A Machine Learning approach for Predicting Corporate ESG Ratings and analysing the impact of Country ESG data on Prediction

MSc Research Project
Data Analytics

Gurpreet Kaur Bhuie
Student ID: x21231061

School of Computing
National College of Ireland

Supervisor: Athanasios Staikopoulos

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Gurpreet Kaur Bhuie
Student ID:	x21231061
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Athanasios Staikopoulos
Submission Due Date:	14/12/2023
Project Title:	A Machine Learning approach for Predicting Corporate ESG Ratings and analysing the impact of Country ESG data on Prediction
Word Count:	4812
Page Count:	16

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	31st January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Machine Learning approach for Predicting Corporate ESG Ratings and analysing the impact of Country ESG data on Prediction

Gurpreet Kaur Bhuie
x21231061

Abstract

This research tries to explain Environmental, Social and Governance (ESG) need and importance in corporate finance. How ESG ratings from different rating agencies differ for the same organization. An examination is made about how having a consistent rating is important for organizations and failing to do this will result in loss or misallocation of millions of dollars. By investigating previous work done in this field the reason for inconsistency is discussed and many other approaches to overcome this inconsistency is also discussed. To address this problem the search, propose a generic model for ESG prediction which can be easily used by companies to predict their ESG rating and get an idea about if their current ESG roadmap is good enough to achieve their targets or not. This research also discusses the impact of country ESG ratings on any company's ESG rating and how it can be included in ESG prediction. For implementing the solution, first both datasets are gathered from sources and cleaned then, both the company and country ESG datasets are combined based by joining them on relevant column. Random forest algorithm is applied to entire dataset and on only ESG dataset to compare the impact of country ESG data. This paper concludes the research thoroughly discussing the results and impact of country ESG datapoints on corporate ESG ratings predictions.

1 Introduction

Due to increase in industrial activities many problems like global warming, pollution and increase in green house gas emission have raised. There is a need to ensure that the companies whether they are from IT, manufacturing, consulting etc follow sustainable development practices. Environmental, Social and Governance (ESG) is one such parameter that help companies quantify their effort put in this direction. In 2004 United Nations published a report named Who Cares Wins[1] that emphasized on sustainable development and different parameters to measure it. This started the trend towards sustainable development and in past two decades organizations started putting in more of effort to ensure that they are following all these guidelines. Investors also want to be associated with the organizations that follows high degree of sustainable development processes.

[1] https://www.unepfi.org/fileadmin/events/2004/stocks/who$_{c}ares_{w}ins_{g}lobal_{c}ompact_{2}004.pdf$

1.1 Motivation and Project Background

ESG ratings are calculated by checking 17 different factors about an organization in areas like Environmental represented by 'E' which assess impact of an organization's operational process on environment and measures taken to neutralize it. Social pillar represented by 'S' asses the HR, data retention and labour policies while Governance pillar 'G' asses the corporate and ethical policies and diversification of board members. These ratings have gained popularity because they have financial impact of company's performance as suggested by Daying and Zi'Ao (2023).
ESG ratings are provided by different agencies and each of them which follows different procedures to calculate ESG ratings. This sometimes may lead to inconsistent ratings and becomes a cause of concern for organization as they may loose investors as explained in a study by Kim and Li (2021) and Stubbs and Rogers (2013). Additionally, the process to calculate ESG ratings is really complex and companies have to relay on third party agencies for yearly rating revision. This is one of the reason predicting ESG ratings has one of the important use case for data analysis. Many studies have been performed to use different ML and mathematical models to predict these ratings and showed good results making it an interesting topic to investigate on. An alternate model may help understand companies their current expected ESG ratings tweak their ESG road map and allocate these fund to correct section. This may result be higher revenue and better ESG ratings.

1.2 Research Question

In addition to a company's ESG performance, a country's ESG data may also impact the ESG ratings. Analysing how this change or impact ESG ratings can be useful to understand long term effect of all the sustainable development practices across an organization. In order to find a solution to this problem statement, this paper will try to propose a generic model to predict ESG ratings and my research question is

RQ: How well random forest algorithm can predict ESG ratings based on historical ESG data of company and country?

In this study, various previous models are studied and the new alternate model is proposed. The developed model can be compared with the actual ESG rating to evaluate it performance in real world scenarios.

1.3 Report Structure

The reminder of the document is structured in following sections 2. Related work that investigates the previous work done on this problem, 3. Methodology specifies the steps performed in this research, 4. Design Specification explains the architecture and process flow of the experiment , 5. Implementation details out how the results are achieved, 6. Evaluation and discussion specifies the results and discuses them, 7. Conclusion and Future Work.

2 Related Work

Though ESG is new concept, it is widely accepted in all industries and now all companies publish their ESG rating as a part of their contribution towards sustainable development. Even after being so important the lack of transparency in calculation of ESG scores and different standards set across different agencies lead to a necessity of simple and consistent method to calculate ESG ratings. Many research have been done in the area before and this section of the paper will discuss previous works and their contribution.

2.1 ESG rating calculation

The importance and relevancy are ESG ratings have been a topic of debate since this concept was introduced. Furthermore, the inconsistency across different agencies lead to lack of trust in these rating and make it is difficult for smaller scale companies to incorporate them. As the world is moving towards sustainable development, ESG scores and investment does play an important role in projecting a company's contribution towards it which in return make investors more comfortable investing with these organizations as discussed by Bhandari et al. (2022) in their research. Their work suggests also talk about an idea to use machine learning techniques to calculate ESG score. However, a more detailed explanation on methodology to predict or calculate score would have been an interesting factor of add but this research helps us to understand how important ESG ratings are in today's corporate financing.

