Nanyang Technological University
Nanyang Business School

BC2406 Analytics I – Visual & Predictive Analytics
Semester 1, 2021

Group Project: Seminar Group 8, Team 8

Machine Learning Proof of Concept for EIU

Developing a new flagship product for EIU through quantitative and qualitative comparative analysis using machine learning and non-machine learning visualization in predicting ESG and GDP of a country

Members:

Name	Matriculation Number
Wong Wei Jun	U1910841D
Lim Qing Rui	U2010816G
Lim Zi Hui	U1911430H
Tan Jin Xuan	U2010840L

Prepared for: Prof. Liu Peng

Table of Contents

Executive Summary

Environmental, Social and Governance (ESG) index is defined in this report as a set of standards that are being used to judge the entity's performance in terms of their sustainability in all 3 areas of environment, social and their governance decision making. Over the year, there has been a higher focus on ESG indexes by different stakeholders such as individual investors and even asset managers as they do look out for each individual company's ESG performance to make their investment decisions. Furthermore, multiple countries are starting to invest more resources towards having sustainable assets and technologies. From the insights derived using modern ESG analytics, the business outcome can improve companies' policies through focusing on variables and help EIU value-add to their analysis in reports through pinpointing the essential variables that are more likely to be of focus in policy making.

Economist Intelligence Unit (EIU) is an organization that was founded in London and its main business activity is to provide research and analysis report for their clients using their team of industry experts, economists, policy analysts and even consultants on multiple financial and non-financial indicators to allow more informed decision making by each individual stakeholder. This report is to introduce a new product into EIU's research and analysis reporting by providing EIU with predictive models that would study various environmental, social and governance factor in each country, identify significant factors of ESG index and ultimately predict the ESG score based on the patterns of the variables. This would allow EIU to offer a much more comprehensive analysis on sustainability for their clients.

This study is based on real world government indicator that was sourced from World Data Bank and Global Risk Profile (GRP). The team has cleaned and prepared the datasets before data exploration was performed to increase the investigative depth and width with regards with the ESG score and variables such as the correlation and coefficient between the independent and dependent variables. Descriptive models such as non-machine learning and predictive models including machine learning models were applied to analyse ESG Score and measure the accuracy of each model through train-test splits and predicted results. The team chose to target the Year 2019 as a time-series horizontal due to dissimilarities between variables and incomplete data in ESG Scores and Ranks in other years.

Among the models, the Linear Regression model has scored the lowest RMSE, and the Support Vector Machine with Linear Kernel has scored the lowest MAPE. NML using Mean, Median and Mode was also analysed and determined to be an unsuitable predictive model due to lack of statistical data obtained from it, the team also attempted to use variable and statistical significance, however, was limited by the lack of unit standardization for comparison. As such, a greater emphasis will be placed over RMSE and MAPE to determine and recommend models in machine learning. The variable importance of the top 4 variables were extracted to give brief insights into variables onto what to take note for countries and justify the differences in ranking of variable in each model through analysis of variable importance calculation.

From the insights provided by each model, the team has derived the top 4 important variables which are mainly used to predict the outcome of ESG Score. These variables are Regulation Quality, Government Effectiveness, Law in the Country, Life Expectancy. Analysis in reports can be used to target these 4 variables to propose better policy and decision making for their consumers to counteract falling behind in ESG Scores and Ranks.

To conclude, the team consolidated our findings and discussed in detail the limitations for the dataset and models for users to understand how each model derived the values and results for suitable error analysis. Further research direction, that was briefly mentioned in this report, could also be carried out such as pilot testing of the models amongst various splits which was not carried out in this report to achieve further insights and to better understand the intricacies of the models and better explain factors that could be the underlying reason behind ESG Scores. Therefore, the team hopes that insights from the predictive models and analysis can help EIU, their consumers such as companies or countries increase ESG Scores and help analyse the accuracy of machine learning models with reproducibility of results across various data themes.

1. Introduction

1.1. Overview

1.1.1. Business Opportunity

Economist Intelligence Unit (EIU) provides forecasting and advisory services through research and analysis. Their reporting includes the various financial indicators such as Gross Domestic Product Growth Rate (CAGR) and inflation rate. EIU also provides non-financial indicators such as political and economic trends and status of different countries. However, EIU only provides qualitative analysis on the non-financial factors to investors. Albeit their analysis is useful, investors may still require quantitative measures to better gauge their risk and returns, resulting in a better decision making on their end. Thus, this report serves to provide and propose various qualitative and quantitative measure of a country's non-financial status using Environmental, Social and Governance (ESG) scores.

1.1.2. Environmental, Social and Governance (ESG) scores

ESG is a set of standards evaluating the country's performance in their sustainable investment integrating economic activity with environment integrity, social concern, and governance systems. ESG serves as a guideline for investors to assess their investment opportunities in respective countries.

1.2. Reasons for Proposal

1.2.1 Shift in focus towards ESG by countries and Cost of problem

In recent years, there is a noticeable global shift in focus towards ESG by more than 100 countries. These countries have pledged carbon neutrality and global GDP 5% CAGR towards sustainability with an increase in spending on smart infrastructure, pharmaceuticals, financial technology as well as a strong focus towards increasing assets in the US and APAC regions through increasing regional asset composition in portfolio. (AllianzGI, 2020) This has led to a growth in ESG securities which is expected to grow 15% year-on-year to over $53 trillion by 2025 (Gurdus, 2021).

1.2.3 Rise in investors' concern about sustainability

An interview was conducted on 73 senior executives from 43 global financial institutional investing firms including the three top largest asset managers namely BlackRock, Vanguard and State Street showed that ESG was almost universally top of mind for these executives. Additionally, in 2006, UN-back Principles for responsible investment (PRI) was launched whereby a large group of investment companies with a total of $6.5 trillion in assets under management signed a commitment to consider ESG matters when they are making their investment decisions (Eccles & Klimenko, 2019).

1.2.3 ESG as a performance indicator

ESG metrics are used by investors to judge how a company may respond to various challenging environments and to look at the preparedness of companies for future crises or unexpected events. The action taken by company to counteract the situation today can have long-lasting implications relating to material social factors. An example would be how companies are reacting to the COVID-19 pandemic and to look at the ESG factors to assess whether companies are prepared for them. (Goldman Sachs Asset Management, 2020). Likewise, investors can look at ESG metrics of countries to determine how prepared are each country in the event of unexpected situation like COVID-19. This allows investors to make a more informed decision on their long-term investing decisions.

1.3. Project Objective and Feasibility

1.3.1. Analytical Problem

The project aims to evaluate current predictive and visualisation models to identify trends and key contributing factors leading to variances in the ESG Scores across the participating countries in the World Data Bank where factors are in relation to ESG Scores. As stated in the proposal, the key objective is to enhance and value-add to ESG Forecast based on upcoming trends of ESG factors and creating new product variety for ESG global outlooks, through a more predictable and accurate data analysis model (by analysing the impact of machine learning versus non-machine learning) to expand the economic and business product variety of EIU. Keeping this in mind, the predictive model accuracy

chosen should reflect the reduction of cost, Statistical Significance and Confidence Interval. The chosen reduction of cost (e.g. 5%) or confidence interval (e.g. 95%) will give the predictive model the lower bound range of the accuracy. Predictive accuracy derived below the lower bound will be considered an ineffective model of prediction in this report, with an evaluation given for each effective and ineffective model.

1.3.2. Project Objectives

In this project, we would be conducting descriptive analysis on the ESG score and the different ESG variables to find out their correlations. Afterwards, the team would be constructing various predictive models using both non-machine learning and machine learning methodologies to provide EIU with a model for predicting ESG scores for other countries by using a set of ESG variables. At the same time, the team would be doing a comparison analysis for both methodologies to evaluate the best model that is the most accurate among all the different models. From the predictive model, EIU would then be able to compute ESG scores for different countries based on the ESG variables and include it into their reports to better value add to the quality of the report that they are providing for stakeholders and investors.

Pilot studies(s) at EIU to evaluate Machine Learning

In current studies, reports and analysis performed by EIU, they are able to provide an integrated approach to macroeconomic analysis with short, medium and long-term forecasts through time series trends and historical data. This descriptive analysis determines increasing or decreasing trends in variables affecting the dependent Y-variable that they are investigating without scenario-based analysis. Through this report, Linear Regression was evaluated to forecast time-series and predictable data for trend analysis and computation due to its low RMSE. This can be performed using various train-test split ratios using historical data collected by EIU and passing data to their target consumers as trials to evaluate direction and policies that their consumers will take up after reading the data from various splits. Machine learning models such as CART/SVM can be used to forecast variable importance or factors that greatly affect the presence of ESG Scores (Y dependent continuous variable). Following the various splits for CART and SVM, EIU can collect data about their Machine Learning through consumer policy implementation and analysis of other machine learning models over the next several years. *(EIU., n.d)*

1.3.3. Project Feasibility

Availability of Data

In this project, it is feasible to construct the predictive models due to the availability of ESG related variable data on the website. There are multiple reputable institutions that provides world government indicator for most countries publicly. However, ESG scores may be more restrictive and are usually only available for purchase from ESG service providers such as ISS or Moody's. Nevertheless, the team was still able to obtain the sufficient ESG score information from a reputable risk management company that also monitors ESG closely within their business activity, to target the following project feasibility needs:

1. **Predictive Need in the ESG Industry:** Institutions and countries lacking behind might not have a benchmark on indicator or unable to determine factors that play a more important weightage in their ESG Scores/ ESG Ranks. The provision of an analytical model helps to predict and target the unknowns in the variables or show countries that are at the forefront of the ESG space to allow laggards to take pre-emptive measures to increase their ESG Scores.
2. **Lack of perfect knowledge:** The current benchmarks in the ESG industry do not allow for prediction of variable comparable to create predictive models for companies or countries. Hence, an analytical model when derived will provide significant insights into models that are suitable for predicting variables allowing target consumers to be more aware of their ESG Ranks and Scores.
3. **Availability of data for train sets and test sets:** There are many ESG datasets allowing for extensive analysis on the current issue of ESG, such as the World Data Bank datasets, RobecoSAM Country Sustainability Ranking, or the ESG Index. In this analysis, we will use data comprising of over 170 countries and the 2019-year data comprising of over 11 variables spanning a total of 2000 datapoints. This allows sufficient points for us to segregate the data into train and test sets to obtain predictive models and analysis.

