Grant Thornton

Financial Services Advisory

Scope Emissions Prediction Model
Linear Models vs Machine Learning Approach

How Grant Thornton can help you to estimate scope emissions for companies that do not report these metrics and incorporate these into banks' Climate & Risk Quantification framework

October 2022

Introduction

In this paper, we focus on the development of a model that predicts scope emissions for any company that cannot currently provide this information.

Our model is built on external data and estimates relationships between these corporate entities' financial information, sector, region and other information, and scope emissions reported by these entities.

The output of the model can be used by both individual entities as well as banks and other financial institutions in order to estimate the emissions structure of their portfolios. In our methodology we compare simple linear models as well as more advanced machine learning techniques.

The best model estimates are achieved using a Linear Mixed Models Methodology.

Industry-wide Challenge

Investors, regulators and leaders are seeking comprehensive disclosures more than ever before with regards to C&E indicators, but the picture is incomplete since many companies don't report these values.

1

Our Response

In order to respond to this challenge we propose a modelling framework using external data to predict the unreported C&E indicators for other companies based on available financial, sector, and region factors.

2

Practical Application

Individual entities can estimate their scope emission structures while banks can understand the emission intensity of their portfolios and support internal modelling and disclosure capabilities.

3

The publication is structured as follows: In the first section, we outline key modelling methodologies considered for the model build. After this, we outline the model build process, key modelling inputs and outputs. The third section aims to select the best model for emission prediction. In the final section, we focus on the practical implementation and deployment of this model. We also seek to conclude and point out potential issues and areas for further development.

Background

Climate & Environmental Data Requirement Challenge

Over the last 3 years we were able to observe a significant increase in regulatory as well as industry demand for addressing challenges associated with climate change and sustainability. Individual entities as well as regulators displayed commitment to face and cope with negative consequences of climate change. Requirements have focused on improving processes, operating models and culture in order to decrease environmental footprints in line with long term carbon neutrality goals. Significant change in financial sector regulation is motivating the banking institutions, as well as individual borrowers, to provide, collect, and, disclose climate and environmental data.

There is an expectation to include this data into companies' and/or banks' decision making and strategies. Availability of this data and processes to collect, store and provide this data is one of the biggest challenges to date. In order to contribute to the solution, we have developed a model that is based on financial and non-financial company inputs and is able to predict company scope emissions.

Considering the key added value of our research, individual entities can use our model to estimate their scope emission structure while banks can understand the emission intensity of their portfolios and support internal modeling and disclosures capabilities.

We would like thank to Solmaz Panahi and Victor Hugo Nagahama, both PhD students from Maynooth University, for their significant methodological and practical contribution to this publication.

Grant Thornton Machine Learning approach and services

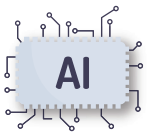
Grant Thornton FSA has significant experience with building and validating IRB, IFRS9 as well as Stress Testing models. Data preparation is essential for each model development where we have a proven record in data analytics and remediation. In our approach we focus on the key benefits and challenges regarding the use of machine learning techniques in regulatory as well as non-regulatory modelling.

How can Grant Thornton help to design Machine Learning strategy and build/validate models and remediate missing or incomplete data

Strategy Design

Determine best use of Machine Learning

- Use of Machine Learning techniques to achieve competitive advantage for bank strategic portfolios
- Cost-benefit function maximisation



Model Build / Validation

Implement machine learning model build or validation

- Use of appropriate Machine Learning techniques
- Model design and implementation
- Validation of existing Machine Learning techniques
- Build of challenger Machine Learning models

Data Solutions

Use machine learning techniques to improve data quality

- Data gathering and processing
- Solutions for historical data remediation
- Solution for missing or incomplete data

Scope Emissions Model - Estimation Approach

In this publication, firstly we define key challenge regarding emission data quality and availability in the context of key regulatory requirements. Consequently, we specify a modelling solution in order to estimate scope emissions for a particular company based on its key financial, sector and geography information. Lastly, we compare the performance of linear vs machine learning based models and point out key modelling challenges, potential benefits and future areas for research in this topic.