The ratings provided by agencies to these company varies across the agencies and this in turn create a pressure on the companies to maintain a consistent rating throughout the agencies which is difficult and may lead to misallocation of company's fund as explained by Chatterji et al. (2016) in their work. Their work provides a detailed explanation of what can be the impact of these in consistent ratings on company review, but their study did not explore the reason behind it. More clarity on this can be gained by the work of Berg et al. (2022) where they explain that the reason behind this could a different method followed by each agency to calculate this rating and categorised this divergence into three main category namely: Scope, Measurement and Weight where different agencies have different scope as in what factors they consider is important, different units of measurement and lastly different weights meaning how important they think a particular metric is for the calculation. They also suggest the one way to standardize this could be to collect data at the source and standardized the process. However, this is not a simple process and companies does not freely provide these data to agencies. Moreover, collecting data only from companies may lead to green washing where they may try to project something that was not done.

Due to this inconsistency and important of ESG score this area became one of the most talked about topic in today's world of data science problems and many studies are being performed to come up with a consistent and easy way to predict these ratings so smaller scale companies can also take advantage from it.

2.2 Machine Learning for ESG rating prediction

Machine learning have been useful in solving many traditional uses cases and can also be used in the space of ESG done by Gupta et al. (2021) in their study. They used regression model to establish relationship between ESG metrics and investments call outs. In their approach, they build two web scrappers and scrapped the data from Yahoo Finance and Sustainalytics website. Furthermore, a regression model was built to analyse the relationship. Though it was one of the good baseline models, but it may not be suitable for everyone as the web scrapping involves legality issues and the guidelines around it are constantly changing. Additionally, they did not clarify why other more dynamic models like random forest was not used for this study which are much more capable in handling categorical data and complex relationships.

Mathematical approach can also be used for predicting the ratings as suggest in the study by García et al. (2020). In this study, financial data from publicly traded companies from EU were used and these companies where group in three to 4 clusters and then a mathematical model is applied to extract information from uncertain context. This model does suggest an alternate approach but lacks in showing promising results when the number of companies are increased in the dataset. Additionally, mathematical models are more complex to implement and are time consuming to execute when dealing with a complex problem like this. An area to investigation would be is to simplify the problem of calculating ESG scores rather than over complicating it.

In D'Amato et al. (2022) approach, the author used random forest algorithm with temporal data from time series and used it to predict ESG ratings. This study clarifies the reason of using this algorithm and explained in detail the steps and methodology used. However, the data used for prediction is based on balance sheet which may have other non ESG data in it. Additionally, they used data from only agency named Thomson Reuters (now Refinitiv) to extract ESG score sample data which may lead to a biased decision as the ratings from other agencies use different parameter. Hence, it becomes important to either take sample data from all the agencies for better coverage or to have an external dataset independent from these agencies like country ESG data.

There are many different methods also that can be used for this purpose. Lin and Hsu (2023) performed an experiment to test the accuracy of four different machine learning model (ELM, SVM, RM, XGBOOST) and found that all these models perform well with an accuracy of about 0.97 during training and about 0.94 during testing phase. This paves the path of research in this area as the suitability of different models are justified by their research.

Another work another study Ang et al. (2023) researched how dynamic company networks can be used to predict ESG ratings and showed some promising results. They performed this study on the companies listed in NYSE and showed that their network-based fixed effect panels model outperformed state of the art XG-Boost model. They also relied on network variables and news-based content to predict ESG ratings. Though this model used one of the advance methods to predict ESG ratings, it is very specific to US market and the author talks about that result may vary in other markets. Additionally, this model comes up with high demand of confidential dataset that is not freely available in market. The need of high computational power is another factor that makes this model not so fit for a general use by small scale organization. This again raises a need for a general model that can be used to predict ESG rating and it easy to build and use.

An interesting study by Krappel et al. (2021) which is divided into two parts also proves the same point. The first part of their study explains the complexity of ESG score and different factor involved in its calculation and second part explains how linear regression is also good enough to predict the accurate ESG ratings and highly complex models like neural network may not be needed. They explained how G (governance) factor is more divergent in nature and may vary across different regions like Europe and Asia. Their study was also focused on data from Refinitiv and can be expanded across different regions.

NLP is another approach that can be used to predict ESG sentiment and in-turn calculate the ratings. Aue et al. (2022) used pre-trained BERT models in their study and used it on new articles to predict ESG ratings. In their approach, BERT models were first used to identify ESG corpus and sentiment analysis is performed on them. After that articles were clustered to derive final prediction. Their study also analysed the trade-off between accuracy and computational power due to high demand of resources needed for this task. It is interesting to note that though they achieved significant results by executing a complex model, they also mentioned that regression models can also be used for this task. Their model uses news article as main dataset, which may lead to biased results as it may contain noise and all the news captured not necessarily be true specially in the era to social media. Additionally, the model uses large amount of data which is not available free of cost and this model showed good results for small cap companies only making it less generic. Hence, having a generic model built on a reliable dataset would be a suitable solution to predicting ESG rating.

In an interesting study, Mooneeapen et al. (2022) analysed how country ESG values may affect the ESG values of a company. They explained how these two values are highly related and one may depend on other. Though they did not used any algorithm to predict ESG values, but this study provided a baseline for the idea to use company ESG data for predicting ESG data.

All the previous work reviewed during this research were focused on predicting ESG score for companies. There were many different techniques which were used in the past to do so but there is still scope to explore the external factors that can affect the ESG scores. With the increase in data points being collected there can be many such usecases can be solved by machine learning and analysing the effect of country ESG data on corporate ESG data can be one such usecase.

2.3 Takeaways from Related work

From the above literature review it is clear that predicting ESG rating is one of interesting usecase in the field of data science and though mentioned work did provided some really interesting solution but is still some scope of investigating external factors which can affect ESG score. Impact of country ESG datapoint on corporate ESG is not explored much. In this study, a random forest algorithm will be used with company and country ESG data for ESG prediction to find this solution.

