Imperial College
London

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

Data Mining in the Mining Industry: A Machine Learning Approach to Predicting Scope 3 Emissions

FINAL REPORT

Author:
Oliver Rogers

Supervisor:
Dr. Ovidiu Șerban

Second Marker:
Dr. Konstantinos Gkoutzis

In partnership with
REFINITIV

Submitted in partial fulfillment of the requirements for the MSc degree in Computing Science of Imperial College London

September 2021

Abstract

To meet the goals of the 2016 Paris agreement companies will need to significantly reduce their carbon emissions. Tracking corporate carbon emissions data will be a key part of measuring and managing this process. However, progress is being held back by incomplete corporate carbon emissions disclosures, particularly for scope 3 emissions. This project focuses on the Mineral Resources sector and shows how machine learning can be used to predict scope 3 emissions using data extracted from public company disclosures. The primary purpose of the model developed by this project is to help fill the existing data gap by estimating the emissions of non-disclosing firms. Additionally, the project also explores novel uses for the model, including year ahead forecasting and anomaly detection, the latter of which is particularly relevant given the inconsistent and often incomplete nature of scope 3 reporting. The project demonstrates the value of taking a sector-specific approach to emissions modelling and the benefits of careful feature selection and engineering. In constructing the model, this project implemented and evaluated a number of different machine learning algorithms and finds that the XGBoost algorithm displays the best performance. The project's best model demonstrates a 23% improvement in MAE compared to the best model published in the literature, this represents an advance in the state-of-the-art of scope 3 emissions modelling.

Acknowledgments

I would like to thank my supervisor Dr. Ovidiu Șerban whose guidance over the evolution of this project has been incredibly valuable. He has been a consistent source of insight and encouragement in our weekly meetings over the last four months. I am very grateful for all his efforts in making this project and the partnership with Refinitiv possible.

I would also like to thank the Refinitiv ESG team for their engagement with the project throughout our ten meetings. The chance to regularly present my findings and get feedback from domain experts has been a great collaborative experience.

Additional thanks go to my second marker Dr. Konstantinos Gkoutzis who provided helpful feedback and advice at the start of the project.

Finally, I would like to thank my family for their continued support and enthusiasm.

Acronyms

CCF Corporate Carbon Footprint.

CDP Carbon Disclosure Project.

CNN Convolutional Neural Networks.

DNN Deep Neural Networks.

EEIO Environmentally Extended Input Output Model.

EIO-LCA Economic Input Output Life Cycle Analysis.

ESG Environmental, Social and Governance.

GHG Greenhouse Gas.

KNN K-Nearest Neighbours.

MAE Mean Absolute Error.

RMSE Root Mean Squared Error.

SDG Stochastic Gradient Descent.

SVM Support Vector Machine.

SVR Support Vector Regression.

TRBC The Refinitiv Business Classification.

WOID World Input Output Database.

WRI World Resources Institute.

Contents

Chapter 1

Introduction

The 2016 Paris Agreement saw 196 state parties sign a legally binding treaty that commits to keeping long term global warming below 2 degrees above pre-industrial levels and sets the additional target of restricting overall warming to 1.5 degrees [1]. This breakthrough multilateral agreement recognises the potentially devastating social, economic and environmental impacts of unmitigated climate change [2]. To keep warming below 2 degrees and to have any chance of meeting the more ambitious target of limiting warming to 1.5 degrees, significant cuts in greenhouse gas emissions are required [3]. As the world's largest emitters, corporations will have to play a central role in these reductions [4]. For markets to efficiently price climate risk and for policymakers to track progress, it is vital that accurate corporate emissions data is made publicly available [5]. Since the founding of the Carbon Disclosure Project (CDP) in 2000 steady progress has been made in Scope 1 and 2 corporate emissions reporting with over half of the world's market capitalisation disclosing some emissions data [6]. However, reporting of scope 3 emissions remains much more limited with only 18% of the MSCI ACWI IMI, a global index with 99% coverage of public equities, reporting scope 3 emissions in March 2020 [7]. The sparsity of scope 3 data is at odds with its overall importance as estimates suggest that it accounts for up to 70% of corporate emissions [8]. This project seeks to develop a machine learning model that is able to predict scope 3 emissions for companies in the Mineral Resources sector using data from public company disclosures. The primary purpose of building the model is to help fill the existing data gap created by companies that do not disclose scope 3 emissions while also exploring auxiliary uses of the model in anomaly detection and year ahead forecasting.

1.1 Justification of Sector Choice

This project focuses on building a sector-specific model rather than a universal scope 3 model. The reasoning for this approach is that while previous attempts to build a universal model have worked well for Scope 1 and 2 emissions [9, 10], they have been less successful when applied to scope 3 modelling [10]. Given the highly varied supply chain contexts of different industries, it is hypothesised that building a more focused sector-specific model might improve model performance. A key mechanism for enhancing performance is predicted to be the ability to include sector-specific features that are particularly relevant to a given sector's scope 3 emissions. For example, in the Mineral Resources sector, this might be production volumes of different materials, while in the automotive industry, it would be features such as the numbers of vehicles produced.

The specific sector this project has chosen to focus on is the Mineral Resources sector as classified by the The Refinitiv Business Classification (TRBC) [11]. The sector includes the following industries; Non-Gold Precious Metals & Minerals, Iron Steel, Aluminium, Specialty Mining Metals, Gold, Mining Support Services & Equipment, Diversified Mining, Construction Materials. Mining Support Services & Equipment companies as well as materials wholesalers are excluded from the analysis as their business model is fundamentally different from all other companies in the Mineral Resources sector that focus on extracting and processing raw materials. The sector is estimated to directly contribute around 4-7 % of global greenhouse gas emissions and its scope 3 emissions have been linked to 28% of global emissions [12]. The significant carbon footprint of the sector makes attempts to improve emissions data coverage particularly valuable. The sector is well suited to the project as it is known to have strong emissions disclosure [13] with many in the sector openly reporting emissions data in line with the Greenhouse Gas (GHG) Protocol standard [14]. In a number of cases, this disclosure extends to scope 3 estimates. Some companies, including Vale, the worlds largest Iron Ore producer, have even set scope 3 targets [15]. The availability of reported data is crucial for building a robust machine learning model and is a key reason why the Mineral Resources sector was chosen to be the focus of this project.

1.2 Emission Scopes Overview

This project will follow the World Resources Institute (WRI) GHG Protocol on corporate emissions [14]. The definitions laid out by this standard for scope 1, 2 and 3 emissions are outlined in Table 1.1. While there still exists a heterogeneity of terminology [16] used to describe and report emissions data, the GHG Protocol is the most widely used and has been adopted by the CDP. The term Corporate Carbon Footprint is often used to describe a firm's emissions, however, it should be noted that the GHG Protocol encompasses all greenhouse gases and not just carbon emissions. These include methane (CH4), nitrous oxide (N2O), hydroflurocarbons (HFCs), perfluorocarbons (PFCs) and sulphur hexafluoride (SF6). While there are important differences in the strength and duration of the warming effect associated with each of these gases, they are converted and combined into a single CO2 equivalent (CO2e) amount. The amalgamation of all GHG gases into a single CO2e figure is useful as it makes it possible to compare the emissions of different companies directly. All references to CO2 in the project can be taken as references to CO2e.

GHG Protocol Scope Definitions		
Scope	**Description**	**Sector Specific Example**
Scope 1	The direct emissions from onsite company operations	Onsite mineral extraction, separation and processing
Scope 2	The emissions associated with the generation of consumed electricity	Purchased electricity and cooling
Scope 3	The emissions associated with indirect upstream and downstream activities in the supply chain	Purchased goods and services (excluding energy), the distribution of mined materials, the use of mined products

Table 1.1: Overview of GHG Protocol scope definitions with mining sector examples. Definitions taken from [14].

The scope 3 category encompasses all of a company's indirect emissions apart from purchased energy which is accounted for in Scope 2. Scope 3 emissions can be further subdivided into 15 categories, eight of which capture upstream supply chain emissions and seven of which capture downstream emissions [17]:

- Upstream - Purchased goods and services
- Upstream - Capital goods
- Upstream - Fuel and energy related activities (excluding scope 2)
- Upstream - Upstream transportation and distribution
- Upstream - Waste generated in operations
- Upstream - Business travel
- Upstream - Employee commuting
- Upstream - Upstream leased assets
- Downstream - Downstream transportation and distribution
- Downstream - Processing of sold products
- Downstream - Use of sold products
- Downstream - End of life treatment of sold products
- Downstream - Downstream leased assets
- Downstream - Franchises
- Downstream - Investments

The relative importance of each category is highly dependent on a company's business model. For example, the position of Materials companies at the top of many value chains means that the emissions captured by downstream scope 3 categories are often more significant than those of upstream categories. This is illustrated by Figure 1.1 which outlines the relative size of each emissions category for five of the world's largest public mining companies. The figure illustrates two important points. Firstly companies can decide not to disclose specific scope 3 categories if they are not relevant to their business model; the grey categories demonstrate examples of these. Secondly, there is some room for interpretation as to how emissions are divided up between the categories. The variable size of 'Processing of Sold Goods' (Category 10) compared to 'Use of Sold Products' (Category 11) is an example of this. As this project predicts an overall scope 3 figure, the split between individual categories is less relevant. However, the 15 categories do provide a helpful framework for thinking about model features; this is discussed in depth in Chapter 3.

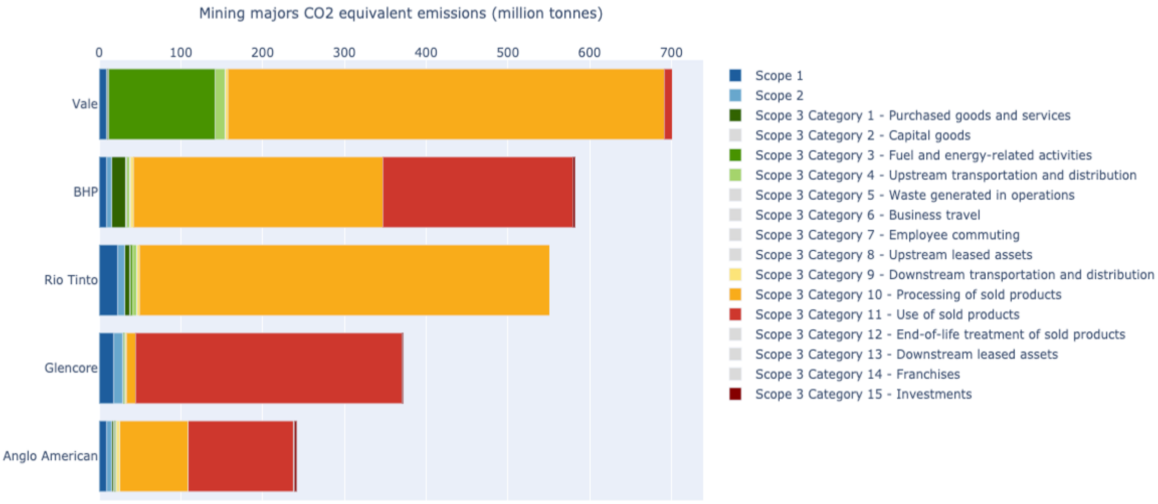


Figure 1.1: Emissions breakdown of five of the world's largest publicly traded mining companies. Data from company filings.

1.3 Motivation

The motivation and research merit justification for creating a model to predict scope 3 emissions is twofold. There is both an economic and policy need for a tool to estimate the scope 3 emissions of companies that do not currently report this metric. The focus on scope 3 is also significant as it recognises that for the majority of companies scope 1 and 2 emissions only account for a small fraction of the overall carbon footprint of a corporation [8]. Indeed, analysing only scope 1 and 2 emissions can result in misleadingly small carbon footprints for corporations [16], and this justifies the focus on scope 3. A wider recognition in the literature of the importance of scope 3 emissions can also be seen in Figure 2.1. This project looks to use machine learning to improve scope 3 emissions data availability and accuracy in the Mineral Resources sector.

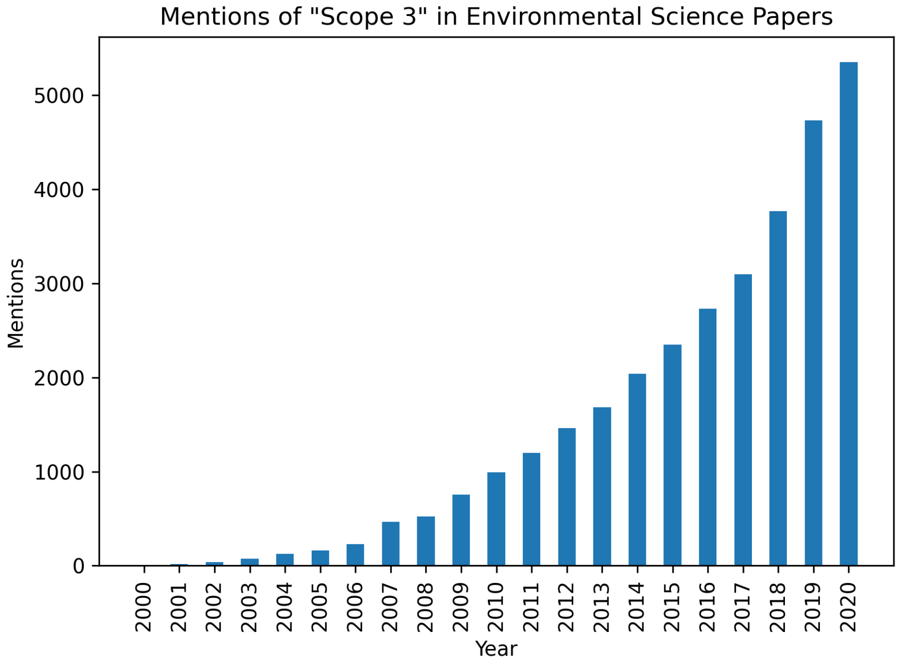


Figure 1.2: Growth in references to 'Scope 3' emissions in environmental science papers between 2000 and 2020. Data from Web of Science.

1.3.1 The Policy Imperative

Improved data coverage makes it easier to implement climate policy and to hold to account companies that are not meeting their climate commitments [5]. Article 4 of the Paris Agreement commits countries to ratchet up their carbon emissions reductions every five years, and emissions data from all three scopes is vital for policymakers setting new emissions targets [1]. These emissions reductions targets will be guided by the reduction pathways outlined in the IPCC special report on 1.5 degrees which suggest current emissions have to be halved by around 2030 with the world reaching net zero by 2050 [3]. In simple terms, to meet this target, the global economy has to go from the current 40 billion tons of carbon released annually to net-zero in the next 30 years [3]. Having reliable data will be central to monitoring and enforcing the transition. This project seeks to demonstrate how a machine learning emissions model can tackle some of the existing data gaps.

A secondary benefit of creating a scope 3 emissions model to improve data coverage is that it can encourage companies to reduce supply chain emissions. In some instances, it is easier for a corporation to cut supply chain emissions than scope 1 or 2 emissions [18]. While different industries and corporations will have different levels of agency over their scope 3 emissions, the size of the mining industry means that some concerted action to tackle indirect emissions could have a significant impact. Getting data on scope 3 emissions could be the catalyst for this process. Additionally, focusing on scope 3 emissions in addition to scope 1 and 2 emissions lowers the incentives for companies to simply outsource their most polluting activities.

1.3.2 The Economic Imperative

> ”Financial regulators, financial institutions, and investors need to have access to the best information and data to measure climate-related financial risks”
>
> (Janet Yellen, US Secretary of the Treasury — May, 2021)

Properly functioning financial markets require reliable emissions data [5]. Without such data, the market cannot accurately assess the opportunities and risks associated with climate change. Such information is crucial for investment, lending and insurance activities [19]. Indeed the continued rapid growth of assets allocated to ESG funds, which is predicted to account for a third of all assets under management by 2025, is placing even more focus on emissions disclosures [20]. While there has been a general push towards increased climate-related financial disclosures with the UK mandating reporting by 2025 and both the EU and US preparing similar legislation [21], the current reality is that emissions disclosure, particularly of scope 3 emissions is highly incomplete. Yet to truly understand a company’s climate risk, you need to consider its whole supply chain, and this is where scope 3 data is vital. This demonstrates the high practical value of the model this project is building.

At the microeconomic level having a grip on scope 3 emissions is important for managing corporate risk. The rising cost and prevalence of carbon prices is one highly visible cost of scope 3 emissions whereby suppliers pass on some of the burden of emissions taxes in the form of higher prices. However, there are also less visible costs associated with supply chain emissions which include regulation costs, financing costs and shifting consumer preferences [22]. Even if a corporation’s direct emissions are low, if the value chain they operate in has very high emissions, then the overall industry is likely to be squeezed by tightening emissions regulations [22]. Furthermore, they face the threat of shifting consumer demand channelling spending towards lower carbon products [22]. This makes it prudent for companies to track and engage with scope 3 emissions. Indeed improving scope 3 data can help companies focus on forcing climate positive supply chain innovation while also improving product use efficiency and end of life processing.

In summary, this section has outlined the tangible policy and economic benefits of improving scope 3 data availability which can help in the transition to a low carbon economy.

1.4 Objectives

Project Objectives

- Select and engineer features that can be used for predicting scope 3 emissions in the Mineral Resources sector
- Implement several machine learning models and identify the best algorithm for prediction scope 3 emissions
- Evaluate model performance and compare the project's sector-specific machine learning approach to other existing models
- Investigate the model's ability to be used for the auxiliary purposes of anomaly detection and year ahead forecasting

Additionally, the project is run as part of an academic partnership with Refinitiv who have provided the data for the project. While not an explicit research objective it is hoped that the findings of the project will be useful to the Refinitiv Environmental, Social and Governance (ESG) team.

1.5 Contributions

- An overview of the importance and research merit of the project, outlined in Section 1.3
- An overview of the machine learning landscape for structured data problems, outlined in Section 2.1
- A review of existing academic and commercial approaches to corporate carbon emissions prediction, outlined in Section 2.2
- The construction of a script to pull and process features from the Refinitiv API
- A curated and engineered a data set of features for predicting scope 3 emissions, outlined in Chapter 3
- A model development pipeline that allows for the comparison of different algorithms and facilitates the testing of hypothesises, outlined in Chapters 4 and 5
- A comparison of model performance against existing models and a demonstration of a reduction in error relative to the state-of-the-art, outline and discussed in Chapter 6

- A demonstration of the performance advantage of using gold standard sector-specific data for modelling scope 3 emissions, outlined and discussed in Chapter 6
- An assessment of the novel auxiliary uses of the model in anomaly detection and year ahead forecasting in Chapter 6

Chapter 2

Background and Related Work

2.1 Machine Learning Background

This section will provide and overview of the Machine Learning algorithms that are relevant to this project. Machine learning is a subset of the field of Artificial Intelligence and can be defined as the process by which a model achieves a higher performance on a given task after undergoing a period of training [23]. The field can trace its routes back to the 1950s and has seen significant developments in the last two decades driven by improvements in computing power, data availability and learning algorithms. This overview will focus on supervised machine learning for multiple regression in which a model draws upon structured labelled training data to learn a function that outputs a continuous real-valued number.

2.1.1 Linear Regression

Linear regression is the foundational supervised machine learning algorithm. In its simplest form, univariate linear regression is a form of statistical learning that involves fitting a line of best fit through the training data points. The model is able to take as its input a single continuous real value input x and outputs a single continuous real value output y. The equation for univariate linear regression can be expressed in the following form [23]:

$$y = mx + c$$

Where m represents the gradient of the line, and c represents the y-axis intersect. Both m and c are real-valued coefficients that will be learned as the model is trained on a data set. The model is trained through the process of minimising a loss function; typically, this loss function is the squared difference of the distance between a predicted value $\hat{y}$ and its real value summed over the whole training set. The equation for the squared loss function can be expressed as [23]:

$$Loss = (\hat{y} - y)^2$$

The process by which optimal values for m and c are actually found is known as gradient descent. The technique requires small steps in the negative direction of the gradient of the loss function to be taken until the minimum is reached. The learning rate hyperparameter α controls the step size at each iteration of the process. The pseudo-code for the process is as follows [23]:

Algorithm 1 Gradient descent algorithm.

```
while local minimum is not yet reached do
        $m \leftarrow m$ - $\alpha \frac{\partial}{\partial m} loss(m)$
end
```

The gradient descent algorithm comes in two main flavours, batch gradient descent and stochastic gradient descent (SGD). In batch gradient descent, the model is trained on the whole data set before the update step is performed, whereas for stochastic gradient descent, the model only uses individual data points in each update step. While is likely to be quicker, it does not come with the same guarantee of reaching the local minimum as you get with batch gradient descent [23]. Univariate linear regression can be easily extended to become multiple linear regression [24], a technique that is well suited to the numerous inputs used by this project to predict carbon emissions. The single input becomes a vector of inputs, and similarly, the weight term, m, becomes a vector of weights.

Elastic Net Regularisation

One additional consideration with multiple regression is the potential need for regularisation of high dimensional input data to avoid overfitting. In this project, elastic net regularisation is implemented, which finds a balance between the Lasso and Ridge regularisation methods [25]. Regularisation penalises complex functions and reduces the complexity of the model, which diminishes its capacity to overfit. One advantage of such regularisation, particularly in the context of this project, is that it often produces a sparse model, which makes it easier to interpret and identify the most important features driving scope 3 emissions.

2.1.2 Nearest Neighbours Regression

Nearest neighbours algorithms are nonlinear and non-parametric models in which no assumption are made about the distribution of the underlying data [26]. The algorithm stores all of the training data points and then inspects them at run time to determine an output. This reactive modelling is why the group of algorithms are sometimes referred to as ’lazy learners’ or, more formally, instance-based learners. A popular form of this algorithm is called K-Nearest Neighbours (KNN) [23]. In the regression form of this algorithm, the value of the k closest neighbours are averaged to produce a regression output. There are several different metrics for measuring distance, the most popular of which include Euclidean distance, Manhattan distance and Minkowski distance [23]. Whatever distance metric is used, a key pitfall with KNN is the curse of dimensionality, which can occur when a significant number of attributes are associated with the training data [23]. The sparseness of high dimensional feature space that makes it difficult to accurately determine the most relevant neighbours. This problem is particularly acute when, as is the case in this project, a small training data set is used. Applying this algorithm to emissions prediction would likely require the number of attributes associated with each corporation to be limited to achieve the best results.