2. Business Objectives

With a predictive model for sustainable development, it will allow companies/countries to be better able to identify governmental movements and effectiveness, social risks, and environmental factors to scale in the industry.

1. **Identify salient variables for ESG growth:** Inputs and variables are explored during predictive and descriptive models to evaluate the weightage in each model and statistical significance. Companies can turn towards essential factors or enhance their current policies to further leverage on the models to create greater business values to consumers and investors.
2. **Assessing risk and pitfalls of current policies in place:** The ability to assess risk and re-evaluate current policies allows companies to reduce risk and hindsight in policy planning. The country/company can identify intervention points of their current policy. An example would be a governmental factor of Regulation Quality affecting ESG Score, and consumers of the report (countries) can implement policies supporting better regulation such as higher transparency, allowing for intervention using this project's recommendations.
3. **Identifying intervention points:** The models describe variable importance or a coefficient which has a weight on determining ESG Scores. These can be translated to when and how a countries/companies should intervene to prevent themselves from trailing behind leading countries or to follow implementations of other leading countries.

2.1. ESG Analytics

As ESG is a pivotal area of focus due to assets growing to USD$53 trillion by 2025 and a compound annual growth rate (CAGR) of 15% Y-o-Y, the analysis and impact of the industry can allow companies to predict and position themselves to become future-ready and socially accepted, by conforming to variables that determine their ESG positions. EIU can leverage on ESG as a new business product and analysis tool to help companies and countries, allowing them to penetrate this new industry through performing greater in-depth machine learning analysis. Such reports serve as a benchmark for various companies and countries to poise themselves at the forefront of the industry and acquire better investor confidence and support from the public. Companies may also choose to invest in countries with greater ESG Scores and ranking to leverage on marketing and government endowments to further grow their businesses.

Enhancing GDP forecast based on upcoming trends of ESG factors and allows for creation of new product variety for ESG global outlooks, through a more predictable and accurate data analysis model (by analyzing the impact of machine learning versus non-machine learning) to expand the economic and business product variety of EIU. It is hence important to identify variables that can impact the ESG Rankings and Scores to allow EIU's consumers to pivot towards factors that can greatly enhance their impact and position in the industry, with a global 20% annual growth rate in ESG data expenditure, and the ESG data industry reaching a net worth of US$1 billion by the end of 2021. *(Kenway N., 2021)*

2.2. Analytics comparison metrics

2.2.1. Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) is defined by standard deviations of the residuals between the testset and the predicted results from the testset data. It measures the errors between the regression line data points, with a smaller model being favorable. RMSE shows how spread out the residuals are, and concentration of the data around the line/hyperplane of best-fit *(StatisticsHowTo., n.d)*. This standardizes the data unit used and provides insights for standard basis of comparison across models.

2.2.2. Mean Absolute Percentage Error (MAPE)

The calculation of mean Absolute Percentage Error (MAPE) is done by standardizing the average absolute percentage errors of forecasts, and similarly to RMSE, MAPE negates negative and positive errors through standardizing errors as a percentage value of the actual data, making it useful in forecasting with a smaller value being more favorable *(Springer., n.d)*.

2.2.3. Statistical/Variable Significance

In this report, Statistical and Variable Significance is used to compare across models to determine whether the model can pick out variables of interest showing significance in determining ESG Scores. The models recommended should be able to pick out variables that are attributed to and provide a weightage in determining ESG Scores, rather than selecting variables based on chance *(Gallo A., 2016)*.

3. Data preparation

3.1. Data sources

Constructing of predictive model requires the team to explore the ESG variables and the actual ESG scores. The primary datasets used in the analytics are ESG_Variables and ESGindex2020. The ESG_Variable's dataset contains 10 ESG variables that are obtained from The World Data Bank. The other datasets contained the actual ESG score for 2020 and are obtained from Global Risk Profile (GRP).

The World Data Bank and Global Risk Profile (GRP) are chosen to obtain our data source because of the reliability and the sufficiency of the data provided by respective organizations.

3.1.1. Data Reliability

The World Data Bank

The World Data Bank (WDB) is an internationally recognized entity that provides various world government indicators dataset. WDB works closely with international statistical community including United Nations and the Organization for Economic Co-Operation and Development (OECD) to compile international data sets and at the same time, develop appropriate template and procedures to ensure that the data are being compiled in a manner that still maintains the reliability and integrity of the statistics. WDB then assembles, analyse, and distribute the data online publicly.

Global Risk Profile

GRP is a leading Swiss company that specializes in third party risk management related services. GRP also monitors local and international regulatory trends through risk indexes such as Global Corruption Index (GCI) and ESG. GRP evaluates 176 countries based on 44 carefully selected datasets that originates from 17 international and reputable organization such as United Nation and World Bank *(Profile, n.d.).*

3.1.2. Data Sufficiency

WDB contains 67 ESG variables over 193 countries across every region and the data spans all the way back from 1960 to 2020. This allows the team to carefully pick out 10 variables out of 67 variables to conduct the analysis.

GRP provides ESG Scores over 176 countries for 2020. This provides enough ESG score to construct the predictive model together with the 10 ESG variables for all 176 countries.

3.2. Data Cleaning and Splitting

There would be a total of 10 ESG variables that would be picked from the 67 variables provided by WDB based on the availability of data and to ensure each variable are independent of each other. Out of the 10 ESG variables, the composition of ESG variables would be 20%, 30% and 50% for Environmental, Social and Governance respectively. This composition is selected based on the same ratio that GRP has used for their computation for the ESG scores. This would ensure that the weightage for each factor would be like the weightage used by GRP to arrive at the ESG scores.

The data for both CSVs used are split into **a 70-30 ratio for train and test sets** for all NML and ML models, with a **set.seed(2004)** function used to control the method of splitting and generation of pseudorandom numbers that mimic the properties of independent generations of a uniform distribution in the interval of (0,1) (0,1) (0,1).

3.2.1. Data Types

All the variables in the datasets are continuous and they would be formatted into integers that would be the most appropriate datatype. Information with regards to the 10 data variables used can be found in **Appendix A**: Data dictionary.

3.2.2. Combination of Datasets

There are a total of 2 Comma-separated values (CSV) files that the team is using to conduct the necessary analysis. The first CSV file namely ESG_Variables contains 10 ESG variables from 193

countries from the period 1996 to 2019. The second CSV file namely ESGIndex2020 contains the ESG score for 176 countries. The datasets were obtained from the World Data Bank and RobecoSAM website, where they were subsequently combined and subset using inner joins and country names in the R Codes. A combination of both data table was carried out and records were adjusted such that countries that does not exist in the ESG Score data table are removed and only variable data from 2019 are obtained in this case. Keeping the NA values, the team has used this dataset for our Classification and Regression Trees (CART) analysis, and dataset labelled 'Dataset for CART.csv'.

3.2.3. Handling Missing Values

From combined dataset, there are a total of 352 missing values. Through observations, all 352 missing values all come from 2 variables in columns named "Carbon" and "Nitrous". The NA values are estimated by using the mean value from previous period for each respective country, where the data was grouped by each country and variable type inclusive, and the mean was obtained and replaced each NA value respectively in accordance to the country which the mean was derived from. This data set has been cleaned and labelled 'Dataset for NML and SVM.csv'.

3.2.4. Removal of Unnecessary Variables

Variables that are not required for our analysis were removed from the data table to ensure that all the information provided in the data tables are essential to the project. Variables that are mainly removed are redundant columns such as "YearCode". All other correlated and ESG Variables were kept as they could provide greater in-depth analysis and insights during data exploration, data manipulation and analytical modelling. Data cleaning might also occur during the modelling process to provide further width and depth with data exploration, and to present further analysis.

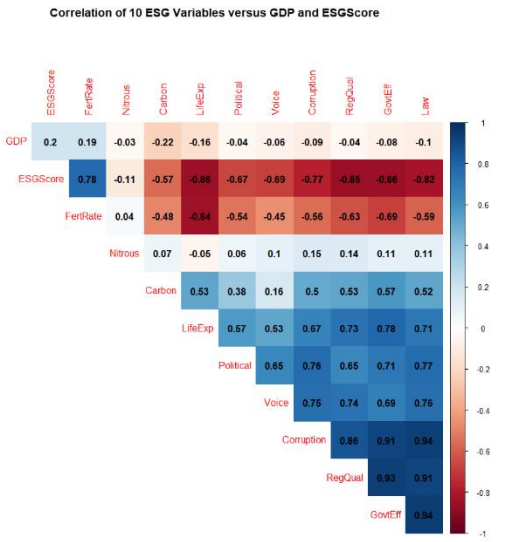
3.2.5 Renaming of variables

All variables were renamed to provide the team with more convenience and to allow RStudio to perform data manipulation with greater ease when carrying out each analysis. The renamed variables and the original name of the variables can be found in Appendix B.

3.3 Data comparation and exploration

A correlation matrix has been plotted from the figure (on the right) based on all data points in the CSV; the following observations can be derived:

1. Law is highly positively correlated to Government Effectiveness variable.
2. Corruption which is a Social factor when compared to Regulation Quality, Government Effectiveness, Law which are Governance factors in ESG are highly positively inter-correlated amongst each other.
3. Fertility Rate and Life Expectancy are highly negatively correlated amongst each other despite being Social factors.
4. Fertility rate is highly positively correlated to Y-variable that we are investigating which is ESG Score.
5. Life Expectancy, Regulation Quality, Government Effectiveness and Law are highly negatively correlated to Y-Variable ESG Score that we are investigating.



4. Non-Machine Learning Methodology

The team has decided to make use of non-machine learning (NML) methodology to compare and to prove that machine learning methodology is much superior in terms of prediction accuracy. In this section, mean, median and mode would be used to conduct a descriptive analysis between the ESG variables and the actual ESG Score. The team would try to use predictive analysis conducted using the mean value of each respective regions.