3 Methodology

The research for data mining can be carried out either is CRISP-DM or KDD. Both methodology can be used for predicting ESG scores however the KDD approach will be better suited for this research as it will allow us to study relationship between country esg data and company esg data while following other steps from data mining. Various steps that are followed are (1) Company historical ESG data collection in csv (2) Country historical data collection from world bank datamart in csv (3) Data Cleaning and pre-processing (4)Data Exploration (5) Feature Selection and Dimensionality reduction (6) Modeling (7)Model Evaluation

3.1 Dataset Selection

Collecting data for ESG is one of the most challenging part of this experiment. This data is being used by agencies for their business and hence its not available freely in market. There are 2 datasets used in this experiment, first is country ESG dataset which

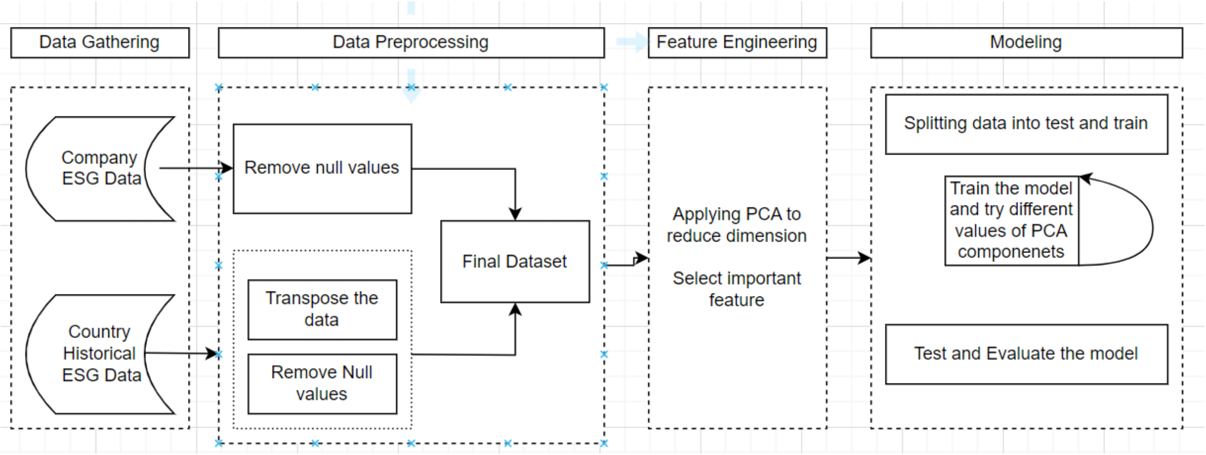
Data Gathering
Data Preprocessing
Feature Engineering
Modeling
Company ESG Data
Country Historical ESG Data
Remove null values
Final Dataset
Transpose the data
Remove Null values
Applying PCA to reduce dimension
Select important feature
Splitting data into test and train
Train the model and try different values of PCA componenets
Test and Evaluate the model

Figure 1: Research Methodology

is sourced from worldbank.org[2] for years between 2013 to 2022. World bank data mart is one of the most trusted datamart and contains most recent data. Data is collected for all countries accross all ESG series available and total number of records are 5259. There are 22 different ESG datapoints for each country under series column and other columns in order are Country Name, Country Code, Series Name, Series Code, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022.

Above dataset is then combined with company's ESG dataset which is originally sourced from bloomberg[3] and available in github for public use. This dataset contains 50460 records for 2407 companies spanning from year 2014 to 2018. It contains columns named ComCode which denotes company code, SEDOL, ConstituentName which is nothing but company name, Country, ISO Code is country code, Exchange Code, Industry Code, Supersector Code, Sector Code, Subsector Code, E denotes the Environmental category values, S denotes the social category values, G denotes Governance category values, Months is the time period for which the data is recorded. Values like exchange code, industry code, supersector code and sector code will help identify the trend of E,S,G values over time.

3.2 Exploratory Data Analysis

In this step, dataset is studied and target variables are plotted to understand its distribution as shows in Figure 1. Using pandas library the data is read into python environment and any required data type conversion was performed like converting E,S,G values datatype to numeric. Statistical analysis is also performed on the dataset and table 3 shows a sample of it. The entire table can be found with code.

Correlation matrix is studied to identify the relationship and their strength and heatmap is plotted to visualize this figure 3.

3.3 Data Prepossessing and cleaning

In order to run the model in both data set it needs to be joined using the logical conditions. In this experiment conditions selected to join these two datasets are data column

[2] https://databank.worldbank.org/source/environment-social-and-governance-(esg)-data
[3] https://www.bloomberg.com/professional/product/esg-data/

column_name	data_type
months	date
Industry Code	numeric
Supersector Code	numeric
Sector Code	numeric
Subsector Code	numeric
e	numeric
s	numeric
g	numeric
sedol	character varying
constituentname	character varying
country	character varying
ISO Code	character varying
Exchange Code	character varying
comcode	character varying