An additional issue with KNN regression is that in its simplest form, it produces a discontinuous output space. To counter this problem, a distance weighting function can be applied to each of the k neighbours. This is, in practice, often a kernel function that produces the intuitive result that neighbours closest to the input get the highest weighting. The value of k is an important hyperparameter; too low a value for k can result in overfitting, and equally too high a value will lead to underfitting [23]. What KNN might be able to do very well, which would be interesting in the context of the project, is to roughly group mining companies by the main type of raw material they extract. If it is the case that the emissions profile of a company is determined by the main mineral they extract then KNN might be able to disaggregate the industry into collections of mining companies with similar extraction profiles.

2.1.3 Support Vector Regression

Support Vector Regression (SVR) [27] is a variation on the popular Support Vector Machine (SVM) model [28] for supervised machine learning. Similarly to SVMs, the SVR algorithm has the idea of a maximum margin separator at the core of the model. The maximum margin separator represents a linear hyperplane that tries to simultaneously minimise the squared regression error and maximise the margin between support vectors [27]. Only data points outside the margin are counted towards the overall error. Support vectors represent the data points that define the position of the margin down the middle of which exists the separating hyperplane. The fitting of the margin is crucial to the model’s ability to generalise to unseen data.

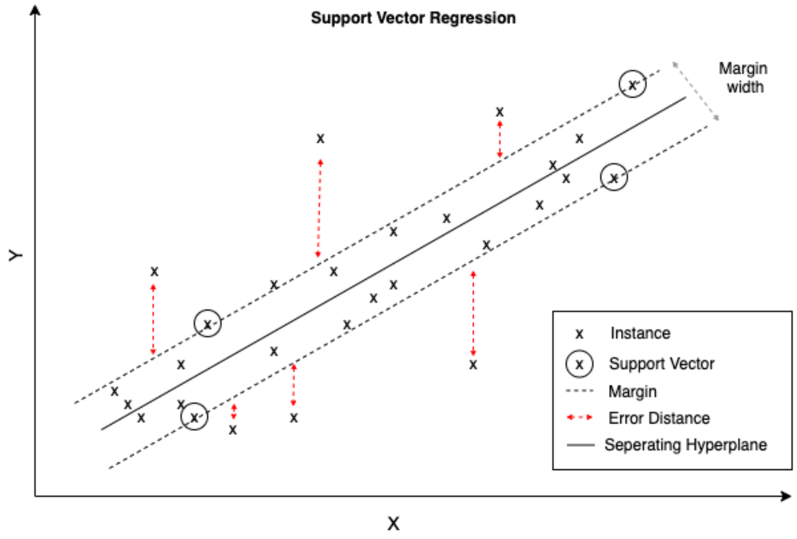


Figure 2.1: Support vector regression diagram. Adapted from [29]

Support Vector Regression represents a type of nonparametric model, but a key part of its attractiveness as a method stems from the fact that it is able to combine beneficial characteristics of parametric and nonparametric approaches [30]. The method is able to effectively handle overfitting like other nonparametric approaches, but it is also able to represent complex functions like parametric methods through the use of the kernel trick [31]. The kernel trick transforms the training data into a higher dimensional feature space so that it is possible to perform linear separation.

2.1.4 Tree Based Algorithms

Tree based algorithms [32] take as input a vector of attributes and output a single value. The tree algorithm aims to sequentially split the feature space to create maximum homogeneity in each partition of the resulting decision tree structure. Decision trees are made up of a branching hierarchy of nodes with leaf nodes representing the end of each branch as shown in Figure 2.2. Each node is subdivided into child nodes according to a splitting criterion that weighs each potential split's contribution to the loss function. When a tree is fully formed, a prediction is achieved by passing an input value down the branch that matches its attributes until a leaf node is reached and a value can be outputted. The decision tree represented in Figure 2.2 splits upon discrete categorical values, but a tree can also handle continuous attributes by choosing a split point within the range of the attribute's values. The type of decision tree required for this project would be a regression decision tree that is able to output a single continuous real-valued output as required to predict emissions. In regression trees, a linear function is fitted to each leaf node the produce a real-valued continuous output. The splitting criteria, usually Gini impurity or information gain, is selected in a greedy manner to attempt to minimise the depth

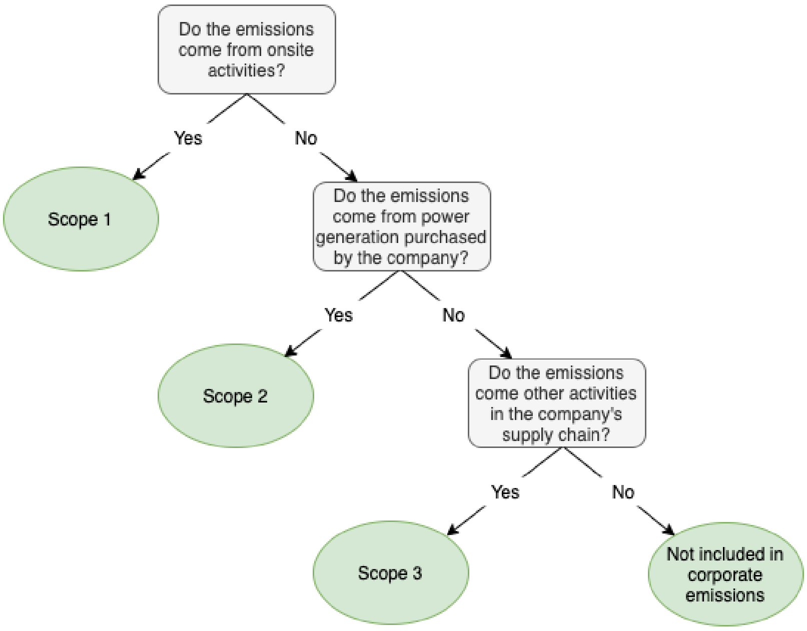

Do the emissions come from onsite activities?
Yes
No
Scope 1
Do the emissions come from power generation purchased by the company?
Yes
No
Scope 2
Do the emissions come other activities in the company's supply chain?
Yes
No
Scope 3
Not included in corporate emissions

Figure 2.2: An example decision tree for classifying corporate emissions.

of the tree and to make it easier to interpret.

Like other learning algorithms, decision trees suffer from overfitting. A technique that is regularly used to try to prevent overfitting is pruning. Pruning involves removing splits from the tree that do not produce a statistically significant reduction in entropy [33]. This often leads to improved model generalisation and because it makes the tree smaller and makes it more interpretable [23]. One key drawback with tree algorithms, particularly with a small data set, is that they can be unstable, meaning that a small change in the input data can significantly change the structure of the tree. While this is a significant drawback for individual trees, it can largely be negated by using ensemble methods such as gradient boosted trees.

Tree Based Ensemble Methods

Ensembles methods such as gradient boosted trees have a strong track record of providing a simple mechanism by which machine learning prediction can be improved [34, 35]. Ensemble methods are effective as by averaging the results of several models, they complement each other by averaging out their individual errors [35]. Boosting is an ensemble technique that sequentially combines decision trees. The method takes advantage of the boosting meta-algorithm to reduce bias by using subsets of the input data to create a number of weak performing predictors that are then combined together to give a final vote. Boosting builds on top of individually weak models by sequentially adding new shallow trees one at a time. Boosting is carried out sequentially and focuses on areas where previous models have shown the greatest prediction errors. A key advantage of Boosting is that it is able to deal well with missing values; this is crucial to this project as there is uneven coverage of many of the ESG metrics in the Refinitiv database. There are a number of different flavours of boosting algorithm, two of the most popular of which are Adaboost [34] and XGBoost [35]. XGBoost in particular is known to demonstrate very strong performance on structured data, as evidenced by the number of Kaggle competitions won using this algorithm.

2.1.5 Deep Learning

Deep learning is a subset of machine learning that builds upon the neural network architecture; the relationship between these overlapping fields of AI is shown in Figure 2.3. Deep learning allows a computer to learn a hierarchy of concepts at different degrees of abstraction, which gives the technique the capacity to model complex relationships between input variables [30, 36]. The term 'deep' refers to the fact that multiple layers of computation are used in the architecture of the algorithm [30]. This layered composition of simple concepts coalescing to represent complex interactions is central to the power of deep learning. Deep learning algorithms have shown impressive capabilities across a range of domains, including speech recognition [37] and image classification [38]. A key attraction of deep learning is that it removes much of the feature engineering required by other machine learning techniques [36]. However, many of the impressive performances of deep learning have come from its application to unstructured data rather than the structured data this project plans to use [36]. It is unclear if deep learning will be advantageous in the regression task of predicting emissions upon which this project focuses. Indeed a considerable difficulty associated with deep learning is its capacity to overfit [30], this is particularly problematic given the small data set size used in this project.

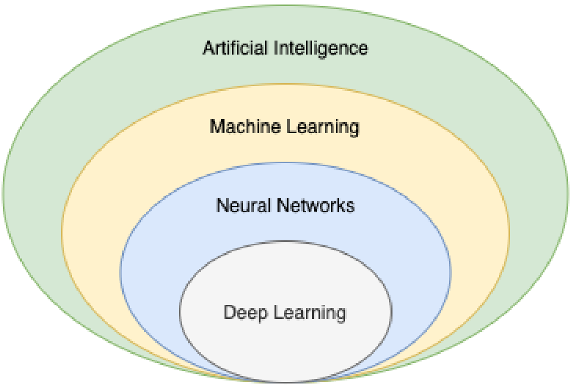


Figure 2.3: Relationship between artificial intelligence, machine learning and deep learning. Adapted from [30].

Two deep learning architectures are potential candidates for this project; Feedforward Neural Networks (also referred to as DNNs) and Convolutional Neural Networks (CNN)s. Additionally, recurrent neural network architectures such as long short-term memory (LSTM) models were initially considered. This project does contain some sequential data in the form of year on year scope 3 emissions data from a single company. However, because of the limited number and length of these series the project does not explore recurrent networks.

Feedforward Neural Networks

Neural networks are composed of layers of linked nodes and are loosely based on the human brain [30]. Each link has an associated weight that governs the sign and the strength of the association between neurons [23]. As depicted in Figure 2.4, neurons pass the sum of their inputs through a predetermined activation function to generate a single output. It is crucial that the chosen activation function is non-linear so that the network can represent non-linear functions [30].

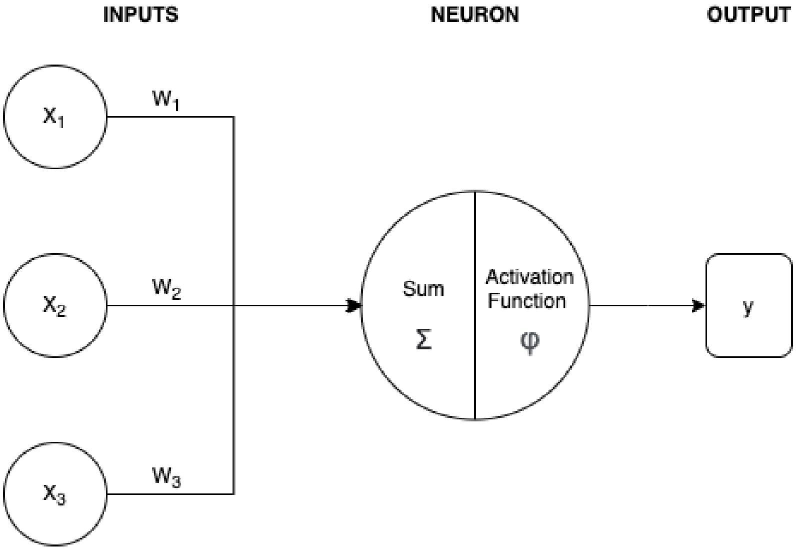


Figure 2.4: The anatomy of a neural network neuron. Adapted from [23].

In its simplest form, a network contains neurons that are connected together in a sequential manner where outputs from the previous layer of neurons serve as inputs to the next layer. Such a feedforward network must contain an input layer and an output layer and can contain any number of hidden layers in between. It has been shown that with a sufficiently large layer of neurons, a feedforward network with one hidden layer can model any continuous function [39]. If a network has more than one hidden layer, then it becomes a deep learning algorithm. The exact structure of a neural network architecture, how many neurons and how many hidden layers should be chosen, is an imprecise science [40]. Indeed it is common to try multiple different structures and use cross-validation to then pick the most promising architecture [23]. One guiding principle is that the size of the network should be proportional to the complexity of the problem and the amount of training data available [30]. The intuition behind this principle is that if large networks are used on simple problems with small quantities of training data, then the network will very likely overfit by storing the value of training examples in its structure rather than learning generalised rules. The relatively small training data set size for this project means that any neural network for predicting scope 3 emissions should be shallow and small in size [41]. One key advantage of using neural networks for this project is that it removes much of the work involved with feature engineering as this type of representational learning allows interactions between features to be engineered internally by the network [36].

Convolutional Neural Networks

CNNs take in input data in the form of multiple arrays and while they have most commonly been used for processing images [36], there have been examples of CNNs being used on more structured data [42]. CNNs are composed of layers that can be broken into two sections, a convolutional layer and a pooling layer. Convolutional layers map onto sections in the previous layer to convolve the input before it is then fed through to the pooling layer[30]. For structured data, a one-dimensional convolutional layer is used. The pooling is then used to reduce the dimensionality of the data by combining multiple data points in the convolutional layer into one single data point. This architecture is demonstrated in figure 2.5. A key reason why CNNs are thought to be so effective is that by limiting the free parameters in the network, the model is forced to perform feature selection which improves its ability to generalise to unseen data [30]. An additional advantage of using CNNs is that convolutional layers have a smaller number of parameters than the fully connected layers of a DNN which makes CNNs easier to train [30].

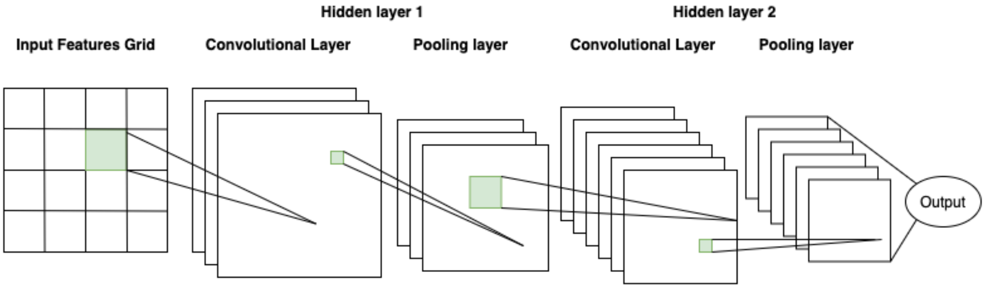


Figure 2.5: Convolutional neural network diagram. Adapted from [40]

2.2 Related Work

This related work section has been structured into three parts. The first subsection gives provides a short general overview and some background context on how machine learning is being used in tackle climate rated problems. The second subsection takes a broad look at the machine learning techniques that have been used to solve structurally similar problems in related fields. This subsection will review the techniques applied to structurally similar tasks in Healthcare, Environmental Science, Economics and Energy. Structurally similar means a multiple regression task using structured data. The rationale for starting with a broad interdisciplinary overview of applied machine learning regression is that it provides much-needed exposure to a broader range of machine learning techniques not visible within the relatively small body of corporate emissions prediction literature [10]. The final subsection examines the existing emissions prediction models that exist in the academic and commercial literature.

2.2.1 Machine Learning and Climate Change

The climate science and AI communities are increasingly recognising that machine learning can be a useful tool in addressing many of the multifaceted challenges that climate change presents [43]. A 2019 paper exploring how AI could be used to tackle climate change counted 2019 Turing Award winner Yoshua Bengio, Deepmind founder Demis Hassabis, Google Brain co-founder Andrew Ng, Institute for Computational Sustainability Director Carla Gomes, former Microsoft Research director Jennifer Chayes and United Nations IPCC Sixth Assessment Report lead author Felix Creutzig among its authors [43]. This momentum is being actively maintained by organisations such as Climate Change AI and Climate Informatics, which are working to promote the application of machine learning to climate change problems. The parameterisation of climate models [44] and the optimisation of carbon friend smart grids [45] are just some of the many ways in which machine learning is already being used to help tackle climate-related issues.

2.2.2 Multiple Regression Machine Learning

Support Vector Regression is an algorithm that appears throughout the reviewed literature and is often used as a baseline model against which more recent modelling techniques have been compared. In healthcare, a 2020 study used machine learning to predict blood pressure and used SVR in conjunction with algorithms such as XGBoost [46]. In Environmental science SVRs were used to predict water quality [47] and wind speeds [48], although again both papers used SVRs as well as other methods. In Economics SVR has been used for revenue prediction in multiple contexts including movie rentals [49], electricity prices [50] and the tourism industry [51]. The technique is most commonly seen in papers that trial a number of candidate algorithms and it is generally outperformed by other more modern methods [46, 48, 50].

The second type of algorithm that appears across the literature reviewed are tree based methods. In particular, gradient boosting ensemble methods such as the XGBoost algorithm, consistently perform well in regression prediction tasks. In agriculture, both random forest and XGBoost techniques have shown promise for predicting crop yields [52]. For predicting movie rentals, it was found that boosted trees and random forest algorithms outperformed SVR [49]. Similarly, a study predicting commercial building energy consumption found that tree based gradient boosting outperformed alternative linear and SVR methods [53]. These examples show that there is a strong precedent for tree based methods performing well at the types of predict that this project is focused on.

The final category of model that regularly appears in the literature are deep learning methods. In particular deep feedforward neural networks. DNNs have seen widespread use in healthcare [36], for example, DNNs were used in an epidemiological model for predicting Covid infection rates [54]. In environmental science, DNNs have performed well in predicting non-corporate emissions [55]. A 2020 overview paper showed how DNNs and cnns have both seen widespread use in applied economics papers [56]. However, deep learning approaches perform best when large quantities of data are available for training these more complex models [30]. The relatively small size of the dataset used in this project is likely to limit the power of potential deep learning approaches.

This section has provided an overview of current multiple regression supervised machine learning techniques in the literature. It has provided an outline of the current state-of-the-art for problems that are structurally similar to the one this project looks to address. By examining a range of fields, including Healthcare, Economics and Environmental Science, it is clear that the most popular techniques are tree ensemble and deep learning methods. This knowledge has been helpful in shaping the design of the project to make sure it is using the latest techniques and methodologies that build off successful precedents in the literature.

2.2.3 Existing Emission Models

Since James Hansen's 1988 congressional testimony popularised the concept of anthropogenic global warming, there have been efforts to estimate greenhouse gas emissions. These estimates have ranged from the level of individual machines [57] to global estimates[58]. Regression models have historically been a popular way to measure emissions and have been used in a number of studies [59, 60]. More recently the field has seen an increased prevalence of machine learning techniques being used to model emissions. For example, a 2016 paper used svr to predict emissions from the production process in the alcohol industry [59]. Additionally, a 2018 paper used svr to predict CO2 emissions for the region of Chongqing. While the most recent papers have demonstrated an increased use of neural networks and other deep learning techniques [61, 55]. Furthermore, it is not just the models but also the data that has become increasingly sophisticated; for example, the carbon TRACE system launched in 2020 feeds satellite imagery into machine learning models to measure real-time global emissions [58]. This introductory subsection has given a general overview of the trends in emissions modelling, and the next subsections will look at the specific topic of corporate emissions modelling.

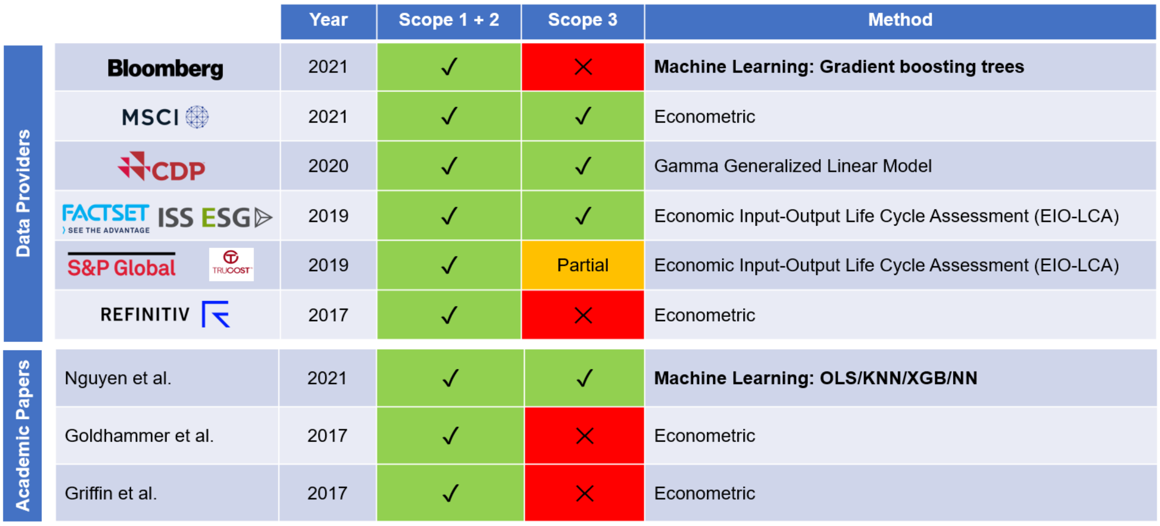
2.2.4 Academic Corporate Emissions Models

The academic literature contains three papers that directly address the topic of estimating corporate carbon emissions using external data. The first two papers were both published in 2017 and use regression analysis to predict emissions for a set of companies; Griffin et al., [9] focuses on the S&P 500 while Goldhammer et al., [62] focuses on European firms in the Chemicals, Construction and Engineering and Industrial machinery sectors. Both papers use a combination of financial and operational features that are made publicly available by company disclosures to predict scope 1 and 2 emissions. However, neither study look to address the challenge of building a model to predict scope 3 emissions. Additionally, neither paper use any machine learning techniques beyond linear regression.