Climate & Environmental data challenge – How to estimate scope emissions

C&E Data Challenge

Methodological approach for emission model design

Our Approach

Comparison of linear vs. machine learning models, key benefits and challenges

Linear vs. ML Models

Application and key areas for further research and development in this areas

Application

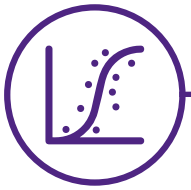
Methodological options

In line with technical progress, faster and more accessible computational power and enhanced options for data storage and manipulation there has been significant progress in modelling methodologies within both industry and academia.
The below information contrasts simple and more advanced modeling techniques we considered and its applicability for solving key modelling challenges.

Linear models

Up to now the use of advanced machine learning techniques in modelling has been limited due to mostly regulatory requirements and in particular, the difficulty in interpreting and explaining the functionality of more complex models. The standard suite of modelling techniques in industry today mainly consists of logistic regressions, linear regressions and decisions trees.

Logistic Regression



Logistic regression is typically used in the context of building probability of default (PD) models within credit risk.
This machine learning technique is commonly used for binary classification problems such as predicting whether default will occur.

Linear Regression

Linear regression is the most commonly used model for time series prediction models. Linear regression allows for a high degree of automation in the development process and are relatively quick to build and easy to understand, adjust and enhance e.g., Error Correction Models, SUR, MLL.

Decision Trees

Some benefits of this approach are ease of interpretability for stakeholders and their non-sensitivity to distributions or statistical instabilities.
However, it requires a lot of manual work compared to other more advanced machine learning techniques.

Machine Learning Techniques

Machine learning is a subset of artificial intelligence which focuses on the use of data and algorithms to automate the process of analytical model building which allows machines to learn from data, identify patterns and make predictions with minimal human intervention. The ability of machine learning algorithms to identify patterns and make predictions improves both with the use of the algorithms and with the quality of the data provided to them. The following sections give an overview of the different types of machine learning techniques. Certain methods can be used in both supervised and unsupervised learning e.g., stacking.

Supervised Learning

The algorithm learns by adjusting its rules through an error function with the goal to minimise/eliminate the error.

Classification
- Neural Networks/Deep Learning
- K-Nearest Neighbours
- Decision Trees
- Support Vector Machines

Regression
- Linear Regression
- Logistic Regression

Ensemble Methods
- Bagging/Random Forests

Unsupervised Learning

The algorithm learns from a training dataset which has no target variable. The goal is to understand the distribution of the data in terms of interpretable patterns, associations and descriptive properties.

Classification
- K-Means Clustering
- Fuzzy C-Means

Pattern Search
- Apriori

Dimension Reduction
- Principal Component Analysis

Reinforcement Learning

The algorithm learns from interacting with the environment rather than from a training dataset and does not require a target variable. The algorithm learns to perform a specific task by trial and error.

- Q-Learning
- State-Action-Reward-Sate-Action (SARSA)
- Deep Q-Networks

C&E data challenge

Significant change in financial sector regulation is motivating banking institutions as well as individual borrowers to provide, collect and disclose climate and environmental data. There is an expectation to include this data into companies/banks' decision making and strategies. Availability of this data and processes to collect store and provide this data is one of the biggest challenges to date.
In order to contribute to the solution we have developed a model that based on financial and non-financial company inputs is able to predict the company's scope emissions.

Regulatory / Industry data requirement

2022 ECB Climate Stress Test Exercise – (H1 2022)

- Evaluation of bank's exposure to sectors (mapping to NACE codes) and Countries of Risk
- Financed greenhouse gas emissions (Scope 1, 2 and 3 GHG emissions)
- Interest, fee and commission income from greenhouse gas intensive industries
- Counterparties' revenues
- Credit Risk Parameters, LTV, Funded Collateral, Collateral NUTS3 location, EPC rating

ESG Disclosures - Pillar III (EOY 2022 - semi-annual thereafter, transition period until 2024)

Ensures institutions are embedding sustainability considerations in their risk management, business models, and strategy and their pathway toward the Paris Agreement goals

- EU Taxonomy aligned financial assets
- Green-Asset Ratio (GAR) on NFRD Corporates and Retail financing (Dec 2023)
- Banking book taxonomy alignment ratio (BTAR) non-NFRD corporates (Jun 2024)

Use of Proxies

Missing GHG emissions (tCO2e) data :

- Scope 1, 2 : Banks may exceptionally use proxies
- Scope 3: Banks can use proxies

Proxy Method Example:

- Economic activity-based emissions
- Physical activity-based emissions
- Average sector-based emissions

Missing EPC Rating for Real Estate Collateralised Exposures:

- Use provisional rating if final is not available
- Use Estimation approach (internal methodology in line with national regulations)
- Report as "Unknown" exceptionally

Proxy Method Example:

- Association of EPC with building period of property
- Association of EPC with footage of property
- Association of EPC with energy costs of building (euro/m^2)

Grant Thornton's view is that proxies can be practically operationalised:

- Using data from external data providers based on representative sample.
- Estimation of proxies for specific Countries /Industries /Size.
- Ensure conservativeness in use of proxies.
- Proxy methodologies expected to be transparent, robust and disclosed in detail.

1 Definition of Scope

There is an increased demand for C&E data for risk management and disclosures by the regulation and industry.
Industry is missing infrastructure, processes and procedures for entities to provide this data.
Reliance on the use of C&E data proxies.

2 Current Solution

Use of the proxies with lack of accuracy and consistency in approach.

Proxy methodologies are still at an early stage and restricted to average based or simple linear models with low prediction power.

3 Our Approach

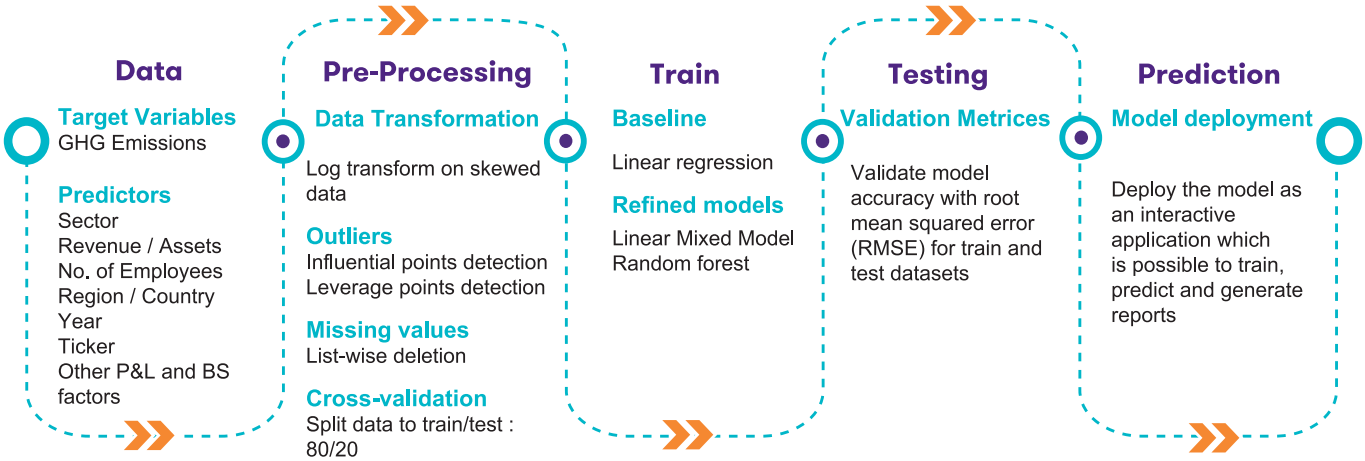
Scope emission prediction model based on corporate data using simple linear models and advanced ML estimation techniques
Tradeof between simplicity and accuracy of proxies estimates modelling solutions.

4 Implementation Solution

We can provide fully integraded R-Shiny based solution for banks and companies in order to work out scope emissions of their portfolios.

Our approach for C&E data modelling estimation

We propose a series of statical learning models to predict the GHG Emission Scope 1 for non-disclosing companies. Unlike current approaches, which tend to construct one model for each industry, our best proposed model make use of all available information while considering the intrinsic effect of each company. The proposed (LMM) model not only uses industries as a factor but also uses longitudinal feature of each companies. We also keep model complexity to a minimum for interpretability purposes. The proposed model also outperforms tree-based models.



Data transformation

The distribution of recorded variables are highly skewed; therefore, we use the log transformation on data to get more symmetrical distribution. For final interpretations, the values are transformed back to the original scale of data.

