DEGREE PROJECT IN MATHEMATICS,
SECOND CYCLE, 30 CREDITS
STOCKHOLM, SWEDEN 2021

Carbon Intensity Estimation of Publicly Traded Companies

Olle Ribberheim

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ENGINEERING SCIENCES

Author

Olle Ribberheim <olle@ribberheim.se>
Applied and Computational Mathematics
KTH Royal Institute of Technology

Place for Project

Stockholm, Sweden
SEB, Kungsträdgården

Examiner

Camilla Johansson Landén
MATHEMATICAL STATISTICS
KTH Royal Institute of Technology

Supervisor

Boualem Djehiche

MATHEMATICAL STATISTICS

KTH Royal Institute of Technology

Abstract

The purpose of this master thesis is to develop a model to estimate the carbon intensity, i.e the carbon emission relative to economic activity, of publicly traded companies which do not report their carbon emissions. By using statistical and machine learning models, the core of this thesis is to develop and compare different methods and models with regard to accuracy, robustness, and explanatory value when estimating carbon intensity. Both discrete variables, such as the region and sector the company is operating in, and continuous variables, such as revenue and capital expenditures, are used in the estimation. Six methods were compared, two statistically derived and four machine learning methods. The thesis consists of three parts: data preparation, model implementation, and model comparison. The comparison indicates that boosted decision tree is both the most accurate and robust model. Lastly, the strengths and weaknesses of the methodology is discussed, as well as the suitability and legitimacy of the boosted decision tree when estimating carbon intensity.

Keywords

Master thesis, financial mathematics, ESG, carbon intensity, carbon emissions, statistical analysis, regression, machine learning, neural networks, boosted decision tree

Sammanfattning

Syftet med denna masteruppsats är att utveckla en modell som uppskattar koldioxidsintensiteten, det vill säga koldioxidutsläppen i förhållande till ekonomisk aktivitet, hos publika bolag som inte rapporterar sina koldioxidutsläpp. Med hjälp av statistiska och maskininlärningsmodeller kommer stommen i uppsatsen vara att utveckla och jämföra olika metoder och modeller utifrån träffsäkerhet, robusthet och förklaringsvärde vid uppskattning av koldioxidintensitet. Både diskreta och kontinuerliga variabler används vid uppskattningen, till exempel region och sektor som företaget är verksam i, samt omsättning och kapitalinvesteringar. Sex stycken metoder jämfördes, två statistiskt härledda och fyra maskininlärningsmetoder. Arbetet består av tre delar; förberedelse av data, modellutveckling och modelljämförelse, där jämförelsen indikerar att boosted decision tree är den modell som är både mest träffsäker och robust. Slutligen diskuteras styrkor och svagheter med metodiken, samt lämpligheten och tillförlitligheten med att använda ett boosted decision tree för att uppskatta koldioxidintensitet.

Nyckelord

Masteruppsats, finansiell matematik, ESG, koldioxidintensitet, koldioxidutsläpp, statistik analys, regression, maskininlärning, neurala nätverk, boosted decision tree

Acknowledgements

I would like to acknowledge the following people for their help and contribution to this thesis:

Oscar Ungsgård, my supervisor at SEB, for helping me throughout the thesis with insights and experience on how to plan and work around the thesis.

Alexander Bea, my colleague at SEB, for giving me insights and inputs on the methodology and machine learning aspects throughout the project.

Boualem Djehiche, my supervisor at KTH, for contributing with guidance and input on what to prioritize and focus on during the thesis.

Acronyms

CSV	Comma Seperated Values
ESG	Environmental, Social, and Governance
LGBM	Light Gradient Boosting Machine
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
WACC	Weighted Average Cost of Capital

Contents

1 Introduction **1**

1.1 Background 2
1.2 Problem 2
1.3 Purpose 2
1.4 Goal 3
1.5 Benefits, Ethics and Sustainability 3
1.6 Methodology 3
1.7 Stakeholders 4
1.8 Delimitations 4
1.9 Outline 5

2 Theoretical Background **6**

2.1 Related Work 6
2.2 The dataset 7
2.2.1 Carbon intensity 8
2.3 Data Analysis 8
2.3.1 Missing Values 8
2.3.2 Dummy Variables 9
2.3.3 Outliers 9
2.3.4 Cook's Distance 10
2.4 Model Selection 10
2.4.1 Linear Regression 11
2.4.2 Bayesian Linear Regression 12
2.4.3 Decision Forest Regression 12
2.4.4 Neural Network Regression 13
2.4.5 Boosted Decision Tree Regression 14

2.4.6 LGBM 14
2.5 Model Validation 15
2.5.1 Training and Test Data Split 15
2.5.2 Root Mean Square Error 15
2.5.3 R-squared 16
2.5.4 Robustness 16

3 Method **17**
3.1 Data Preparation 17
3.2 Model Implementation 18
3.3 Model Comparison 19

4 Results **20**
4.1 Scatter Plots of Continuous Variables 20
4.2 Robustness 22
4.3 Accuracy 23

5 Discussion & Conclusions **24**
5.1 Discussion 24
5.1.1 Model Performance and Evaluation 24
5.1.2 Methodology Evaluation 26
5.1.3 Further Improvements and Continued Studies 27
5.2 Conclusions 28
5.2.1 Model Usefulness and Implementation 28
5.2.2 Final words 28

References **29**

Chapter 1

Introduction

Environmental, Social, and Governance (ESG) is an ever-increasing area of interest for institutional investors, legislators, corporate management, and consumers in their decision making. Everyone wants to make a positive impact with their decisions but still, there are dissensions and discrepancies regarding what really is sustainable and what could be considered ESG aligned or not. Up until recently, when EU published its taxonomy on sustainable finance, there had been no established frameworks on ESG nor explicit definitions of sustainability. Additionally, regulations differ between countries and regions, where some jurisdictions are more prone to adopt ESG policies and legislations whereas others are not.

Today, investors are looking for ways to track the carbon footprint of their investments and how ESG-aligned their portfolio is. Publicly traded companies in some countries, mostly western, have greater obligations in publishing ESG data, such as carbon emissions, water usage in critical areas, and equal pay. This creates the possibility to quantify the ESG footprint of a portfolio by determining how much of each ESG factor each position is attributable to. However, this becomes problematic when the portfolio contains stocks of companies which do not publish these metrics. Hence, there is a need for methods and models to estimate the ESG footprint of publicly traded companies which do not measure and/or publish ESG data.

This thesis is written for a bank who wishes to offer its clients a tool that the clients can measure ESG impact themselves and how their investment decisions affect their overall ESG impact. For example, how much carbon emissions or water usage one specific position is attributable to.

1.1 Background

To be able to offer a comprehensive ESG analysis tool for clients' portfolios, the carbon intensity of publicly traded companies will have to be estimated initially. Due to the low number of companies which publish their carbon intensity and the structural differences between regions and industries, this estimation will be difficult, complicated, and run the risk of being inaccurate to the point of being misleading in some sectors due to lack of comparable companies. Although, as more and more jurisdictions start to require ESG information from their companies, the number of companies which publish data will eventually increase and make the estimations more manageable and accurate.

1.2 Problem

Around 3,000 out of 40,000 publicly traded companies are publishing relevant ESG data, which makes it challenging for an investor to accurately determine quantities such as carbon emission or water consumption attributable to their portfolio. Even though there are several companies trying to gather this data, and pushing companies which do not publish it to publish it, the majority of the publicly traded companies do not even measure the data themselves.

Additionally, since the companies which do publish their data are often in western jurisdiction and certain industries, companies in other jurisdictions and industries becomes even more complex to estimate due to the low number of comparable companies for those industries and regions.

1.3 Purpose

The purpose of this thesis is to develop a mathematical or statistical model to estimate the carbon emissions of companies which do not publish the data of their emissions. The model will be deployed in a larger ESG product at a bank to help thier clients estimate their own ESG alignment and carbon emissions attributable with their investment decisions and positions. This will help clients make more informed decisions and help them align their decisions with the interest of their stakeholders. Additionally, this will direct fund flows and financing to more ESG-aligned companies

and hence, will incentivize companies to be more ESG-aware as that would decrease their cost of capital.

1.4 Goal

The goal of this thesis is to develop a scalable, trustworthy, robust, and as accurate model as possible which estimates carbon emissions of companies which do not measure or publish their carbon emissions on their own. The model will use discrete data, such as the region and industry it is operating in, as well as publicly available financial data such as revenue, employees etc.

1.5 Benefits, Ethics and Sustainability

With this thesis, clients and institutional investors will be more aware of the ESG impact attributable to their positions in their portfolios, which in turn will help them make more well-informed decisions regarding their investment strategy and responsibilities towards their stakeholders and society.

When investors direct their investments towards more ESG-aligned companies with smaller carbon intensity, the Weighted Average Cost of Capital (WACC) of those companies will decrease.[2] A lower WACC will bring multiple benefits to a company, for example, it will be easier to finance current operations and new projects.

1.6 Methodology

The methodology will consist of three parts: data analysis, model implementation, and model validation.

The data analysis will be performed on the provided data. Firstly, which features to be used for the models are chosen. Secondly, missing data or wrong data will be managed and corrected. Lastly, outliers and influential data points are identified and handled.

The model implementation will be based on a literature review on what suitable models there are for this specific data set and objective. Based on previous research and how

models have performed, a number of models which will be chosen to be implemented and created based on the data set.

The model validation will be performed by splitting the data set in a training and test data set, consisting of 90% and 10% of the data points in the original data set. The models will be created based on the training data set, and will then be run on the test data set to compare the accuracy of the trained model with the actual values. Additionally, the robustness of the models will be tested by comparing their accuracy and explanatory value for different number of outliers removed.

1.7 Stakeholders

There are two main stakeholders for this project. Firstly, it is the bank and more specifically, the risk and valuation services team and sustainable banking team. These teams will operate and sell the products and services which will utilize the model developed in this thesis. The better the estimation model becomes, the more accurate and trustworthy the products will be, and thus also easier to sell.

Secondly, the clients of the bank are the second stakeholders, which constitutes of institutional investors, such as pension funds, and larger corporations. When purchasing the products based on this model, they expect the data to be accurate and trustworthy.

1.8 Delimitations

This thesis will be limited i mainly three ways.

Firstly, when declaring what variables and/or features that will be used, only the data provided by the same company which has provided the carbon intensity data set will be used. This is due to three reasons. Firstly, there is supposed to be a strong causality between the provided data features and carbon intensity, which would make these specific features appropriate to use. Secondly, there would be a tremendous amount of work to manually look up each company. As there are roughly 40,000 companies to get data from, it is not an option for the scope of this thesis to manually get it. Thirdly, the bank has already been using these features when estimating the carbon intensity before, and would prefer to continue to do so as they have not got the time nor interest

to get data manually.

Secondly, only data from the reporting will be used. This is mainly due to the fact that corporations are actively trying to decrease their carbon intensity, which results in old data not accurately describe the corporation and their carbon intensity today. Moreover, since corporations who have published carbon intensity data before most likely still is do so, there will be data points from a later date already in the data set which would be of more interest.

Thirdly, there will be a limited number of methods and models to estimate carbon intensity with, and there might be other models which would yield similar or better results than the ones used in this thesis. However, according to previous literature on the topic, these models are among the most promising ones and the probability that models not used in this thesis would perform on a better accuracy or greatness robustness is very low.