Table 1: ESG data table data definition

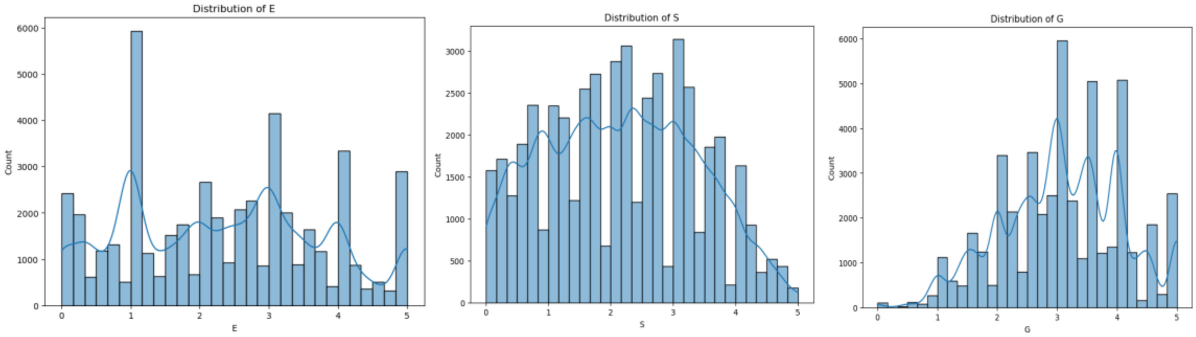

Figure 2: Distribution of E,S,G values

Table 2: Statistical Analysis of Target variables

	e	s	g
count	46744	46744	46744
mean	2.384177648	2.182158138	3.097225312
std	1.399589013	1.212352365	1.001837839
min	0	0	0
25%	1	1.2	2.4
50%	2.4	2.2	3.1
75%	3.4	3.1	3.9
max	5	5	5

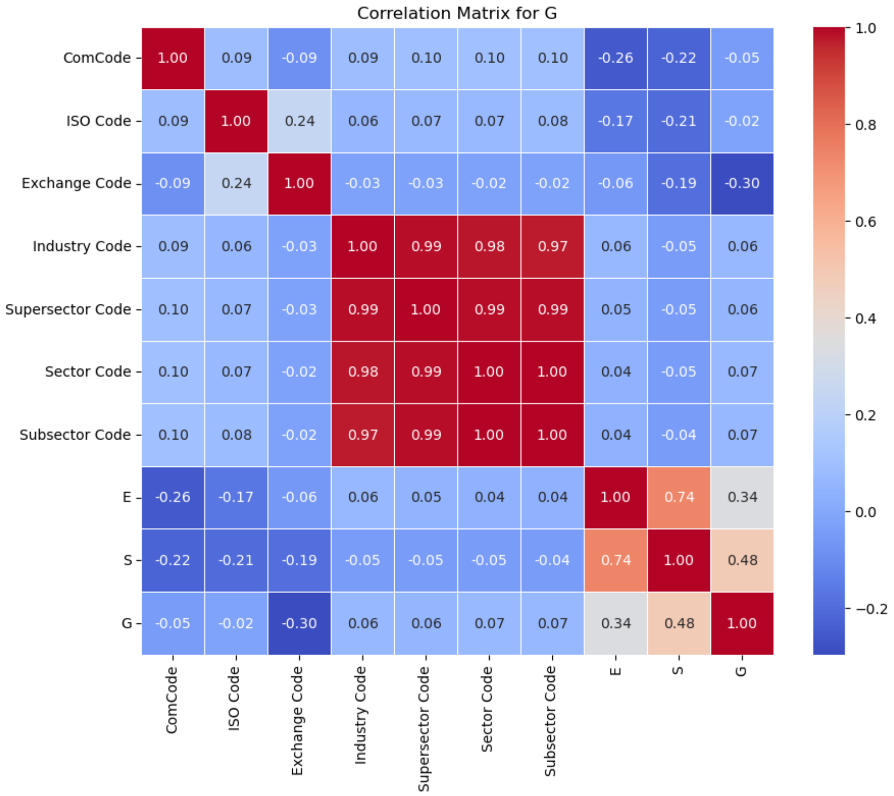

Figure 3: Enter Caption

and country code. One challenge is to ensure that both datasets have the correct columns available to join. While the company ESG data have country code and date column, the Country ESG data only had country column and data for all the years where separated in different column. To overcome this issue, years and series data were transformed using python code and the new dataset looked had columns in below order Country Name, Country Code, Year, Access to clean fuels and technologies for cooking, Access to electricity, Annualized average growth rate in per capita real survey mean consumption or income, Cause of death, Children in employment, Fertility rate, Gini index, Government expenditure on education, Hospital beds, Income share, Labor force participation rate, Life expectancy at birth, Literacy rate, adult total, Mortality rate, People using safely managed drinking water services, People using safely managed sanitation services, Population ages 65 and above, Poverty headcount ratio at national poverty lines, Prevalence of overweight, Prevalence of undernourishment, School enrollment, Unemployment rate.

Once the data is transformed the null values are removed as there where only few records with null value. This action is performed for both Country ESG and Company ESG data. The cleaned data is then inserted into database Postgresql in this case using python code. For this first the connection is established with Postgresql database and tables with required column names and data type are created using python code. The table name for company esg data is `esg_data` and table 1 below explains the data definition of this table. Country esg data is `country_esg_data_new` and described in table 2 below. Once both tables are available in DB is was analyzed for joining condition and another table named `country_info` with only country name, iso code is created which will act as a bridge table for both these tables. In order to fetch all the records available for company ESG data a left outer join is performed on `esg_data` and `country_info` on the column 'ISO Code' and then `country_esg_data_new` is joined on 'Country Name' and year. This sql fetched all the records a for company esg data and joined them with their matching country and yearly data for country ESG data.

The result set contains 48246 records which will be the final dataset on which the model will be created to ensure the country esg datapoints are also analysed. This data is then exported into excel for ease of reproducibility of experiment. The complete data have columns in order comcode, sedol, constituentname, country, ISO Code, Exchange Code, Industry Code, Supersector Code, Sector Code, Subsector Code, e, s, g, months, country, iso code, country name, Country Name, Country Code, Year, Access to clean fuels and technologies for cooking, Access to electricity, Annualized average growth rate in per capita real survey mean consumption or income, Cause of death, Children in employment, Fertility rate, Gini index, Government expenditure on education, Hospital beds, Income share, Labor force participation rate, Life expectancy at birth, Literacy rate, adult total, Mortality rate, People using safely managed drinking water services, People using safely managed sanitation services, Population ages 65 and above, Poverty headcount ratio at national poverty lines, Prevalence of overweight, Prevalence of undernourishment, School enrollment, Unemployment rate.