The most recent paper published in 2021 by Nguyen et al., from here on referred to as the Otago paper, is particularly notable for two key reasons [10]. Firstly, it is the first published attempt to use a model to estimate scope 3 emissions from external data. Secondly, it is the first paper to use machine learning to estimate corporate emissions from external data. The results of the paper show that the best single algorithm for predicting emissions is the XGBoost algorithm. The paper used features that can be placed in one of five categories which are; scale of operations, business model, technology, energy information and business environment[10]. The scale of operations categories included six features; revenue, employees, total assets, net plant property and equipment, intangibles and leverage [10]. The business model categories contained two features; a categorical GICSector classification feature and a gross margin feature [10]. The energy categories includes three features; fuel intensity, energy produced and energy consumed [10]. Finally the business environment category contains two categorical features; a country level income group feature and a country level CO2 law feature [10]. These features helped inform the feature selection used in this project, however, one crucial difference is that this project has tried to include sector specific features that are not present in the Otago model as it is designed to be a sector universal model.

2.2.5 Commercial Corporate Emissions Models

There are six commercial corporate emissions estimation models. This large number of proprietary commercial models is good evidence of the value and existing demand for estimated emissions data. These are outline in Figure 2.6. However, the commercial nature of these models means that full documentation is often restricted to paying customers.



		Year	Scope 1 + 2	Scope 3	Method
Data Providers	Bloomberg	2021	✓	✗	**Machine Learning: Gradient boosting trees**
	MSCI	2021	✓	✓	Econometric
	CDP	2020	✓	✓	Gamma Generalized Linear Model
	FACTSET ISS ESG	2019	✓	✓	Economic Input-Output Life Cycle Assessment (EIO-LCA)
	S&P Global TRUCOST	2019	✓	Partial	Economic Input-Output Life Cycle Assessment (EIO-LCA)
	REFINITIV	2017	✓	✗	Econometric
Academic Papers	Nguyen et al.	2021	✓	✓	**Machine Learning: OLS/KNN/XGB/NN**
	Goldhammer et al.	2017	✓	✗	Econometric
	Griffin et al.	2017	✓	✗	Econometric

Figure 2.6: Overview of existing emissions prediction models.

Refinitiv model

One of the first commercial solutions was a set of three models released by Refinitiv in 2017. These models are examined in detail as they are used as baselines for this project. Refinitiv has three patented carbon data estimation models [63], the CO2 model, the Energy model and the Median model. The models estimate a 'Total Emissions' figure which represents a combined scope 1 and 2 total. Scope 3 emissions are not covered, although it is feasible that, while imperfect, the methodology could be extended to this related problem. Which model is applied depends on data availability. The first model, the CO2 model, is used if a corporation has previously disclosed emissions data [63]; given that the primary goal of this project is to predict the emissions of companies that have never previously disclosed emissions data, this model is less relevant. The second model, the Energy Model, calculates the ratio of Energy Consumed to Total Employees of the chosen company and then gives the company a percentile rank compared to the rest of the industry. The companies in the industry that report CO2 emissions are then used to create a distribution of CO2 Emissions to Total Employees, and the previously calculated percentile rank is used to place the chosen company at a specific point in this distribution and hence estimate its total CO2 emissions [63]. The same process is repeated with net sales instead of the number of employees, and the resulting CO2 emissions estimate is averaged with the previous estimate obtained using the number of employees. The third model, the Median model, calculates an industry median ratio for CO2 Emissions to Net Sales and for CO2 Emissions to Total Employees and these two ratios are then applied to the selected company to produce two CO2 estimates that are then averaged to give a single final estimate [63]. Refinitiv has not published accuracy scores for the models; however, the models have been recreated in Python to access their performance so that they can act as baselines for this project.

MSCI model

In 2016 MSCI published the methodology underpinning their carbon estimation models, which focused on scope 1 and 2 emissions[64]. This offering was augmented with a new scope 3 model as outlined in a 2021 MSCI report [65], however, the methodology behind the scope 3 model has not been published so this section instead gives an overview of the preexisting MSCI scope 1 + 2 emissions models. MSCI have three models for estimating scope 1 + 2 emissions, the first of which, the Production Model, is designed to be applied to power generating electric utilities. The model is feed with data on the power generation by fuel types and applies the relevant emission factor to estimate scope 1 emissions. This model is interesting as it shows a precedent in the literature for building models for specific sectors, which is the approach this project has persued. The second model, the company specific carbon intensity model, is similar to the Refinitiv CO2 model in that it can only be used on companies that have previously released CO2 emissions data. This data is used to calculate an emissions intensity ratio for the company (Revenue/CO2) that is used to make an estimate for future emissions given future revenues. Finally, for companies that have no disclosed historical data, the 'Industry specific intensity model' is used. The model calculates an emissions intensity for each of the 156 sub-industries of the Global Industry Classification (GICS) Standard. Corporate revenues combined with the relevant sub-industry emissions intensity are used to estimate corporate emissions. The MSCI research paper on the models report a median absolute error of 7% and that 60%of the predicted data points fell in the region of -33% to +50% of actual emissions. Overall, the model was found to overestimate predicted emissions.

Bloomberg model

The most recently released commercial emissions model was developed by Bloomberg. This model uses gradient boosted decision trees to estimate scope 1 and 2 emissions, but as with the Refinitiv models, it does not attempt to predict scope 3 emissions. The model is trained on Bloomberg ESG and financial data [66]. The performance of the model is reported using RMSE and is benchmarked against two baseline models. The first baseline model simply predicts mean industry emissions, while the second baseline model is a Gamma generalised linear model trained using industry-specific data that uses a greedy approach to feature selection, choosing features by their ability to reduce negative log-likelihood [66]. The Bloomberg model is the only existing commercial model to apply machine learning to predict corporate emissions.

Economic Input Output Life Cycle Analysis Models

Economic Input Output Life Cycle Analysis (EIO-LCA) Models are used by the data providers Trucost and ISS ESG in estimating carbon emissions. EIO-LCA model seeks to capture the full associated energy and material requirements of different economic activities. A key benefit of using an EIO-LCA approach is that the approach considers the whole supply chain of a particular good or service. Trucost, a data provider owned by S&P Global, use a modified EIO-LCA called an environmentally extended input-ouput model (EEI-O), which specifically focus on the environmental supply chain impact of a particular good or service [67]. However, Trucost's EEI-O models are only used to estimate upstream scope 3 emissions, while downstream emissions are estimated through bespoke consultancy projects. ISS ESG, a data provider, owned Deutsche Boerse, use EIO-LCA models to map all scope 3 emissions [68]. One drawback of such an approach, as noted by ISS ESG, is that the EIO-LCA databases upon which the model relies provide sector averages which means that while the model might perform well at an aggregate level, it is not as suited as a machine learning approach for picking up on differences in scope 3 between companies in a sector.

Carbon Disclosure Project (CDP) Model

The CDP model takes a slightly different approach to modelling scope 3 emissions compared to the other models [69]. It is the only model that looks to provide individual estimates for each of the fifteen categories of Scope 3 emissions. However, the individual category models are only used if it is not possible to come to a figure using a bottom-up approach drawing on internal company data. In this way the CDP model is a hybrid model that uses both internal and external data. The approach uses a Gamma Generalised Linear model with two input features revenue and activity data. The model uses data of other companies in the same sector to estimate a revenue scaled emissions intensity. This approach is acknowledged to work best in what the CDP terms 'homogeneous sectors' where the product and processes are relatively similar between companies [69].

GHG Protocol and Quantis model

The GHG Protocol Quantis model is not included in 2.6 as it is distinct from the other models in that it uses internal rather than external company data to estimate scope 3 emissions [70]. While the other models outlined in this subsection have been tools for company outsiders to estimate corporation emissions, this is a tool used by the corporate accounting function and requires detailed input to help them come up with a figure for scope 3 emissions. The model takes a bottom-up approach in which it requests user input data on employee numbers, facilities, waste, purchased goods and services, transport and distribution, business travel and sold product usage [70]. While this project aims to produce a general-purpose prediction model that is able to estimate scope 3 emissions without access to internal company data, the GHG Quantis tool is still instructive in that it suggests which corporate attributes might be the most important in determining scope 3 emissions.

2.3 Legal, Social, Ethical and Professional Considerations

The European Commission High-Level Expert Group on Artificial Intelligence released a report in 2019 that put forward seven requirements for Ethical AI [71]. These seven requirements were:

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Societal and environmental well-being
- Accountability

Fairness is a key consideration. Because the model is built using historical data, it will naturally try to extrapolate historical trends into the future. This could be unfair if a company makes a significant emissions reduction push in a given year which means the model and the historical data is has been trained on could significantly overestimate emissions. This can have a real-world impact; for example, the overestimated model prediction might prevent investment into a company if the prediction is used by a fund manager to influence their investment decision. Such limitations need to be studied in depth and then clearly publicised if the model is deployed in any form.

Transparency is another important consideration for this project. A key component in getting stakeholders to trust the model is being able to explain its results [71]. Some model techniques are more transparent than others. In normalised linear regression, the feature coefficients represent one interpretable way of explaining model results. Feature importance experiments are still possible with tree based methods; however the mechanisms become harder to understand [72]. Finally, explainability for deep learning remains challenging, which can create tension between stakeholders as such techniques often yield very strong results, but this has to be traded off against decreased explainability.

Technical Robustness is also relevant to this project. Decision tree, k-nearest neighbours and deep learning algorithms all have the potential to exhibit unstable results [72]. Small changes in the input data set can result in significant changes in the final prediction. This could be a problem if the model is deployed, as the small data set size means that predictions for a given company have the potential to change dramatically as the data set is expanded over time. This would likely be alarming to anyone using the model to inform their decisions. As scope 3 disclosures and the volume of data increases over time, this should help improve model stability.

Chapter 3

Data

3.1 Data Source

The data was obtained from the Refinitiv Workspace product as part of an academic partnership program with Imperial. Alongside their extensive financial data set, Refinitiv also offers a range of over 450 ESG related metrics covering over 10,000 companies (80% of global market capitalization) with data going back over 15 years [73]. Importantly this offering includes data on scope 3 carbon emissions. All the data provided by Refinitiv is in a structured as opposed to an unstructured format. Here structured data is defined as any data that can be stored in a spreadsheet, whereas unstructured data is defined as data, such as images, that cannot be stored in a spreadsheet. A script was built in Python to pull the data using the Refinitiv Data Platform API.

While other previous studies, such as the 2021 Nguyen paper [10], use data from a number of different providers, this project just uses data from Refinitiv. There are two reasons for this, first is that Refinitiv data is exhaustive, so other data sources are of limited value. Secondly, as the project is built with the Refinitiv ESG team in mind, just using Refinitiv data makes any final model more accessible and easier to maintain for this team.

3.2 Data Preprocessing

For the 2019 financial year, Refinitiv provides data on 71 number companies that published scope 3 emissions data in the Mineral resources sector. This represents 4% of public Minerals Sector companies with revenues of over $100m. The actual % of disclosing companies is likely to be slightly higher as Refinitiv only collects data on the firms which are included in its ESG coverage. It is also interesting to note that these 4% of disclosing firms have a combined market capitalization of over 25% of the Mineral Resources sector. This suggests that there is a bias towards larger companies in the scope 3 data set; this is not entirely surprising given that larger companies will likely be under more scrutiny and have more resources available to devote to providing more in-depth financial reporting. To augment the data set size, we also use historical scope 3 emissions figures from previous year's disclosures. This historical data goes back to 2007, the first year that any company in the sector published a scope 3 emissions total. The number of disclosing firms has risen steadily over time. Using this methodology, we get to a total of 627 scope 3 data points. The data set size is suitable for standard machine learning algorithms but is very small for a deep learning project. Additionally, a key issue with scope 3 disclosures is data quality, and some disclosures are likely to represent incomplete company reporting of their scope 3 emissions which will require further data cleaning [74].

Data cleaning			
Cleaning	**Scope 3 data points**	**Unique firms**	**Criteria**
Uncleaned	627	120	None
Level 1	464	112	Above 1,000 tonnes
Level 2	203	65	Above 0.2 of scope 1 + 2
Level 3	160	51	Above 0.3 of scope 1 + 2

Table 3.1: Number of distinct data points and firms at each level of data cleaning

The data provided by Refinitiv does not distinguish between partial and full scope 3 disclosures. As a result, some of the data points represent incomplete disclosures where the company has only released data on a subset of its relevant scope 3 categories. Because of its ease of calculation, partial disclosure tends to focus only on business travel. Identifying partial versus full disclosure is a difficult task, given simply an emissions figure. This project's approach to the problem has been to set up a number of levels of data cleaning where progressively more rigorous criteria are applied to the data. Level 1 cleaning removes any data points below 1,000 tonnes as a selective manual inspection of company disclosures reveal that emissions totals below this figure were all examples of partial disclosure. Level 2 cleaning removes any scope 3 data points that are less than a fifth of combined scope 1 + 2 emissions; this level was in line with guidance from domain experts in the Refinitiv ESG team. Level 3 cleaning removes any scope 3 data points that are less than 0.3 of combined scope 1 + 2 emissions. There is an inherent trade-off in the data cleaning process; as you increase the level of data cleaning you remove more instances of incomplete reporting but you also increase the chances of inadvertently removing legitimate full disclosure data points. For example OceanaGold has a full disclosure scope 3 emissions total of under 0.3 of scope 1+2 and would be removed by level 3 cleaning [75].

One more additional rule is applied at all levels; if a scope 3 data point for a single company increases by more than 100% from year to year without at least a 50% increase in revenue (allows for mergers) or a change in TRBC Activity classification (allows for change in business focus) then it is removed. The reasoning behind this is that such a large jump in scope 3 emissions without a correspondingly large jump in revenues or a change in the business model seems highly likely to be the result of a change in scope 3 emissions reporting. More specifically, it is likely to represent the movement of the company from partial to complete scope 3 reporting. One final additional benefit of such data cleaning is that if any data collection errors exist in the underlying data, for example where the reported values are being incorrectly fed into the Refinitiv database, then such cleaning mechanisms have the potential to remove some of these errors.

3.3 Feature Selection

In the process of selecting features, it is important to have an overriding theory to guide selection. The challenge of feature selection in the project has been guided by previous work [10, 62, 9] and a careful consideration has been given to the likely drivers of each of the 15 categories of scope 3 emissions outline in Section 1.2. Particular attention paid to the categories were singled out in Section 1.2 as being important in the Mineral Resources Industry. The features used in this project to predict scope 3 emissions were hand-selected from a choice of over 1000 financial and operating metrics and the 450 available ESG metrics. In total 82 features were extracted from Refinitiv. A full list of features can be found in the Appendix A. This represents the most comprehensive selection of features of any of the previously published emissions models and is the first to select features for the sole purpose of predicting scope 3 emissions.

Such as a large number of features were selected, knowing that a significant number would not actually be used in many of the modelling experiments. A correlation and availability analysis experiment in combination with a manual inspection of the features was conduction to extract the most important features. This is discussed in Subsection 3.2. This selection included many features that could be used as universal feature across different sectors. For example, features like revenue and number of employees are likely widely applicable. However, the advantage of the sector-specific approach that this study takes is that it is able also to include sector-specific features such as mineral production volumes. Indeed, certain features could clearly be used as a proxy for a specific scope 3 categories as shown by Figure 3.1.

The features can broadly fit into three categories that were all hypothesised to have an impact on scope 3 emissions. Such a grouping of features builds on similar work in previous corporate emissions prediction papers [10, 62, 9]. The first category is features are related to the size of the business. These include features such as

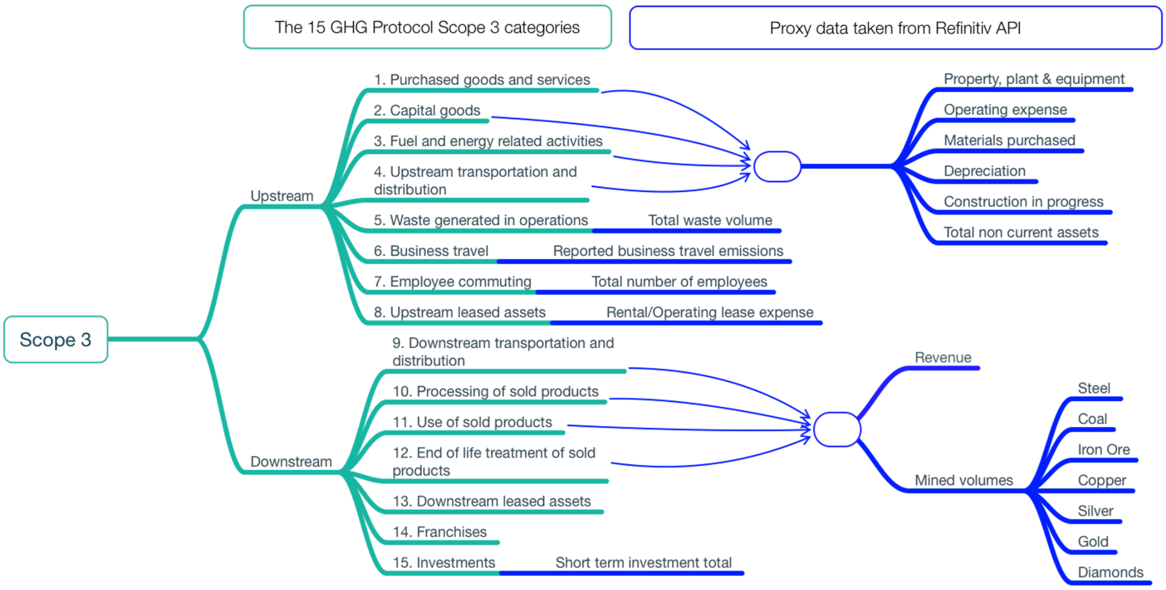

The 15 GHG Protocol Scope 3 categories
Proxy data taken from Refinitiv API
Scope 3
Upstream
1. Purchased goods and services
2. Capital goods
3. Fuel and energy related activities
4. Upstream transportation and distribution
5. Waste generated in operations
Total waste volume
6. Business travel
Reported business travel emissions
7. Employee commuting
Total number of employees
8. Upstream leased assets
Rental/Operating lease expense
Property, plant & equipment
Operating expense
Materials purchased
Depreciation
Construction in progress
Total non current assets
Downstream
9. Downstream transportation and distribution
10. Processing of sold products
11. Use of sold products
12. End of life treatment of sold products
13. Downstream leased assets
14. Franchises
15. Investments
Short term investment total
Revenue
Mined volumes
Steel
Coal
Iron Ore
Copper
Silver
Gold
Diamonds

Figure 3.1: Example of hypothesised mapping between scope 3 categories and specific features.

Revenue, production volumes and the number of employees. It is hypothesised that there will be a positive correlation between company size and scope 3 emissions. For example, if you take two companies that produce the same products using the same processes, all other things being equal, the larger company will have higher scope 3 emissions [10, 62, 9]. The second set of features concern a companies business model; these include features such as margin, leverage, asset age and capital intensity. It is hypothesised that there are certain business models that have higher scope 3 emissions profiles. For example, low margin businesses rely on higher volumes to drive profitability, and high volumes are likely to be correlated with higher scope 3 emissions [10, 62, 9]. The final category of features seeks to capture the environmental context of the company [10]. These draw heavily upon the Refinitiv ESG metrics to determine if a firm is a climate leader or laggard. The hypothesis is that companies with the best environmental metrics, climate leaders, will have smaller size and business model adjusted emissions. While there is some overlap between the categories, the 3 categories are a helpful framework for breaking down the features. A full list of all the features is displayed in the modelling stage can be found in Section 3.5, additionally, the Appendix A includes a hypothesised overview of the connection between each feature and scope 3 emissions. Finally, there are also some notable absent production volume features. For example there are a number of commodities such as aluminium that Refinitiv do not provide data on. These would have been included if the data was readily available.

3.4 Feature Engineering

The objective of feature engineering is to manipulate the existing features so that they become better proxies for scope 3 emissions. This project introduces several engineered features. The first of these is commodity price adjusted revenue. Revenues are hypothesised to be linked to downstream scope three emissions. The higher the revenue, the higher the volume of goods sold, the higher the downstream value chain emissions. However, it is possible to link revenues more closely to output by adjusting for commodity prices. This is done by normalising revenues against the previous years Refinitiv Core Commodity CRB Index Non-Agriculture Livestock price index. This also has the additional benefit of adjusting for the effects of inflation. The reason that the previous year's commodities prices are used to adjust revenue is that commodities are commonly sold on year ahead futures. As Figure 3.2 shows, the Refinitiv Core Commodity CRB Index Non-Agriculture & Livestock price index exhibits significant commodity prices changes over time.