1.9 Outline

In chapter 2, the background to the project and theoretical background of the models used will be presented. In chapter 3, the methodology and how the project was carried out will be presented, and in chapter 4, the result from the project and the final model will be presented. Lastly, chapter 5 will be a conclusion and analysis of the project, discussing what its strengths and weaknesses are and potential improvements.

Chapter 2

Theoretical Background

2.1 Related Work

Adamowski and Christina Karapataki [1] found in their paper that artificial neural network accurately predicts water consumption in Nicosia, and does so significantly better than the alternative regression methods used in the paper. Although the datasets used in that paper differ from the ones to estimate carbon intensity by both frequency, nature, and data type, it suggests procedures on how to implement an artificial neural network and how it outperforms other regression models.

Manna et. Al. [10] compared boosted decision trees with other regression methods in their paper and concluded that boosted decision trees generally outperform other regression tecniques, such as ordinary decision/random forest or linear regression. They tried to predict flight delays at a daily basis with different sequential features available beforehand and found both higher accuracy and robustness in the method.

Karim, Albitar, and Elmarzouky [3] found that there is a strong causality between carbon intensity and capital expenditure in British companies. Furthermore, while British carbon emissions has successively declined in the UK in recent years, it has mainly been attributable to decrease in corporate capex.

2.2 The dataset

The dataset consists of all publicly traded companies in the world, their carbon intensity (if reported), and key metrics derived from their income statement, balance sheet, and annual report. The key metrics included are listed in table 2.2.1.

Feature	Data type	Description
Country	String	Country of which the company is noted in
Currency	String	Currency of which the company's stock is noted in
Exchange	String	The code of the exchange on which the company's stock is noted on
Industry	String	Primary industry the company is operating in
Issue Type	String	Type of security
Region	String	Region the country belong to
Sector	String	Primary sector the company is operating in
Nace Code	Float	Sector ID number, will not be used
Total Revenue	Float	Total revenue of the group
Capex to revenue	Float	Capex over total revenue
Gross PPE to revenue	Float	Gross PPE over total revenue of the group
Log Market Cap 3y Look Back	Float	Log of average market cap over last three years
Total employees/revenue	Float	Total full-time employment equivalents over total revenue
Sub-industry	String	Sub-industry the company is operating in
Sub-sector	String	Sub-sector the company is operating in

Table 2.2.1: List of all available features, what type of data the feature is, and corresponding description. Features marked in **bold** will be used when implementing the models.

2.2.1 Carbon intensity

Carbon intensity is a measurement of a companies carbon emission in relation to economic activity in the form of revenue generated. Carbon intensity is defined as follows

$$\text{CarbonIntensity} = \frac{\text{Totalgreenhousegasemissionequivalents}}{\$\text{MUSDinrevenue}}. \tag{2.1}$$

The total greenhouse gas emission equivalents include both the emissions from the company's direct operation and the emissions from its electricity usage. The greenhouse gas emission equivalents contain all types of greenhouse gases, and is measured in CO_2 equivalents weighted by potency.

2.3 Data Analysis

2.3.1 Missing Values

Some data points in the data set may not be complete and have missing values, possibly due to wrongly inputted data, errors in measurement or due to other circumstances. A regression model cannot use incomplete data points for neither training nor testing; these data points need to be modified or removed. There are several methods to handle missing values, these are

- Listwise deletion
- Last observation carried backward
- Conservative imputation
- Multivariate imputation

Listwise deletion is a technique where each data point with some missing values is deleted. Hence, it is the most ”right” technique since it does not manipulate the data set in any way, however, there may be structural reasons for why some values are missing and some not. For example, if there is a form about obesity and people have to fill in their weight and waist measurements, obese people might to a higher extent not fill in their measurements. When evaluating the data, the data would be skewed if we remove the missing data, since the missing data is mostly from obese people.

Last observation carried backward is a method where a missing value is filled by taking the last available value in the time series. This method is not applicable in this

specific case since it requires a time series to get older data, and in this case there are no historical values available for each company.

Conservative imputation is when missing values are filled with a pre-determined value. Hence, this method skews the data set and is not scientifically motivated. To use conservative imputation, the chosen pre-determined value to fill the missing values with must be carefully motivated and taken into account when performing the further data analysis.

Multivariate imputation consists of multiple different methods and techniques to estimate the missing value based on the rest of the data set. Some common examples are taking the mean of the specific value from the data points which are not missing the value, or using a regression model based on the remainder of the data set to estimate the missing values.

2.3.2 Dummy Variables

When performing a regression on a data set with discrete underlying variables, such as text strings or classification integers, one needs to expand the underlying variable into multiple binary variables, one for each possible discrete value. For example, if one of the underlying variables for a regression model is country of origin and each country would be assigned an integer, the model would misinterpret one country to be higher ranked than another, or that there is some sort of numerical relation between the countries, even though there is not. Therefore, a binary dummy variable is created for each discrete value of a variable, which takes the value 1 if the underlying variable is the specific value and 0 otherwise.

2.3.3 Outliers

Outliers [6] are observations which are more influential to the model than other data points. These data points are identified by abnormally large residuals or high leverage and should be investigated to see why they are deviating from the rest of the data points. Outliers may indicate inadequacies in the model, such as faulty measurements or incorrect data. If this is the case or if an outlier involves certain circumstances not relevant to the situation, then it should be corrected or deleted from the data set. However, if no such characteristic features exist, the point might be more important than the rest of the data since it may impact many key model properties.

When testing models, the model will perform better on the test data set, as described in 2.5.1

2.3.4 Cook's Distance

Cook's distance [7] is a method used to detect outliers and measures the outliers' influence on the regression model. It is defined as

$$D_i = \frac{(\hat{\beta_{(i)}} - \hat{\beta})'(X'X)(\hat{\beta_{(i)}} - \hat{\beta})}{(k-1)MS_{Res}}, \quad i = 1, \ldots, n, \tag{2.2}$$

where $\hat{\beta_{(i)}}$ is the least-squares estimator of the ith estimation based on all observation except the ith one and the mean squared residual

$$MS_{Res} = \frac{SS_{Res}}{n-k-1}. \tag{2.3}$$

Cook's distance of x_i depends on two things that can contribute to an increased value of D_i, firstly the location of x_i in the space of x. Secondly, the effect of the residual.

The first expression reflects how far point x_i is from the rest of the data. This is important when determining the parameter estimates, predicted values, and standard errors, etc. However, some points might be far from the rest of the data set but will lie on, or close to the regression model passing through the remaining sample points and will therefore have little to no effect on the regression coefficients. The next expression reflects how well the model fits the ith observation y_i. Data points which surpass the threshold

$$D_i > \frac{4}{n-k-1}, \tag{2.4}$$

are considered as outliers.

2.4 Model Selection

Previously, a linear regression model has been used to estimate the carbon emissions at the bank. Although it is a reasonable model to use for the estimation, there are several different types of regression models which works in different ways and have different strengths and weaknesses. Below some of the more interesting will be listed, together with how they work and their respective strengths and weaknesses.

2.4.1 Linear Regression

Linear regression [6] is a model which is based on the assumption that there exists a linear relation between the dependent variable and each of the independent variables, such that the dependent variable is equal to a linear sum of all independent variables. It is derived as

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \tag{2.5}$$

where y is the dependent variable, x_j are the independent variables and β_j are the coefficients, $j = 1, ..., k$. For a given data set, the equations for the data points can be written in matrix form as

$$Y = X\beta + \epsilon, \tag{2.6}$$

where

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, X = \begin{bmatrix} 1 & x_{1,1} & \ldots & x_{1,k} \\ 1 & x_{2,1} & \ldots & x_{2,k} \\ 1 & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \ldots & x_{n,k} \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}. \tag{2.7}$$

ϵ is an error term which is defined by the difference between the true value of Y and the estimated $\hat{Y}$, such that $\epsilon = |Y - \hat{Y}|$. ϵ_i is assumed to follow a normal distribution N(0,σ^2) with constant variance σ^2 and are independent of each other.

$\hat{Y}$ is defined as the predicted Y and are calculated as

$$\hat{Y} = X\hat{\beta}, \tag{2.8}$$

were $\hat{\beta}$ is defined as the least-squares estimators which are the coefficients that minimize the sum of squares of the residuals SS_{Res},

$$\begin{aligned} \sum_{i=1}^{n} \epsilon_i^2 = (y - X\beta)'(y - X\beta) = y'y - \beta'X'y - y'X\beta + \beta'X'X\beta = \\ = y'y - 2\beta'X'y + \beta'X'X\beta \end{aligned} \tag{2.9}$$

since $\beta'X'y$ is a 1 x 1 matrix, or a scalar, and its transpose $(\beta'X'y')' = y'X\beta$ is the same

scalar. The least-squares estimators must satisfy

$$\left.\frac{\partial S}{\partial \beta}\right|_{\hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0 \Rightarrow X'X\hat{\beta} = X'y. \tag{2.10}$$

By multiplying both sides by $(X^TX)^{-1}$ the least-square estimator of β is obtained

$$\hat{\beta} = (X'X)^{-1}X'Y. \tag{2.11}$$

2.4.2 Bayesian Linear Regression

Bayesian linear regression [9], is similar to ordinary linear regression, as described in section 2.4.1, in notation and variables. The main difference is that we introduce a prior probability distribution over the model parameters **w**, and treat the noise precision parameter β as a known constant. $p(\mathbf{t}|\mathbf{W})$ is given by

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T\phi(\mathbf{x}_n, \beta^{-1}) \tag{2.12}$$

where **X** are the explanatory variables, **t** the corresponding response variables, ϕ are the basis functions, and $\mathcal{N}$ is the normal distribution. The corresponding conjugate prior is therefore given by the following gaussian distribution of the form

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0) \tag{2.13}$$

where $\mathbf{m}_0$ and $\mathbf{S}_0$ are the mean and the variance, respectively. To compute the posterior distribution, which is proportional to the product of the likelihood function and the prior, it is possible to use the standard result of a normalized gaussian distribution. With some deriving, one can find that the posterior function is

$$\begin{aligned} p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \\ \mathbf{m}_N &= \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\phi^T\mathbf{t}), \\ S_N^{-1} &= S_0^{-1} + \beta\phi^T\phi. \end{aligned} \tag{2.14}$$

2.4.3 Decision Forest Regression

A decision tree or forest [4] is a technique where a tree is formed by dividing data based on criteria or questions. The source data set is split in smaller sub data sets in

each step. The name decision tree comes from the fact that it begins as one root which splits into several branches lastly ending in leafs, where an estimation is made. The decision tree is simple in its nature, and the factors which determine its accuracy and characteristics are in which order the features are regarded and split based upon and how many layers deep the tree is before the leafs come. Given a set of data, the tree is updated and ”trained” to minimize the errors of the estimation. A random forest is multiple random trees and generated in a random manner.

2.4.4 Neural Network Regression

An artificial neural network [4] is a computing system which consists of units, called nodes, inspired by neurons in the human brain. Each node can transmit impulses to other nodes, which perform a non-linear transformation, called propagation functions, on the sum of received inputs and in turn send out new impulses to other nodes. Neural networks are trained by giving them inputs and corresponding targets, where a network adapts its connections and weights to adapt the inputs to the corresponding target.