3.4 Evaluation Techniques used

The evaluation techniques used in this experiment are as follows:

- **Mean Absolute Error:** Measures the deviation between actual and predicted values.
- **Mean Squared Error:** Measures the square of deviation between actual and predicted values. This magnifies the small errors.
- **R squared:** Explains the fit of the model and ranges from 0-1 where 1 being perfect fit

4 Design Specification

The framework of this research can be divided into three main layers as shown in figure3. Presentation layer which is the top most layer showing all the visualization. Middle layer is the business layer where all the data transformation and implementation of model is performed. Last layer is database layer and data is present in CSV format in this layer after being extracted from postgresql database. For the ease of reproducibility data is stored in different tables in postgresql and then joined into a single table to create final dataset.

Random Forest Regressor is used for modeling in this experiment. It is an ensemble method and works by building multitude trees and predicts the average output for each tree.

5 Implementation

This section explains implementation which can be divided into three main phases like Data preparation, Feature selection, Fitting the model. Below is the detailed description of the implementation:

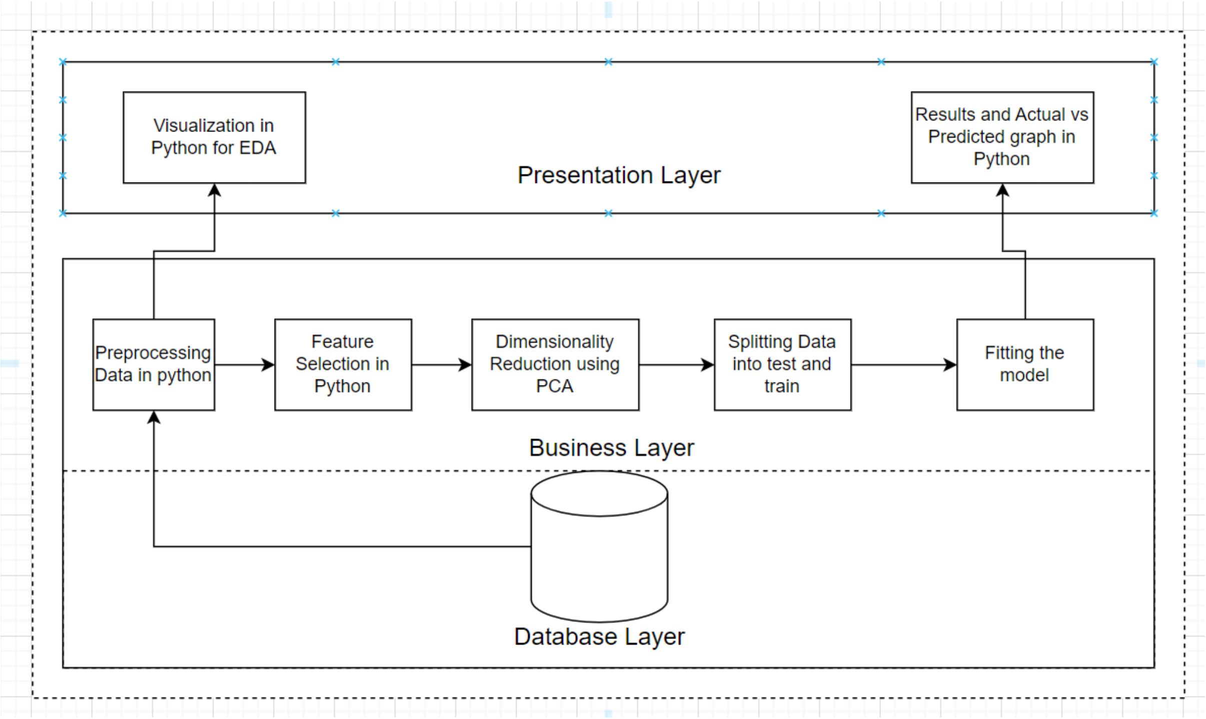
Visualization in Python for EDA
Presentation Layer
Results and Actual vs Predicted graph in Python
Preprocessing Data in python
Feature Selection in Python
Dimensionality Reduction using PCA
Splitting Data into test and train
Fitting the model
Business Layer
Database Layer

Figure 4: Project Design and Process Flow

5.1 Tools used

In this experiment, excel was used to maintain initial dataset. To combine the datasets sql query was used to in postgresql database. For coding and visualization, python environment in jupyter notebook was used which provides all the required libraries for model implementation and visualization of results.

5.2 Data Preparation

Once data is cleaned and null values are removed as explained in section 3.2 comes the next step where the columns that are not relevant are dropped in order to avoid over-fitting. For example column like constituent name, ticker symbol etc. Post that categorical features like company code, industry code, sub-sector code, sector code are converted into numerical column by label encoding technique. This will help retain the relationship of these features with target variables.

5.3 Dimensionality Reduction

Principal Component analysis is used here for dimensionality reduction and most important features were selected. The threshold is set to 10 components based on the Elbow chart where the values starts to saturate Figure 5 and Figure 6. Post that data was spit into test and train 80:20 ratio.