Model predictors

We tested many possible predictors and found a model that has good accuracy. We choose minimal models to ensure simplicity and interpretability. Final model for scope 1 relies on company-specific predictors such as company sales, for getting a sense of its size, number of employees, tangible assets, Industry sector and regional (or country) predictors, that provides insights into the local operating environment. Since data has a time-series feature (longitudinal), we included year as well. Our analysis proves that companies' sales and sector are the most important variable that can capture the most variance in the data.

Model training and validation

A predictive model should generate accurate predictions for the data that is used to train the model (observed data) but also for new observations that it has never seen before. It is possible to simulate this scenario by splitting the data in two sub datasets: the training set which is used to train the model and the test set which is the unseen data. To train model, we split disclosing companies' data into train and test samples with 80 /20 ratio and stratified sampling to force the distribution of the target variable among the different sectors to be the same. The quality of the model can be evaluated using one of many error metrics which is based on the difference of the actual value and its prediction.

Modelling options summary

We opt for three different models: Linear regression as a baseline, Linear mixed model to deal with our longitudinal data, and Random Forest model. Following table gives as summary of these models.

	Linear Model (LM)	Linear Mixed Models (LMM)	Random Forest
How the model works	Given a list of predictors, find a line which minimise the sum of distance between the actual values and their corresponding points.	It is an extension of LM, which assumes that observations for the same groups are correlated. Useful for longitudinal and hierarchical data.	Based on a sample of data, a decision tree generates splits minimising the error between the actual value and the mean value of this group. This process is repeated many times and the final prediction is the average of the prediction of the individual trees.
Pros	• The relationship between variables is interpretable • Provides a confidence measurement for the results • There are many extensions and variations that make it possible to model the phenomenon of the data		• Requires little data preparation • Simple to understand the method idea • Robust for outliers
Cons	• Only captures linear relationships • Sensitive for outliers • Can require many data preparation • Requires to check the model assumptions (diagnostics)		• Can overfit to the data • Can be unstable • Hard to interpret predictions for big trees

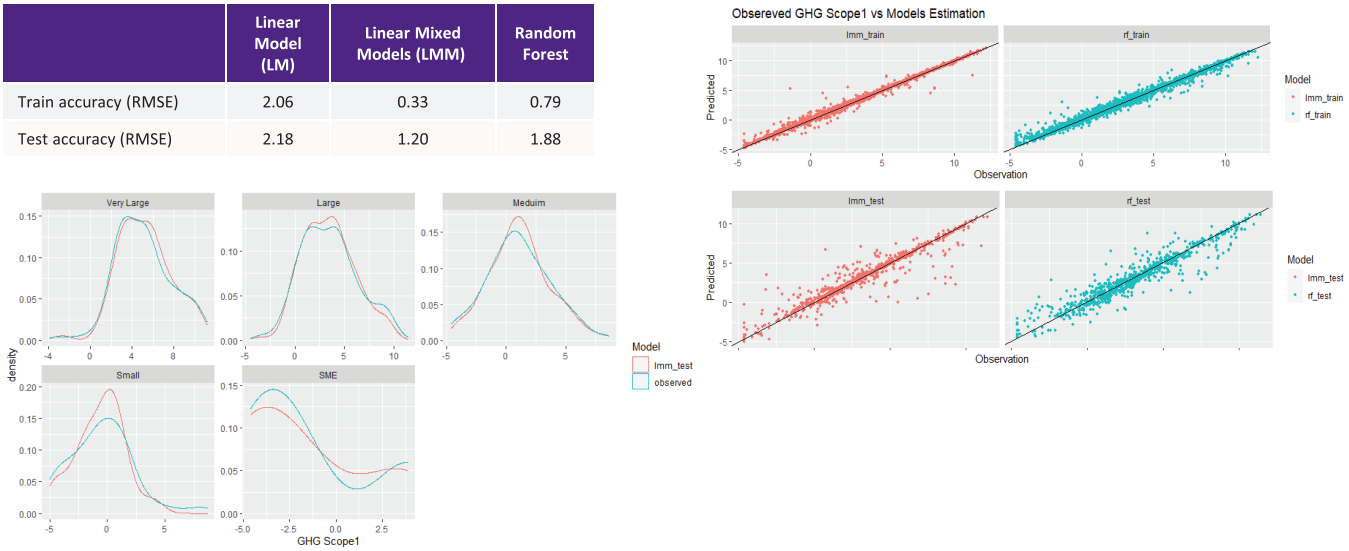
Model performance

Linear mixed models (also known as hierarchical models) is an extension of the general linear model (GLM) that considers the presence of fixed and random effects. These models are suitable to deal with longitudinal data or repeated measurements. The ESG data includes a group of companies on whom multiple individual observations of GHG emission over time is recorded for which the LMM is appropriate.