A neural network can be described as a series of functional transformations. Construction of M linear combinations of input variables $x_1, \dots x_D$ gives

$$a_j = \sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)} \tag{2.15}$$

where $j = 1, \dots M$ and the superscript (1) indicates that the corresponding parameters are in the first layer of the network. a_j are quantiles, also known as activations, each transformed by a differentiable, non-linear activation function $h(\cdot)$. These quantiles correspond to the output values of the basis functions for linear models used for regression and classification, described in equation 2.16.

$$y(\mathbf{x}, \mathbf{w}) = f\left(\sum_{j=1}^{M} w_j \phi_j(\mathbf{x}) \right) \tag{2.16}$$

$f(\cdot)$ is a non-linear activation function in the case of classification. For regression cases, $f(\cdot)$ is the identity. $\phi_j(\mathbf{x})$ are non-linear basis functions. Often, the non-linear functions $h(\cdot)$ are chosen to be sigmodial functions. Combining equations 2.15 and 2.16, one get

$$a_k = \sum_{j=1}^{M} w_{kj}^{(2)} z_j + w_{k0}^{(2)} \tag{2.17}$$

where $k = 1, \ldots, K$ are the number of outputs. Apart from equation 2.17, this corresponds to the second layer of the network. Lastly, to define the output y_k, the activation functions need to be chosen, which depends on the type of problem the network should solve.

To train and further improve the neural network, more data with the corresponding label is needed. With the data vectors $\{\mathbf{x}_n\}$ and target vectors $\{\mathbf{t}_n\}$, one wants to minimize the error function

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} ||\mathbf{y}(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n||^2. \tag{2.18}$$

2.4.5 Boosted Decision Tree Regression

A boosted decision tree [4] works in the same way as a decision tree, as described in 2.4.5, but with gradient boosting. Gradient boosting is a technique which forms a model based on an aggregation of multiple smaller and weaker models. Specifically for decision trees, this method is called *boosted decision trees*.

One example of boosting is AdaBoost, which works in the following way

1. Set the data weighting coefficients $\{w_n\}$ by setting $w_n^{(1)} = \frac{1}{N}$ for $n = 1, \ldots, N$.
2. For a given classifier $y_m(\mathbf{x})$, we want to adjust the weights in order to minimize the weighted error function $J_m = \sum_{n=1}^{N} w_n^{(m)} I(y_m(\mathbf{x}_n) \neq t_n)$. Do this for $m = 1, \ldots, M$.
3. Evaluate the quantities $\epsilon_m = \frac{\sum_{n=1}^{N} w_n^{(m)} I(y_m(\mathbf{x}_n) \neq t_n)}{\sum_{n=1}^{N} w_n^{(m)}}$, and calculate $\alpha_m = \ln\{\frac{1-\epsilon_m}{\epsilon_m}\}$,
4. Update the weights: $w_n^{(m+1)} = w_n^{(m)} \exp\{\alpha_m I(y_m(\mathbf{x}_n) \neq t_n)\}$,
5. Make new predictions, given by $Y_m(\mathbf{x}) = \text{sign}(\sum_{m=1}^{M} \alpha_m y_m(\mathbf{x}))$.

2.4.6 LGBM

Light Gradient Boosting Machine (LGBM), is a gradient boosting framework that uses tree based learning algorithm. Instead of growing the tree horizontally, it grows new

leafs first and then new layers vertically. This improves speed and is more data light than conventional methods, making it more suitable for larger data sets.

LGBM is increasing in popularity and is hence of interest to try in this thesis. However, LGBM has had problem with overfitting which implies it needs quite large data sets to ensure robustness and high accuracy. Nevertheless, it will be implemented and tried out to compare its prediction to the ones of other models.

2.5 Model Validation

2.5.1 Training and Test Data Split

When training and testing the accuracy of regression and machine learning models, to ensure that the model does not overfit and is still accurate, the data set is divided in two smaller data sets, one for the model to train on and one to test the accuracy. The model will therefore not have seen the data in the test data set when training on the train data set. This will give a measure of the accuracy and explanatory value of the model. The model created for the bank ultimately will be created based on all available data.

2.5.2 Root Mean Square Error

Root Mean Square Error (RMSE) [5] is defined as the root of the arithmetic mean of the square of the difference between the actual value and the estimated value, that is

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{1}^{n}(\hat{y}_i - y_i)^2}. \tag{2.19}$$

RMSE is a measure of the accuracy of the model and is suitable for regression of continuous estimated values, but not for classification problems, since it does not make sense to calculate a mean square error of labels. The smaller the RMSE, the more accurate the model.

2.5.3 R-squared

R^2 [8] is a measure of how much of the deviancy in the response variable that can be attributed by deviances in the explanatory variable. That is

$$\begin{aligned} \bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i, \\ \mathrm{SS}_{\mathrm{tot}} &= \sum_i (y_i - \bar{y})^2, \\ \mathrm{SS}_{res} &= \sum_i (y_i - f_i)^2 = \sum_i e_i^2, \\ R^2 &= 1 - \frac{\mathrm{SS}_{\mathrm{res}}}{\mathrm{SS}_{\mathrm{tot}}}, \end{aligned} \tag{2.20}$$

where $\bar{y}$ is the estimate of y, y_i is a data point of the response variable. $\mathrm{SS}_{\mathrm{tot}}$ is the total sum of squares and $\mathrm{SS}_{\mathrm{res}}$ is the residual sum of squares. R^2 will take a value between 0 and 1, where 1 indicates a perfect explanatory value while 0 is zero explanatory value.

2.5.4 Robustness

In this thesis, robustness of the model is defined as small changes in accuracy, i.e RMSE and R^2, when changing the data. In this case, it is of interest to see how the accuracy change when more or less outliers are removed from the data set. Thus, RMSE and R^2 of a robust model does not change much when outliers are removed, and controversially, the model does not perform significantly worse when outliers are present.

Chapter 3

Method

3.1 Data Preparation

The data from the bank was provided in a Comma Seperated Values (CSV) format, which was uploaded to python into a pandas data frame using the pandas library. The data was then filtered so only the data points of the issue type ”Shares” was used, which is the only issue type which is of interest for this specific product. Since the data set is relatively small, with only 2260 data points, the risk of overfitting for each category of feature becomes high if the data is bucketed in too small sets. Therefore, all categorical data features apart from *country* and *sector*, which are the variables with the highest explanatory value for the carbon intensity of the companies, were removed. The *country* and *sector* features were then divided into dummy variables, as described in 2.3.2.

After all features of interest was saved down, missing values in the data set needed to be handled. A multivariate imputation was used since the data set is small and we do not want to lose the information in the entries with missing values. The multivariate imputation method, as described in 2.3.1, used is the *IterativeImputer* function from the sklearn library, which makes a regression to estimate the missing values. The data set was then split in two sets, one with the companies which have reported their emission data and one with the companies which have not. Cook’s distance, as described in 2.3.4, was calculated to find the most influential data points in the model. Even though data points should ideally not be removed, three variants of the data set were created to both be able to compare the robustness of the models and determine

the accuracy given different level of ”conservativeness” when removing data points. The first variant used all data points, i.e no outliers were removed. In the second one, the four most influential data points were removed since these represented states of companies which are highly unlikely or misguiding. These states might have been correct but at the same time, might just as well have been false. One example is a company where it had more than 3,000 employees per M$ revenue, which is equivalent in this case of the company having a revenue of about $300 per employee, far lower the rest of all companies. This might have been a legitimate case for this specific company, but due to the high influence on the model the datapoint was removed in this data variant. In the third variant, a lower threshold for outliers were used and a total of 12 outliers were removed. These included companies which had remarkable financials, such as negative revenue and high capex or PPE.

3.2 Model Implementation

The data set with the corresponding reported carbon intensity measures were divided in a test set and training set, where 90 percent of the data points were in the training set and 10 percent in the test set. Each model described in 2.4 was implemented using different python libraries. From the library sklearn, the functions Linearregression, BayesianRidge, and tree were imported and used for the linear regression, bayesian linear regression and decision tree models, as described in 2.4.1, 2.4.2, and 2.4.3. The neural network was built with tensorflow, in the same way as described in 2.4.4. It consisted of three layers with 33, 16 and 1 node respectively. The boosted decision tree was built with the xgboost library, as described in 2.4.5 and when implementing the boosted decision tree, an optimal depth was calculated by implementing the model multiple times and training the model on the training data set and then running it on the test data set. The optimal depth was the depth which corresponds to the lowest RMSE and highest R^2. Additionally, the LGBM regressor from lightgbm was tested along the other models. This was repeated 1,000 times for each model and each variant of the data set, saving the mean of the 1,000 iterations.

3.3 Model Comparison

After each model was trained, it was run on the test data set and the performance was evaluated by calculating RMSE and R^2. It was then compared to the same measures from the other data set variants.

Chapter 4

Results

4.1 Scatter Plots of Continuous Variables

In figure 4.1.1, 4.1.2, 4.1.3, 4.1.4, and 4.1.5, one can see the scatter plots of carbon intensity to the different continuous variables. To summarize, the outliers which were removed in the data set variants greatly reduced the spread in the data, but not enough to see any clear trends.

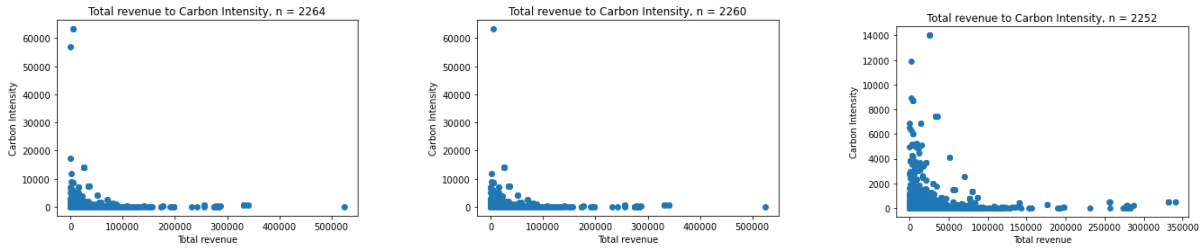

Figure 4.1.1: Plots over total revenue to carbon intensity for the three different variants of the data set. Larger plots can be found in appendix A.1.

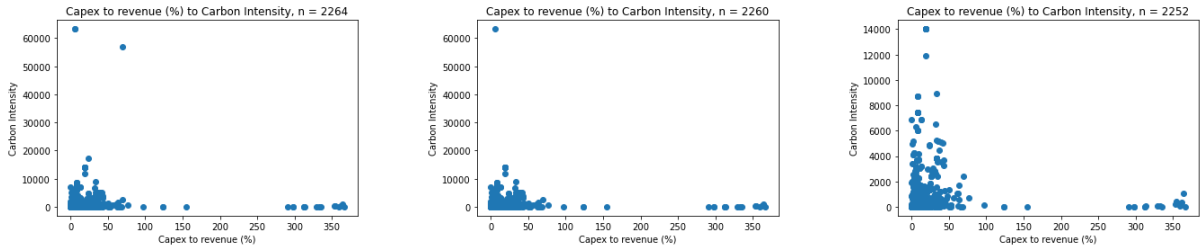

Figure 4.1.2: Plots over total revenue to capex over revenue for the three different variants of the data set. Larger plots can be found in appendix A.2.