5.4 Model Training

Random Forest Regressor is initialized and random state is set to 42 for reproduciblity. Model is then trained on the train dataset obtained after PCA transformation. There are

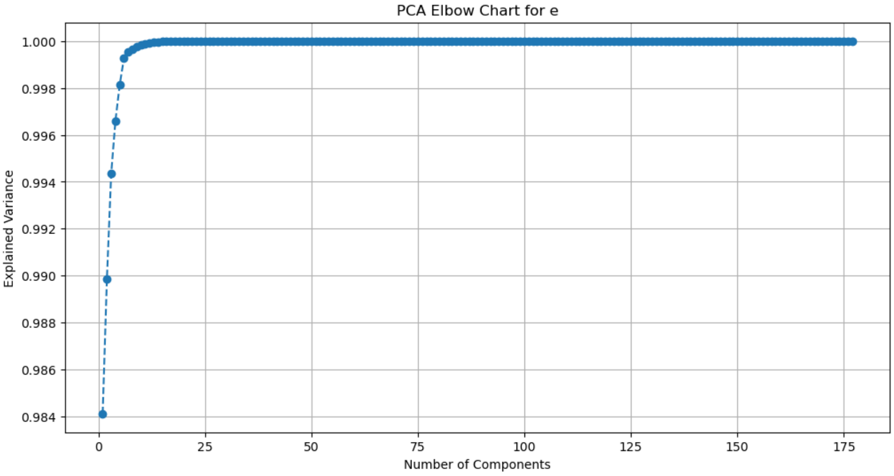

Figure 5: PCA Optimal components by Elbow Chart

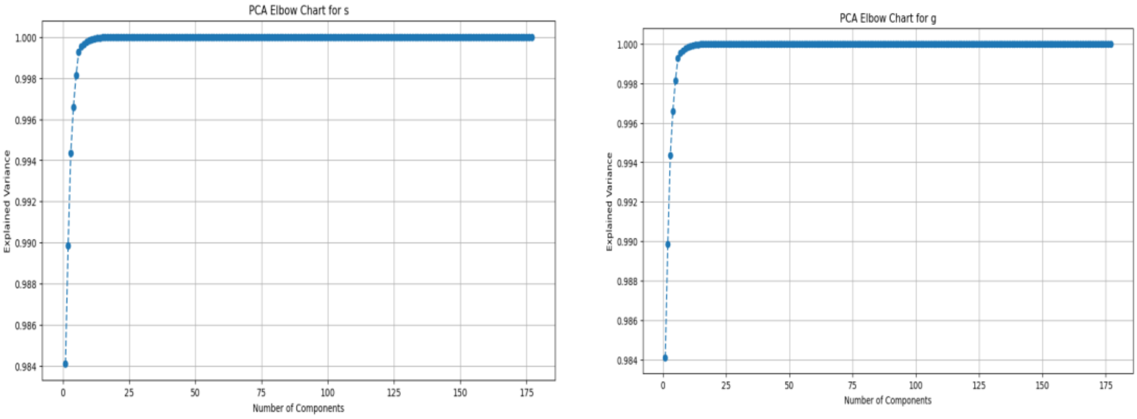

Figure 6: PCA Elbow chart for S and G

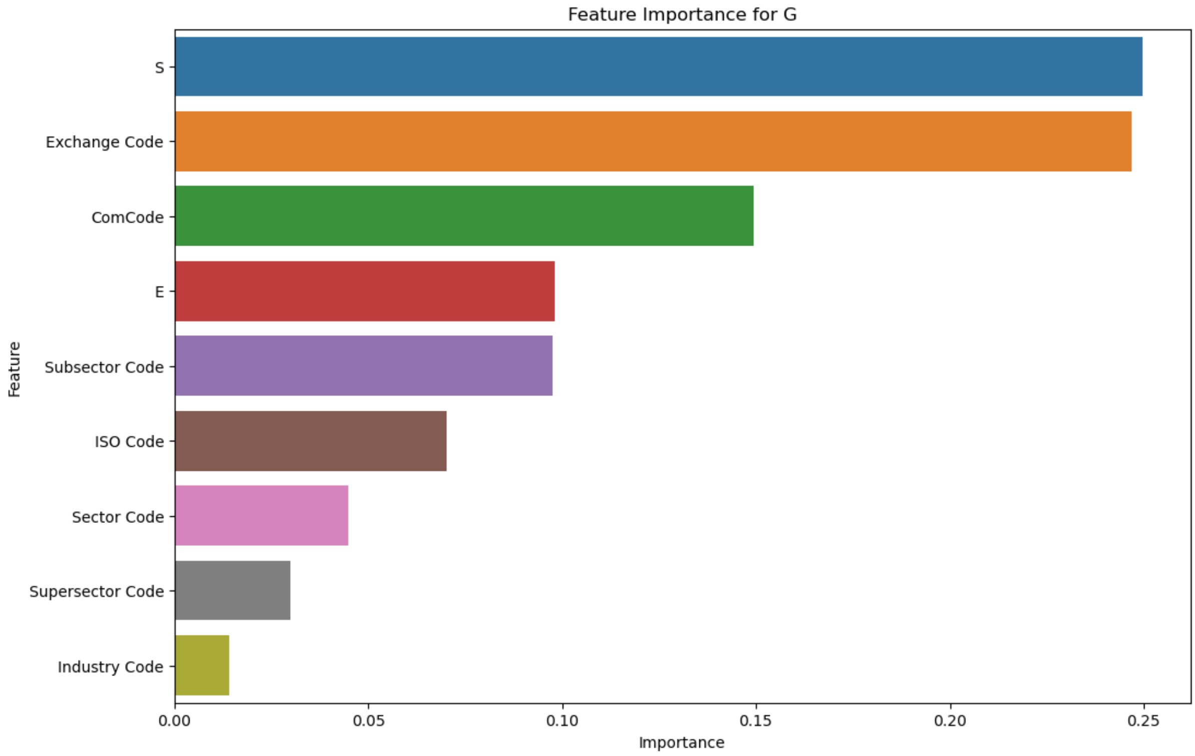

Figure 7: Feature importance for G

three target variables namely e,s and g, so the model implementation is iterated 3 times meaning once for each target variable to predict e,s,g.

5.5 Feature Importance

The feature importance bar chart is plotted to visualize which features are most important for prediction in their descending order of important as seen in figure 6.