4.1 Description Analysis

4.1.1. Mean, Median, Mode

In this area, we investigate the variables across 2019 in general against ESGScores to check for anomalies in the data. To understand the distribution for each variable, box plots were generated. Box plots provide a visualisation of the distribution of data through 5 values: minimum, first quartile, median, third quartile, and the maximum. All boxplot can be found under **appendix C**.

For easier reference to the figures, a summary table is shown below.

Variable	Minimum	Q1	Median	Q3	Maximum	Mean
Rule of Law	-2.32	-0.65	-0.06	0.6	2.02	0.02
Voice and Accountability	-2.13	-0.74	0.05	0.85	1.59	0.02
Political Stability and Absence of Violence	-2.65	-0.54	0.015	0.7	1.66	0.01
Government Effectiveness	-2.02	-0.69	0.035	0.64	2.22	0.05
Regulatory Quality	-2.36	-0.7	-0.06	0.66	2.16	0.04
Control of Corruption	-1.72	-0.76	-0.08	0.67	2.16	0.02
Life Expectancy at Birth	54.33	68.04	73.55	77.86	84.36	72.86
CO2 Emissions	0.03	0.71	3.66	7.07	36.99	4.94
Nitrous Oxide Emissions	0.01	0.24	0.47	0.59	3.74	0.57
Fertility Rate	0.92	1.66	2.39	3.44	6.82	2.59
GDP	-6.78	1.31	2.67	4.99	18.72	2.92

Based on the table above, the following can be derived:

1. Law has a maximum to minimum range of 4.34 and a median of -0.06. As seen in Figure 4.1.1.1, there are no outliers, and the distribution of Law appears to be slightly positively skewed with a higher mean of 0.02. The interquartile range (IQR) of 1.25 is relatively small compared to the range of the box plot, thus suggesting that most of the countries achieved a Law index between -0.65 and 0.6.

2. Voice has a maximum to minimum range of 3.72 and a median of 0.05. As seen in Figure 4.1.1.2, there are no outliers and the distribution of Voice is negatively skewed, with a lower mean of 0.02. The IQR of 1.59 appears to be in proportion with the range of the data, with the median being in the middle of the IQR, suggesting that the middle 50% of the data is evenly spread.

3. Political has a maximum to minimum range of 4.31 and a median of 0.015. As seen in Figure 4.1.1.3, there is an outlier of –2.65 and the distribution of Political Stability is negatively skewed, with a lower mean of 0.012. The IQR of 1.24 is relatively small as compared to the range of the data, thus suggesting that the variability of the middle 50% of data is smaller, between –0.54 and 0.64.

4. GovtEff has a maximum to minimum range of 4.24 and a median of 0.035. As seen in Figure 4.1.1.4, there are no outliers and the distribution of GovtEff is slightly positively skewed, with a higher mean of 0.05. The IQR of 1.33 is also relatively small as compared to the range of the data, thus suggesting that the variability of the middle 50% of data is smaller, between –0.69 and 0.64.

5. RegQual has a maximum to minimum range of 4.52 and a median of –0.06. As seen in Figure 4.1.1.5, there are no outliers and the distribution of RegQual appears to be slightly positively skewed with a higher mean of 0.04. The IQR of 1.36 is relatively small compared to the range of the data, thus suggesting that the variability of the middle 50% of data is smaller, between –0.7 and 0.66.

6. Corruption has a maximum to minimum range of 3.88 and a median of –0.08. As seen in Figure 4.1.1.6, there are no outliers and the distribution of Corruption is positively skewed, with a higher mean of 0.016. The IQR of 1.43 is relatively small as compared to the range of the data, thus suggesting that the variability of the middle 50% of the data is smaller, between –0.76 and 0.67.

7. LifeExp has a maximum to minimum range of 30.03 and a median of 73.56. As seen in the figure on the left below, there are no outliers and the distribution of LifeExp is negatively skewed, with a lower mean of 72.86. The IQR of 9.83 is small as compared to the range of the data, thus suggesting that the variability of the middle 50% of the data is smaller, between 68.04 and 77.86.

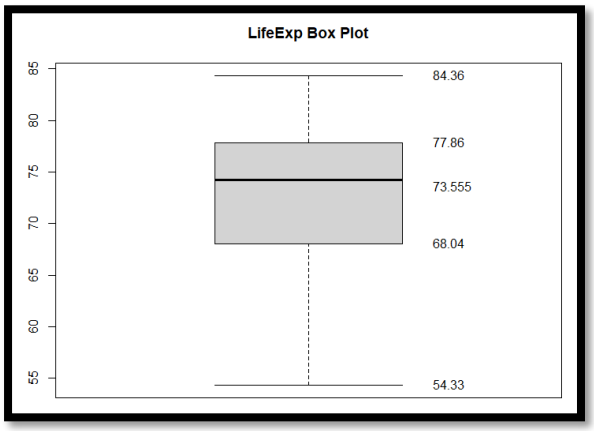


Figure showing LifeExp Box Plot

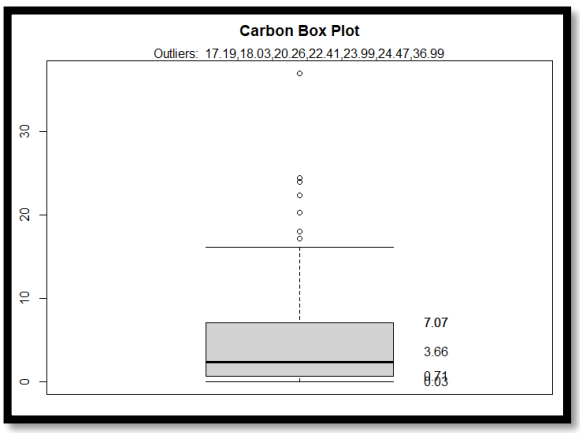


Figure showing Carbon Box Plot

8. Carbon has a maximum to minimum range of 36.96 and a median of 3.66. As seen in the right figure above, there are 7 outliers ranging from 17.19 to 36.99 and the distribution of Carbon is positively skewed, with a higher mean of 4.94. However, the mean may be influenced by the presence of outliers and hence, using IQR will be a better representation of the variability of the data. The IQR of 6.36 is small compared to the range of the data, thus suggesting that the variability of the middle 50% of the data is smaller, between 0.71 and 7.07.

9. Nitrous has a maximum to minimum range of 3.73 and a median of 0.47. As seen in Figure 4.1.1.7, there are 13 outliers ranging from 1.13 to 3.74 and the distribution of Nitrous is positively skewed, with a higher mean of 0.57. However, given the presence of the outliers, the mean is likely to be skewed as well. Hence, the IQR will be a better representation of the variability of the data. The IQR of 0.35 is small compared to the range of the data, thus suggesting that the variability of the middle 50% of the data is compacted between 0.24 and 0.59.

10. FertRate has a maximum to minimum range of 5.9 and a median of 2.395. As seen in Figure 4.1.1.8, there is an outlier of 6.82 and the distribution of FertRate is positively skewed, with a higher mean of 2.59. The IQR of 1.78 is small compared to the range, thus suggesting that the variability of the middle 50% of the data is smaller, between 1.66 and 3.44.

11. GDP has a maximum to minimum range of 25.5 and a median of 2.67. As seen in Figure 4.1.1.9, there are 3 outliers ranging from -6.78 to 18.72, and the distribution of GDP is positively skewed, with a higher mean of 2.92. The IQR of 3.68 is relatively small compared to the range of the data, thus suggesting that the variability of the middle 50% of the data is compacted between a smaller range of 1.31 to 4.99.

4.1.2. Limitation of the Box plot

In the case of our dataset, box plots are useful for comparing the performance of different countries for each variable, as well as understanding the distribution of the data. However, box plots do not draw any meaningful comparisons for the performance against each variable as the characteristic of each variable varies. It is not advisable for predictive analysis to be drawn from such descriptive variables as they do not provide accurate precision, recall and accuracy, due to outliers affecting the quartiles as shown in the Nitrous boxplots where data can be clustered at one end of the quartiles. Boxplots do not allow for us to perform train and test sets for variables or obtain predicted data based on test sets, resulting in lack of data analysis in prediction depth. The stability of the model through boxplot is affected by the presence of outliers, which will be shown when the team predicts the mean values of ESG Scores across regions, resulting in high RMSE and MAPE showing that outliers result in skewed data.

4.2. Predictive Analysis

4.2.1. Mean Value of Regions

In this predictive analysis, the team would be splitting the data set of each respective region namely Asia, Africa, Oceania, Europe, South America, and North America to 70% and 30%. Subsequently, the ESG Score mean value of the 70% countries would be computed and served as the predicted value for the remaining 30% countries. Boxplot for the 70% data set for all 6 regions is plotted to observe the variability of the ESG scores between different countries within the same region. **Appendix C (Figure 4.2.1.2 to 4.2.1.6)** provides the boxplot result for all regions. Subsequently, Root mean square error (RMSE) is used to measure the accuracy of the prediction model.

Computation of Predicted ESG Value using Mean Over Region

Region	Total number of countries per region	Predicted mean value
Asia	40	53.61
Africa	50	66.39
Europe	43	23.66
North America	21	42.14
Oceania	10	50.64
South America	12	44.60

Measurement of Accuracy of Model using RMSE

Region	RMSE
Asia	14.19
Africa	7.67
Europe	12.81
North America	15.09
Oceania	23.51
South America	8.55

From the RMSE result provided, the average RMSE across all 6 regions is **13.64**. The MAPE of all regions displayed an average of **0.3398**, or **33.98%**. In addition, statistical and variable significance was unable to be determined from the mean, median and mode from the descriptive analysis. This will serve as a benchmark for the comparison analysis between NML and ML methodology.

4.2.2. Limitations of NML (Mean, Median and Mode Boxplot) Model

With the variances in RMSE across all regions, this could show that boxplot is not an excellent indicator in substituting as a predictive model and analysis. This could be due to difference in ESG policies across different countries and regions and should not be a benchmark used to segregate the data derived. The RMSE will subsequently be used to compare across other Machine Learning Models to analyse suitability in EIU's business lines and future reports. The high MAPE of the variables displayed showed that there could be possibility of model overfitting, or that data was skewed for certain variables such as Carbon and LifeExp, resulting in high MAPE across regions and an overall high MAPE.