Modelling options performance assessment

To decide on the right model to estimate GHG emissions, we looked at the root mean squared error (RMSE) matrices. It gives us an idea of the average distance between the observed data and the predicted values. The results show that our proposed model (LMM) consistently has better RMSE than the baseline model and tree-based model.

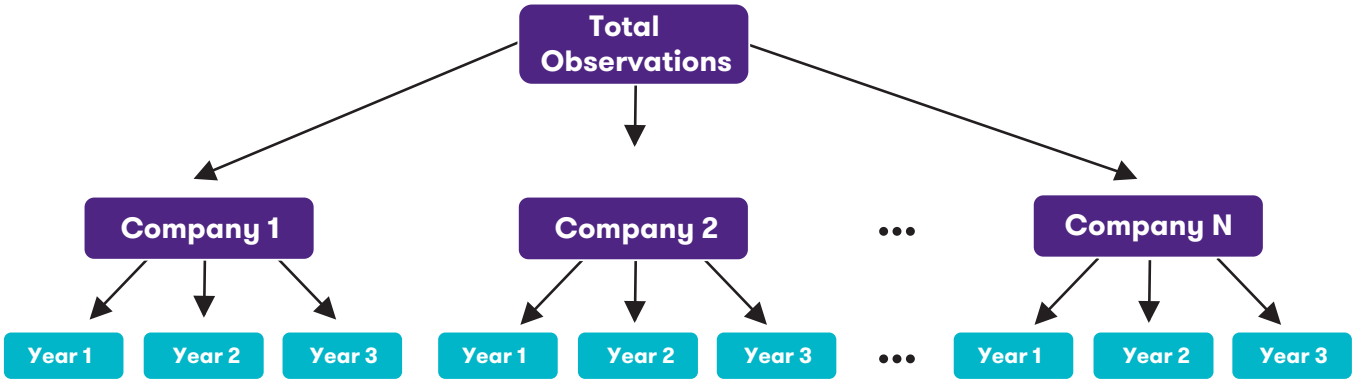
	Linear Model (LM)	Linear Mixed Models (LMM)	Random Forest
Train accuracy (RMSE)	2.06	0.33	0.79
Test accuracy (RMSE)	2.18	1.20	1.88



The scatterplot of predicted versus observed GHG Scope 1 shows that the RF and linear mixed models are both accurate although the LMM has a better performance.

The density of GHG Scope 1 for observed and tested values, shows that the model can capture general behaviour of data. However, the predictions are less precise for SMEs and small firms.

Linear mixed model detail



Hierarchical data insights into two important features:

- Within-group variability : can examine how consistent a company's GHG emissions are from year to year.
- Between-group variability : can examine the degree to which GHG emission pattern vary from one company to another.

We formulate linear mixed model as following:

$$\begin{gathered} GHG_{scope1} = (\beta_0 + b_{ticker}) + \beta_1 \log(Sales) + \beta_2 \log(Assets) + \beta_3 \log(Employees) + \\ \beta_4 Year + \beta_5 Sector + \beta_6 Region + \epsilon \\ \epsilon = N(0, \delta) \\ b_{ticker} = N(0, \tau) \end{gathered}$$

Where the ticker term includes a unique effect for each company, the effect term is an additive term to the fixed intercept. As a result, for each company, the model estimates fixed values for predictors coefficients and random intercepts (intrinsic effect of company).

It was expected that the LMM performed better since the data is longitudinal and the other models cannot easily capture the within company correlation. So far, we only have data from 2019 to 2021 and it is expected that in the future when re-training the model with more data collect across time, the LMM will be even more accurate when comparing to other models.

Best model results

Since the LMM is a statistical method, it is possible to understand the relation between the financial factors and emission though the associated regression coefficients in the model. The following table shows fixed coefficients and the model generates additional terms specific for each company that gives us the specific intercepts for each company.