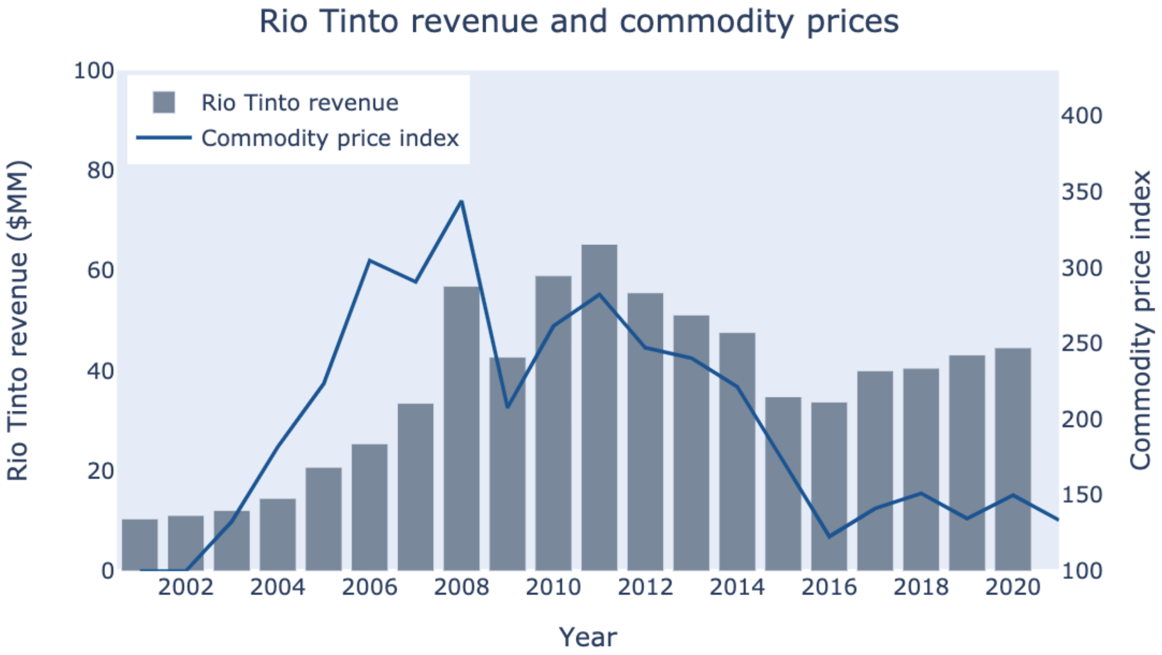


Figure 3.2: Change in the Refinitiv Core Commodity CRB Index Non-Agriculture & Livestock price index over last 20 years. Data pulled from Refinitiv Workspace.

To test the effectiveness of the engineered feature, the coefficient of determination (R2 score) was measured against that of the simple revenue feature. This experiment was performed at various levels of data cleaning. The results are summarised in the table 3.2 and these results suggest that the adjusted revenue only shows superior performance at the highest level of data cleaning. However, the relative similarity between the two revenue features at lower levels of cleaning combined with the strong theoretic underpinning justifies this feature's inclusion in the modelling stage.

Data Cleaning Level	Unadjusted Revenue R2 Score	Adjusted Revenue R2 Score
Level 1	0.49	0.47
Level 2	0.66	0.63
Level 3	0.81	0.86

Table 3.2: Correlation with adjusted and unadjusted revenues at varying levels of data cleaning

Of the 82 features selected from Refinitiv, 25 are Boolean ESG features. These have been engineering into four composite features; upstream features, general features, downstream features and combined features, as shown in figure 3.3. In this context, general features are taken to be features that are directly related to either upstream or downstream scope 3 categories. The reasoning behind this engineering is that while individual categories could have a negligible impact on scope 3 emissions. Combined, they start of give an indication of a firms environmental credentials which could be taken as a proxy for scope 3 emissions focus and action. All features exhibit the same polarity meaning that a value of 'true' is taken to be a positive environmental action. Each feature is given an equal weight with true features receiving a score of one and false features receiving a score of zero. The features are then combined into a single score where a higher overall value is taken to be a proxy of a more environmentally conscious firm. One concern with these features is that if there is a bias in which only the larger companies release enough data in their annual report for these features to be measured, then maybe the feature is actually a poor proxy for company size rather than environmental responsibility.

Two other engineered features are asset age and adjusted energy. These features were created by combining two pre-existing features. This is the sort of feature engineering that a neural network can do internally [30]. However, for the other algorithms used in this project, such engineering remains useful. Adjusted energy is simply the total energy consumed by a company multiplied by (1- % of that energy produced by renewable sources). While energy use is counted in Scope 2 emissions, the adjusted energy use figure gives an indication of the environmental conditions in which the firm operates, which is relevant to scope 3 emissions. However, when tested, the adjusted energy figure showed consistently lower R2 scores than the unadjusted energy figure, which suggests this feature should not be included in the modelling stage.

Asset age is calculated by dividing total fixed assets by depreciation and has been used in several previous corporate emissions models [10, 9]. The hypothesised relationship to scope 3 is that companies with a lower asset age are likely to have newer equipment that is assumed to be more environmentally friendly. Both asset age and the combined categorical features are invariant to company size, so they cannot simply be evaluated looking at coefficients of determination. Instead, they are examples of a moderating feature whose impact can only be measured when used in conjunction with other features that are directly proportional to scope 3.

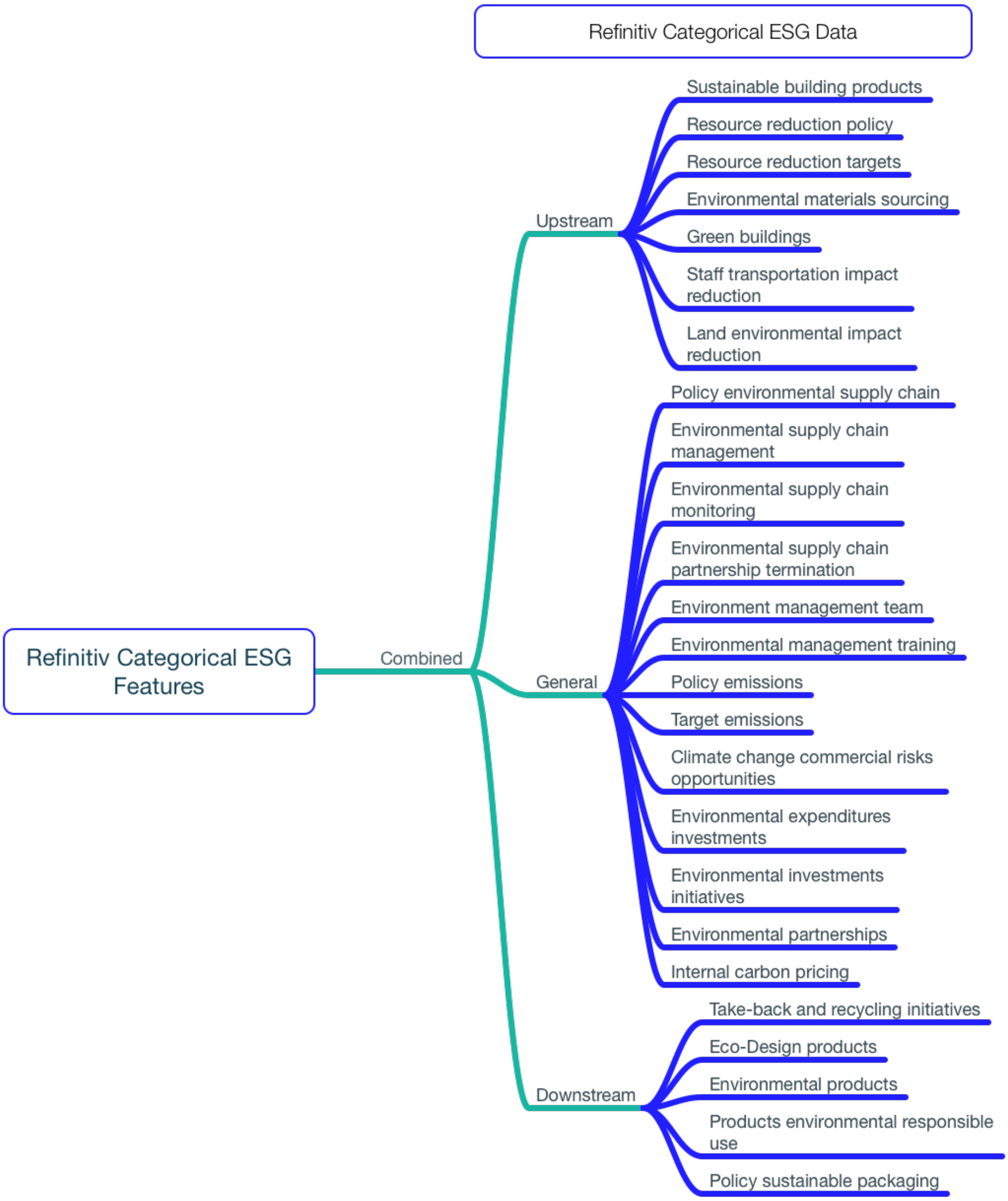

Refinitiv Categorical ESG Data
Refinitiv Categorical ESG Features
Combined
Upstream
Sustainable building products
Resource reduction policy
Resource reduction targets
Environmental materials sourcing
Green buildings
Staff transportation impact reduction
Land environmental impact reduction
General
Policy environmental supply chain
Environmental supply chain management
Environmental supply chain monitoring
Environmental supply chain partnership termination
Environment management team
Environmental management training
Policy emissions
Target emissions
Climate change commercial risks opportunities
Environmental expenditures investments
Environmental investments initiatives
Environmental partnerships
Internal carbon pricing
Downstream
Take-back and recycling initiatives
Eco-Design products
Environmental products
Products environmental responsible use
Policy sustainable packaging

Figure 3.3: Grouped Boolean variables.

The final feature engineering experiment tried in this project was the inclusion of polynomial features. It was hypothesised that such features would capture potential nonlinear relationships in the data. Experiments were conducted using second and third-degree polynomials; however, the results showed only degraded rather than improved performance. As a result polynomial features are not included in the modelling stage.

3.5 Modelling Features

There are important trade offs to consider when deciding which features should be selected for modelling. On the one hand, including more can features allow higher levels of nuance in the model, on the other hand too many features can introduce problems such as the curse of dimensionality and increases the chances of the model relying on coincidental relationships [23]. Another consideration is the fact that some machine learning algorithms are better at handling a large number of features than others. For example XGBoost is able to easily handle missing values and irrelevant features [35] whereas k nearest-neighbors struggles with a high dimensional feature space [26]. With this in mind the project creates three different sets of features for different scenarios. The first set of features is the core features list which include five basic features that are universal sector agnostic features that could be readily used by those without access to premium data. The second set of features are high availability gold standard features (GS) that contain some sector specific features and premium ESG data from Refinitiv. High availability means that the feature is available for over 95% of companies. The third and final list of features includes sector specific and gold standard data and includes lower availability features. This third list is designed for the XGBoost algorithm that can handle missing data. Full list of all features and their hypothesised link to scope 3 emissions can be found in appendix A. The references in the list indicate which features have been used in previous emissions modelling papers.

All GS Features

- TRBC industry [63]
- Total CO2
- Leverage % [10]
- Intangibles [10, 9]
- Categorical combined score
- Categorical upstream score
- Categorical general score
- Categorical downstream score
- Revenue adjusted for commodity price
- Total fixed assets [10]
- Property, plant & equipment [10]
- Operating expense
- Gross margin [62, 9, 10]
- Year [10]
- Industry group [63]
- Activity [63]
- Scope 1
- Scope 2
- Employees [63, 10]
- Total waste (Tonne)
- Energy use total (Gigajoule) [63, 10]
- Capital intensity [62, 10]
- Material expenses
- RD expense
- Construction in progress
- Depreciation [9]
- Operating lease payments total
- Exchange country
- Cement (Tonne)
- Steel (Tonne)
- Coal (Tonne)
- Crude oil (Barrel)
- Natural gas (Cubic Meter)
- Combined oil and gas (BOE)
- Iron ore (Tonne)
- Copper (Tonne)
- Manganese ore (Tonne)
- Zinc (Tonne)
- Nickel (Tonne)
- Titanium slag (Tonne)
- Silver (Troy Ounce)
- Gold (Troy Ounce)
- Diamonds (Carat)
- Platinum (Troy Ounce)

All GS Features Continued

- Non-Hazardous waste
- Hazardous waste
- Environmental expenditures
- Water withdrawal total
- Water discharged
- Cement CO2 emissions
- Environmental pillar score
- Resources use score
- Renewable energy use ratio
- Emissions reduction target percentage
- Emissions reduction target year
- Environmental provisions
- Sustainable building products
- Resource reduction policy
- Resource reduction targets
- Environmental materials sourcing
- Green buildings
- Environmental products
- Eco-Design products
- Take-back and recycling initiatives
- Products environmental responsible use
- Policy sustainable packaging
- Renewable energy use
- Policy environmental supply chain
- Environment management team
- Environment management training
- Staff transportation impact reduction
- Environmental supply chain management
- Environmental supply chain monitoring
- Environmental supply chain termination
- Land environmental impact reduction
- Policy emissions
- Target emissions
- Climate change commercial risks opportunities
- Environmental expenditures investments
- Environmental investments initiatives
- Environmental partnerships
- Internal carbon pricing
- Assets age [10, 9]

Core Features

- Revenue [63, 9, 62, 10, 69]
- Total fixed assets [10]
- Property, plant & equipment [10]
- Operating expense
- Gross margin [62, 9, 10]

High Availability GS Features

- TRBC industry [63]
- Total CO2 [63]
- Leverage % [10]
- Intangibles [10, 9]
- Categorical combined Score
- Categorical upstream Score
- Categorical general Score
- Categorical downstream Score
- Revenue commodity adjusted
- Total fixed assets [10]
- Property, plant & equipment [10]
- Operating expense
- Gross margin [62, 9, 10]

Chapter 4

Design

This section outlines the design of the modelling project. It demonstrates how the project draws upon and extends previously published work and introduces elements of novelty.

4.1 Scope 3 Focus

The first key design decision made by this project was to focus on modelling scope 3 emissions exclusively. Several papers have focused on modelling scope 1+2 emissions [62, 9], and one paper has tried to build a universal model for scope 1,2 and 3 [10]. However, it was hypothesised that this project could achieve improved performance over the universal model by focusing exclusively on scope 3. While there is a sizable overlap, the drivers of scope 3 emissions are not completely the same as the drivers for scope 1 and 2. For this reason, being able to select features that are directly relevant to scope 3 rather than trying to select universal features is hypothesised to lead to improved model performance.

4.2 Choice of Models

The 2021 Otago paper showed that machine learning techniques were able to improve emission modelling performance compared to the simple regression techniques used in previous studies [62, 9]. Drawing from these existing emissions modelling studies and the wider research done in Section 2.2.2, this project implements the following models in the following order:

- Simple regression
- Elastic Net regression
- K Nearest Neighbors regression
- Support Vector regression
- AdaBoost
- XGBoost
- Feed-Forward Neural Network
- Convolutional Neural Network

Each of these models have been outlined in detail in Chapter 2.1. To the best of the author's knowledge, this is the first time that the svr and cnn algorithms have been used in corporate emissions modelling.

4.3 Experiment Design Overview

For each model, two experiments are run. The first use the basic core features, while the second uses the larger selection of gold standard features. The core feature experiments are designed to test the performance of the model trained on readily available sector agnostic features. The gold standard feature experiment demonstrates the model's performance when trained on sector-specific data from a premium data provider. The list of basic core features is consistent between all experiments. The list of gold standard features is consistent across all experiments apart from for XGBoost where a number of additional lower availability features are included as this model can handle missing values.

The one exception to this setup is the KNN experiment. Here two experiments are run, both using the basic core features. The difference between the two experiments comes from the data used in each instance. For the first experiment, the data set only allowed one scope 3 data point from each company. The second experiment used the standard data set used by the rest of the models, which allows multiple historical data points from the same company. The motivation for these two experiments was to see how the model performed when it did and did not have access to historical data points for a given company. This experiment allows you to answer if, given historical data points, is it preferable to use KNN over other models. No experiments with the longer list of gold standard features were performed with KNN because this method is particularly susceptible to the curse of dimensionality [72].

4.4 Baseline Overview

In order to be able to track the performance of our modelling experiments, it is important to have baselines against which the models can be compared. Two sets of baseline were used for this project. The first highly simplistic baseline used is simply the mean and median of the scope 3 data set. The second baseline consists of the Refinitiv emissions models outlined in Section 2.2.3. These models were recreated in Python and validated for correctness against combined scope 1 + 2 emissions pulled from the Refinitiv Workspace API. These models were then fed with scope 3 data rather than scope 1 + 2 data and used to predict scope 3 emissions. It is important to note that these models were not designed for scope 3 emissions modelling, so sub-optimal performance is expected. However, given the overlap in drivers between scope 1+2 and scope 3 emissions, the models are still expected to have some predictive power. Furthermore, in the absence of any other suitable published model baselines, weak baselines are better than no baselines. The Refinitiv Median and Energy models are the two primary baselines for this project. As discussed in 2.6 the Refinitiv CO2 model uses previous company CO2 emissions as a feature and thus is only a suitable baseline for the KNN historical data included experiment. Other commercial scope 3 models exist as outlined in Figure 2.6, however they sit behind paywalls and with no institutional access option for Imperial students.

There is one published academic paper that has built a universal emissions model that can be used for scope 3 modelling [10], this is not used as a baseline as some of the data it uses is not available from Refinitiv or online. This external data was published as part of the paper but as the model only looked at scope 3 emissions up to 2017 this external data is not available for any scope 3 data point after 2017. This project looks at data points up to 2021 which includes a significant number of scope 3 emissions disclosures released after 2017. To be precise, 184 of the 597 data points in the data set have been released since 2017 (31%). While the model is not used as a baseline, a direct comparison in performance of the two studies is conducted using data up to 2017. This is explained in more detail in Section 5.6.

4.5 Error Metrics

An important design decision is choosing which error metrics will be used for optimising experiments. This project looks at three error metrics to compare results. These are Mean Absolute Error, Root Mean Square Error and the Coefficient of Determination (R2). While all three metrics are used for comparing different algorithms, the mean absolute error metric is used for the purpose of parameter tuning.

Mean absolute error represents the average difference between the true and the predicted values. A key strength of this metric is that it is intuitive to interpret. However, a weakness is that it can be significantly affected by a small number of very large errors.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|e_t|$$

Root Mean Squared Error represents the standard deviation of the prediction errors. The advantage of RMSE is that it penalises large errors. Interpreting this error metric is less straightforward than MAE but again, the lower the value the better the model.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}e_t^2}$$

The R2 Coefficient of Determination represents the proportion of the variance in the dependent variable y that is explained by the independent variable. The results range between 0 and 1, with 1 being a model with perfect predictive power. It can be calculated according to the following equation [72]:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

Where:

$$TSS = \text{Total Sum of Squared} = \sum(y_i - \bar{y})^2$$

$$RSS = \text{Explained Sum of Squared} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Chapter 5

Implementation and Experimentation

5.1 API Interface

The implementation of the project starts with the Refinitiv Data Platform API. To access the API an app key had to be created attached to a Refinitiv account. The Refinitiv account access was provided free of charge as part of an academic partnership between Refinitiv and Imperial. The script to interact with the API is written in Python, and this language is used across the whole project. The API was first queried for a list of all the unique Reuters Instrument Codes (RICs) of companies with a revenue of over $100M in the TRBC Mineral Resources Business Sector. The second query to the API then retrieved the list of 82 features used for modelling that are associated with each company in a given year. This request is broken down into a number of smaller requests to be rate limit compliant. The final query to the API retrieves a time series of values for the Refinitiv commodity index that is used as part of the feature engineering process to normalise revenue with respect to commodity prices. Once extracted from the API, the pandas data analysis library is used to manipulate the data.

This script focuses on usability and is designed to be easily extendable and modifiable. For example, all of the complexity of the interface with the API has been abstracted away behind a get data function. The user enters a start and end year and returned from the function is a DataFrame of all scope 3 data points and their associated features for that time period. One huge advantage of interfacing with the Refinitiv API is that it will be able to dynamically adapt to industry changes and this will be reflected in the data. For example, if there are new entrants into the market, then this will be picked up by the first call to the API without the user having to think about such complexities.

5.2 Baseline Construction

The Refinitiv Energy, Median and CO2 models were recreated in Python for this project. The implementation followed the model description published by Refinitiv [63]. The reason the models had to be recreated is that Refinitiv uses the models for scope 1 + 2 predictions rather than scope 3 prediction. In order to use the models for scope 3 prediction, they had to be recreated and fed with scope 3 data. To validate that the recreated models had been correctly implemented, released scope 1 + 2 predictions from Refinitiv were used as a testbed against which the recreated model performance was tested.

5.3 Implementation of Models

A general overview of the machine learning pipeline implemented in this project can be seen in figure 5.1. Previous sections have outlined what data is extracted from Refinitiv and how it is processed 3; this section now looks at how the models are trained upon the data and used to make predictions.

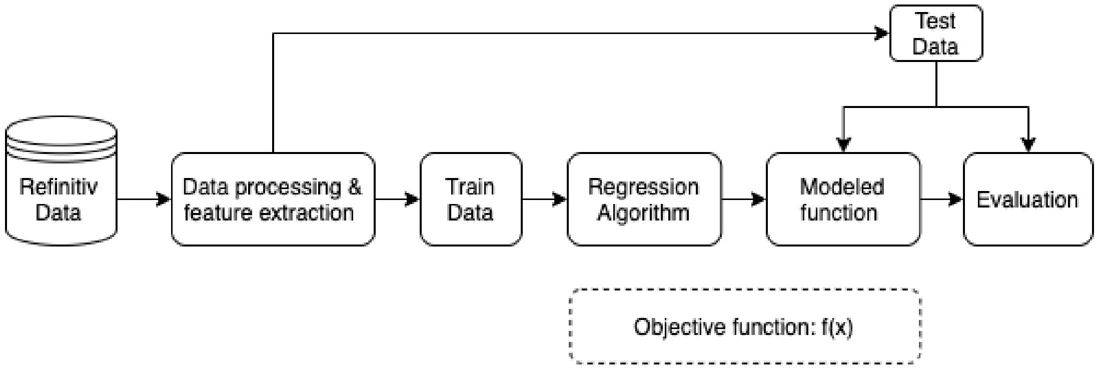


Figure 5.1: Machine learning pipeline. Adapted from [76].

This project uses the open-source Scikit-learn, XGBoost, Tensorflow and Keras libraries for implementing the machine learning algorithms used in the experiments. To fit the model on the data, the target variable and the model appropriate features are passed into the model. For all models, apart from XGBoost, any rows with missing values must be dropped in advance of fitting the model on the data as these algorithms cannot handle missing data. It was decided to drop the data rather than trying to impute approximate values for reasons of transparency. This is why the features used for models that cannot handle missing values are all high availability features so that as few Scope 3 data points as possible are discarded. Additionally, categorical values must be one hot encoded, and a standard scalar is applied to continuous features.