The usage of mean value of each region to predict the ESG scores for countries within the same region is a method considered to be a simple sample mean prediction method to figure out trends and to predict outcomes about the larger population. This method could be used to obtain ESG scores for countries that do not have any available ESG Scores and yet stakeholders may still wish to find out the ESG score of that country. However, this method has a huge flaw in terms of the sheer number of assumptions that are being made when ESG score of another country is used to predict the ESG score of other countries. Albeit being in the same region, countries may differ greatly from each other in terms of their environmental, social and governance factors. For example, even though Singapore and Afghanistan both belongs to the Asia region, we know for a fact that the ESG of Afghanistan is not comparable with Singapore due to the huge gap in governance effective, terrorism rate and even social decisions made by both countries. This could either over or understate the ESG score of countries and the result may not be useful in decision making for investors that seeks accurate and reasonable prediction.

5. Machine Learning Methodology

5.1. Introduction

The team has decided to develop three machine learning models to predict ESG Score as a continuous variable and compare the accuracy of machine learning methodologies with the non-machine learning used. In this section, Linear Regression, Classification and Regression Trees (CART) and Support Vector Machines (SVM) would be used to perform a predictive analysis between predicted ESGScore and actual ESGScore. To avoid the overfitting issue and accurately evaluate the models, the team applied 70-30 train-test split on ESGScore. Train set is used to predict the ESGScore and test set is used for testing the predicted ESGScore.

5.2. Linear Equation and Linear Regression

5.2.1. Linear Equations

Linear equation is an algebraic equation in which provides the visualisation of the relationship between two variables. The team conducted the linear equations to identity the relationships between ESGScore and 10 ESG variables. Linear regression was used in this analysis as the Y variable was continuous, in compared to logistic regression which is usable on categorical Y-variables.

Data visualization provided a better view of the relationship for each variable. Plot function and ggplot are used to generate the data visualization in this report. All plot and graph can be found under **Appendix D.**

Firstly, the team ran the linear equation for each ESG variables individually using the train set. The ESG variables are Rule of Law (Law), Voice and Accountability (Voice), Political Stability and Absence of Violence (Political), Government Effectiveness (GovtEff), Regulatory Quality (RegQual), Control of Corruption (Corruption), Life expectancy at birth (LifeExp), CO2 emissions (Carbon), Nitrous oxide emissions (Nitrous) and Fertility rate (FertRate).

Based on all the figure generated, all the variable showed a negative relationship with ESGScore except Fertility Rate.

Variable	Relationship (-/+)	Coefficient
Rule of Law	-	-16.47
Voice and Accountability	-	-13.22
Political Stability and Absence of Violence	-	-15.42
Government Effectiveness	-	-17.71
Regulatory Quality	-	-17.05
Control of Corruption	-	-15.5
Life Expectancy at Birth	-	-2.292

CO2 Emissions	-	-1.674
Nitrous Oxide Emissions	-	-7.608
Fertility Rate	+	12.01

The following observations could be derived from Linear Regression and plotting of variables when compared to ESG Scores:

1. **Rule of Law:** With reference to Figure 5.1.1.1, there is a negative relationship between ESGScore and Law. The law data point formed a decreasing gradient in the graph plotted. An increase in a unit of Law would decrease ESGScore by -16.47 units.

2. **Voice and Accountability:** With reference to Figure 5.1.1.2, there is a negative relationship between ESGScore and Voice. The law data point formed a decreasing gradient in the graph plotted. An increase in a unit of Voice would decrease ESGScore by -13.22 units.
3. **Political Stability and Absence of Violence**: With reference to Figure 5.1.1.3, there is a negative relationship between ESGScore and Political. The law data point formed a decreasing gradient in the graph plotted. An increase in a unit of Political would decrease ESGScore by -15.42 units.

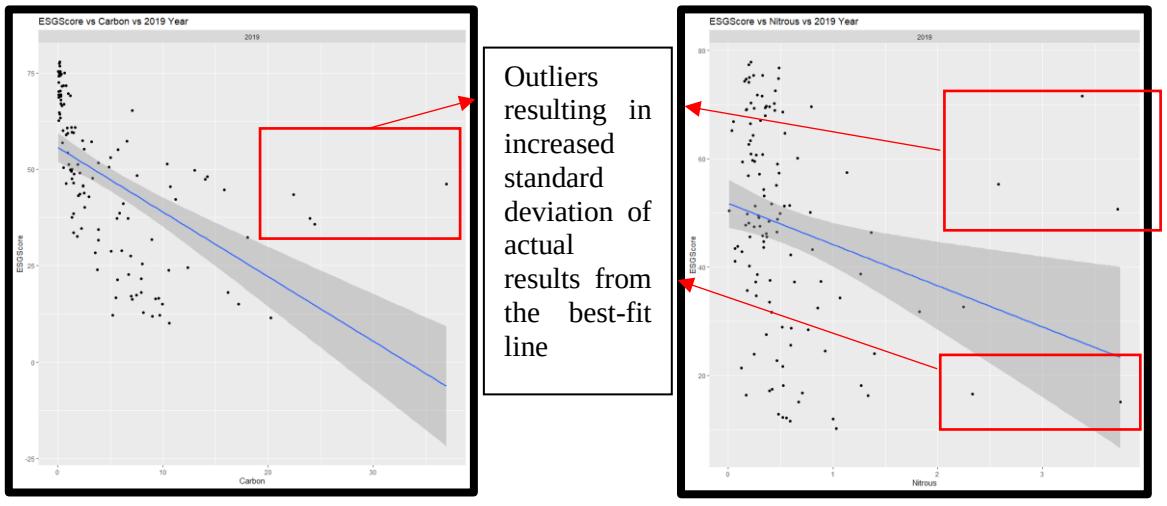
4. **Government Effectiveness:** With reference to Figure 5.1.1.4, there is a negative relationship between ESGScore and GovtEff. The law data point formed a decreasing gradient in the graph plotted. An increase in a unit of GovtEff would decrease ESGScore by -17.71 units.

5. **Regulatory Quality**: With reference to Figure 5.1.1.5, there is a negative relationship between ESGScore and RegQual. The law data point formed a decreasing gradient in the graph plotted. An increase in a unit of RegQual would decrease ESGScore by -17.05 units.

6. **Control of Corruption**: With reference to Figure 5.1.1.6, there is a negative relationship between ESGScore and Corruption. The law data point formed a decreasing gradient in the graph plotted. An increase in a unit of Corruption would decrease ESGScore by -15.5 units.

7. **Life expectancy at birth**: With reference to Figure 5.1.1.7, there is a negative relationship between ESGScore and LifeExp. The law data point formed a decreasing gradient in the graph plotted. An increase in a unit of LifeExp would decrease ESGScore by -2.292 units.

8. **CO2 emissions**: With reference to Figure 5.1 there is a negative relationship between ESGScore and Carbon. The law data point formed a decreasing gradient in the graph plotted. An increase in a unit of Carbon would decrease ESGScore by -1.674 units. There are some outliers shown in Figure 1 but there is no requirement to remove the outliers (shown in the red box) because the outliers are less influential to the overall shape of plots. *(Image shown below)*

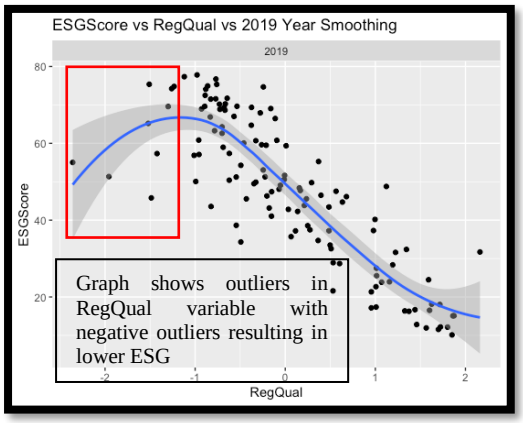


9. Nitrous oxide emissions: With reference to Figure 5.2, there is a negative relationship between ESGScore and Nitrous. The law data point formed a decreasing gradient in the graph plotted. An increase in a unit of Nitrous would decrease ESGScore by -7.608 units. The NA value of Nitrous variable has been replaced by the mean value of the previous year's Nitrous value and there are 7 outliers shown in Figure 2 in the red box *(image shown above)* which are considered as influential outliers, resulting in increased standard deviation shown in the grey shaded area surrounding the regression line.

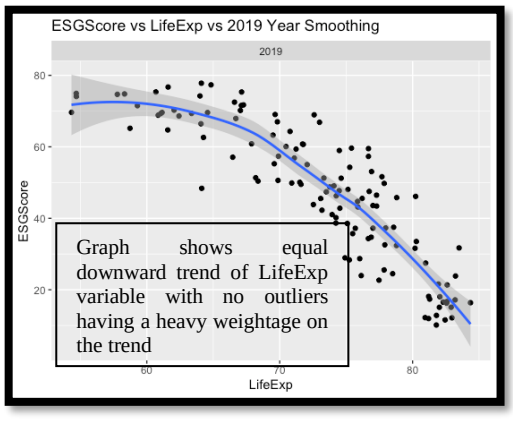
10. Fertility Rate: With reference to Figure 5.1.1.8, there is a positive relationship between ESGScore and FertRate. The law data point formed a increasing gradient in the graph plotted. An increase in a unit of FertRate would increase ESG score by 12.01 units.