6 Evaluation

Evaluation is one of the most important phase in any machine learning problem as it helps understand the performance of the model and check whether its working as intended or not. Multiple iteration of the model were executed to analyze which model is giving the best results.

6.1 Case Study 1: Experiment with only ESG data for companies

As discussed before in section 3.3 the final dataset was the combination of both company ESG dataset and Country's ESG dataset. In the first experiment, the model was executed only on ESG dataset from companies to analyze the performance. This will act like a baseline model which can be used to compare the model performance with country ESG. Table 3 below shows evaluation metrics for this experiment. All the parameters are calculated for E,S,G separately and the accuracy lies around 99 percentage for prediction

Table 3: Evaluation parameters for ESG data of companies

Target Variable	MAE	MSE	RMSE	R-squared	Adjusted R-squared	Explained Variance
E	0.03	0.011	0.10658	0.994282	0.9942765	0.9942817
S	0.031	0.013	0.11469	0.991029	0.9910209	0.9910291
G	0.033	0.013	0.11582	0.986502	0.9864898	0.9865023

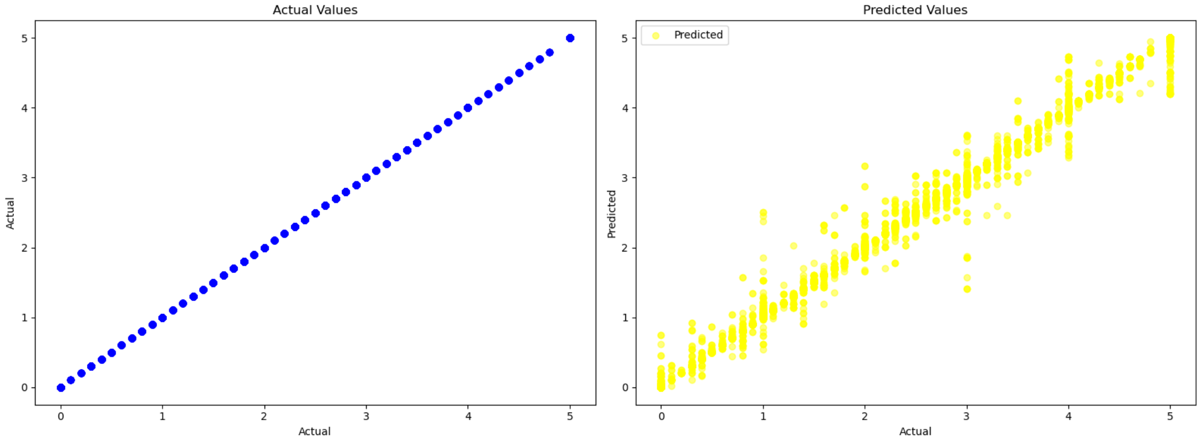

Figure 8: ESG values Actual vs Predicted

E and S values while for G its around 98 percentage. The actual versus prediction graph shows the same results in a graphical pattern in Figure 7.

6.2 Case Study 2: Experiment with ESG data for companies and countries

In the next phase, experiment was performed on the bigger dataset of with combination of ESG data of companies and country. In this experiment, PCA was applied and elbow chart was plotted to check the optimal number of components. The graph shows that the optimal number of components should be around 10, so multiple iteration were executed with 8,10,12 components in PCA and results were observed. The table below shows the value of evaluation metrics. By looking at the values in table 4,5 and 6, it is clear that best results are shown when PCA component = 10. Additionally, Figure shows a plot for actual versus predicted values.

Table 4: Evaluation metrics for PCA component=8

Target Variable	PCA	MAE	MSE	RMSE	R-squared	Adjusted R-squared	Explained Variance
e	8	0.0124338	0.00346	0.058824	0.998222	0.9982205	0.9982221
s	8	0.0101438	0.0019	0.043584	0.9987172	0.9987160	0.9987172
g	8	0.0119546	0.002564	0.050638	0.9974783	0.9974761	0.9974785

Table 5: Evaluation metrics for PCA component=10

Target Variable	PCA	MAE	MSE	RMSE	R-squared	Adjusted R-squared	Explained Variance
e	10	0.0129163	0.0041738	0.0646051	0.9978553	0.9978530	0.9978553
s	10	0.0104514	0.0016363	0.0404521	0.9988949	0.9988937	0.9988949
g	10	0.0102791	0.0019314	0.0439479	0.9981006	0.9980985	0.9981006

Table 6: Evaluation metrics for PCA component=12

Target Variable	PCA	MAE	MSE	RMSE	R-squared	Adjusted R-squared	Explained Variance
e	12	0.0118262	0.0030731	0.0554353	0.9984210	0.9984189	0.9984210
s	12	0.0090022	0.001366	0.0369569	0.9990776	0.9990764	0.9990777
g	12	0.0108554	0.0018691	0.0432328	0.9981619	0.9981595	0.9981620