A key pitfall in machine learning, particularly when trained on a small data set, is overfitting. Overfitting occurs when a model learns the training data set too closely and as a result is unable to generalise to unseen data. The more complex a model, the more capacity that model has to overfit. While different models have different techniques for preventing overfitting, they are all essentially trying to do the same thing, to reduce the complexity of the model relative to the size of the data set. This project implements a number of different techniques for addressing overfitting. For example Ridge and Lasso regularisation are used for regression modelling [25], varying the size of K is used in KNN modelling [33], varying tree max depth is used in boosted gradient tree algorithms [40] and dropout [77] is used in neural networks.

Additionally, to detect if overfitting is occurring, five-fold cross-validation is applied to each of the models. Given the relatively small data set size, it is more difficult to draw robust conclusions from a single train test split of the data. This is where k fold cross-validation is helpful. In five-fold cross-validation, the data is split up into five sets. The data is split randomly in an attempt to maximise the chance that each set is as representative as possible of the larger population. Four of the sets are used to train the model, and one set is used for testing. This is then repeated until each set has been used as the test set. The model performance can then be averaged across each of the five rounds [23]. This technique is beneficial as the full data set is used for testing while maintaining the crucial separation between training and test data at each stage. This technique allows you to use all of the data as validation data at some point which will give you a better indication of the model's ability to generalise than using a simple single train test split. In order to ensure consistent results between model runs, a random seed is specified for splitting the data into a test and train set.

Hyperparameter tuning is another area where cross-validation is important. This project uses both grid search and the Hyperopt Bayesian optimisation library for tuning hyperparameters. The mean absolute error metric was used to guide the optimisation process. Grid search was used for the Elastic Net, SVR and Adaboost models, where the moderate number of hyperparameters facilitated the and exhaustive search. Hyperopt was used to speed up the process for the more computationally demanding XGBoost tuning process. When accessing the performance of different hyperparameter combinations, five-fold cross-validation is used to once again guard against overfitting. Indeed, given that the values of the hyperparameters play a crucial role in determining a models ability to overfit, this process is an important one. For the KNN algorithm, instead of using grid search or Hyperopt, error metrics were manually inspected while altering the value of K in what is known as informally known as an elbow plot experiment. Additionally, a manual trial and error approach to hyperparameter tuning was taken for both the deep learning algorithms. A full list of the chosen hyperparameters and hyperparameter search space can be found in Appendix A.

After the hyperparameters were chosen, the predictive power of each model was evaluated again using cross-validation. One additional alteration was made to each model before its final performance was evaluated. This was to restrict the domain of the predictions so that negative predictions were not possible. While this was rare, there were a small number of negative predictions made by the linear regression and XGBoost models. Here, real-world domain knowledge was applied to restrict scope 3 emissions to positive values as negative values are not possible. Ideally, the models would have learnt this automatically without having to be coerced and it is acknowledged that restricting the domain artificially inflates the performance of the algorithm.

5.4 Anomaly Detection

The primary purpose of the models constructed in this project is to predict scope 3 emissions. However, these models can also have a very useful secondary purpose which is in anomaly detection. A key problem with scope 3 emissions data is that some companies only release partial scope 3 figures. If a scope 3 value is significantly lower, for example, by order of magnitude, than is estimated by the model, then perhaps the disclosed scope 3 data point is only a partial one. This application of the model is of particular interest to a data provider like Refinitiv, for whom having accurate and well-labelled data is very important. This auxiliary experiment is implemented by plotting predicted vs published scope 3 data points and investigating areas of large discrepancy.

5.5 Year Ahead Forecasting

An extension of the primary use of the model is to use it for the year ahead scope 3 emissions forecasting. Analysts often release projected financial and operating metrics for public companies. These projected metrics can then also be used to predict scope 3 emissions. While such a forecasting model will be restricted by the limited number of features that are predicted by analysts, the same modelling process can be applied. Because of the small number of features deep learning techniques are not applied to the task of year ahead forecasting. A manual inspection of analysts forecasts available on Refinitiv showed that four of the five core features are regularly forecast by analysts; these are shown below and are used as the features in the forecasting experiments. This extension of the model is not the primary focus of the project but instead is implemented as a proof of concept.

Forecasting Features

- Revenue

- Total fixed assets
- Operating expense
- Gross margin

5.6 Otago Comparison

As outlined in Section 2.2.3, there is one model published in the academic literature that has previously tried to estimate scope 3 emissions [10]. A key part of assessing the performance of this project is to compare it to the current state-of-the-art in the literature. In order to facilitate this comparison, the model from the Otago paper had to be recreated. Like this project, they pulled data from Refinitiv for their model and combined this with several other additional data sets. Because of data licensing, the Refinitiv data was not shared as part of the data and code published alongside the paper. However, the ReadMe did include field names of each of the features they had pulled from Refinitiv and from these it was possible to reconstruct the Refinitiv part of the data. Then the Refinitiv data had to be merged with the other three external data sets. This included one data set from the Internal Energy Agency and two data sets from the World Bank. Once the full data set used by the Otago paper had been recreated, it was then cut down to just companies in the Mineral Resources sector. A mirror data set was then created, which contained exactly the same target scope 3 data points, but instead of the Otago features, it included the features used in this project. The two data sets with the same target variables and the specific features used by each projects allowed for a fair comparison between the two modelling approaches. The best model from this project would then be compared to the best single algorithm model from the Otago paper. The best single algorithm model was taken from the Otago paper, rather than a meta learner, as this project focuses on single algorithms for reasons of explainability and methodological transparency.

Chapter 6

Results and Discussion

6.1 Results

6.1.1 Scope 3 Model Results

Lev	Error	Mean	Med	Ref Eng	Ref Med	Ref CO2	OLS Core	OLS GS	EN Core	EN GS	KNN His	KNN NHis	SVR Core	SVR GS	Ada Core	Ada GS	XGB Core	XGB FGS	NN Core	NN GS	CNN Core	CNN GS
L1	MAE	68.0	42.5	55.2	34.1	16.6	31.0	23.7	30.8	22.0	13.8	21.8	43.2	24.6	19.1	16.8	15.7	11.8	25.1	15.8	43.4	25.1
L1	RMSE	123.0	130.0	154.6	101.5	51.7	65.6	49.4	65.7	47.0	38.4	48.2	129.1	57.0	46.2	38.9	40.8	38.6	69.3	37.7	129.5	58.4
L1	R2	0	Neg	0.11	Neg	0.89	0.7	Neg	0.7	Neg	0.89	Neg	Neg	Neg	0.84	0.27	0.88	0.88	0.67	Neg	Neg	Neg
L2	MAE	125.8	93.5	111.7	59.1	42.2	48.7	45.0	48.5	41.0	25.2	30.7	93.8	47.6	38.8	21.4	30.3	19.4	45.6	32.5	95.1	53.7
L2	RMSE	172.2	192.8	249.9	127.7	84.7	81.4	70.8	81.6	65.1	51.3	57.1	191.5	97.8	68.4	45.7	59.0	44.8	88.3	69.5	195.4	105.2
L2	R2	0	Neg	0.05	Neg	0.85	0.77	0.17	0.77	0.31	0.89	Neg	Neg	Neg	0.81	0.74	0.86	0.92	0.71	0.38	Neg	Neg
L3	MAE	145.4	116.3	96.8	69.6	52.7	56.6	40.3	56.4	40.8	34.3	42.3	116.1	61.6	45.6	23.8	41.6	22.3	58.6	28.8	118.5	65.4
L3	RMSE	187.1	216.1	178.3	128.8	94.9	87.2	68.1	87.7	70.5	62.1	69.5	213.2	109.3	73.4	39.2	71.4	50.5	101.2	49.2	218.2	115.3
L3	R2	0	Neg	0.38	0.14	0.83	0.76	Neg	0.76	Neg	0.85	Neg	Neg	Neg	0.81	Neg	0.83	0.92	0.69	Neg	Neg	Neg

Table 6.1: Table of experimental results of model performance at each level of data cleaning. Units are Mt CO2e. Models shown in following order: Mean Baseline, Median Baseline, Refinitiv Energy Model, Refinitiv Median Model, Refintiv CO2 Model, Regression Core Features, Regression Gold Standard Features, Elastic Net Core Features, Elastic Net Gold Standard Features, KNN Historical Data Included, KNN No Historical Data, SVR Core Features, SVR Gold Standard Features, AdaBoost Core Features, AdaBoost Gold Standard, XGBoost Core Feature, XGBoost Full Gold Standard Features, Neural Network Core Features, Neural Network Gold Standard Features, CNN Core Features, CNN Gold Standard Features.

Cleaning Level	Best Model	MAE	RMSE	R2
Level 1	XGBoost Full Gold Standard Features	11.8	38.6	0.88
Level 2	XGBoost Full Gold Standard Features	19.4	44.8	0.92
Level 3	XGBoost Full Gold Standard Features	22.3	50.5	0.92

Table 6.2: Best model results at each cleaning level.

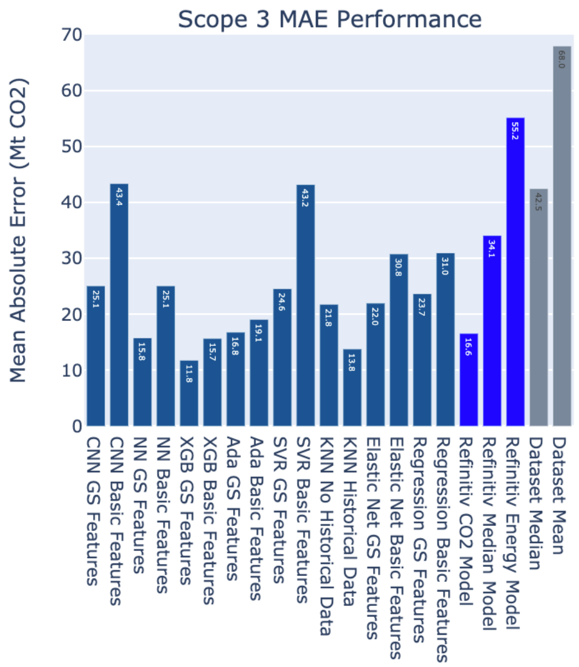


Figure 6.1: MAE results at level 1 data cleaning.

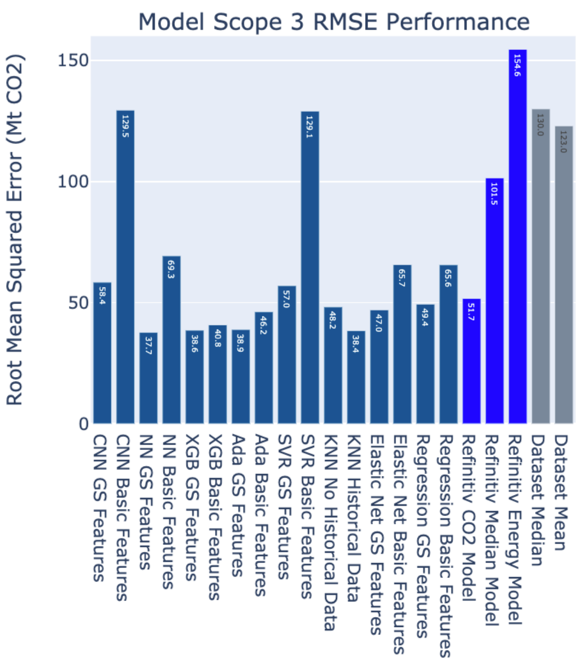


Figure 6.2: RMSE results at level 1 data cleaning.

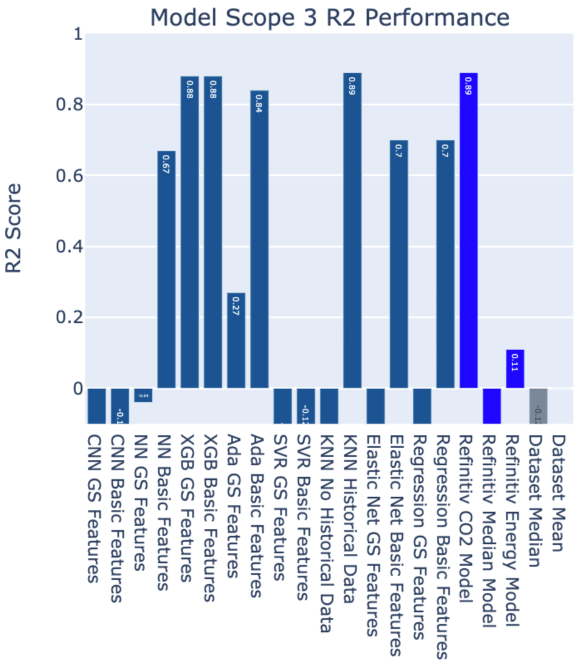


Figure 6.3: R2 results at level 1 data cleaning.

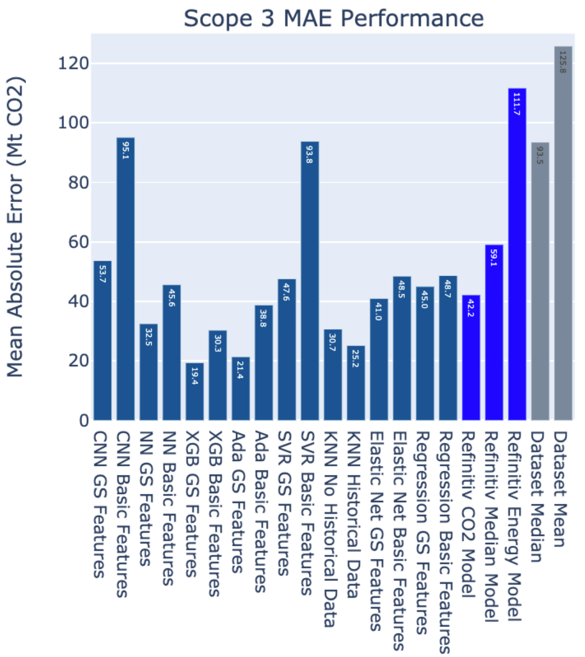


Figure 6.4: MAE results at level 2 data cleaning.

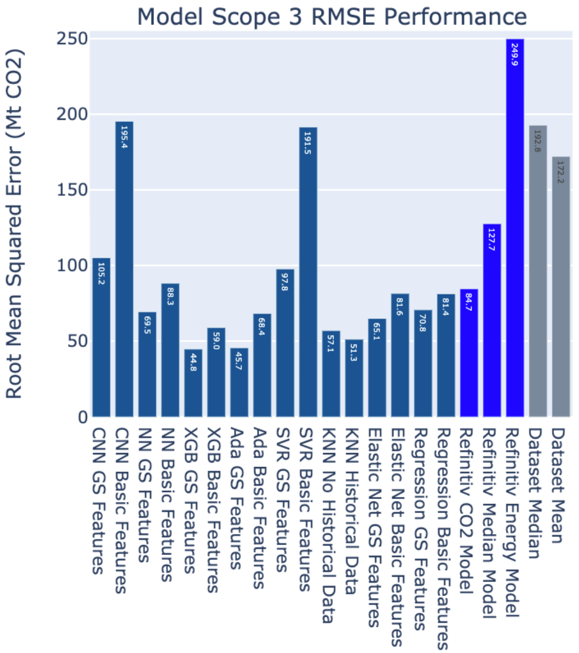


Figure 6.5: RMSE results at level 2 data cleaning.

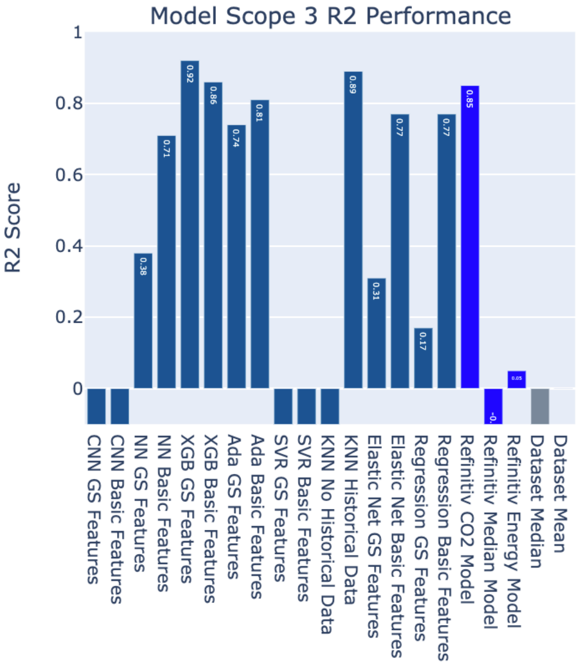


Figure 6.6: R2 results at level 2 data cleaning.

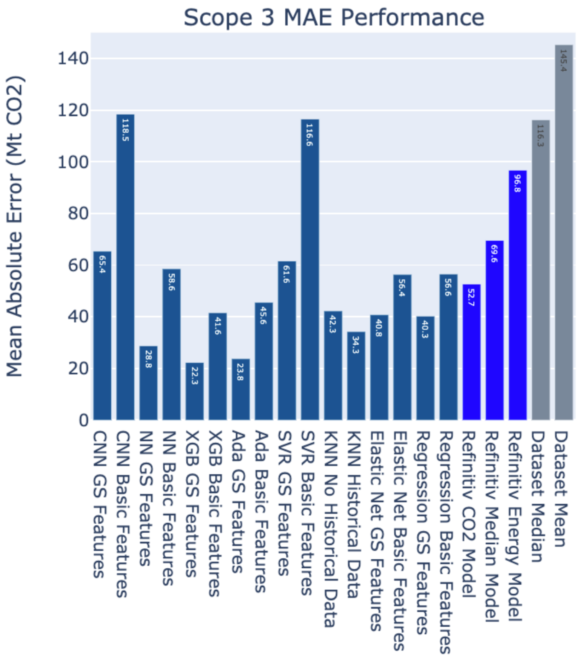


Figure 6.7: MAE results at level 3 data cleaning.

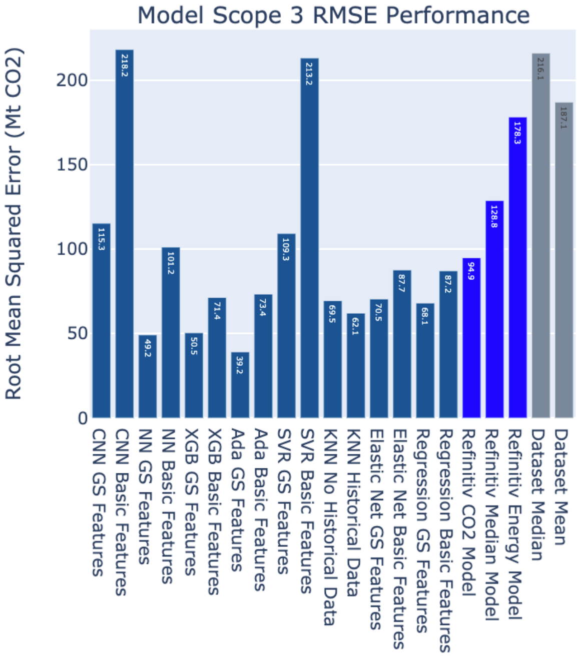


Figure 6.8: RMSE results at level 3 data cleaning.

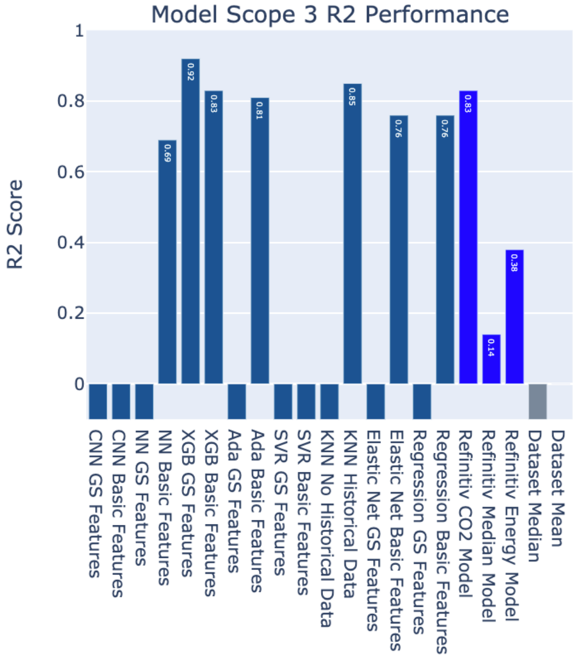


Figure 6.9: R2 results at level 3 data cleaning.

6.1.2 Otago Comparison Results

Otago Model Comparison Results			
Error	**Project Model**	**Otago Model**	**Improvement (%)**
MAE	11.4	14.9	23.5
RMSE	36.6	51.9	29.5
R2	0.77	0.59	30.5

Table 6.3: Table comparing the results of the best model from this project to the best single model from the state-of-the-art Otago paper.

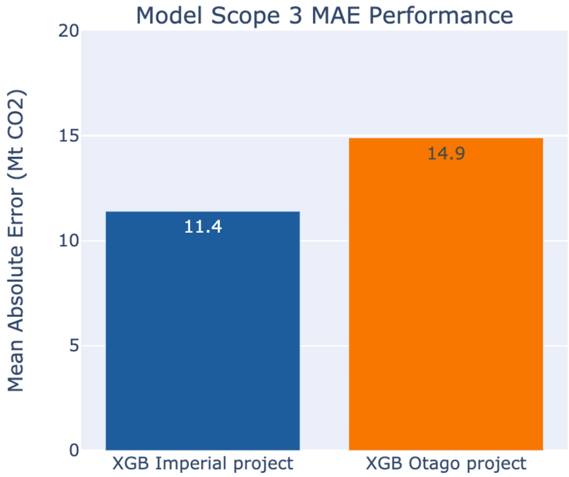


Figure 6.10: MAE comparison with Otago paper.