Additional Regression Methodology using Smoothing

The team also attempted to plot smoothing of variables in the base year 2019 comparing variables across regions and to itself compared to ESG Scores. **(Appendix D, Figure 5.1.1.11. to 5.1.1.17.)** This allowed the team to visualise if there were any other trends across the variables, negating the outliers. Smoothing takes into account variables at the point in time in the X-axis which functions to reduce noise of the variables and reduce weightage of clustered variables in analysis or pick outliers if there are any. This helps to determine if further analysis can be obtained in the data instead of a linear trend. *(Huber J., 1977)* To perform smoothing of variables for comparison, the team used to "loess" method in the ggplot2 library, which provided a weightage for each value of the variable being analysed in compared to the relative ESG Score. In the analysis of RegQual compared to ESGScore, we can see that ESGScore is highest at a negative Regulation Quality value, which is an anomaly given that ESG is about better Governance, and thus better regulation quality. However, we can observe outliers in the left graph (red box) which could show the presence of outliers affecting the smoothing parameter. In the right graph, it is better used for smoothing as it shows an equal distribution and direction of variable trend of Life Expectancy in compared to ESG Scores. Hence, the team decided that smoothing was an analysis methodology that could be used for time-series forecast to pick up outliers if ESG Scores over a greater range of years were given, which would allow EIU to perform a time series forecast.



ESGScore vs RegQual 2019 weighted smoothing



ESGScore vs LifeExp 2019 weighted smoothing

5.2.2. Linear Regression

The team chose to conduct Linear Regression to predict ESG score as ESG score is a continuous Y-dependent variable based on the 10 independent ESG X-variables. Linear regression would estimate the coefficient of each variable which represents the statistical relationship between ESG score and ESG variables. **(Appendix D, Figure 5.1.1.1. to 5.1.1.10.)**

Firstly, the team ran the basic Linear Regression model on the trainset using the lm() function. The team used the summary function to check the statistical significance using P-value of each variable **(Figure 5.1.2.1 in Appendix)** to determine the statistically significant variables with an acceptable value of less than **0.01 statistical significance.** The P-value showed that "Law", "Corruption", "Political", "GovtEff", "RegQual", "Nitrous", "FertRate" and "GDP" are insignificant with ESGScore but the team

chose to include the insignificant variables in Linear Regression model to maintain the percentage of the composition of ESG variables.

Using the ESG Scores in 2019, we derived a linear regression equation for ESG Score which shows that **ESG Score = 71.95897 + 0.50107*(Law) – 3.70967*(Voice) – 1.22374*(Political) – 3.01566*(GovtEff) – 3.20658*(RegQual) – 0.01605*(Corruption) – 0.45347*(LifeExp) – 0.35267*(Carbon) – 1.33539*(Nitrous) – 3.07652*(FertRate).**

The Adjusted R-Squared value is 0.8796 which means that 87.96% of the variability of outcome data can be explained by the Linear Regression model, showing a good fit of variables onto the linear regression model in predicting the ESG Scores. The team used the Linear Regression model to run the predictions on the testset and calculated RMSE based on the predicted ESG score and testset ESG score. The RMSE for Linear Regression model is **6.901092.** The MAPE observed between predicted results and testset data was **0.1381269 or 13.81%** in percentage error, which was significantly lower than regional MAPE in boxplots.

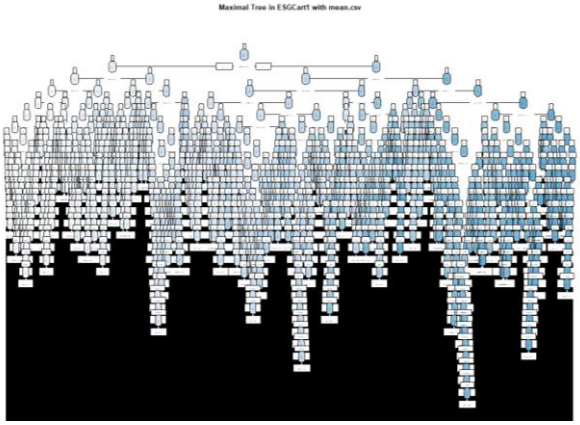
5.3. Classification and Regression Trees (CART)

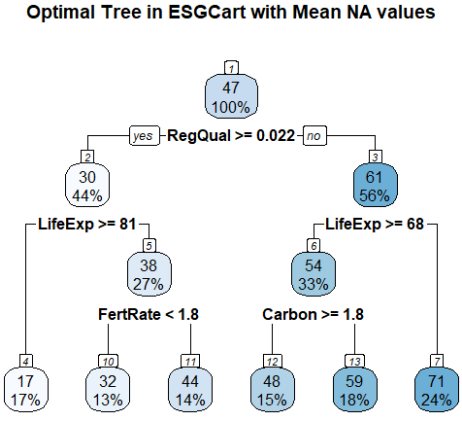
The second machine learning methodology that the team has decided to make use of is CART. This CART model would be used to predict the ESG score by using the 10 ESG variables as the input into the model. CART model is also able to identify the variables that are important in the prediction of ESG score. In Continuous Y, CART leverages on a partitioning approach to segregate variables according to the importance and weightage in affecting the ESGScore (Continuous Y). Depending on the variable importance, CART determines the minimum split based on the 1 Standard Error (1 SE) approach to obtain the best trees to display. There are many trees which can be statistically equivalent in terms of errors, and CART scales the trees down to the nodes which fulfil the SE rule where the simplest tree will perform well in predicting the ESG variables which plays a higher weightage in ESG Score. Based on the model's performance, the data is then transposed to predicting the values based on the testset to determine the outcome variables. The Root Mean Squared Error is then determines using the rmse() function in the metrics library to obtain significance of the model and evaluate the performance metrics of CART in both using the mean dataset and surrogate values.

5.3.1. CART using Mean Values

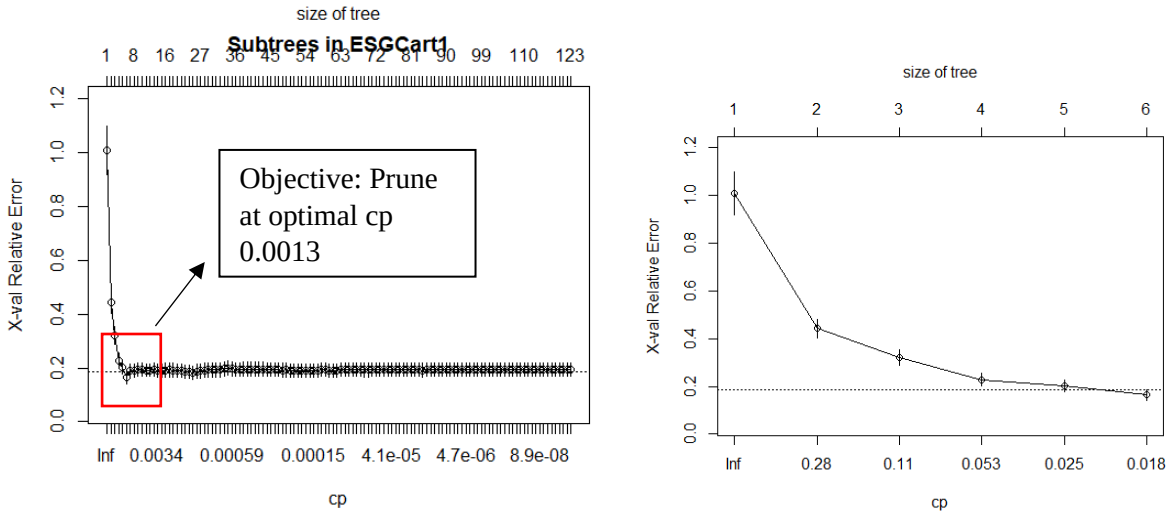
The first data table that would be used to conduct the CART would be the same data table that are used in linear regression model whereby the null values for each variable was replaced with the mean values grouped by country.

Firstly, the team would be constructing the CART model using ESGScore as the continuous Y-value and 10 ESG variables as the X-values. Subsequently, the tree would be grown to its maximum by setting the complexity parameter (cp) to 0. Then the optimal complexity parameter is computed at cp = 0.00130, and the tree would be pruned to its optimal size as shown on the right below (Unpruned trees shown on the left).





As seen from the result of the pruned tree, only “RegQual”, “LifeExp”, “FertRate”, “Carbon” are the decision rules used to carry out the binary split. The unpruned tree results in overplotting and is undesirable as it does not show statistical significance or evaluative properties in determining optimal variables in the Machine learning methodology. Model overfitting could be seen from the unpruned tree. The optimal CP region which CV error is below the CV error cap is 0.00130, and hence used to derive the optimal tree with variables in the tree (right image) at tree 5. The nodes for the optimal model are attached in **Appendix K.**



After pruning, we can see that there are lesser nodes in the optimal model as displayed on the right. The pruning sequence displays nodes that are statistically important in determining the optimal tree and following the 1 SE rule. The R printed optimal tree can be seen in **Appendix E**.

Variable Importance function is used to show significant variables which affects ESG score. The variables are “RegQual”, “LifeExp”, “GovtEff”, “Law”, “FertRate”, “Corruption”, “Voice”, “Carbon”, “Political” and “Nitrous”. Variables used in plotting CART with mean values regression tree and node split (printcp) can be found in **Appendix L.**

RMSE of the predicted result from the pruned CART model was computed and an RMSE score of **8.417404** was derived. The MAPE value obtained from our analysis showed CART to have a value of **0.1550, or 15.50%** between the predicted data and test set results.