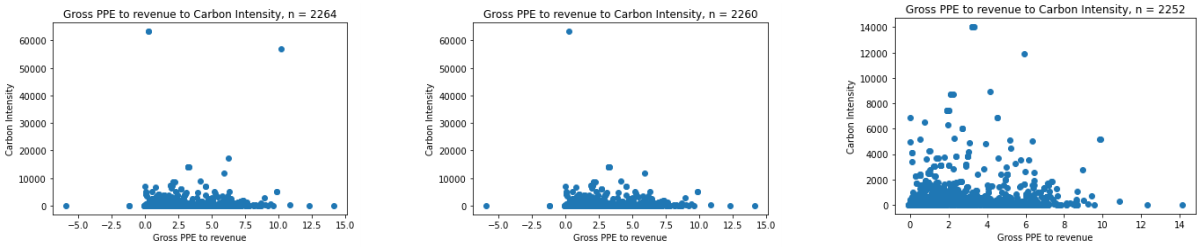

Figure 4.1.3: Plots over total revenue to PPE for the three different variants of the data set. Larger plots can be found in appendix A.3.

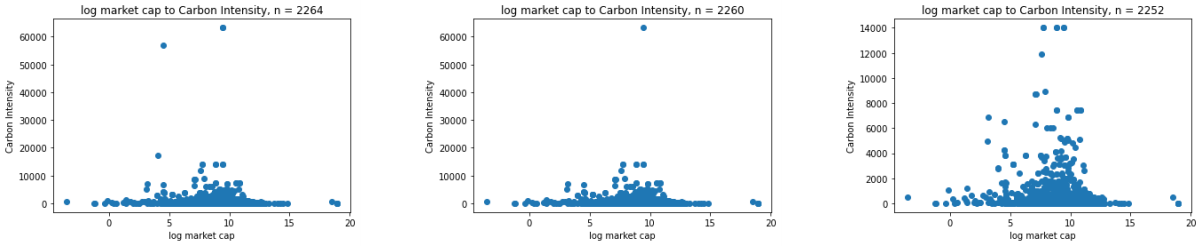

Figure 4.1.4: Plots over total revenue to market cap over revenue for the three different variants of the data set. Larger plots can be found in appendix A.4.

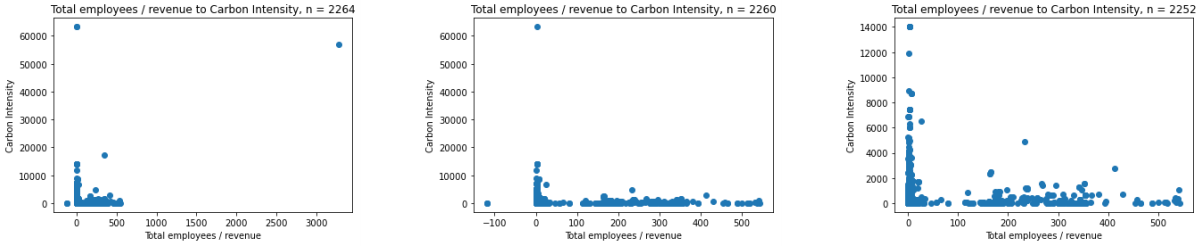

Figure 4.1.5: Plots over total revenue to number of employees over revenue for the three different variants of the data set. Larger plots can be found in appendix A.5.

4.2 Robustness

The results for the robustness are presented in figure 4.2.1 and 4.2.2. While multiple models perform exceptionally well when they are based on the data set variant where the most outliers are removed, the boosted decision tree performs by far the best when based on the data set variant where no data points where removed.

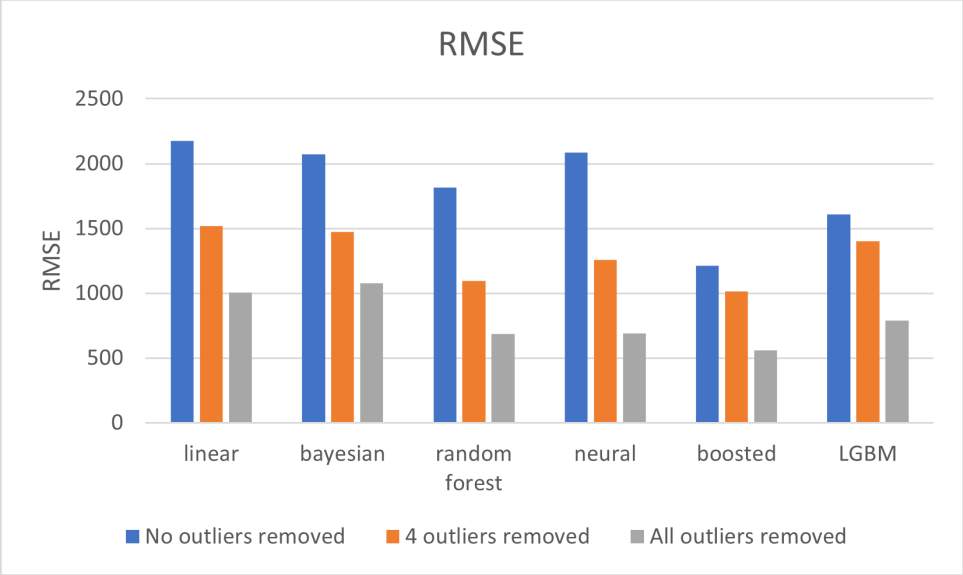

Figure 4.2.1: RMSE for different models and data set variants.

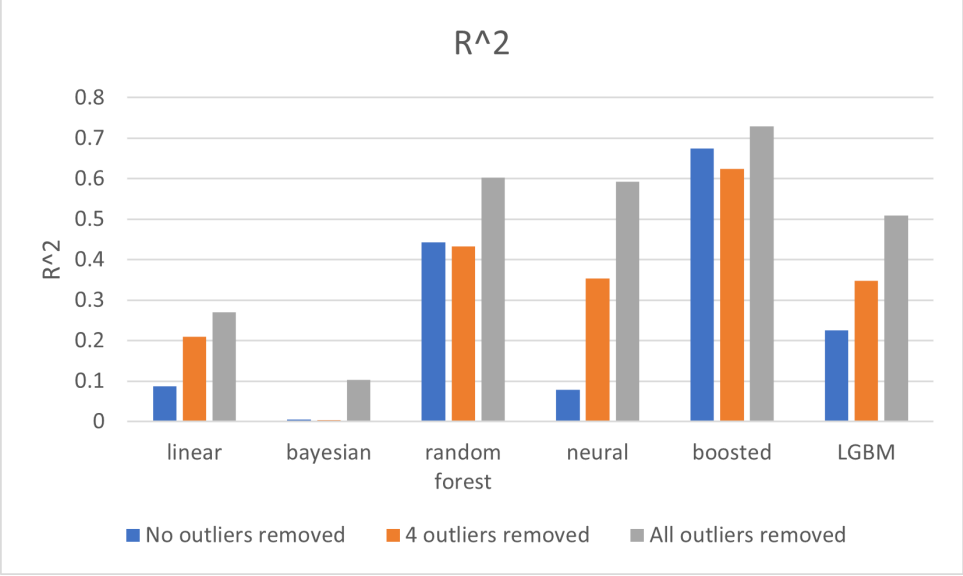

Figure 4.2.2: R^2 for different models and data set variants.

4.3 Accuracy

The result for the depth optimization of the boosted neural network is shown in table 4.3.1. It is shown that a depth of 50 gives the most accurate model. The result from different models on the test data set is presented in table 4.3.2. Similar to the result of the robustness, the boosted decision tree has the best accuracy.

Depth		
Depth	RMSE	R^2
1	1181.19	0.202
5	1167.69	0.513
10	967.58	0.640
50	904.87	0.664
100	981.42	0.652
1000	945.47	0.648

Table 4.3.1: The precision of the boosted decision tree for different depths.

Model performance on test data						
	All data points		4 outliers removed		12 outliers removed	
Model	RMSE	R^2	RMSE	R^2	RMSE	R^2
Linear	2175.3	0.0866	1519.9	0.209	1006.5	0.2701
Bayesian linear	2072.1	0.0056	1475.4	0.0035	1079.1	0.1035
Random forest	1817.3	0.4426	1096.1	0.4327	683.3	0.6032
Neural network	2083.9	0.0783	1257.4	0.353	688.21	0.593
Boosted decision tree	1214.0	0.6742	1016.1	0.6244	560.85	0.7298
LGBM	1610.2	0.2254	1402.0	0.3482	789.92	0.5092

Table 4.3.2: RMSE and R^2 for the different data set variants and models.

Chapter 5

Discussion & Conclusions

5.1 Discussion

5.1.1 Model Performance and Evaluation

As seen in section 4, boosted decision tree is the model with the lowest RMSE and highest R^2 for all data set variants. This indicates that it is the model which will be implemented at the bank, since it will have the lowest estimation errors and thus, be the most trustworthy model.

Both the linear and especially the bayesian models performed poorly, even when the twelve outliers were removed. This is probably due to two main reasons. Firstly, these models assume that there are linear relationships or gaussian distribution, which not necessarily is the case for this data. These assumptions make these models outright inappropriate if the underlying data do not follow the assumptions on which the models are based. Secondly, these models and methods are not especially powerful in regard to prediction value. Since the data is multifaceted and scattered between many industries and regions, more powerful methods would be needed in this case, which clearly shows in the results.

The random forest performs exceptionally well considering its simplicity. Although it is outperformed by its boosted variant, it still performs better than the other models. This could indicate that this type of problem is optimal for decision trees to solve, which would explain its high prediction power even though the model is relatively simple, outperforming both the more complex neural network and the LGBM regressor.

The main difference between the random forest and its boosted counterpart is the performance on the data set variants where no, or few, outliers were removed. This indicates that apart from higher accuracy, the boosted counterpart is far more robust for extreme outliers and companies whose carbon emissions are harder to estimate. Another strength with the random forest is the simplicity and easy implementation. When implementing a model as a part of a larger architecture, it will need to be as easy to understand and change as possible. Both the neural network and LGBM is hard to understand and explain while also complicated to update and change when implemented, especially as staff turnover means the people who implemented it initially might not be in the organization any longer. However, whereas both the neural network and LGBM are too complicated to implement, the pros of the decision tree compared to the boosted decision tree are out-weighed. Primarily due to the bad robustness of an ordinary decision tree and the fact that it is quite simple to boost a decision tree, which would increase both accuracy and especially robustness significantly.

Both the neural network and LGBM regressor are complex methods but performed more poorly than the decision forest and boosted decision tree. Both these methods are by their complex nature unsuitable to implement in the product portfolio at the bank, as described above. However, whereas the LGBM regressor does not distinguish itself and performs rather poorly, the neural network is the model which benefits from the largest gains in accuracy when removing outliers. If another variant of the data set with even more outliers removed would had existed, the neural network would probably outperform the boosted decision tree. This demonstrates a low robustness of the model, but high accuracy given ”good” data. Nevertheless, the neural network becomes unsuitable for this application due to the low robustness, but it is still a powerful tool which could be considered in the future for similar application or if there would be more data.