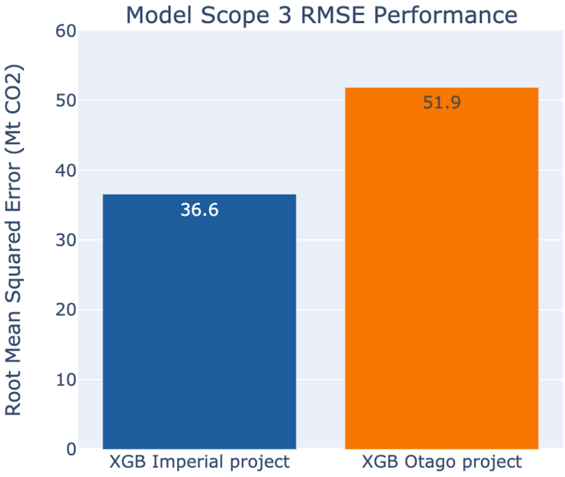


Figure 6.11: RMSE comparison with Otago paper.

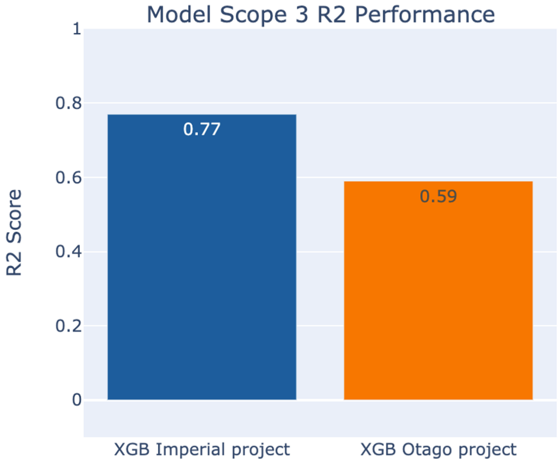


Figure 6.12: R2 comparison with Otago paper.

6.1.3 Anomaly Detection Results

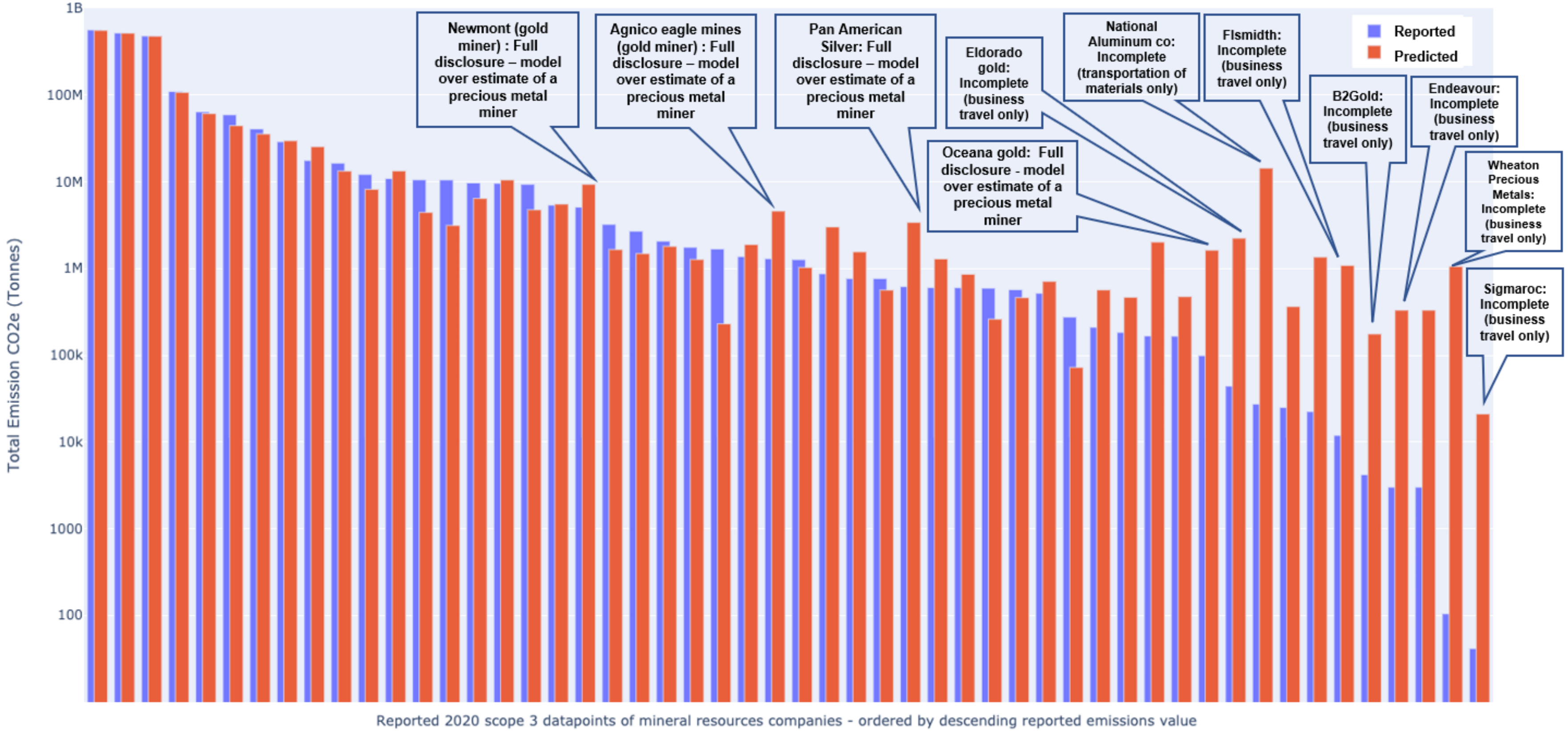


Figure 6.13: Scope 3 data points published by Mineral Resources companies in 2021 for the 2020 financial year compared against model predictions. Labels explain discrepancies between reported and predicted results. The two primary causes of discrepancies are incomplete reporting of scope 3 emissions and model overestimation of precious metal miners. Predictions were made using the XGBoost model trained on level 2 cleaned data using full GS features.

6.1.4 Year Ahead Forecasting Results

Lev	Error	Mean	Med	Ref Eng	Ref Med	Ref CO2	OLS	EN	KNN H	KNN NH	SVR	Ada	XGB
L2	MAE	125.8	93.5	111.7	59.1	32.2	51.0	50.8	23.2	28.3	93.8	40.4	30.4
L2	RMSE	172.2	192.8	249.9	127.7	84.7	85.9	86.1	48.5	51.5	191.5	67.0	60.3
L2	R2	0	Neg	0.05	Neg	0.85	0.74	0.73	0.89	Neg	Neg	0.82	0.86

Table 6.4: Year ahead forecasting results conducted on level 2 cleaned data using selected forecasting features.

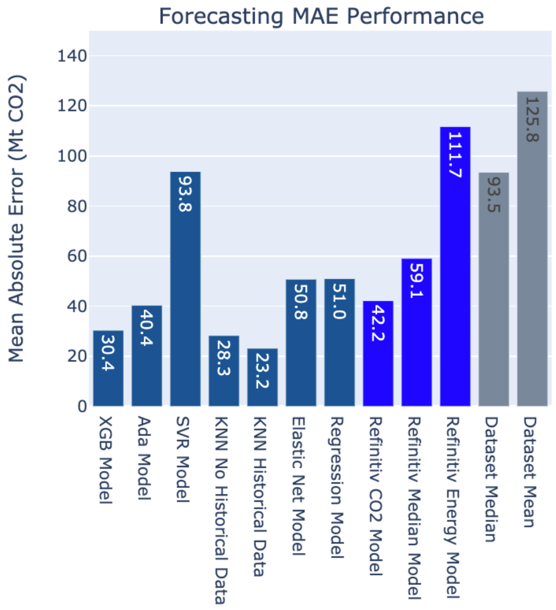


Figure 6.14: MAE results for year ahead forecasting. Level 2 cleaned data used.

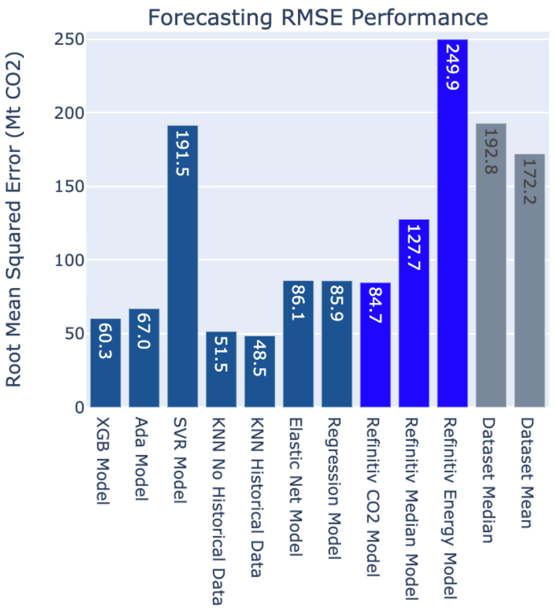


Figure 6.15: RMSE results for year ahead forecasting. Level 2 cleaned data used.

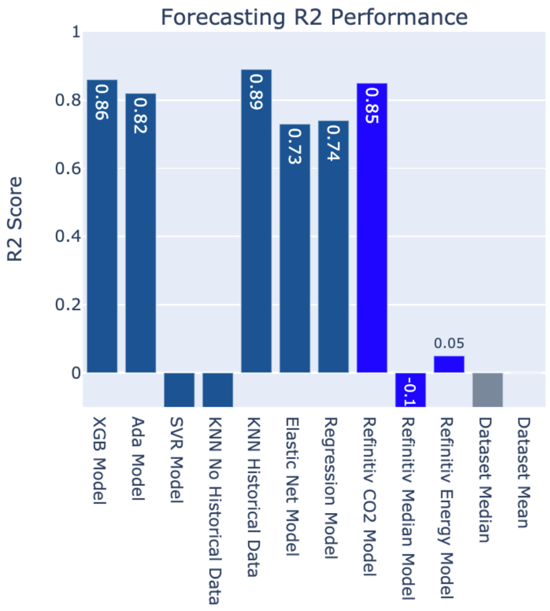


Figure 6.16: R2 results for year ahead forecasting. Level 2 cleaned data used.

6.2 Evaluation and Discussion

This chapter critically evaluates the experimental results outlined in Section 6.1.

6.2.1 Absolute Results

This section focuses on the performance of each of the models at the three levels of data cleaning. Before accessing the performance of each model in detail, it is instructive to highlight three general findings that hold across all the results. Firstly, all of the machine learning models implemented by this project demonstrate an improvement against the project baselines when trained on gold standard data. Secondly, there is consistent empirical support for the hypothesis that the gold standard industry-specific features (GS) outperform the core basic features. Lastly, XGBoost is the best performing algorithm across the three data cleaning levels. This section will now go on to individually evaluate and discuss the performance of each of the algorithms.

6.2.2 Individual Model Performance

Regression Models

The regression experiments perform reasonably well and demonstrate a small outperformance of the baselines both in terms of MAE and RMSE. The experiments with elastic net regularisation show further slight improvements in performance for the experiments that use the GS features at level 1 and level 2 cleaning. This is as expected, the larger number of features used in these experiments compared to the basic feature experiments means that the elastic net regularisation has more scope to combat overfitting and multicollinearity in the data [25]. However, the basic core feature elastic net experiments do not exhibit any performance benefit when compared to simple regression which suggests that overfitting and multicollinearity is minimal for experiments using this reduced size feature set. One final element of the regression results are worth discussing; this is the fact that the R2 scores for the gold standard feature models are much lower than for the core features model. This can be explained by the fact that a number of the GS features are size invariant moderating features are not directly correlated with scope 3 emissions but instead are used to scale the features that are directly correlated with scope 3. For example, say a company has a strong Categorical Combined Score, indicating that it is an environmentally friendly company. This score in itself does not directly correlate with scope 3 emissions as it gives no indication of the size of the firm; hence it is a size invariant feature. Instead, this feature is useful once a rough scope 3 figure is calculated from features that do directly relate to scope 3 emissions, such as revenue. The Categorical Combined Score can then be used to scale the estimated scope 3 prediction figure, hence why it is termed a moderating feature.

KNN Models

The KNN results support the hypothesis that including historical data points from the same company in the training set will improve the performance of this algorithm. This can be seen from the fact that the ’Historical Data’ experiment consistently outperforms the ’No Historical Data’ experiment. The reason that having a data set that includes historical data points makes such a difference for KNN is that, unlike other algorithms, the model does not learn from all the data but instead only focuses on the 3 data points that are closest in the feature space. The KNN experiment with no historical data points outperforms the four standard baselines and marginally outperforms the regression experiments. The KNN experiment with historical data points takes the Refinitiv CO2 model as a baseline and demonstrates improved performance. As this project looks to build a model to predict scope 3 emissions for companies that do not currently disclose scope 3 emissions, the KNN Historical Data experiment is interesting but not relevant unless you are looking to predict emissions for a company that has previously released emissions. This model is much more applicable when it comes to year ahead forecasting, as discussed in Subsection 6.2.9.

SVR Models

The SVR results for both the core and gold standard features perform worse than the regression models. The SVR basic features model performs particularly poorly and and fails to compete with the Refinitiv Median model benchmark. Support vector methods are known to perform poorly with noisy data [33] and this is thought to help explain the relatively poor performance seen in this experiment. Furthermore, kernel methods struggle with feature selection so it is possible that a single bad feature could be adversely effecting the model [33]. This is in line with other studies in the literature that have found similar levels of relative performance from SVR models compared to other machine learning algorithms [46].

Tree Based Models

Both of the tree-based algorithms, Adaboost and XGBoost, demonstrate strong performance. A key reason for this strong performance is thought to be the ability of tree-based algorithms to effectively perform internal feature selection [40]. XGBoost with the gold-standard features is consistently found to be the best model. This is hypothesised to be in large part related to the model’s ability to handle missing values. This allows the model to be fed with an expanded list of features which in turn allows the model to use these additional features to make more nuanced predictions. That said it is also worth discussing one potential risk of feeding such a large number of features into the model and that is the risk of spurious correlations. The more features you have, the more chance the model picks up on statistical relationships that are the result of chance rather than any underlying relationship between variables and this might damage the model's ability to generalise. To guard against this possibility, the technique of five-fold cross-validation employed by this study should provide a robust assessment of the model's ability to generalise.

Deep Learning

Of the two deep learning techniques trialled in this project, feed-forward neural networks demonstrate superior performance compared to convolutional neural networks. The feed-forward neural network demonstrates an improvement over all the baselines and all the other models apart from the tree-based algorithms. This strong performance is somewhat surprising considering the relatively small size of the data set. It suggests that this approach holds a lots of promise as number of disclosures increases over time leading to a larger data set size. The CNN, on the other hand, demonstrates relatively poor performance. This poor performance is thought to be the result of the small data size as such experiments are usually performed with orders of magnitude more data [30].

6.2.3 Impact of Data Cleaning

Across all three levels of data cleaning, the relative performance of different models remains very similar. This replicated relative performance allows robust conclusions to be drawn regarding the relative merits of different models. One element that does change between each of the data cleaning levels is that the size of the absolute error increases. This can be explained by the fact that the average data set emissions increases at each additional level of cleaning as it is most commonly the smallest data points that are being removed. It is interesting to note that while the absolute error increases with more stringent data cleaning, the ratio of best model MAE to dataset mean emissions actually decreases as show by Figure 6.17. This could be because the data cleaning removes some of the noise from the data set, which allows the models the better capture the signal. The fact that the relative performance of the Refinitiv baselines models versus the data set mean and median models also improves with data cleaning supports the idea that increased data cleaning improves the signal to noise ratio. The trade-off here is that as you implement more stringent data cleaning, the chance you also end up removing legitimate scope 3 data points also increases.

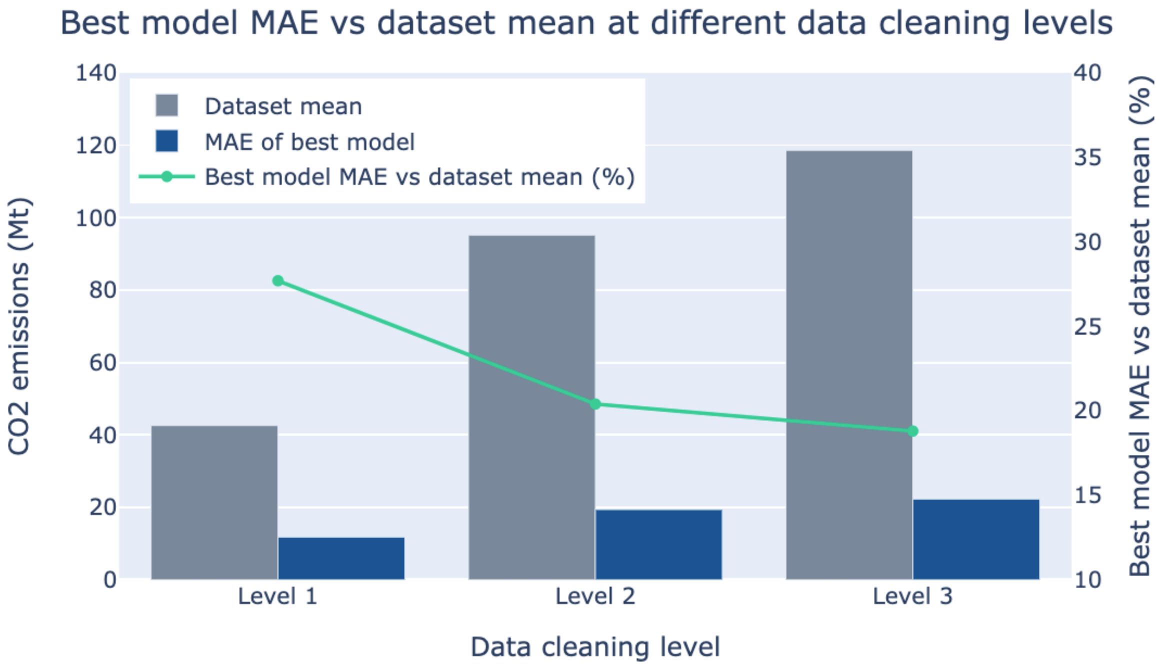


Figure 6.17: Plot showing how MAE of the best model and dataset mean scope 3 emissions evolve at each level of data cleaning.

6.2.4 Basic vs Gold Standard Feature Performance

The gold standard features consistently demonstrate better performance than the universal core basic features. This is thought to be because the gold standard features contained features that were more closely linked to both the specific sector and scope 3 emissions. Indeed these results strongly support the use of premium sector-specific data in scope 3 modelling. However, there is a trade-off inherent in this approach. By using gold standard sector-specific data, you are sacrificing the universal applicability of the model. It makes it more difficult to adapt the model to new sectors, it can act as a further barrier to explainability and it can only be used by those with access to premium data.

6.2.5 Additional Performance Considerations

This project has focused on three main quantitative error metrics; however, several additional factors are worth considering when evaluating the best choice of model. These include implementation complexity, computational complexity, interpretability, stability and the guiding principle of Occam's razor. These additional considerations make it clear that while the feed forward neural network model shows marginally better performance than the regression model it may in fact be preferable to just use the regression model. This is because deep learning models are more complex to implement, take longer to train and are less interpretable. Indeed interpretability is of key importance when the model is used by outside stakeholders, and this can explain the advantages of the simple elegance of the Refinitiv baseline model.

A final dimension upon which the various models can be assessed is applicability. The availability of the features determines the applicability of the model, assuming the model cannot handle missing values. As this project has used only highly available features (over 95% availability), the models in the project are thought to have a strong level of applicability. Particularly in comparison to the baseline Refinitiv Energy Model which relies on a corporate ’Energy use’ feature which limits its applicability to the minority of firms in the data set that report this figure. This factor is also an additional argument in favour of XGBoost as the best overall model as its ability to handle missing values gives the model effective universal coverage.

6.2.6 Results Disaggregation

This section has focused on discussing aggregate results metrics, however, it is also instructive to examine disaggregate model performance. Figure 6.18 shows a box plot of individual model predictions for the best XGBoost model trained on the L2 cleaned dataset. The plot illustrates a couple of key points, firstly it suggests that the model exhibits slight tendency to over rather than underestimate scope 3 figures. Secondly it shows that there are still a number of data points for which the model does particularly poorly at predicting. For example the five data points that represent a negative 100% error are all data points for which the model outputs a scope 3 value of zero. Furthermore, there are an additional 10 outliers with a percentage error of over 500% that are not shown in the plot but also represent significant over estimations.

It is revealing to look at where the model performs well and where the model performs less well. By comparing actual vs predicted scope 3 data points, there are a couple of insights that can be made. Firstly, and perhaps unsurprisingly, the performance of the model is strongest for companies where lots of other very similar companies also exist in the data set. For example, the two sub industries with the most data points are ’Diversified Mining’ and ’Iron & Steel’ and the model tends to perform well when predicting a scope 3 value for these companies. However, the model exhibits worse performance when applied to predicting scope 3 values for the ’Specialty Mining Metals’ sub industry for which there are only six available data points.