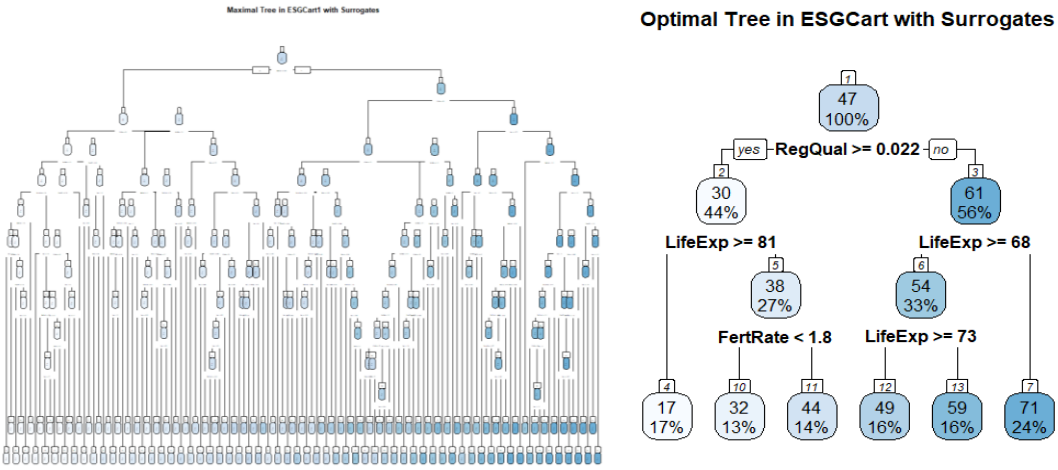
To summarize decisions rules and the nodes in pruning for all CART models used herewithin, the team used rpart.rules(CART2) to display the arguments as shown below:

0: No extra information
1: Number of observations in the node
2: Class models: Classification rate (ncorrect/nobservations), Poisson and exp models: number of events
3: Class models: Misclassification rate
4: Class models: Probability per class
5: Class models: Like 4 but don't display the fitted class
6: Class models: Probability of second class only
7: Class models: Like 6 but don't display the fitted class
8: Class models: Probability of the fitted class
9: Class models: Probability relative to all observations
10: Class models: like 9 but display the probability of the second class only

5.3.2. CART using Surrogate Values

Aside from using mean value to replace the null value and to construct the CART model using the estimated value. The team has decided to also make use of CART’s feature of automatically replacing null values in the data table using surrogate values.

Firstly, the team would be constructing the CART model using ESGScore as the continuous Y-value and 10 ESG variables as the X-values. Subsequently, the tree would be grown to its maximum by setting the complexity parameter (cp) to 0. The unpruned tree also produces an undesirable result as it results in a model overfitting. The X-relative error that results from the unpruned tree can be found in both **appendix G**. Then the optimal complexity parameter is computed at cp = 0.012, and the tree would be pruned to its optimal size as shown on the right. While the unpruned tree result can be seen on the left.



Like CART with mean, we obtained a cleaner RPlot of the trees, with the exception that CART uses surrogate values to analyse the nodes, resulting in a cleaner tree before pruning is performed. The optimal CP region which CV error is below the CV error cap, which is 6 in this instance, and hence used to derive the optimal tree with variables in the tree (right image). The lower CV Error Cap shows the sensitivity of the model in relation to unknown variables or NA values, creating a disparity in pruning sequence due to error tolerance. The R printed optimal tree can be seen in **Appendix F.**

As seen from the result of the pruned tree, only "RegQual", "LifeExp", "FertRate" and "LifeExp" are the decision variables used to carry out the binary split. . The nodes for the optimal model are attached in **Appendix M.** Variables used in plotting CART with mean values regression tree and node split (printcp) can be found in **Appendix N.** The significant variables that affect ESG Score in this CART model is similar to the previous CART model, however, this CART model does have "Carbon" and "Nitrous" as significant variable.

RMSE of the predicted result from the CART model is computed and it gave a RMSE score of **9.016454**. A higher MAPE to CART using the mean values was obtained, with MAPE of **0.1640**, or **16.40%** obtained for the surrogate CART model between predicted data and test set data.

5.3.3. Additional CART Model

Another cart model was constructed and the variables namely "Carbon" and "Nitrous" was removed because these variables have a full null value throughout the column, and the team has decided that both variables may not be useful in conducting the predictive analysis for ESG scores.

The result that was produced is the same as the above-mentioned CART model with the same decision rules and the same RMSE score. This proves that the two null variables may not have any impact on the decision, model prediction and classification error.

5.3.4. CART Model Evaluation

From the 2 CART models that were constructed for this project, both CART models produce slightly different decision rules and result. Furthermore, the accuracy of both CART models differs too. The CART model with imputation in 5.2.1 gave a better RMSE result as compared to CART model using surrogate values in 5.2.2. This is coherent with a study which showed that with about 10% missing data

in the training set, single imputation already shows a better predictive performance than just using surrogate splits. *(Feelders, 1999).* In the ESG variable dataset, there are a total of 352 null values which is equivalent to 352/1936 = 18% of the data set that would be used for the CART model. Therefore, it would be much more ideal to estimate and impute the mean values into the dataset before conducting the CART analysis to provide a more accurate prediction. Despite this, CART is known to be able to handle missing values using surrogates, which are best fit data using similar variables. If there are too many missing variables, EIU can consider using the model with surrogates in ESG Trend and Variable Analysis as including mean variables will lead to skewed data analysis *(Steinberg D., 2013)*.

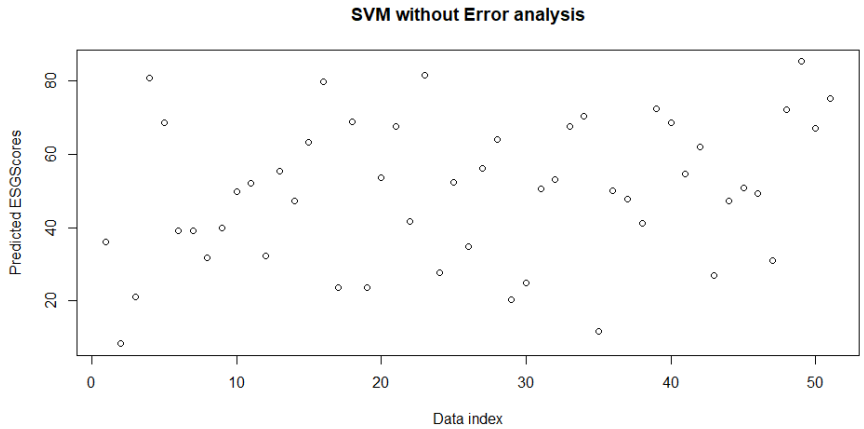
5.4. Support Vector Machines (SVM) with Linear Kernel

Support Vector Machines (SVM) provides an in-depth analysis of variables by visualizing the independent variables (X-variables) to predict the dependent variable (Y-variable) in a hyperplane dimension. Given 10 variables, we will use the variables to investigate ESGScore using a '10-dimensional plot' to find the best fit hyperplane that can classify the variables using the best-fit regression model similar to linear regression. The position and orientation of the hyperplane is influenced by Support Vectors, data points that are closer to the hyperplane. The support vectors will assist in maximizing the margin of the classifier.

Removal of the support vectors will change the position of the hyperplane, with a cost function and allowing data scientists to set misclassification errors for scenario-based analysis. These are the points that help us build our SVM. An in-depth model will be provided by RStudio to continuously re-sample and cross-validate the dataset according to dataset variables input into the model. It is also notable that SVM is unable to handle NA values, hence we will use the dataset with mean values in place of NA values to test the model.

5.4.1. SVM without Scenario/Error based analysis

The team approached SVM firstly using a 'best-fit' option where we analysed the 10 variables against ESGScore. In this model, a k-fold cross validation was performed on trainset data to resample and cross-validate the model 10 times and resample/repeat the procedure 3 times. This would obtain the best trainset model using cross-validation method for testing of the 30% split testset. *(Rpubs, n.d)* The predicted data using the test set was plot into the following graph, showing a generalised equal spread of data over all ESG Scores range. **(Results available in Appendix H.1)**



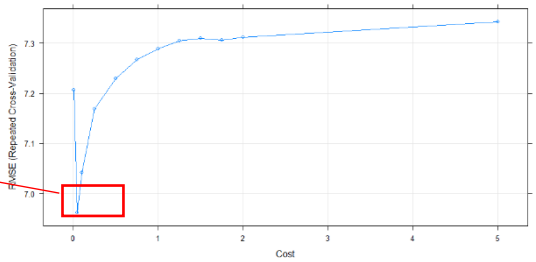
Compared with the testset data itself, the predicted test set derived a RMSE value of **8.449**. Upon comparing the resampling results with the trainset model, the team derived RMSE values of **7.309** and a model R-squared value of **0.866. (Appendix H.2)** The values displayed low RMSE and high R-squared values showing lower error and higher accuracy of the model while providing a good fit for the test set data. The MAPE displayed without scenario-based analysis is **0.1429**, or **14.29%. (Appendix H.3)** Thus, we can rely on MAPE to show that the SVM model shows forecastability of data and prediction of ESG Variables across 2019 Year. A summary of the predicted data from the testset can be found in **Appendix H.1** and variable importance with ranking found in **Appendix J.1**.

5.4.2. SVM with Scenario/Error based analysis

The second approach takes into account a predictive algorithm based on misclassification cost of the model whereby we factor in errors using the tuneGrid and tuneLength function, which lets us decide which values the main parameter will take and tells the algorithm to try different default values for the main parameter. *(Rpub., n.d)* The limits for misclassification error have been set from 0 to 5, based on possible error classification that can exist within data analysis. *(Stack Exchange., 2019)* This approach allowed the team to derive trainset models with the lowest RMSE with the values specified. In this approach we derived the lowest RMSE value of **6.961** with the R-Squared value of **0.878** with a misclassification error limit of 0.05. A summary of the predicted data from the testset can be found in **Appendix H.4.**

```
Pre-processing: centered (8), scaled (8)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 112, 113, 113, 113, 113, 112, ...
Resampling results across tuning parameters:

  C     RMSE      Rsquared   MAE
  0.00       NaN        NaN       NaN
  0.01  7.206762  0.8741166  6.054635
  0.05  6.961713  0.8785369  5.846449
  0.10  7.040905  0.8758744  5.909992
  0.25  7.169141  0.8725611  5.998483
  0.50  7.229250  0.8705441  6.036293
  0.75  7.267271  0.8693605  6.068724
  1.00  7.288946  0.8686638  6.090620
  1.25  7.305041  0.8680847  6.102290
  1.50  7.309046  0.8679600  6.104129
  1.75  7.305980  0.8681058  6.104942
  2.00  7.311696  0.8679971  6.110221
  5.00  7.342268  0.8671746  6.131088

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was C = 0.05.
```

Value with lowest RMSE and best model fit, used in SVM scenario analysis computation.