To summarize, the boosted decision tree is both the most accurate and robust model. Additionally, while not being the simplest model to implement, it is simple enough for the bank implement and for other developers to understand, whereas an ordinary random forest is not robust enough.

5.1.2 Methodology Evaluation

Data Preparation

In the data preparation, there were many alternative ways of performing the preparation. There were a limited number of features available and a relatively small data set, which confined the possibilities in the choice of data. As described in the delimitations, in section 1.8, it would not be possible to add any more features to the data set. However, since ESG is a growing area of topic and as demand for such data increases, the supplier of the data might be incentivized to add more features in the future, which would open up for the bank to reconsider what features are used for the model and what not.

Three different data set variants were used when training and evaluating the models to determine the robustness of them. Ideally, it would be desirable that the models handle the outliers and influential data points with little to no effect on the prediction of other data points. Since influential data points have a large effect on the model, this is not the case and the existence of outliers will decrease the accuracy and explanatory value. Furthermore, there are different reasons why companies do not report their carbon intensity, which might be due to regulations, PR, capabilities, or other factors. These are the companies which are going to be estimated, and while a company might be considered an outlier among the companies which are reporting the data, the probability that there are no other companies like them is very low. Thus, by removing outliers, information about these companies are lost which would have been very helpful in the estimation, but since we cannot see that information loss in practice now, one might be inclined to remove outliers for the seemingly better RMSE and R^2, which would be wrong in this case unless the data is wrong.

The purpose of this thesis is, as described in section 1.3, to develop a model for the bank to use to estimate carbon emissions. Whether or not the outliers will be removed and if so, to what extent, when the model is implemented at the bank is unknown, and thus, one would seek a model which is as accurate as possible regardless of how aggressive the outliers are removed. In this case, when 12 outliers were removed, the RMSE was only marginally better than the one of the neural network and random forest. However, the majority of the increase errors can be derived to the outliers which were removed, and thus the bank will be recommended to not remove any outliers at all.

Model Implementation

In this thesis, many of the models are based on pre-existing libraries and/or packages for python which provides simple but state of the art algorithms for the different techniques. I would not have been able to produce equal or better algorithms myself, especially given the time frame and scope of this thesis. Since many of these library are open source projects with extremely talented and skilled developers working on enhancing the capabilities of the model, any model I would have developed myself would at best be at par with these ones and therefore, it would not be worth to consider.

Model Comparison

There are several different ways of comparing the prediction, where RMSE and R^2 are only two of these. For the given purpose, the main objective is to minimize the estimation errors, and whether there really is a causal relationship between the used features and the carbon intensity or not is secondary. R^2 may therefore not be the most interesting measure for this thesis, but RMSE is measuring the estimation error, exactly what is of interest for the bank. Although the prediction error can also be measured in different ways. One example of another way to quantify the error is Mean Absolute Error (MAE). RMSE, compared to MAE, ”penalizes” large errors by giving them larger weights. What is more appropriate in this case is up to the product owner, but if the bank would estimate carbon intensities far from the real value of the companies which is a stakeholder of the bank, there might be conflict of interest between divisions where the reputation or relation of the bank is endangered due to this error. There is therefore desirable to minimize the largest errors at the cost of slightly larger errors of the other data points, at least in this case.

5.1.3 Further Improvements and Continued Studies

There are several improvements which can be done to this project. Firstly, many of the models can be fine-tuned for a specific task, and with more time, there is room for improvement by fine-tuning them even more. Secondly, as more and more companies will report their carbon intensity and the data sets will get more complete, the most optimal model and methodology might change. Hence the methodology and results should be evaluated at a regular basis as more data become available. Lastly, there

might be other methods, techniques, and available features to estimate the carbon intensity which will improve the estimation. In this scope, the methodology was limited to six techniques, however, there are many more which may, even if unlikely, increase the prediction performance.

5.2 Conclusions

5.2.1 Model Usefulness and Implementation

The final model is a great improvement from previous estimations at the bank and will hence be implemented after this thesis is completed. It will be automated and updated once a quarter when new corporate data becomes available.

5.2.2 Final words

Hopefully, the insights and methodology in this thesis will be of interest and of use to other students and companies in a similar situation. As the knowledge of machine learning, especially neural networks, advances, both new and current techniques and methodologies will become better and prediction performance will increase.

Bibliography

[1] Adamowski, Jan and Karapataki, Christina. “Comparison of Multivariate Regression and Artificial Neural Networks for Peak Urban Water-Demand Forecasting: Evaluation of Different ANN Learning Algorithms”. In: *Journal of Hydrologic Engineering* (2010). ISSN: 14413523. DOI: 10.1061/.ASCEHE.1943-5584.0000245. URL: https://ascelibrary.org/doi/pdf/10.1061/%5C%28ASCE%5C%29HE.1943-5584.0000245.

[2] Anthony C. Ng, Zabihollah Rezaee. “Business sustainability performance and cost of equity capital”. In: *Journal of Corporate Finance* (2015), VOL. 34, 128–149. DOI: https://doi.org/10.1016/j.jcorpfin.2015.08.003.

[3] ATM Enayet Karim Khaldoon Albitar, Mahmoud Elmarzouky. “A novel measure of corporate carbon emission disclosure, the effect of capital expenditures and corporate governance”. In: *Journal of Environmental Management* (2021).

[4] Bishop, Christopher. *Pattern Recognition and Machine Learning*. 1st ed. Springer-Verlag New York, 2006. ISBN: 978-0-387-31073-2.

[5] Chai, T. and Draxler, R. R.: “Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature, Geosci. Model Dev., 7, , , 2014.” In: (2014), pp. 1247–1250. DOI: https://doi.org/10.5194/gmd-7-1247-2014.

[6] Douglas C. Montgomery Elizabeth A. Peck, G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. 5th ed. Wiley, 2012. ISBN: 978-0-470-54281-1.

[7] José A.Díaz-Garcı, Graciela González-Farías. “A note on the Cook’s distance”. In: *Journal of Statistical Planning and Inference* (2004), VOL. 120, 119–136. DOI: https://doi.org/10.1016/S0378-3758(02)00494-9.

[8] Kasuyaas, Eiiti. “On the use of r and r squared in correlation and regression”. In: *Ecological Research* (2018), Pages 235–236. DOI: 10.1111/1440-1703.1011.

[9] Minka, Thomas P. “Bayesian linear regression”. In: (September 29, 1999).

[10] Suvojit Manna, et Al. “A statistical approach to predict flight delay using gradient boosted decision tree”. In: *2017 International Conference on Computational Intelligence in Data Science(ICCIDS)* (2017). DOI: 10.1109/ICCIDS.2017.8272656.

Appendix - Contents

A Larger Scatter Plots **32**

A.1 Carbon intensity to total revenue 32

A.2 Carbon intensity to Capex 34

A.3 Carbon intensity to PPE 36

A.4 Carbon intensity to Market Cap 38

A.5 Carbon intensity to Employees over revenue 40

Appendix A

Larger Scatter Plots

A.1 Carbon intensity to total revenue

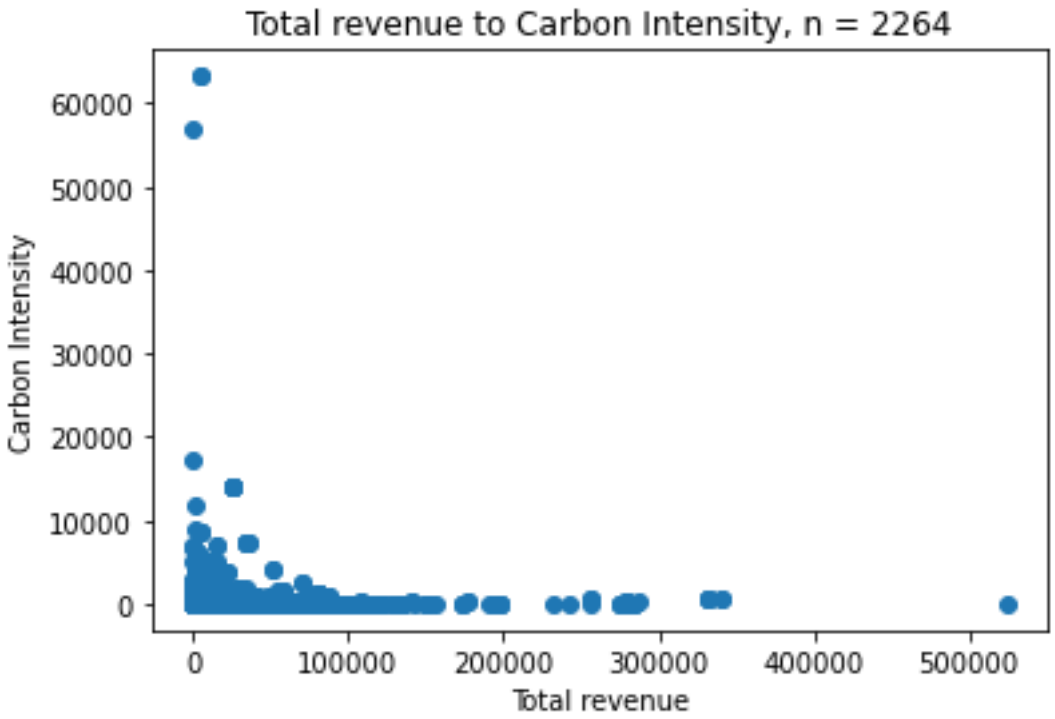

Figure A.1.1: Carbon intensity to total revenue with no outliers removed.

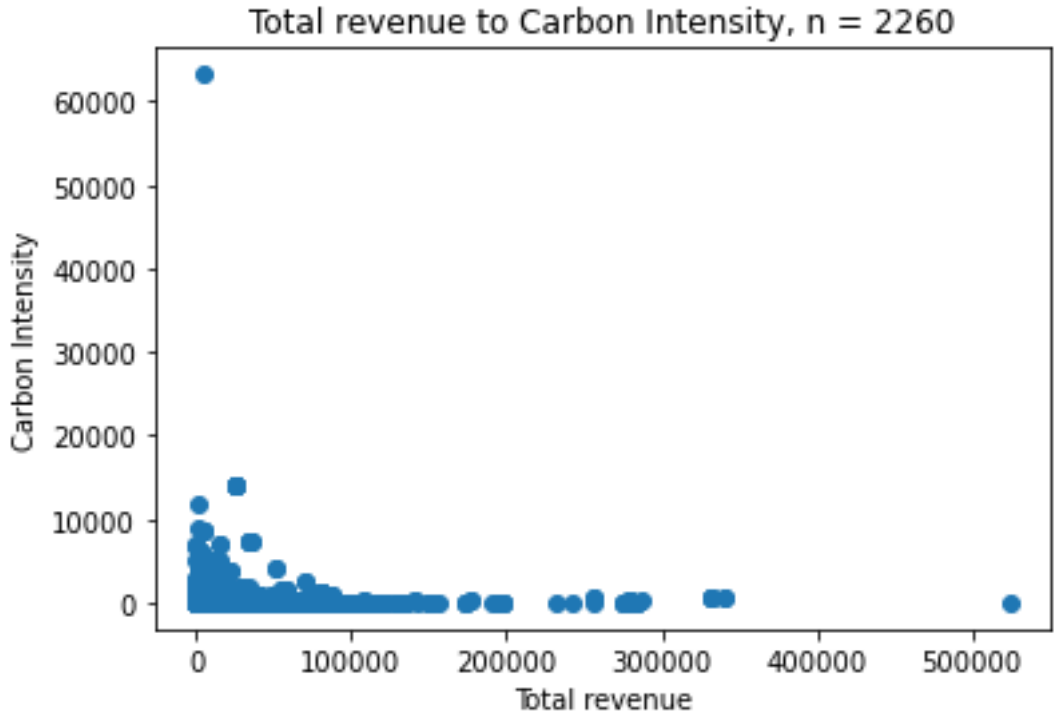

Figure A.1.2: Carbon intensity to total revenue with four outliers removed.