Fixed Coefficients

Intercept*	Year	log Number of employees	log Sales/revenue	log Assets
Alpha	●	●	●	●

Sector *							
B	C	D	E	F	G	H	I
●	●	●	●	●	●	●	●
J	K	L	M	N	Q	R	S
●	●	●	●	●	●	●	●

Region *				
Asia Pacific	Central Europe	Western Europe	South America	Middle East
●	●	●	●	●
Southeast Europe	Southern Europe	Northern Europe	North America	Eastern Europe
●	●	●	●	●

***For each categorical variable the model consider a baseline level which, A is the baseline sector and Africa is baseline region. For a company from Africa in sector A the value is fixed intercept.**
For companies from other regions and sectors the coefficients are added to the baseline intercept.

Model interpretation

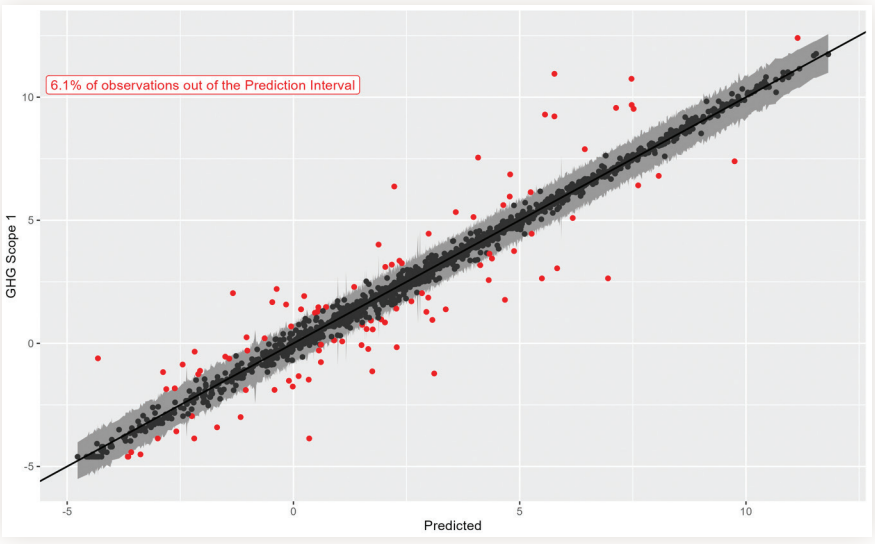
The estimated effect of revenue is 0.46, that means if a company increases the revenue by 1%, then the emission increases by 0.46%.

Companies from NACE sector D (Electricity, gas, steam and air conditioning supply) produce more emission than sector A (Agriculture, forestry and fishing). In fact, the difference is 10.34 kt CO2e on average.

For a company that has not reported the emission in a specific year, its predicted value has a median absolute error of 4.71 kt CO2e* or on average, 694.07 kt CO2e.

Prediction Intervals (PI)

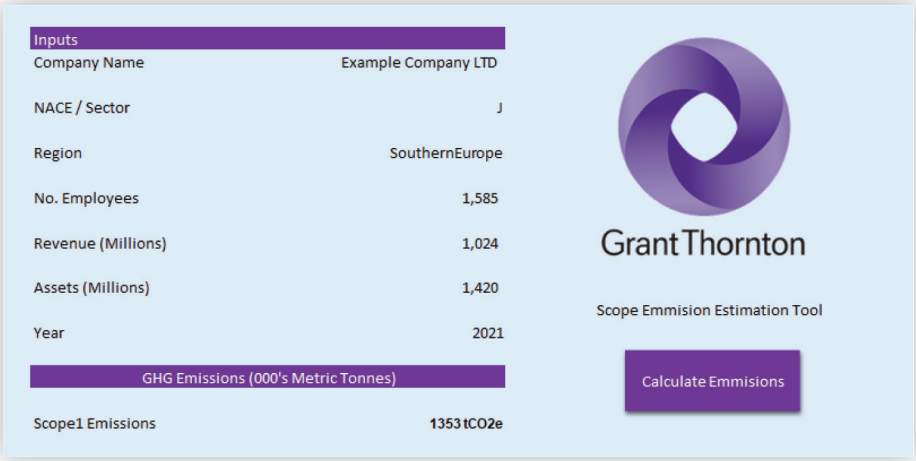
Since the model is an extension of a linear regression model, it is possible to obtain the PI that is the interval which a new observation will fall given a probability level. Only 6.1% of the new observation are not contained in the 80% PI, so we conclude the model as a good prediction accuracy.