Subsequently, this model was selected by RStudio to test the testset data and it derived a RMSE value of **8.231**. **(Appendix H.5)** The results of predicted data are reflected in **Appendix H.4** and variable importance with ranking found in **Appendix J.2**. This showed that with scenario-based analysis, we would obtain lower RMSE from our predicted values, but gave the team insights into reasons leading to lower RMSE, possibly due to the low misclassification errors set within the datasets. The trainset without scenario-based analysis could be found in the error-based analysis if the team had expanded the misclassification error limits. An insight into SVM could be seen that Errors would increase if misclassification error limits increase, this leads to a higher RMSE in other models were predicted. Another reason leading to higher RMSE in the first SVM model could be due to data overfitting as well as the complexity of the second model, which took several minutes to run, resulting in additional model complexity to derive the optimal model when compared to the first model. *(IBM, 2021)* The MAPE displayed with scenario-based analysis is **0.1313**, or **13.13%** which is the lowest amongst all models. **(Appendix H.6)** This shows suitability of SVM as a forecasting methodology and disproves the overfitting of the SVM scenario-based analysis model, as MAPE has the ability to point out models that over-forecast through high MAPE values *(Lewinson E., 2020)*.

5.5. Limitations of the Machine Learning Models

<u>Linear Regression</u>

Linear regression is a machine learning algorithm that is limited only to linear relationship between the dependent (ESG Score) and independent variables (ESG variables). Even though most of the 10 variables that are chosen for the prediction of ESG score for this project has linear relationship with the dependent variable. In the future, EIU may want to include additional ESG variables into the equation to provide a more accurate and timely prediction. In this case, when the variables do not have a linear relationship with ESG Score, or if EIU decides to analyse the variables through a Categorical Y-dependent variable such as countries, it would not be able to provide a good prediction model and these variables would be considered statistically insignificant. In that scenario, logistical regression should be implemented which is not discussed in greater detail in this report.

Additionally, linear regression assumed that the variables are independent of each other. Depending on the type of variables that are chosen for the prediction for ESG score, in some cases, these variables might not be in complete independent of each other. One example is “Rule of Law” and “CO2 Emissions” variable, it can be argued that the rule of law may affect the CO2 emission within the country as some countries may have very strict environmental laws that controls the CO2 emission. This can be seen in the United States with federal policies regulating carbon emissions *(C2ES., n.d)*.

CART

CART models are particularly sensitive to missing data (MPH, 2017) and this can also be observed from the accuracy of the CART model that has no missing data (mean value imputation) which is better than the CART model that has missing data. Therefore, when variables that has many missing values are being applied into constructing the CART model, it would not be able to predict results as accurately.

Another drawback of using CART models is that any slight changes in input feature of the CART model can have a big impact on the predicted outcome which is undesirable, and it will make the tree to be very unstable such that a small change in the training dataset can create a completely different tree. This is because the split is dependent on the parent split (first split feature), so when the first split feature changes, the entire tree structure will change *(Molnar, 2021).* As seen from the team's results, the node splitting between CART using mean and CART using surrogates were significantly different especially at the subset of the LifeExp node, resulting in accurate reporting of variable if inaccurate data was provided, and this can limit countries/companies using the variable in executing proper policies according to the model selected.

SVM

Support Vector Machine, despite being able to work with scenario and error-based calculations, requires low NA values as this can suppress the accuracy and error provided in the analysis. This was shown we the team tried to work with the Machine Learning CSV file with NA values. If null values are in place, this can result in skewed data and models generated, and hence manual data filling is required to be obtained and filled into the Null values. Taking the boxplot into consideration, with data for variables that are right-skewed, such as Fertility rate, a mean value to replace null values would not lead to accurate data exploration and models generated by the plot. Due to the complexity of the model, time is required to train the model to obtain various analysis and train outputs as suitable for the user's needs or to further understand the model with regards to error acceptability and obtaining the model required for analysis.

In the **non-scenario-based analysis**, we were able to obtain the best-fit model from SVM, but it does not generate the cut-off error limit or display the error in which we could accept the minimum cut-off for the data to be generated. Hence, the user utilising the model would be subjected to hindsight bias, not factoring in various scenarios or errors resulting in fluctuations in ESG Scores prediction, resulting in lack of accuracy in data comparison. In **scenario-based analysis**, this requires additional information about erroneous data or scenario-based analysis (continuous or categorical scenarios) to be categorized into numerical form, providing greater insights. This allows greater accuracy for the model to point the exact model to use and test the data using the best-fit trainset detailed. This could result in overfitting of data onto models causing larger variance in errors as shown in RMSE of scenario-based analysis compared to non-scenario-based analysis. *(EliteDataScience., n.d.)*

6. Recommendations and Statistical Comparisons

Models	RMSE	MAPE	Variable/Statistical Significance in Predicted Data
Non-Machine Learning (Mean value with region)	13.63678	33.98%	NA
Linear Regression	6.901092	13.81%	**Variables exhibiting P-Value<0.001:** Voice, Political, GovtEff, RegQual, LifeExp, Carbon, Nitrous, FertRate **Refer to Appendix 5.1.2.1 for Variable Importance values**
CART (Mean value)	8.417404	15.50%	**Variable Ranking according to descending root node value:** RegQual, GovtEff, Law, LifeExp, Voice, Corruption, FertRate, Carbon, Nitrous, Political **Refer to Appendix I.1 for Variable Importance values**

CART (Surrogate)	9.016454	16.40%	**Variable Ranking according to descending root node value:** RegQual, LifeExp, GovtEff, Law, FertRate, Corruption, Voice, Political **Refer to Appendix I.2 and I.3 for Variable Importance values**
SVM with Linear Kernel (Non-Error based analysis)	8.449	14.29%	**Variable Ranking according to descending R-Squared variable importance value:** RegQual, LifeExp, GovtEff, Carbon, Law, FertRate, Corruption, Voice, Political, Nitrous **Refer to Appendix J.1 for Variable Importance values**
SVM with Linear Kernel (Scenario/Error based analysis)	8.231	13.13%	**Variable Ranking according to descending R-Squared variable importance value:** RegQual, LifeExp, GovtEff, Law, FertRate, Corruption, Voice, Political **Refer to Appendix J.2 for Variable Importance values**

6.1. Differences between Non-Machine Learning (NML) and Machine Learning (ML)

The difference between NML and ML is that machine learning algorithms will improve when more data are being fed into the model, whereas non-machine learning algorithms does not improve with more data. While ML uses algorithms to parse data, learn from the data and make informed decisions based on what it has learned *(Grossfield, 2020)*. NML predicts using a fixed mathematical technique. Non-machine learning does not usually have predictive values as the presence of outliers/anomalies can results in skewed data, and hence only provide descriptive value. The predictive analysis brought about by NML can only occur when a low variance between data occurs such as RMSE reflecting itself as a similarly low value when compared to the ML models (see above table). This results in NML becoming a more descriptive analysis and providing lesser value as real-life datasets are often skewed, and hindsight bias accompanied with reduction in detection complexity can occur when we use NML models to justify future data. ML, on the other hand, is able to leverage on mathematical solutions and models to account for errors and test for variable-model fit through multi-variable and plane analysis, resulting in better predictive modelling and prescriptive analysis (SVM Model). *(Cheng Q., et al., 2018)*

6.1.1. Statistical/Variable Comparison

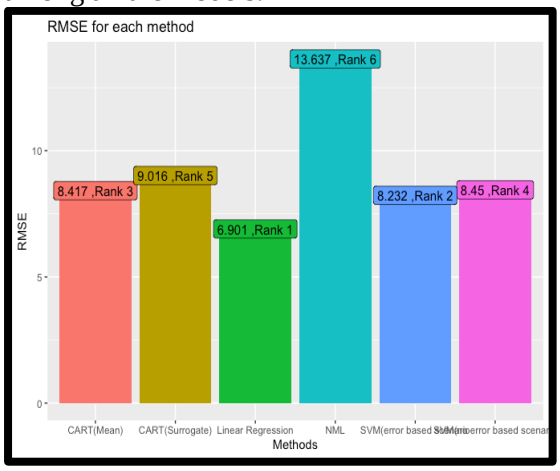
Through a quantitative analysis, we managed to obtain variables that affect the determination of accuracy and outcome of the models. Linear Regression provided P-Value of less than 0.001 for 8 out of the 10 variables, excluding Corruption and Law variables. Linear regression is also unable to rank variables according to the importance in determining ESG Scores unlike CART and SVM. However, Law consistently appeared for CART and SVM, and this discrepancy was a point-of-inquiry for the team. According to CART with mean values, "RegQual", "LifeExp", "FertRate" and "LifeExp" nodes which were used to carry out the binary split showed similar rankings of the top 4 variables of Regulation Quality, Life Expectancy, Government Effectiveness and Law as the main determinants of ESG Scores in Countries. This was also similar to the top 4 variables in variable importance as given by SVM, which smooths the error between outcome and predictor and calculates the R-Squared values for the variable subsequently. *(GitHub., n.d)* As the **statistical computation of variable importance is not standardised between models**, with NML not being able to quantify statistically important variables, Linear Regression classifying variables by P-Value, CART ranking variables according to closer proximity to root node or if the variable is a surrogate node that has a higher value and SVM categorizing and ranking variables according to the error measure in predicted and testset data, the statistical variable comparison can only be qualified up till variables that are reflected differently in the models through their rankings. We will illustrate this with a table as seen below:

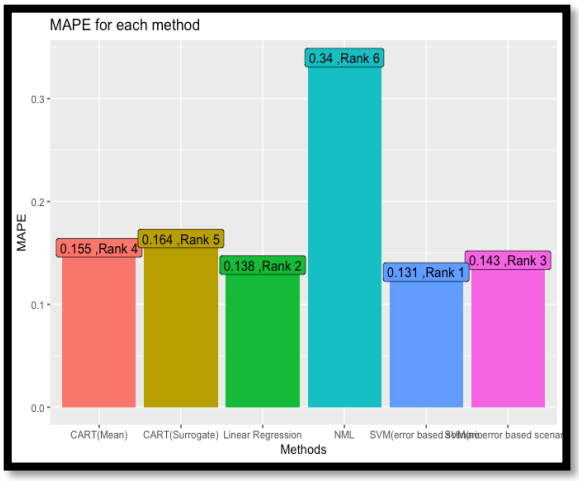
Model	NML (Boxplot)	Linear Regression	CART (Mean)	CART (Surrogates)	SVM (Scenario analysis)
Able to show statistical quantification?	NA	YES	YES	YES	YES
Variable chosen by:	NA	Lower P-Value of predicted data	• Closer proximity to root node OR • If the variable is a surrogate node that has a higher value		Error measure in predicted and testset data
Top 4 variables (Ranked)	NA	NA	1) RegQual 2) GovtEff 3) Law 4) LifeExp	1) RegQual 2) LifeExp 3) GovtEff 4) Law	1) RegQual 2) LifeExp 3) GovtEff 4) Law

Refer to Appendix fig. 5.1.2.1., Appendix I and J on variable importance and statistical significance value

6.1.2. RMSE Comparison

As observed from the RMSE comparison bar chart below ***(left image)***, the RMSE of NML model is exceptionally high at 13.63678 which shows that the standard deviation of the residuals is relatively higher, and that the data is not as concentrated around the line of best fit as compared to the other models. On the other hand, the 5-machine learning produces a lower RMSE score with a difference of about 2.11 to 1.33. Amongst the models, Linear regression models performed the best with the lowest RMSE score of 6.901092 following by SVM with linear kernel (Scenario/Error based analysis), CART (mean value), SVM with linear kernel (Non-Error based analysis) and CART (Surrogate). Therefore, in terms of accuracy measurement using RMSE, **Linear Regression** produced the most accurate prediction among all the models.