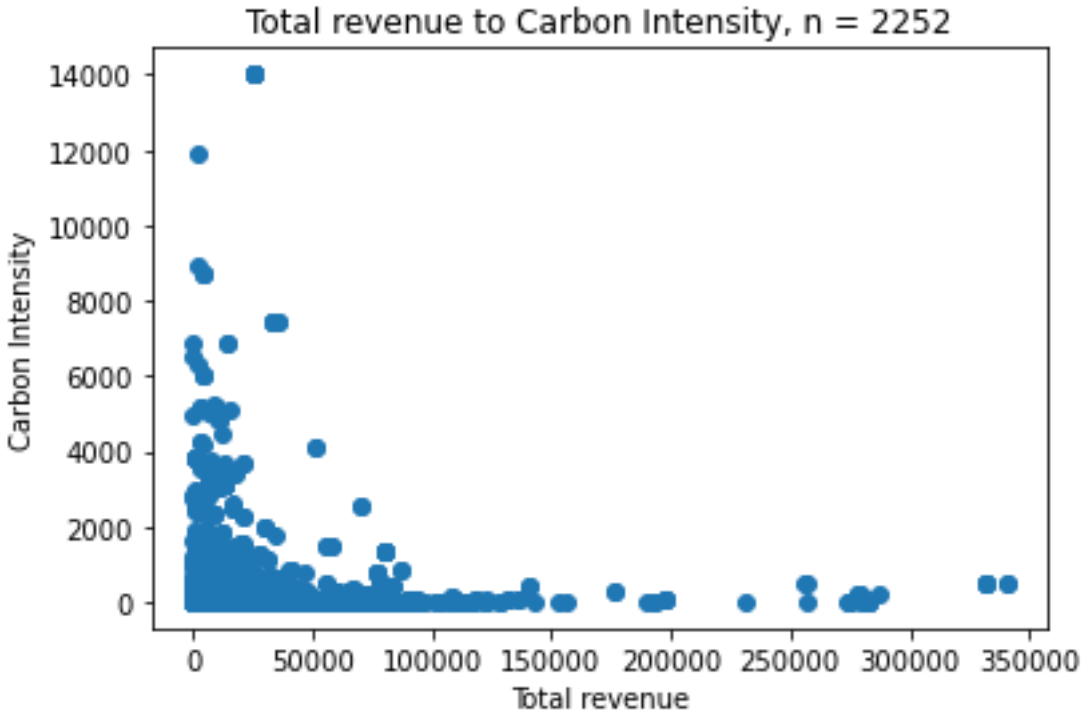

Figure A.1.3: Carbon intensity to total revenue with twelve outliers removed.

A.2 Carbon intensity to Capex

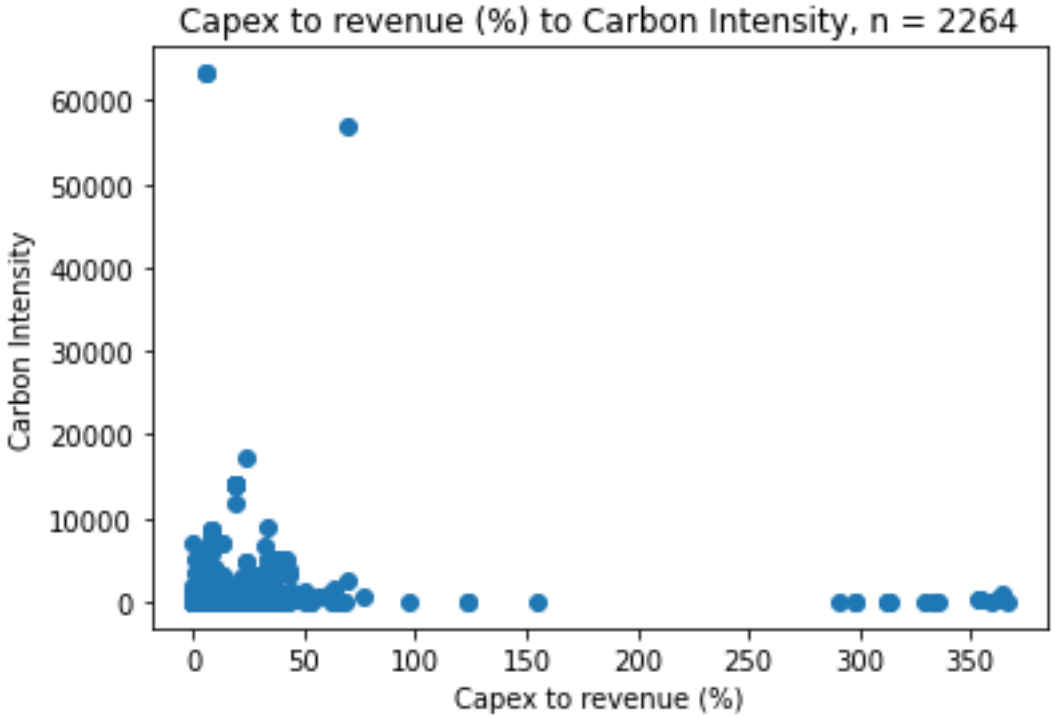

Figure A.2.1: Carbon intensity to capex over revenue with no outliers removed.

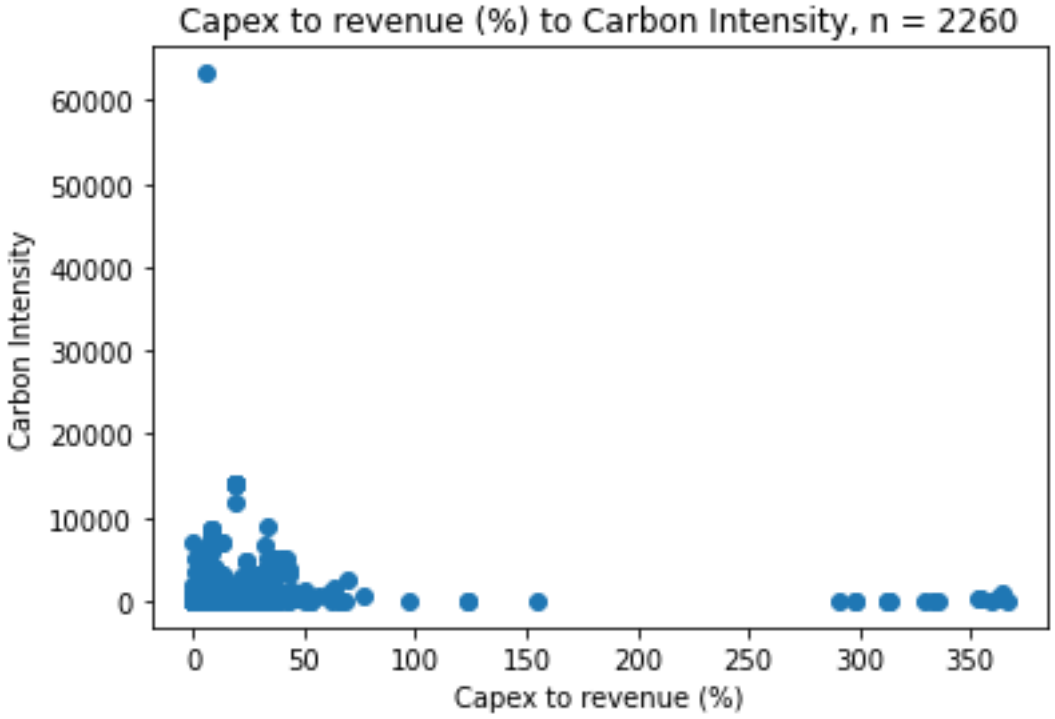

Figure A.2.2: Carbon intensity to capex over revenue with four outliers removed.

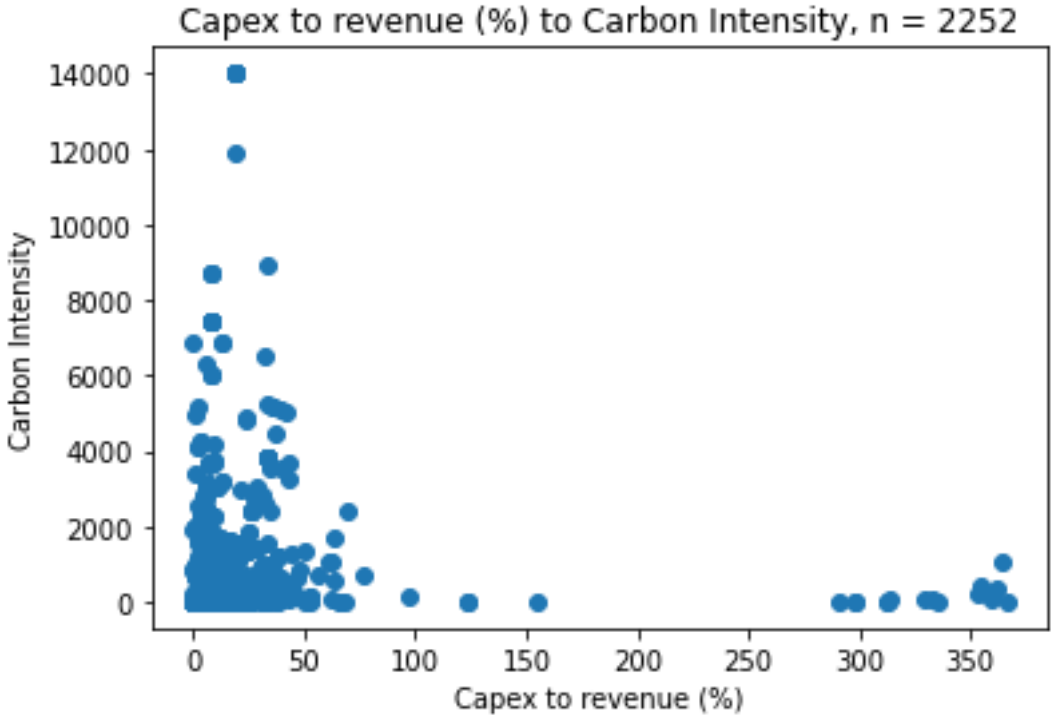

Figure A.2.3: Carbon intensity to capex over revenue with twelve outliers removed.