This disaggregation of model performance has provided two key insights into the model’s performance. The first is that the model has uneven performance within the different sub industries of the Mineral Resources sector, and this is largely dictated by the availability of data on different activities. Secondly, the model can miss nuances

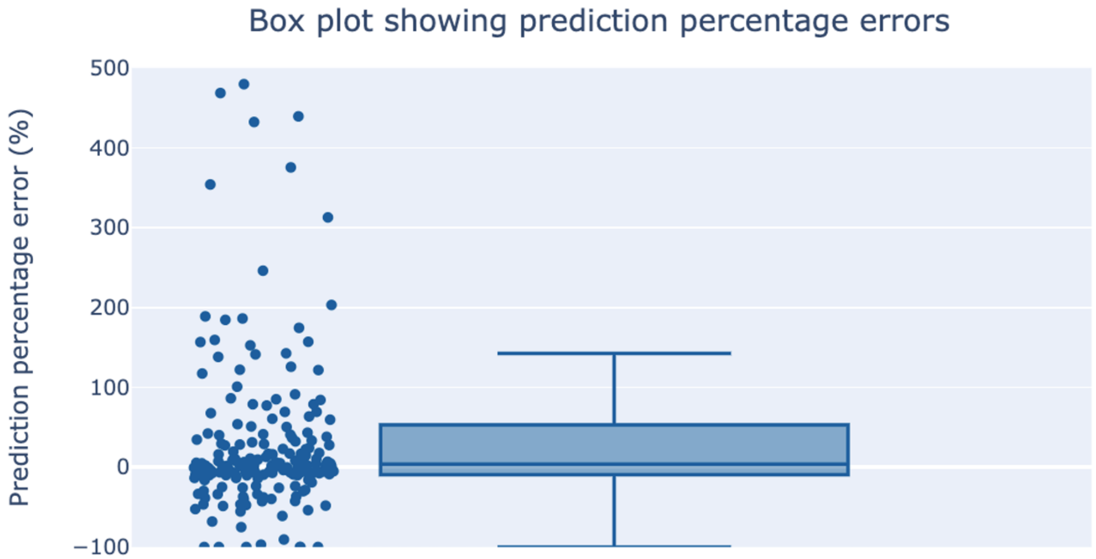


Figure 6.18: Box plot of the percentage errors of all predictions made by the best XG-Boost model on the L2 cleaned dataset.

between firms in the same sub-sector, and this means that caution should be applied if the model is being used to compare individual firms.

6.2.7 Comparison to State-of-the-Art

The best performing XGBoost model from this project was compared to the state-of-the-art model published in the literature[10]. Full details of the comparison experiment setup can be found in Subsection 5.6. The results show that the model developed by this project outperforms the Otago model in each of the three error metrics. MAE drops from 14.9 to 11.4 Mt of CO2e, a 23.5% improvement. RMSE drops from 51.9 to 36.6 Mt CO2e, a 29.5% improvement. The R2 score increases from 0.59 to 0.77 a 30.5% increase in performance. Both papers use the same modelling algorithm, so the key difference in the performance is due to the features used by each of the models. The superior performance of the model in this project adds weight to the conclusion that using gold standard sector-specific data is crucial to model performance.

The Otago paper looked to build a universal Scope 3 model. This is undoubtedly an attractive goal as it means the model can then be used on a wide variety of sectors and has much broader applicability than a sector-specific model. However, this project has demonstrated the performance benefit of taking a sector-specific approach in the case of the Mineral Resources sector. It is hypothesised that a sector-specific approach will also be of value in measuring scope 3 emissions in other sectors, although further research is required to confirm this.

6.2.8 Anomaly Detection

A key issue with scope 3 emissions data is that there are many instances of incomplete reporting. One potential way to tackle this is to use the model developed in this project to flag data points where the model prediction is much higher than the reported figure. This then allows an analyst to manually inspect data points to see if it is an example of incomplete reporting. This type of anomaly detection is important for building confidence in the data.

Figure 6.19 shows how model predictions can identify partial disclosure. The global mining company Anglo American went from partial in 2009 to full disclosure in 2010. The model predicted value for 2009 (141M) is orders of magnitude above the company report value (15k), which suggests, as confirmed by company disclosures, that the Anglo American only released partial scope 3 figures in 2009. The company moves to full scope 3 emissions disclosure in 2010, and here, you can see that model predict value (208M) is much closer to the actual reported value of (177M). While not the main focus of this report, this serves as a proof of concept that the model could be used for anomaly detection.

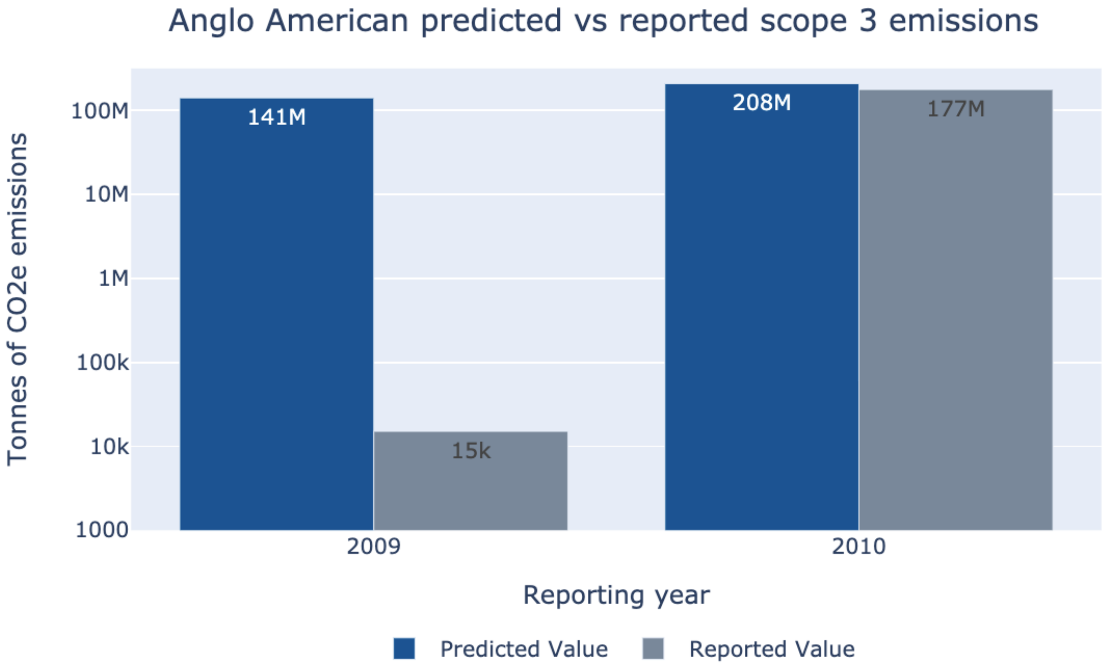


Figure 6.19: Model predicted vs reported values for Anglo American scope 3 emissions. Company moves from partial disclosure in 2009 to full disclosure in 2010.

6.2.9 Year Ahead Forecasting

The year ahead forecasting results demonstrate that it is possible to forecast scope 3 emissions. The model is constrained to using the data provided by analysts forecasts for a company. As expected, the limited number of features means that the model has a coarse accuracy relative to predictions made using company published financial data. Furthermore, there is an additional level of uncertainty introduced by the inevitable inaccuracies of analysts forecasts. By definition, this technique is also limited to companies for whom analysts forecasts are available; this is not the case for all companies, especially smaller companies. The results also reveal an interesting result in terms of which model is best to use for year ahead forecasting. The results indicate that if historical scope 3 data points are available for the company for which you are trying to estimate emissions, then a KNN model with historical data points included should be used. However, if you are trying to forecast emissions for a company for which there are no historical data points, then the XGBoost model is preferable.

6.2.10 Data Quality

Model performance is highly dependent on the quality of the underlying data, and a discussion on data quality can help put the model performance into perspective. As scope 3 data relates to activities outside of a company's direct control and measurement, it is inherently more uncertain than scope 1 and 2 data [14]. Challenges around the opaque visibility of supply chain activities, data collection and protocol interpretation are defining features of scope 3 data [14]. Given these difficulties, is it perhaps unsurprising that some companies only focus on categories of scope 3 emissions that are easy to calculate, like business travel. This highlights two key points that should be considered when working with scope 3 data. Firstly there is considerable uncertainty around even the company published figures, and secondly, it is important to distinguish between complete and incomplete scope 3 reporting. This project has done its best to guard against including incomplete figures in its modelling by running multiple experiments at different levels of data cleaning. However, the project cannot do anything to alter the uncertainty inherent to published scope 3 figures. Attempt to characterise the difficulty and uncertainty in the measurement and calculation of scope 3 figures puts the average errors achieved by this project in a favourable light. Ultimately when the underlying data contains large error margins, this will feed through into the error margins of any attempt to model the data.

6.2.11 Concept Drift

Concept drift refers to the idea that the relationship between model features and the target variables can change over time. In the context of this project, corporate climate emissions commitments are likely to be a key driver of concept drift. Figure 6.20 outlines the climate commitments of five global mining companies, which, if they are to be met, will require major operational evolution within these companies.

The rate at which clean technologies are and will have to advance to meet the climate commitments means that the model will constantly have to evolve its understanding of the relationships between features and the target variable. It should be considered when using the model, particularly for the year ahead predictions, that it has been trained on historical data and learnt relationships between features and the target variable that are unlikely to stay constant over time. This is an example of where the knowledge of human sector expert can come in to integrate any knowledge of recent evolution in supply chain operations to moderate the predictions of the model.

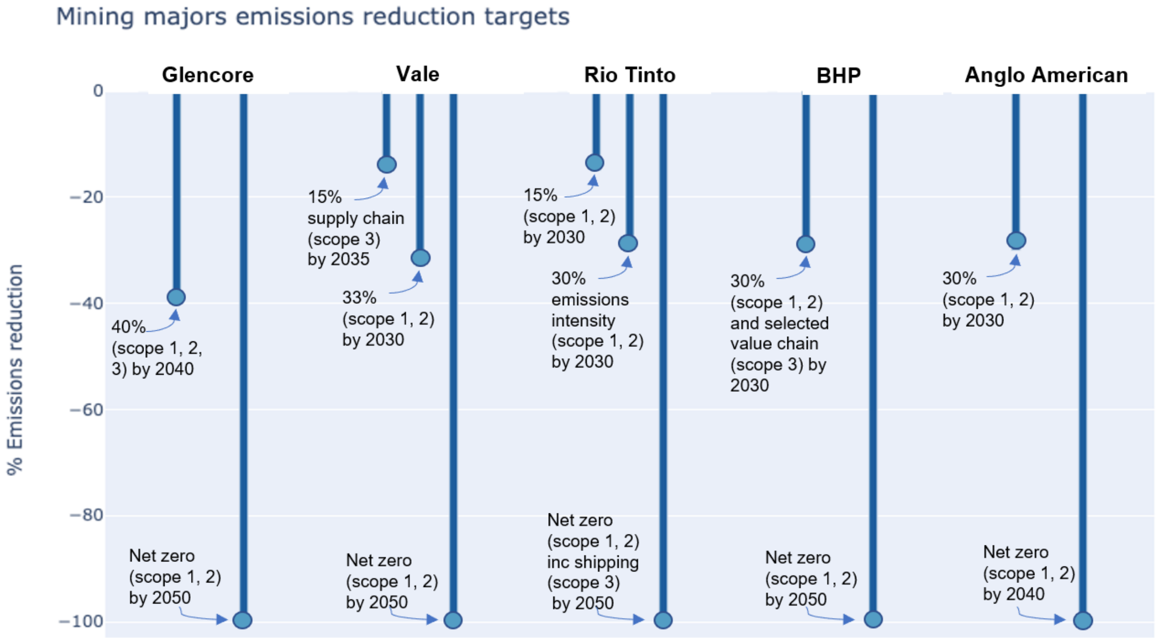


Figure 6.20: Emissions reduction targets of five of the major global publicly traded mining companies. Inspired by chart in [78].

6.2.12 Evaluation of Experiment Design

There are two elements of the experimental design that are useful to evaluate in more detail. The first is the choice of the sector. The Mineral Resources sector is an example of a largely homogeneous sector in which companies use similar processes to produce similar goods. There is far more similarity in the processes and products of different copper mining companies than there is between different consumer goods companies. It is likely that this project has benefited from the relative homogeneity of the mining sector as it is easier for models to learn general relationships that hold over a number of different firms. This is important to consider if a similar approach is adopted to build a model for a different sector.

The second critique of experimental design is the inclusion of historical data points in the data set. By including all historical data points from a given company in the data set, the homogeneity of the underlying data is further increased. This is likely to improve the performance of the model artificially and is acknowledged as a limitation of the methodology. However, given the highly limited nature of available scope 3 data, this was a trade-off that had to be made.

6.2.13 Feedback Loop

The final part of this evaluation section looks to the future to consider how scope 3 emissions are likely to evolve in the coming years and what this means for the models created by this project. This project has been conducted at a time when scope 3 emissions disclosure is still in its early stages, but there are a number of indications that disclosure levels are likely to increase in the future [74]. This increase in disclosures is being driven by increased pressure on firms to respond to climate change-related issues. It is hypothesised, as outlined by Figure 6.21, that this pressure will lead to a positive feedback cycle of increasing disclosures. The increase in the quantity and quality of data [74] that this would bring about is a very positive thing for attempts to model scope 3 emissions as data quantity and quality are two key current limiting factors. While the field of scope 3 emissions modelling is in its infancy and the performance of models remains somewhat coarse, the promise of increasing quantities of data and an increased focus on scope 3 emissions holds great promise for the future evolution of scope 3 modelling.

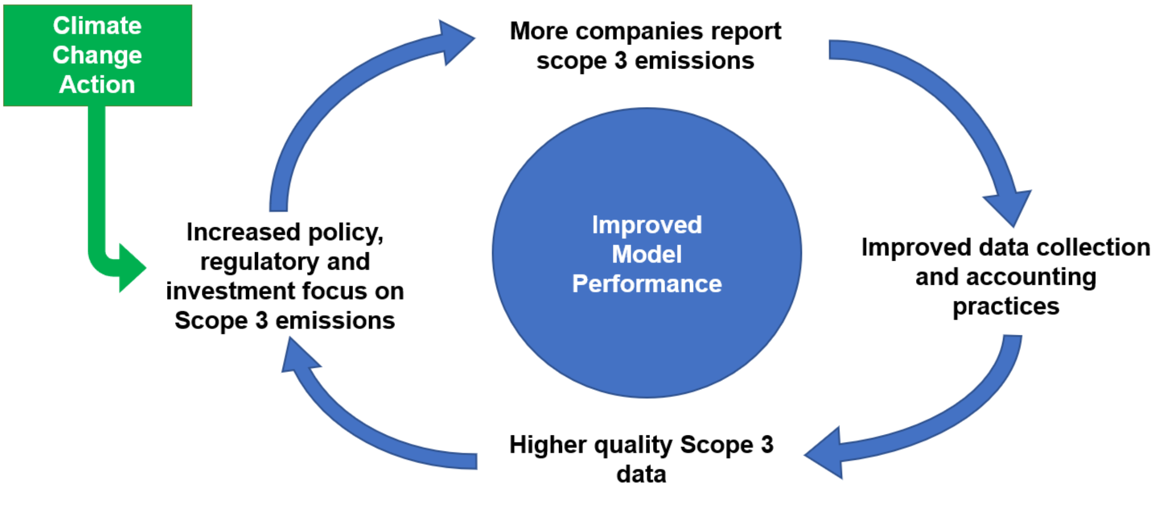

Climate Change Action
More companies report scope 3 emissions
Improved data collection and accounting practices
Higher quality Scope 3 data
Increased policy, regulatory and investment focus on Scope 3 emissions
Improved Model Performance

Figure 6.21: Theorised positive feedback loop for scope 3 emissions disclosure.

Chapter 7

Conclusion and Future Work

7.1 Summary

This project laid out four primary objectives against which its attainment can be summarised and its performance measured. The first objective was to select and engineer features for predicting scope 3 emissions in the Mineral Resources sector. The project has achieved this objective and identified over 80 features for this purpose. Additionally, four different sets of features have been outlined for use in different modelling contexts. The results indicate that sector-specific features drawn from gold standard data improve modelling performance.

The second objective of the project was to implement and identify the best algorithm for predicting scope 3 emissions. The project compared eight different algorithms, and the results clearly show that the XGBoost algorithm is the best for the model's primary purpose of scope 3 emissions prediction. For the auxiliary model use in year ahead forecasting, both XGBoost and KNN offer promising results. The results suggest KNN modelling should be used if historical scope 3 data for a given company is available.

The third objective was to evaluate the model performance against the state-of-the-art from both academic and commercial modelling solutions. A number of the existing commercial scope 3 models sit behind corporate paywalls, which prevented a performance comparison. However, this project was able to compare its results against a baseline of repurposed Refinitiv Scope 1 and 2 models and the state-of-the-art in the academic literature. The results of the model's built in this project show considerable improvement over the Refinitiv baselines, although this was expected given they were designed for scope 1 and 2 emissions. More significantly, the project also demonstrates an over 20% improvement in MAE compared to the existing state-of-the-art in the academic literature.

The fourth objective of the project was to evaluate the ability of the model to be used for the auxiliary purpose of anomaly detection and year ahead forecasting. This objective was achieved, and it was demonstrated that the model can be used for both additional purposes.

Overall the project has been successful in meeting its four original objectives and provides a useful framework for future work on this important topic. As the proverb states, 'what gets measured gets managed', and improving the availability of scope 3 data is an important activity to help combat climate change. With the prospect of increasing scope 3 disclosures and heightened focus from stakeholders on scope 3 emissions, this line of research looks to have a promising future.

7.2 Future Work

There are a number of possible extensions to this project that would be interesting to pursue if more time was available.

7.2.1 Data Augmentation

A key constraint on the project has been the quantity of available data. This project has used Refinitiv as a data source, and Refinitiv only collects scope 3 data on the firms covered as part of its ESG metrics. While this represents 80% of the global market cap, it is likely that scope 3 data points exist outside of the Refinitiv coverage. One way in which data quantities could be improved is to employ web scraping to collect scope 3 data from firms that are not covered by Refinitiv. The lack of standardised scope 3 reporting means that the companies that release scope 3 data often display it in different formats and locations. This would make the web scraping task challenging but very worthwhile if it can increase the quantity of available scope 3 data. Two geographic regions to focus on are Asia and South America for which there are few current data points in the Refinitiv data.

An additional area in which web scraping could be valuable is for production volumes. There are a number of materials, including aluminium, magnesium, silicon, graphite and lithium, for which Refinitiv do not provide production volumes. These production volumes are often available in mining company disclosures and could be scraped. These such data points would be valuable additional features in the scope 3 model.

7.2.2 Trialing a Sector-Specific Approach on Another Industry

This project has taken a sector specific-approach to building a scope 3 emissions model for the Mineral Resources sector. The model has demonstrated improved performance over an existing universal model, and this suggests that there is value in building sector-specific models. To build confidence in the hypothesis that sector-specific modelling results in better scope 3 emissions predictions, it would be useful to test a similar approach on other sectors to see if the improvement in performance can be replicated.

7.2.3 Individual Scope 3 Category Prediction

Refinitiv is planning to start disclosing scope 3 data broken down by each of the fifteen scope 3 categories. Currently, the scope 3 data provided by Refinitiv exists only as an aggregated figure of all the categories. A future version of this project could look at building a model of individual scope 3 categories. This narrows the scope of the task, which could result in improved model performance. Additionally, it would allow for the construction of a hybrid model in which different techniques are combined. For example, it would then be possible to combine machine learning models for some categories with emissions factor-based models for other categories.

7.2.4 Model Explainability

When using a scope 3 emissions prediction model, a key question that stakeholders are likely to have is how does the model come to its prediction. This project has primarily focused on optimising model performance, and further work could be done on model explainability. There are a number of techniques that appear promising for such work. For example, SHAP (Shapley Additive Explanations) [79] can be used to measure the contribution of an individual feature to a final prediction. Such an investigation would make the model more acceptable to users and would reveal which features are crucial to model performance and hence should be the focus of future scope 3 emissions modelling attempts.

7.2.5 Scenario Testing and Stress Testing

A final area in which further research would be valuable is stress testing. Such research could, for example, explore the model's ability to adapt to a change in a companies business strategy or operations. For example, the model's ability to model emissions for a company pre and post-merger would be an interesting experiment and one in which the Refinitiv ESG team expressed particular interest. This could be done through the small number of examples of mergers and changes in business focus that exist in the published data. An additional way this could be tackled is through the generation of artificial data that simulate such events.

