.063	3.853	6.363	5.836
	MSE	1.18e11	6.7e12	1.04e12	1.7e14	1.39e13	6.59e14
	RMSE	2812572	2337678	856114	11710663	1799811	22795369
	R-Squared	0.537	0.335	0.457	0.443	0.199	-0.078
In-Sample	Theil's U	0.110	0.232	0.144	0.139	0.560	0.266
	MAPE	0.162	0.196	0.194	0.226	0.307	0.345
	MSE	1.3e10	1.62e12	1.33e11	2.37e13	8.12e12	1.6e14
	RMSE	113758	1269721	363803	4859953	2771077	12617638
	R-Squared	0.943	0.833	0.919	0.920	0.439	0.753
Bias-Variance	Bias	0.802	0.885	0.833	1.368	2.143	2.936
	Variance	0.015	0.011	0.015	0.016	0.036	0.049

Table B.16: Results of Random Forest model within 'GICS Sector Name' classified subsets for 'Total' emissions (2/2).

B.11 Region

	Metric	Total	Scope 1	Scope 2	Scope 3
Out-of-Sample	Theil's U	0.448	0.492	0.440	0.694
	MAPE	2.106	3.898	1.659	325.155
	MSE	3.79e13	3.59e13	8.16e11	9.83e15
	RMSE	5529459	5589376	826897	82771540
	R-Squared	0.490	0.445	0.390	0.202
In-Sample	Theil's U	0.175	0.183	0.168	0.415
	MAPE	0.163	0.240	0.204	0.783
	MSE	7.05e12	6.73e12	1.43e11	4.38e15
	RMSE	2652724	2593231	377920	65935968
	R-Squared	0.894	0.889	0.889	0.592
Bias-Variance	Bias	0.721	1.281	0.983	5.390
	Variance	0.010	0.017	0.010	0.077

Table B.17: Results of Random Forest model using 'Region' instead of 'Country'.