A.3 Carbon intensity to PPE

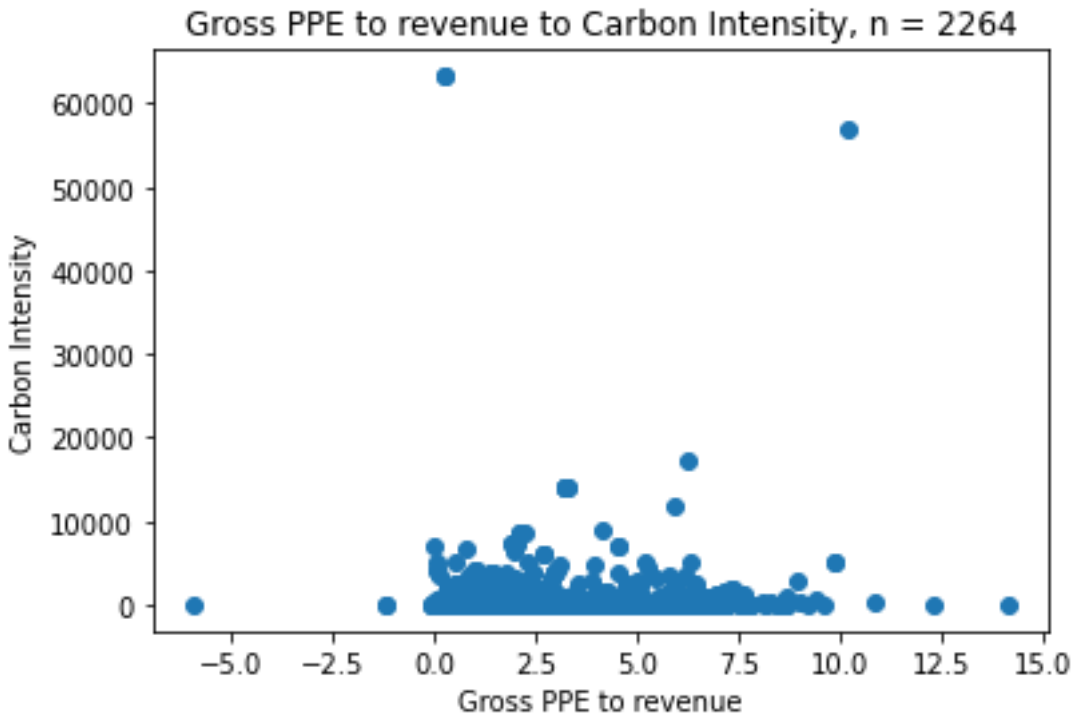

Figure A.3.1: Carbon intensity to PPE over revenue with no outliers removed.

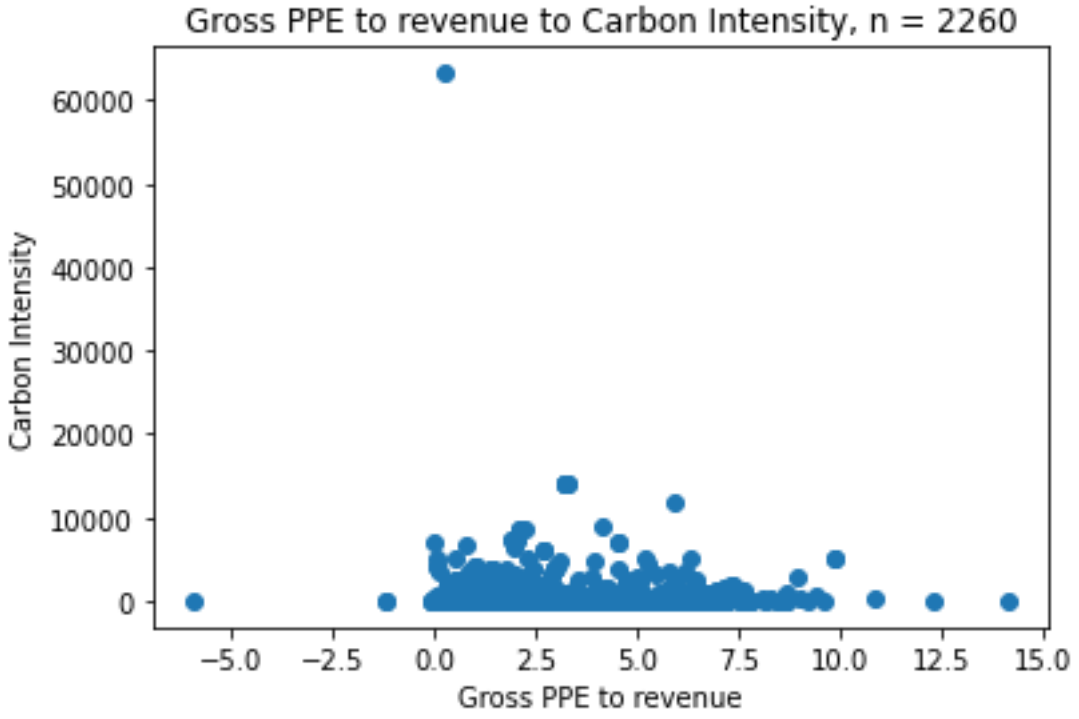

Figure A.3.2: Carbon intensity to PPE over revenue with four outliers removed.

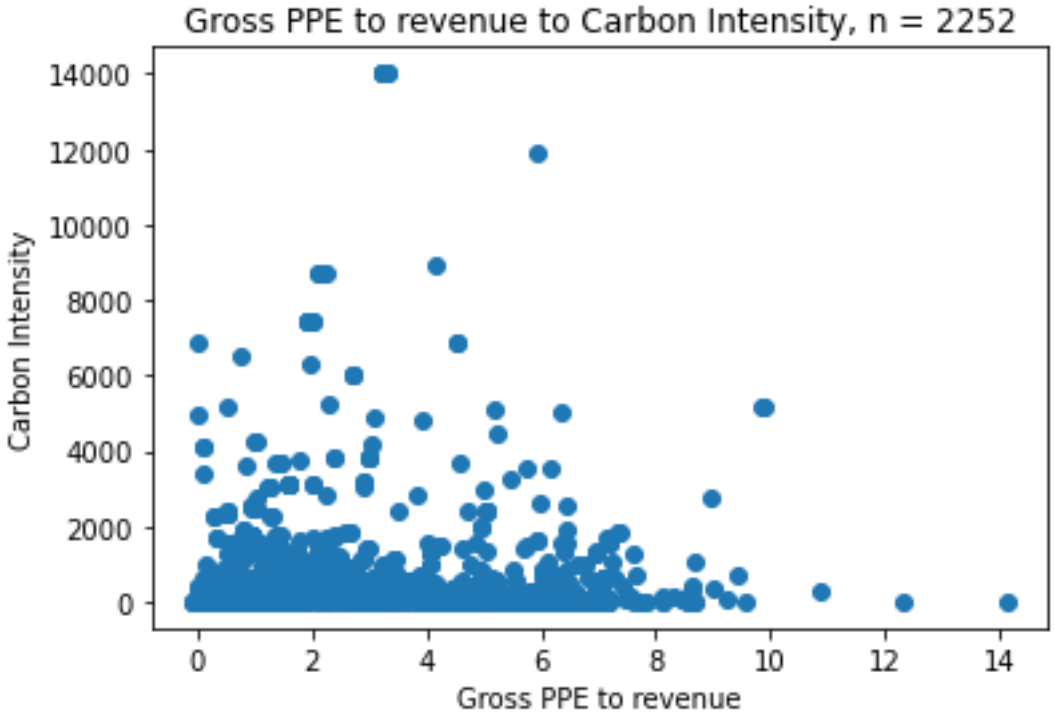

Figure A.3.3: Carbon intensity to PPE over revenue with twelve outliers removed.

A.4 Carbon intensity to Market Cap

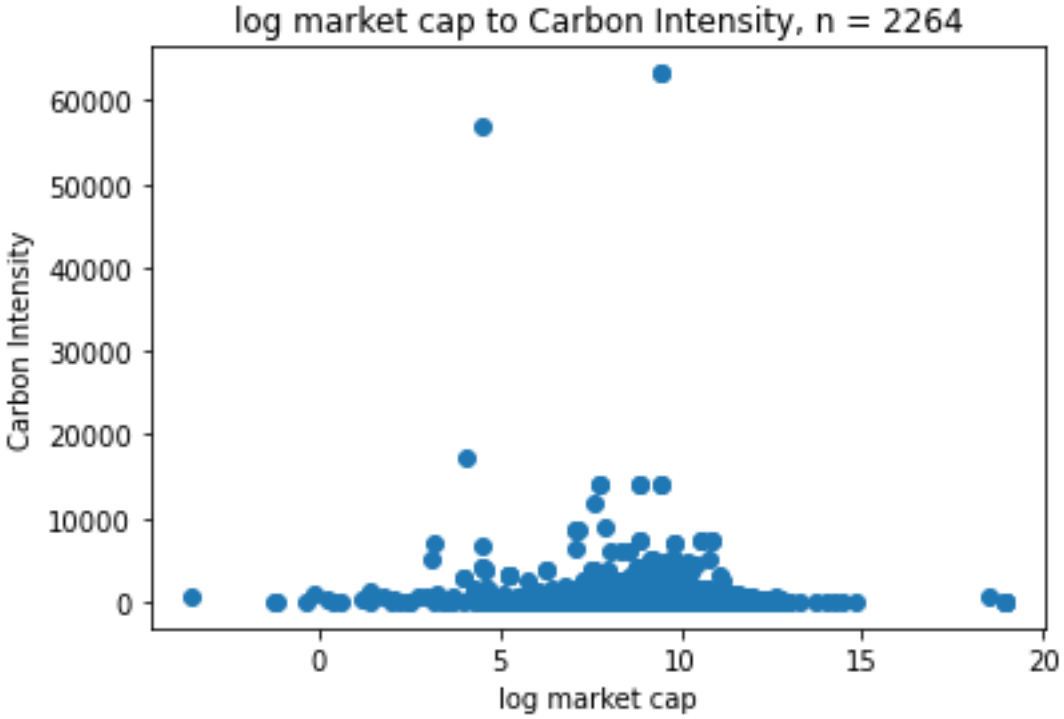

Figure A.4.1: Carbon intensity to market cap with no outliers removed.

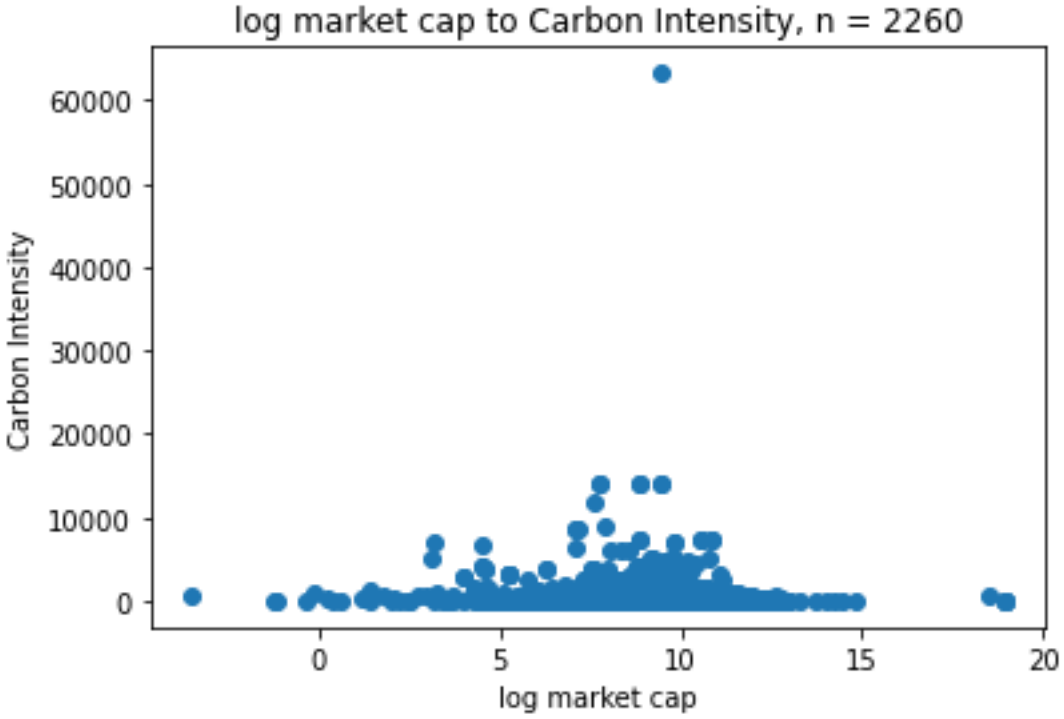

Figure A.4.2: Carbon intensity to market cap with four outliers removed.