Bibliography

[1] Agreement P. Paris agreement. In: Report of the Conference of the Parties to the United Nations Framework Convention on Climate Change (21st Session, 2015: Paris). Retrived December. vol. 4. HeinOnline; 2015. p. 2017. 1, 5

[2] Masson-Delmotte V, Zhai P, Pirani A, Connors S, Péan C, Berger S, et al. Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press: IPCC; 2021. 1

[3] Allen M, Babiker M, Chen Y, Coninck Hd, Connors S, Diemen Rv, et al. Global Warming of 1.5° C. Summary for Policymakers. IPCC; 2018. 1, 5

[4] Krabbe O, Linthorst G, Blok K, Crijns-Graus W, Van Vuuren DP, Höhne N, et al. Aligning corporate greenhouse-gas emissions targets with climate goals. Nature Climate Change. 2015;5(12):1057–1060. 1

[5] TFCD. Recommendations of the Task Force on Climate-related Financial Disclosures. [Accessed 16/05/2021]: Task Force on Climate Related Financial Disclosures; 2017. Available from: `https://assets.bbhub.io/company/sites/60/2020/10/FINAL-2017-TCFD-Report-11052018.pdf`. 1, 5, 6

[6] Bloomberg M, Ultermann A, Buberl T, Starace F, Solomon D, Mizuno H, et al. Financing the Low-Carbon Future A Private-Sector View on Mobilizing Climate Finance. [Accessed 15/05/2021]; 2019. Available from: `https://data.bloomberglp.com/company/sites/55/2019/09/Financing-the-Low-Carbon-Future_CFLI-Full-Report_September-2019.pdf`. 1

[7] Baker B. Scope 3 Carbon Emissions: Seeing the Full Picture. Available. MSCI: [18/07/2021]; 2020. Available from: `https://www.msci.com/www/blog-posts/scope-3-carbon-emissions-seeing/02092372761`. 1

[8] Matthews HS, Hendrickson CT, Weber CL. The importance of carbon footprint estimation boundaries. ACS Publications; 2008. 1, 4

[9] Griffin PA, Lont DH, Sun EY. The relevance to investors of greenhouse gas emission disclosures. Contemporary Accounting Research. 2017;34(2):1265–1297. 2, 19, 26, 27, 29, 32, 33, 34, 73, 74

[10] Nguyen Q, Diaz-Rainey I, Kuruppuarachchi D. Predicting corporate carbon footprints for climate finance risk analyses: A machine learning approach. Energy Economics. 2021 3;95:105129. 2, 17, 19, 24, 26, 27, 29, 32, 33, 34, 36, 42, 56, 73, 74

[11] Economic sector Business sector Industry group Industry Activity PermID ® TRBC Hierarchical ID. Refinitiv: [Accessed 19/05/2021];. Available from: `https://www.refinitiv.com/content/dam/marketing/en_us/documents/quick-reference-guides/trbc-business-classification-quick-guide.pdf`. 2

[12] Delevingne L, Glazener W, Gregoir L, Henderson K. Climate risk and decarbonization: What every mining CEO needs to know. McKinsey: [16/05/2021]; 2020. Available from: `https://www.mckinsey.com/business-functions/sustainability/our-insights/climate-risk-and-decarbonization-what-every-mining-ceo-needs-to-know`. 2

[13] Andrew J, Cortese CL. Carbon disclosures: Comparability, the carbon disclosure project and the greenhouse gas protocol. Australasian Accounting, Business and Finance Journal. 2011;5(4):5–18. 2

[14] WBCSD WRI. The greenhouse gas protocol. A corporate accounting and reporting standard, Rev ed Washington, DC, Conches-Geneva. 2004. 2, 3, 58

[15] Hume N. Vale to set 'Scope 3' emission targets. Financial Times: [Accessed 15/05/2020]; 2019. Available from: `https://www.ft.com/content/00392aec-152c-11ea-8d73-6303645ac406`. 2

[16] Hertwich EG, Wood R. The growing importance of scope 3 greenhouse gas emissions from industry. Environmental Research Letters. 2018 10;13(10):104013. 2, 4

[17] GHG Protocol, Carbon Trust. Technical Guidance for Calculating Scope 3 Emissions. [Accessed 03/06/2021]; 2013. Available from: `https://ghgprotocol.org/sites/default/files/standards/Scope3_Calculation_Guidance_0.pdf`. 3

[18] Lee KH. Integrating carbon footprint into supply chain management: the case of Hyundai Motor Company (HMC) in the automobile industry. Journal of cleaner production. 2011;19(11):1216–1223. 5

[19] Jung J, Herbohn K, Clarkson Carbon Risk, Carbon Risk Awareness and the Cost of Debt Financing. Journal of Business Ethics. 2018;150. 6

[20] Bloomberg Intelligence. ESG assets may hit $53 trillion by 2025, a third of global AUM. Bloomberg: [15/08/2021]; 2021. Available from: `https://www.bloomberg.com/professional/blog/esg-assets-may-hit-53-trillion-by-2025-a-third-of-global-aum/`. 6

[21] Nauman B. US may join Europe in mandating climate risk disclosures. Financial Times: [Accessed 21/05/2021]; 2021. Available from: https://www.ft.com/content/77a8292d-2e7f-43a1-9062-2e639c1e6b2a. 6

[22] Porter ME, Reinhardt FL, Schwartz P, Esty DC, Hoffman AJ, Schendler A, et al. Climate business— business climate. Harvard Business Review. 2007;1. 6

[23] Russell S, Norvig P. Artificial intelligence: a modern approach. 2002. 9, 10, 11, 13, 15, 31, 40

[24] Freedman DA. Statistical models: theory and practice. Cambridge University Press; 2009. 10

[25] Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology). 2005;67(2):301–320. 10, 40, 51

[26] Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician. 1992;46(3):175–185. 11, 31

[27] Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V. Support vector regression machines. Advances in neural information processing systems. 1997;9:155–161. 11

[28] Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995;20(3):273–297. 11

[29] Drucker H. Improving regressors using boosting techniques. In: ICML. vol. 97. Citeseer; 1997. p. 107–115. 12

[30] Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep learning. vol. 1. MIT press Cambridge; 2016. 12, 14, 15, 16, 18, 29, 53

[31] Vapnik VN. An overview of statistical learning theory. IEEE transactions on neural networks. 1999;10(5):988–999. 12

[32] Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. CRC press; 1984. 12

[33] Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. vol. 1. Springer series in statistics New York; 2001. 13, 40, 52

[34] Freund Y, Schapire RE. Experiments with a new boosting algorithm. In: icml. vol. 96. Citeseer; 1996. p. 148–156. 13, 14

[35] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016. p. 785–794. 13, 14, 31

[36] LeCun Y, Bengio Y, Hinton G. Deep learning. nature. 2015;521(7553):436–444. 14, 16, 18

[37] Zhang Z, Geiger J, Pohjalainen J, Mousa AED, Jin W, Schuller B. Deep learning for environmentally robust speech recognition: An overview of recent developments. ACM Transactions on Intelligent Systems and Technology (TIST). 2018;9(5):1–28. 14

[38] Minaee S, Boykov YY, Porikli F, Plaza AJ, Kehtarnavaz N, Terzopoulos D. Image segmentation using deep learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021. 14

[39] Funahashi KI. On the approximate realization of continuous mappings by neural networks. Neural networks. 1989;2(3):183–192. 15

[40] Géron A. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media; 2019. 15, 16, 40, 52

[41] Pasupa K, Sunhem W. A comparison between shallow and deep architecture classifiers on small dataset. In: 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE); 2016. p. 1–6. 16

[42] Ryan M. Deep learning with structured data. Simon and Schuster; 2020. 16

[43] Rolnick D, Donti PL, Kaack LH, Kochanski K, Lacoste A, Sankaran K, et al. Tackling Climate Change with Machine Learning. arXiv preprint arXiv:190605433. 2019 6. 17

[44] Gentine P, Pritchard M, Rasp S, Reinaudi G, Yacalis G. Could machine learning break the convection parameterization deadlock? Geophysical Research Letters. 2018;45(11):5742–5751. 17

[45] Karagiannopoulos S, Aristidou P, Hug G. Data-driven local control design for active distribution grids using off-line optimal power flow and machine learning techniques. IEEE Transactions on Smart Grid. 2019;10(6):6461–6471. 17

[46] Huang JC, Tsai YC, Wu PY, Lien YH, Chien CY, Kuo CF, et al. Predictive modeling of blood pressure during hemodialysis: a comparison of linear model, random forest, support vector regression, XGBoost, LASSO regression and ensemble method. Computer Methods and Programs in Biomedicine. 2020;195:105536. 17, 52

[47] Hafeez S, Wong MS, Ho HC, Nazeer M, Nichol J, Abbas S, et al. Comparison of machine learning algorithms for retrieval of water quality indicators in case-II waters: a case study of Hong Kong. Remote sensing. 2019;11(6):617. 17

[48] Li L. Geographically weighted machine learning and downscaling for high-resolution spatiotemporal estimations of wind speed. Remote Sensing. 2019;11(11):1378. 17

[49] Ahmad IS, Bakar AA, Yaakub MR, Muhammad SH. A survey on machine learning techniques in movie revenue prediction. SN Computer Science. 2020;1(4):1–14. 17, 18

[50] Lago J, De Ridder F, De Schutter B. Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms. Applied Energy. 2018;221:386–405. 17

[51] Chen R, Liang CY, Hong WC, Gu DX. Forecasting holiday daily tourist flow based on seasonal support vector regression with adaptive genetic algorithm. Applied Soft Computing. 2015;26:435–443. 17

[52] Storm H, Baylis K, Heckelei T. Machine learning in agricultural and applied economics. European Review of Agricultural Economics. 2020;47(3):849–892. 17

[53] Robinson C, Dilkina B, Hubbs J, Zhang W, Guhathakurta S, Brown MA, et al. Machine learning approaches for estimating commercial building energy consumption. Applied energy. 2017;208:889–904. 18

[54] Pinter G, Felde I, Mosavi A, Ghamisi P, Gloaguen R. COVID-19 pandemic prediction for Hungary; a hybrid machine learning approach. Mathematics. 2020;8(6):890. 18

[55] Bakay MS, Ağbulut Electricity production based forecasting of greenhouse gas emissions in Turkey with deep learning, support vector machine and artificial neural network algorithms. Journal of Cleaner Production. 2021;285:125324. 18

[56] Nosratabadi S, Mosavi A, Duan P, Ghamisi P, Filip F, Band SS, et al. Data science in economics: comprehensive review of advanced machine learning and deep learning methods. Mathematics. 2020;8(10):1799. 18

[57] Saleh C, Dzakiyullah NR, Nugroho JB. Carbon dioxide emission prediction using support vector machine. In: IOP Conference Series: Materials Science and Engineering. vol. 114. IOP Publishing; 2016. p. 012148. 18

[58] Gore A, McCormick G. We Can Solve the Climate Crisis by Tracing Pollution Back to Its Sources. A New Coalition Will Make It Possible.. [Accessed 22/05/2021]; 2020. Available from: https://medium.com/@algore/we-can-solve-the-climate-crisis-by-tracing-pollution-back-to-its-sources-4f535f91a8dd. 18

[59] Saleh C, Chairdino Leuveano RA, Ab Rahman MN, Md Deros B, Dzakiyullah NR. Prediction of CO2 emissions using an artificial neural network: The case of the sugar industry. Advanced Science Letters. 2015;21(10):3079–3083. 18

[60] Kadam P, Vijayumar S. Prediction Model: CO 2 Emission Using Machine Learning. In: 2018 3rd International Conference for Convergence in Technology (I2CT). IEEE; 2018. p. 1–3. 18

[61] Acheampong AO, Boateng EB. Modelling carbon emission intensity: Application of artificial neural network. Journal of Cleaner Production. 2019;225:833–856. 18

[62] Goldhammer B, Busse C, Busch T. Estimating corporate carbon footprints with externally available data. Journal of Industrial Ecology. 2017;21(5):1165–1179. 19, 26, 27, 32, 33, 34, 73

[63] Refinitiv ESG carbon data and estimate models. Refinitiv: [Accessed 14/05/2021];. Available from: `https://www.refinitiv.com/content/dam/marketing/en_us/documents/fact-sheets/esg-carbon-data-estimate-models-fact-sheet.pdf`. 20, 32, 33, 39, 73

[64] Shakdwipee M, Lee LE. Filling the blanks: comparing carbon estimates against disclosures msci ESG Research Issue Brief; 2016. 21

[65] Bokern D, Hadjikyriakou P, Klug AP. Scope 3 Carbon Emissions Estimation Methodology MSCI ESG Research LLC; 2021. 21

[66] Bloomberg. A Bloomberg Professional Services Offering Distributional Greenhouse Gas Emissions Estimates Data Challenges And Modeling Solutions. [Accessed 15/05/2021]; 2021. Available from: `https://www.bloomberg.com/professional/sustainability-data-solutions-greenhouse-gas-ghg-emissions/?bbgsum-page=DG-WS-PROF-BLOG-POST-107272&mpam-page=21140&tactic-page=429986`. 21

[67] Trucost. Trucost Environmental Register Methodology FAQs (Downloaded from website). [01/09/2021]; 2019. Available from: `www.trucost.com`. 22

[68] ISS ESG. ISS ESG Methodology Carbon Footprinting. [01/09/2021]; 2019. Available from: `https://www.issgovernance.com/esg/climate-solutions/climate-analytics/`. 22

[69] CDP. CDP Full GHG Emissions Dataset, Technical Annex IV: Scope 3 Overview and Modelling. [01/07/2021]; 2020. Available from: `www.cdp.net`. 22, 33

[70] Quanits, GHG Protocol. Documentation of the data and calculations to support the Greenhouse Gas Protocol Scope 3 Screening Tool . [21/05/2021]; 2017. Available from: `https://quantis-suite.com/Scope-3-Evaluator/`. 22

[71] High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworth AI. European Comission: [17/08/2021]; 2019. Available from: `https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai`. 23

[72] James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. vol. 112. Springer; 2013. 23, 35, 37

[73] An overview of Environmental, Social and Corporate Governance - ESG . Refinitiv: [Accessed 19/05/2021];. Available from: `https://www.refinitiv.com/en/financial-data/company-data/esg-data`. 24

[74] Ducoulombier F. Understanding the Importance of Scope 3 Emissions and the Implications of Data Limitations. The Journal of Impact and ESG Investing. 2021. 25, 60

[75] OceanaGold. OceanaGold 2020 Sustainability Report. [Accessed 01/07/2021]; 2021. Available from: `https://oceanagold.com/wp-content/uploads/2021/06/OceanaGold-2020-Sustainability-Report.pdf`. 25

[76] Wang J, Cully A, Rei M. Introduction To Machine Learning. Imperial College MSc Computing Course Notes. London: Unpublished; 2021. 39

[77] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research. 2014;15(1):1929–1958. 40

[78] Hume N. Miners face up to climate challenge. Financial Times: [01/08/2021]; 2021. Available from: `https://www.ft.com/content/8469ef8b-86a1-4260-a280-2a18ed19b2ef`. 59

[79] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Proceedings of the 31st international conference on neural information processing systems; 2017. p. 4768–4777. 63

Appendix A

Appendix

Full Overview of Features

	Feature	Relationship to scope 3
1	TRBC industry [63]	Business model
2	Total CO2	Company output and Environmental focus
3	Leverage % [10]	Business model
4	Intangibles [10, 9]	Environmental focus - higher intangibles often means better branding and R&D
5	Categorical combined score	Environmental focus
6	Categorical upstream score	Environmental focus
7	Categorical general score	Environmental focus
8	Categorical downstream score	Environmental focus
9	Revenue adjusted for commodity price	Company output
10	Total fixed assets [10]	Company size
11	Property, plant & equipment [10] CO2	Company size
12	Operating expense	Company output
13	Gross margin [62, 9, 10]	Business model
14	Year [10]	Environmental focus - more focus in recent years
15	Industry group [63]	Business model
16	Activity [63]	Buisness model
17	Scope 1	Company output and Environmental focus
18	Scope 2	Company output and Environmental focus
19	Employees [63, 10]	Company size
20	Total waste (Tonne)	Company size and output
21	Energy use total (Gigajoule) [63, 10]	Company size and output
22	Capital intensity [62, 10]	Business model
23	Material expenses	Company output
24	R&D expense	Environmental focus
25	Construction in progress	Current business situation
26	Depreciation [9]	Company size
27	Operating lease payments total	Business model
28	Exchange country	Environmental focus
29	Cement (Tonne)	Company output
30	Steel (Tonne)	Company output
31	Coal (Tonne)	Company output
32	Crude oil (Barrel)	Company output
33	Natural gas (Cubic Meter)	Company output
34	Combined oil and gas (BOE)	Company output
35	Iron ore (Tonne) CO2	Company output
36	Copper (Tonne)	Company output
37	Manganese ore (Tonne)	Company output
38	Zinc (Tonne)	Company output
39	Nickel (Tonne)	Company output
40	Titanium slag (Tonne)	Company output
41	Silver (Troy Ounce)	Company output
42	Gold (Troy Ounce)	Company output
43	Diamonds (Carat)	Company output
44	Platinum (Troy Ounce)	Company output
45	Non-Hazardous waste	Company size
46	Hazardous waste	Environmental focus
47	Environmental expenditures	Environmental focus

48	Water withdrawal total	Company size and business model
49	Water discharged	Company size and business model
50	Cement CO2 emissions	Company size and business model
51	Environmental pillar score	Environmental focus
52	Resources use score	Environmental focus
53	Renewable energy use ratio	Environmental focus
54	Emissions reduction target percentage	Environmental focus
55	Emissions reduction target year	Environmental focus
56	Environmental provisions	Environmental focus
57	Sustainable building products	Environmental focus
58	Resource reduction policy	Environmental focus
59	Resource reduction targets	Environmental focus
60	Environmental materials sourcing	Environmental focus
61	Green buildings	Environmental focus
62	Environmental products	Environmental focus
63	Eco-Design products	Environmental focus
64	Take-back and recycling initiatives	Environmental focus
65	Products environmental responsible use	Environmental focus
66	Policy sustainable packaging	Environmental focus
67	Renewable energy use	Environmental focus
68	Policy environmental supply chain	Environmental focus
69	Environment management team	Environmental focus
70	Environment management training	Environmental focus
71	Staff transportation impact reduction	Environmental focus
72	Environmental supply chain management	Environmental focus
73	Environmental supply chain monitoring	Environmental focus
74	Environmental supply chain termination	Environmental focus
75	Land environmental impact reduction	Environmental focus
76	Policy emissions	Environmental focus
77	Climate change commercial risks opportunities	Environmental focus
78	Environmental expenditures investments	Environmental focus
79	Environmental investments initiatives	Environmental focus
80	Environmental partnerships	Environmental focus
81	Internal carbon pricing	Environmental focus
82	Assets age [10, 9]	Environmental focus and business model

Table A.1: Overview of all features extracted from Refinitiv Workspace. Where a feature has been used in a previous academic emissions model it has been cited. The relationship to scope 3 column outlines how a feature relates to a business characteristic that impacts scope 3 emissions. Discussion of business characteristics can be found in Section 3.3.

Chosen Model Parameters

Feature set	Search Space	L1 clean	L2 clean	L3 clean
Elastic Net Regression				
Core features	alpha: [0.00001, 0.0001, 0.001, 0.01, 0.1, 0.0, 1.0, 10.0, 100.0]	100.0	100.0	100.0
	l1_ratio: [0, 1, 0.1]	0.0	0.0	0.0
GS features	alpha: [0.00001, 0.0001, 0.001, 0.01, 0.1, 0.0, 1.0, 10.0, 100.0]	10	0.1	0.1
	l1_ratio: [0, 1, 0.1]	0.6	0.6	0.1
KNN Regression				
No Historical	k: [1, 20, 1]	3	3	3
	weights: [’uniform’, ’distance’]	’distance’	’distance’	’distance’
Inc Historical	[1, 20, 1]	3	3	3
	weights: [’uniform’, ’distance’]	’distance’	’distance’	’distance’
SVR Regression				
Core features	kernel: [’poly’, ’rbf’, sigmoid’]	’ploy’	’ploy’	’poly’
	C: [100, 50, 10, 1.0, 0.1, 0.01, 0.001]	’100’	’100’	’100’
GS features	kernel: [’poly’, ’rbf’, sigmoid’]	’sigmoid’	’sigmoid’	’sigmoid’
	C: [100, 50, 10, 1.0, 0.1, 0.01, 0.001]	’100’	’100’	’100’
AdaBoost				
Core features	n_estimators: [20, 50, 100]	20	50	100
	learning_rate = [0.0001, 0.001, 0.01, 0.1, 1.0]	0.0001	0.0001	0.1
GS features	n_estimators: [20, 50, 100]	20	10	100
	learning_rate = [0.0001, 0.001, 0.01, 0.1, 1.0]	1.0	0.1	0.1
XGBoost				
Core features	eta: [0.001, 0.05, 0.01,0.1,0.2,0.3,0.4,0.5]	0.05	0.5	0.5
	max_depth: [3,4,5,6,7]	7	7	7
	min_child_weight: hp.uniform(1,3)	1.77	1.53	1.83
	subsample: hp.uniform(0.5, 1)	0.62	0.93	0.98
	colsample_bytree: hp.uniform(0.5, 1)	0.88	0.84	0.60
	gamma: hp.uniform(’gamma’, 0, 0.5)	0.07	0.43	0.46
GS all features	eta: [0.001, 0.05, 0.01,0.1,0.2,0.3,0.4,0.5]	0.3	0.5	0.3
	max_depth: [3,4,5,6,7]	6	7	7
	min_child_weight: hp.uniform(1,3)	0.92	1.35	1.25

	subsample: hp.uniform(0.5, 1)	1	0.75	0.99
	colsample_bytree: hp.uniform(0.5, 1)	1	0.89	0.66
	gamma: hp.uniform('gamma', 0, 0.5)	0.02	0.28	0.47
Feed-Forward Neural Network				
Core features	num hidden layers:	4	4	4
	size of hidden layers:	[128, 64, 32, 16]	[128, 64, 32, 16]	[128, 64, 32, 16]
	num dropout layers	v	4	4
	size dropout layers	[0.2, 0.2, 0.2, 0.2]	[0.2, 0.2, 0.2, 0.2]	[0.2, 0.2, 0.2, 0.2]
	activation function:	'relu'	'relu'	'relu'
	initial learning rate:	0.005	0.005	0.005
	learning rate optimiser:	'Adam'	'Adam'	'Adam'
	epochs:	100	100	100
	batch size:	32	32	32
GS features	num hidden layers:	4	4	4
	size of hidden layers:	[128, 64, 32, 16]	[128, 64, 32, 16]	[128, 64, 32, 16]
	num dropout layers	v	4	4
	size dropout layers	[0.2, 0.2, 0.2, 0.2]	[0.2, 0.2, 0.2, 0.2]	[0.2, 0.2, 0.2, 0.2]
	activation function:	'relu'	'relu'	'relu'
	initial learning rate:	0.005	0.005	0.005
	learning rate optimiser:	'Adam'	'Adam'	'Adam'
	epochs:	100	100	100
	batch size:	32	32	32
Convolutional Neural Network				
Core features	Order of hidden layers	[Conv, Pool, Dense, Drop, Dense, Drop]	[Conv, Pool, Dense, Drop, Dense, Drop]	[Conv, Pool, Dense, Drop, Dense, Drop]
	size of conv layers:	[128]	[128]	[128]
	size of pooling layers:	[3]	[3]	[3]
	size of dense layers:	[64, 32]	[64, 32]	[64, 32]
	dropout size:	[0.2, 0.2]	[0.2, 0.2]	[0.2, 0.2]
GS features	Order of hidden layers	[Conv, Pool, Conv, Pool, Conv, Pool Dense, Drop, Dense, Drop]	[Conv, Pool, Conv, Pool, Conv, Pool Dense, Drop, Dense, Drop]	[Conv, Pool, Conv, Pool, Conv, Pool Dense, Drop, Dense, Drop]
	size of conv layers:	[64, 128, 256]	[64, 128, 256]	[64, 128, 256]
	size of pooling layers:	[3, 2, 2]	[3, 2, 2]	[3, 2, 2]
	size of dense layers:	[128, 32]	[128, 32]	[128, 64]
	dropout size:	[0.2, 0.2]	[0.2, 0.2]	[0.2, 0.2]

Table A.2: Full table of chosen parameters for each model used in this project.