Application and key areas for further development

The output of our research is two fold. From a theoretical perspective, we believe that we have contributed to the better understanding on methodologies which can be used for data proxies estimation. From a practical perspective, we have developed a model that can be used by users to estimate emissions for particular companies as well as for banks to estimate emissions for their non retail portfolios.

Model Application in R-Shiny



Areas for further research and development

The proposed model is easy enough to be explained to non-technical audiences.
Although there is room for improvement in the LMM.

- From explanatory analysis, we know there is more hierarchal behaviour in the data. For example, the between sector variations is observed, but due to a lack of observations we couldn't fit random effects for sector. One solution might be to re-group the sectors with similar behaviour to get more populations per sectors. To know which sectors can be grouped together needs some domain knowledge to have reasonable results.
- We know that energy consumption is an indicator used to predict emissions. We didn't use it as a predictor in our model since many companies didn't disclose energy consumption. We suggest using missing value imputations techniques before training a model.
- Since the predictors have different ranges and units, it is useful to normalise the data in the pre-processing stage.
- It makes interpretations easier at the end.

References

1. Bloomberg's Greenhouse Gas Emissions Estimates Model: A Summary of Challenges and Modeling Solutions
3. Serafeim, George and Velez Caicedo, Gladys, Machine Learning Models for Prediction of Scope 3 Carbon Emissions (June 1, 2022). Harvard Business School Accounting & Management Unit Working Paper No. 22-080, Available at SSRN: https://ssrn.com/abstract=4149874 or http://dx.doi.org/10.2139/ssrn.4149874

Contacts

Our team would be delighted to discuss your challenges and opportunities in any aspect of climate risk.
Our services are flexible and efficient, designed to facilitate and support your business model.
Our highly qualified Quantitative Risk team provides support to financial institutions across the full spectrum of risk measurement and modelling strategies, including the development, deployment and validation of key models and risk measurement methodologies in regulatory capital, stress testing and IRB, IFRS9 and bank risk modelling. Team has experience implementing machine learning techniques in the context of credit risk modelling as well as a keen interest in emerging trends within the machine learning space.

Contact us today to discuss.

Partner group

Dwayne Price, Partner
Financial Services Advisory
T +353 1 436 6494
E dwayne.price@ie.gt.com

Amanda Ward, Partner
Financial Services Advisory
T +353 1 433 2440
E amanda.ward@ie.gt.com

Stavros Ioannou
Managing Partner
CEO Grant Thornton Cyprus
T +357 22 600 103
E stavros.ioannou@cy.gt.com

Frankie Cronin, Partner
Financial Services Advisory
T +353 1 646 9044
E frankie.cronin@ie.gt.com

Brian O'Dwyer, Partner
Financial Services Advisory
T +353 1 433 2538
E brian.odwyer@ie.gt.com

Nuala Crimmins, Partner
Financial Services Advisory
T +353 1 483 8577
E nuala.crimmins@ie.gt.com

Sustainability model development group

Lukas Majer, Director,
Quantitative Risk,
ESG Modelling
T +353 1 646 9006
E lukas.majer@ie.gt.com

Mark Perry, Director
Quantitative Risk,
ESG Modelling
T +353 1 408 6909
E mark.perry@ie.gt.com

Andreas Spyrides, Director
Quantitative Risk,
ESG Modelling
T +357 2 260 0270
E andreas.spyrides@cy.gt.com

Catherine Duggan, Director
Head of Sustainability IE
T +353 1 433 2535
E catherine.duggan@ie.gt.com

Janice Daly, Director
Head of Sustainable Finance
T +353 87 237 5946
E janice.daly@ie.gt.com

Phanis Ioannou, Manager
Quantitative Risk,
ESG Modelling
T +357 2 260 0296
E phanis.ioannou@cy.gt.com

Cian Greenwood, Consultant
Quantitative Risk,
ESG Modelling
T +353 1 680 5805
E cian.greenwood@ie.gt.com

Offices in Dublin, Belfast, Cork, Galway, Kildare, Limerick, Longford, Nicosia, and Limassol.

grantthornton.ie @GrantThorntonIE Grant Thornton Ireland

Grant Thornton

grantthornton.ie