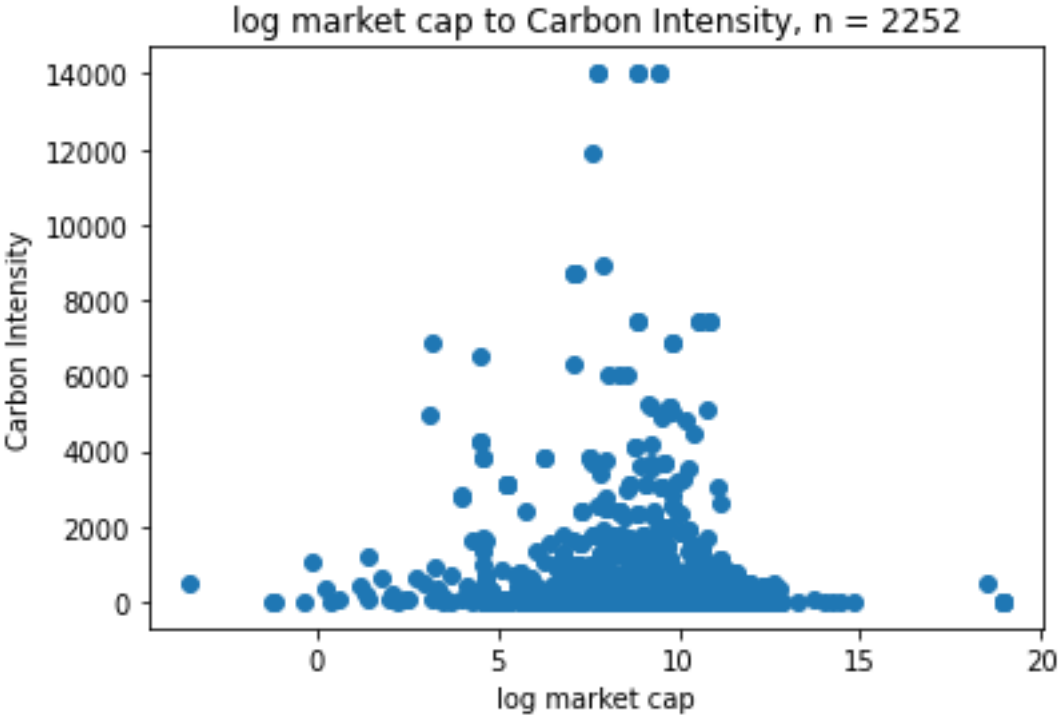

Figure A.4.3: Carbon intensity to market cap with twelve outliers removed.

A.5 Carbon intensity to Employees over revenue

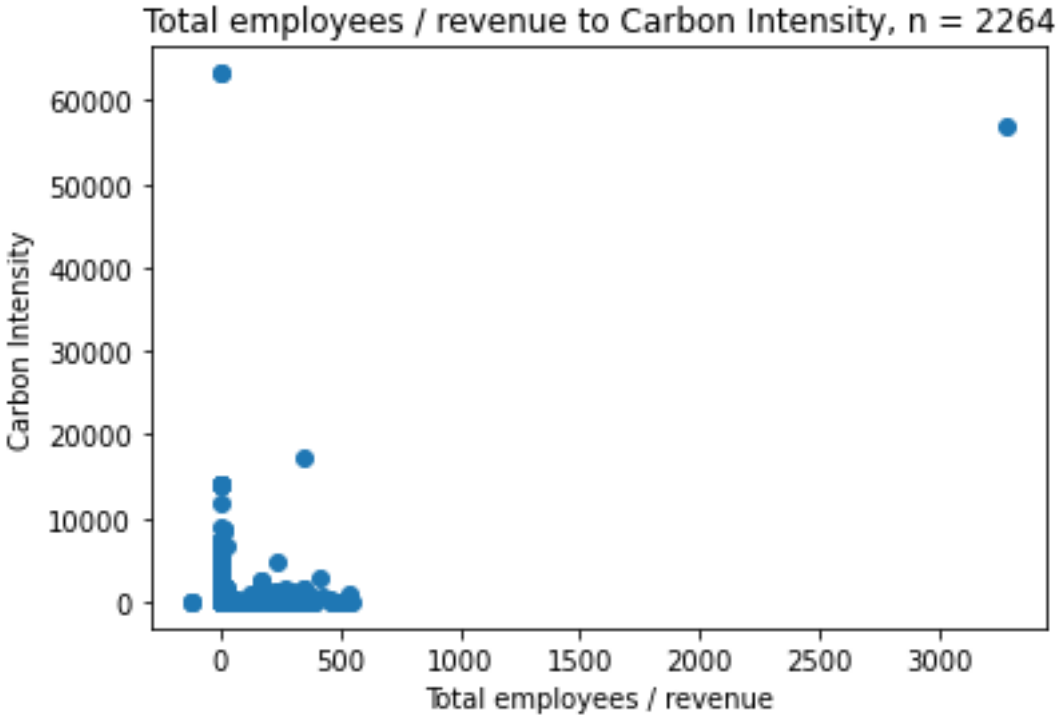

Figure A.5.1: Carbon intensity to employees over revenue with no outliers removed.

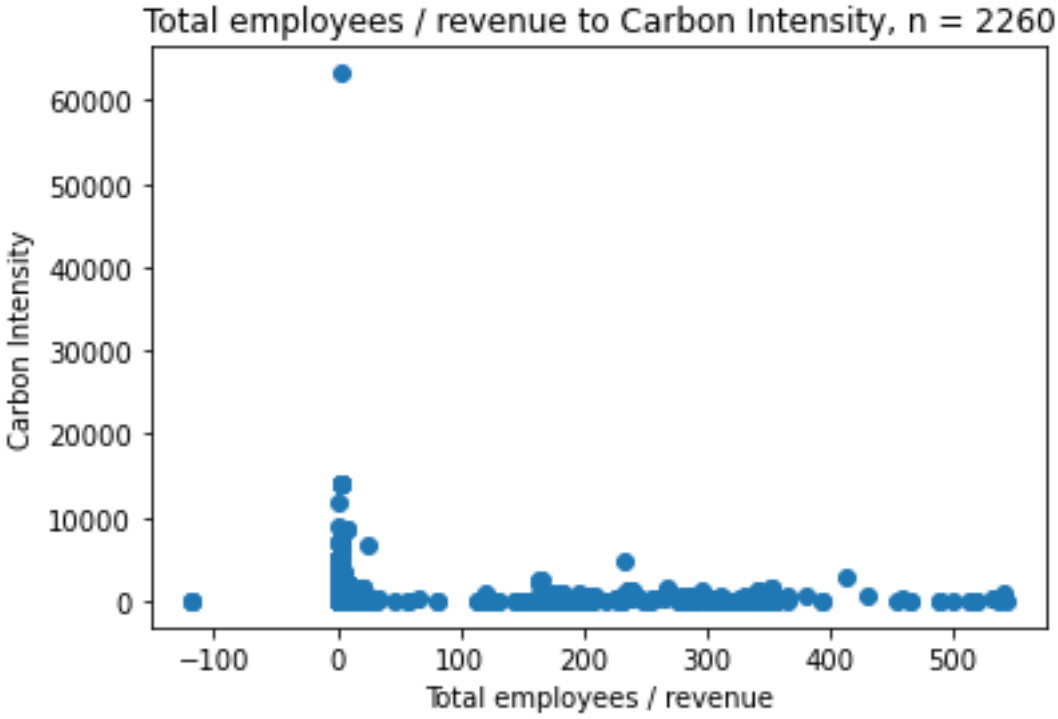

Figure A.5.2: Carbon intensity to employees over revenue with four outliers removed.

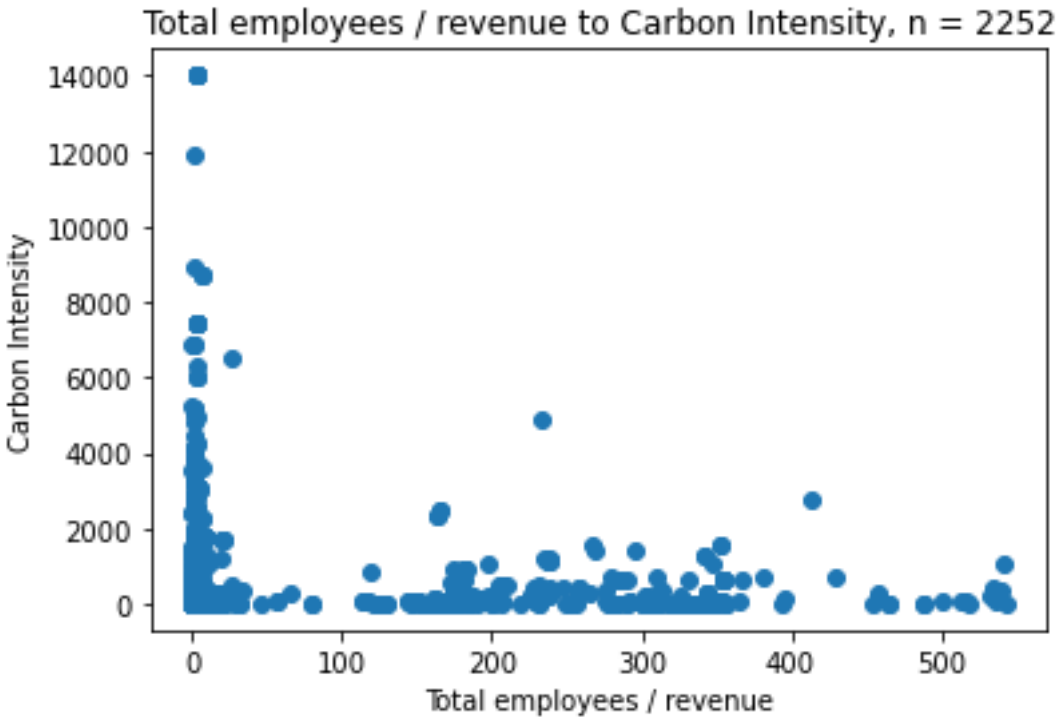

Figure A.5.3: Carbon intensity to employees over revenue with twelve outliers removed.

TRITA TRITA-SCI-GRU 2021:206