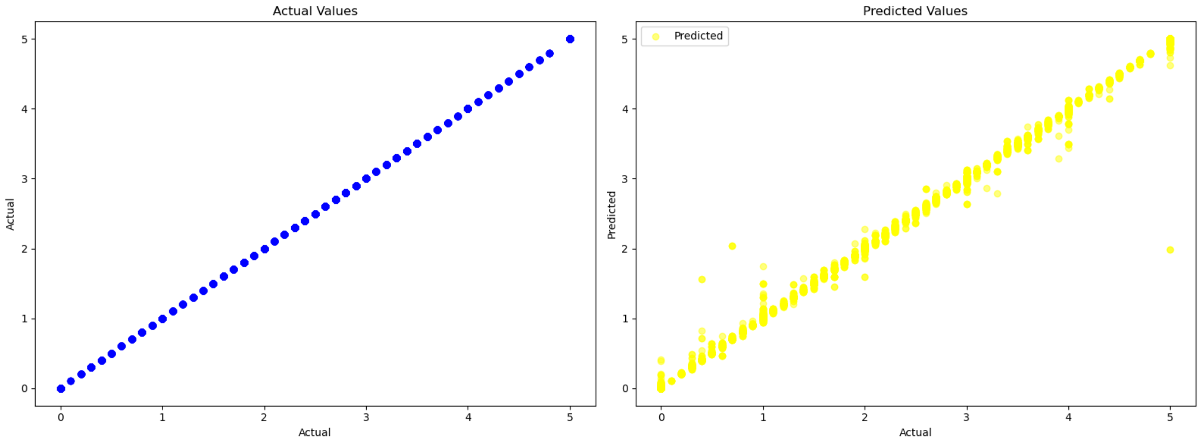

Figure 9: Actual vs Predicted values with country ESG data

6.3 Discussion

In this research, Random Forest regression technique is used to predict ESG scores first on historical ESG data of companies and secondly on combined dataset of companies and countries. Both the experiments have been successful and shows promising results and models shows high level of accuracy. So, the objective of predicting ESG ratings have been met successfully. However, another objective of this study is to analyze the impact of country ESG data on corprate ESG ratings. From the results it is clear that the model accuracy is almost same with or without country ESG datapoints. So, relationship between country ESG data on corporate ESG can not be clearly established as corporate ESG ratings can be predicted successfully to a high degree of accuracy with or without country data. One of the reason behind this could be that most of the companies these days operate from different countries and possibly their ESG rating is calculated based on their overall global performance on ESG and hence neutralizing any factor related to country ESG on them.

This experiment used a diverse set of data where companies from all regions were selected which overcame the limitations of previous work like Aue et al. (2022) and Krappel et al. (2021) where the dataset used was not diverse in nature and was limited to US market.

7 Conclusion and Future Work

This research was focused on predicting corporate ESG scores and analysing if country ESG impact this prediction or not. To examine this the historical data for ESG was collected and random forest regression model was used to predict the corporate ESG ratings. This model acted as a baseline result for the study. In the next phase, country ESG data was also collected from world bank data mart. Random forest regressor is used again to predict ESG ratings in the overall combined dataset and PCA is used for dimensionality reduction in this model. Both the experiment shows model accuracy to be around 99 percentage which is quite promising. After carefully examining the results of both the experiments it can be concluded that country ESG does not play any significant role in predicting corporate ESG because model was able to successfully predict ESG rating without country ESG data. Additionally, ESG ratings can be successfully predicted using historical data and random forest regression shows promising results in this experiment.

The data used in this study was from different regions, but it can be expanded to most recent data as well. Additionally, this search can also be expanded to include other factor like economy or social reforms in a country and analyse if that impact corporate ESG ratings in that region.

References

Ang, G., Guo, Z. and Lim, E.-P. (2023). On predicting esg ratings using dynamic company networks, *ACM Trans. Manage. Inf. Syst.* **14**(3).
URL: *https://doi.org/10.1145/3607874*

Aue, T., Jatowt, A. and Färber, M. (2022). Predicting companies' esg ratings from news articles using multivariate timeseries analysis.

Berg, F., Kölbel, J. F. and Rigobon, R. (2022). Aggregate Confusion: The Divergence of ESG Ratings*, *Review of Finance* **26**(6): 1315–1344.
URL: *https://doi.org/10.1093/rof/rfac033*

Bhandari, K. R., Ranta, M. and Salo, J. (2022). The resource-based view, stakeholder capitalism, esg, and sustainable competitive advantage: The firm's embeddedness into ecology, society, and governance, *Business Strategy & the Environment* **31**(4): 1525–1537.

Chatterji, A., Durand, R., Levine, D. I. and Touboul, S. (2016). Do ratings of firms converge? implications for managers, investors and strategy researchers, *Strategic Management Journal* **37**(8): 1597–1614.

Daying, Y. and Zi'Ao, Y. (2023). Discovering variation financial performance of esg scoring through big data analytics, *2023 Asia-Europe Conference on Electronics, Data Processing and Informatics (ACEDPI)*, pp. 141–150.

D'Amato, V., D'Ecclesia, R. and Susanna, L. (2022). Esg score prediction through random forest algorithm, *Computational Management Science* **19**(2): 347–373.

García, F., González-Bueno, J., Guijarro, F. and Oliver, J. (2020). Forecasting the environmental, social, and governance rating of firms by using corporate financial performance variables: A rough set approach, *Sustainability* **12**(8).
URL: *https://www.mdpi.com/2071-1050/12/8/3324*

Gupta, A., Sharma, U. and Gupta, S. K. (2021). The role of esg in sustainable development: An analysis through the lens of machine learning, *2021 IEEE International Humanitarian Technology Conference (IHTC)*, pp. 1–5.

Kim, S. and Li, Z. F. (2021). Understanding the impact of esg practices in corporate finance, *Sustainability* **13**(7).
URL: *https://www.mdpi.com/2071-1050/13/7/3746*

Krappel, T., Bogun, A. and Borth, D. (2021). Heterogeneous ensemble for esg ratings prediction.

Lin, H.-Y. and Hsu, B.-W. (2023). Empirical study of esg score prediction through machine learningmdash;a case of non-financial companies in taiwan, *Sustainability* **15**(19).
URL: *https://www.mdpi.com/2071-1050/15/19/14106*

Mooneeapen, O., Abhayawansa, S. and Mamode Khan, N. (2022). The influence of the country governance environment on corporate environmental, social and governance (esg) performance, *Sustainability Accounting, Management and Policy Journal* **13**(4): 953–985.

Stubbs, W. and Rogers, P. (2013). Lifting the veil on environment-social-governance rating methods, *Social Responsibility Journal* **9**(4): 622–640.
URL: *https://doi.org/10.1108/SRJ-03-2012-0035*