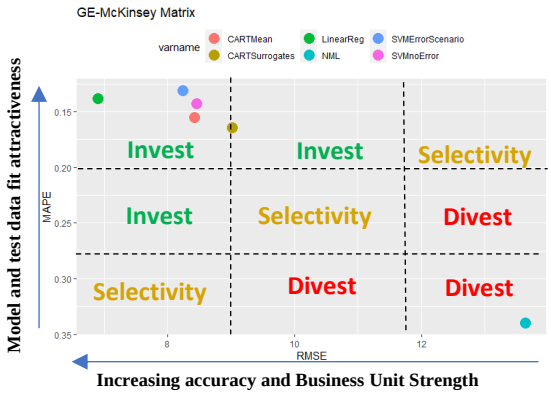
6.1.3. MAPE Comparison

As can be seen in the MAPE comparison bar chart above ***(right image)***, MAPE of the NML model has an exceptionally high MAPE value of 33.98% which meant that the predicted results of the NML model differs from the actual value at an average of 33.98%. Whereas the 5 machine learning models came close to each other with a difference of about 3.27% to 0.68%. Amongst the models, SVM with linear kernel (Scenario/Error based analysis) has the lowest MAPE percentage of 13.13% which meant that SVM with linear kernel (Scenario/Error based analysis) predicted results that defer with the actual values by the least margin among all the models. Coming very close to that is linear regression model that stands at 13.81%, followed by SVM with linear kernel (Non-Error based analysis), CART (mean value) and CART (surrogate). Thus, in terms of accuracy measurement using MAPE, **SVM with linear kernel (Scenario/Error based analysis)** produced the most accurate prediction among all the models.

6.2. Metrics evaluation and Model Selection

Between MAPE and RMSE, MAPE has a huge drawback in which MAPE puts a heavier penalty on negative errors than on positive errors because the percentage error (Actual value – Predicted value) cannot exceed 100% for forecasts that are too low *(Lewinson, 2020)*. This will result in biases in that it will systematically select a model that has low forecasts *(Tofallis, 2015)*. This meant that when the actual value is lower than the predicted value, it would give a higher MAPE than when the predicted value is lower than actual value, even though both values are the same.

Statistical significance and variable importance are not good indicators to compare between models as it does not provide a standardised unit of measure. It should only be used to evaluate predicted results obtained intra-models from test sets and thus the variables importance as reflected in the model selected should indicate to consumers of EIU such as companies or countries as to the direction of policies that they should focus on.



The GE-McKinsey Matrix ***(Left Image)*** focuses on comparing various models through 2 parameters that can determine the feasibility of usage for strategy, and segregate models based on potential investment opportunities due to high growth. In this comparison, the team has categorised the 6 models using the matrix to determine the best model based on RMSE and MAPE. The range of the matrix is determined based on Invest, Selectivity and Divest, which suggest respective excellent to unacceptable suitability of the model in predicting future data. From the matrix, we can see that Linear Regression excels at the efficient frontier, followed by SVM then CART and the model suggesting that NML is a bad indicator in predicting values. EIU, countries or companies can hence pivot data analysis towards using the models suggested by the matrix.

Therefore, 2 metrics are chosen to measure the accuracy of all 6 models. The team has concluded that RMSE and MAPE are the best metrics for predictive accuracy because RMSE gives a relatively large penalty when there are exceptionally large errors in the prediction *(JJ, 2016)*. The team strives to ensure that the model is able to predict ESG Scores as accurately as possible which makes these large residual errors to be undesirable for us while providing excellent model-data fit using MAPE. **Linear Regression** is hence recommended for EIU to test their ESG data variables (assuming it is continuous) to predict future values as it is on the comparable efficient frontier for lowest RMSE and MAPE *(see GE-McKinsey Matrix)*. The **other ML models (CART/SVM)** can be used when EIU wants to investigate variables that play a more weighted role in ESG Scores.

6.3. Usage of machine learning for forecasting and intelligence in other companies

Some of EIU's competitors include IBM and Oxford Economics. Given that these companies are also in the business of providing forecasting and advisory services vying for market share in the analytics and data market, it is crucial for EIU to differentiate themselves by providing the most accurate and valuable forecasts to customers.

IBM provides Multimodal Predictive Analytics and Machine Learning Solutions to help their customers to make better predictions about their business processes and operations. Using their proprietary software, IBM is able to perform data preparation, automation and analytics to generate powerful factors for decision making. This assists consumers with risks and mitigations in policy and decision implementation, while optimizing operational decisions. *(IBM., n.d.)* Oxford Economics uses artificial intelligence and machine learning in forecasting economic factors such as GDP, which was proven to be more accurate than conventional statistical methods *(Thompson, 2021)*. However, it is noteworthy that the nature of the COVID-19 pandemic for machine learning algorithms to carry out forecasts accurately as the timeframe may be too short resulting in an insufficiency of data collection and statistics.

6.4. Conclusion and the Value of Machine Learning to EIU

In conclusion, Machine Learning can help EIU, countries or companies stay abreast of the competition and leverage on predictive models to help describe trends or future variables that play a pivotal role in ESG Scores. Machine learning can be used to visualise geospatial data through comparative basis to find out and change the positioning of the countries in relation to other countries or regions that are doing better to benchmark themselves to increase their Scores in the ESG space, or as a time-based series forecast. As the world continues to grow towards a ESG focused industry, information becomes a critical source of advantage that can help countries strategize and develop better plans to attract investors. In order to fully utilize information, EIU can use Linear Regression to accurately forecast trends using a time-based forecast analysis methodology, and CART/SVM to forecast variable importance. This creates **a new business line**, expanding the target market that EIU can serve and the consumer base with an increased quantity of people who read ESG Reports while serving an underutilized segment of lack of ESG data and rapid evolvement of analysis globally. EIU would be able to market reports on ESG data with minimal errors and provide timely data for mitigations across countries based on errors in predictive methodology, capitalizing on the 20% annual growth rate in ESG data expenditure, expected to reach a net worth of US$1 billion by the end of 2021. *(Kenway N., 2021)*

7. References

AllianzGI. (2020). Active is: Anticipating what's ahead 2021 outlook: regional views. Retrieved from https://sg.allianzgi.com/-/media/allianzgi/ap/singapore/pdf/en/market-insights/202012-sg/202012-2021-outlook-regional-view-sg.pdf

Eccles, R. G., & Klimenko, S. (2019). Shareholders are getting serious about sustainability. Retrieved from https://hbr.org/2019/05/the-investor-revolution

Feelders, A. (1999). Handling missing data in trees: surrogate splits or statistical imputation ? Retrieved from https://webspace.science.uu.nl/~feeld101/pkdd99.pdf

Goldman Sachs Asset Management. (2020). COVID-19 and the rising importance of the 'S' in ESG. Retrieved from https://www.gsam.com/content/gsam/us/en/institutions/market-insights/gsam-connect/2020/COVID-19_and_the_Rising_Importance_of_the_S_in_ESG.html

Grossfeld, B. (2020). *Deep learning vs. machine learning: a simple way to learn the difference.* Retrieved from https://www.zendesk.com/blog/machine-learning-and-deep-learning/

Gurdus, L. (2021). ESG investing to reach $1 trillion by 2030, says head of iShares Americas as carbon transition funds launch. Retrieved from https://www.cnbc.com/2021/05/09/esg-investing-to-reach-1-trillion-by-2030-head-of-ishares-americas.html

JJ. (2016). *MAE and RMSE — Which Metric is Better?* Retrieved from https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d

Lewinson, E. (2020, November 1). *Choosing the correct error metric: MAPE vs. sMAPE*. Retrieved from Towards Data Science: https://towardsdatascience.com/choosing-the-correct-error-metric-mape-vs-smape-5328dec53fac

Molnar, C. (2021). *Interpretable Maching Learning.* Retrieved from https://christophm.github.io/interpretable-ml-book/index.html

MPH, Z. C. (2017). Evaluation of classification and regression tree(CART) model in weight loss predictionfollowing head and neck cancer radiationtherapy. Retrieved from https://reader.elsevier.com/reader/sd/pii/S2452109417302294?token=517FA47744C343703C869125CC6510FA9948A764A90BD4BC8D5874B6A608411FF4C0009F9EF589F16F26CF57263A3E04&originRegion=eu-west-1&originCreation=20211029045215

Profile, G. R. (n.d.). *About Us*. Retrieved from https://globalriskprofile.com/about/

Root Mean Squared Error Versus Mean Absolute Error. (2018, July 1). Retrieved from GitHub: https://jmlb.github.io/flashcards/2018/07/01/mae_vs_rmse/

Thompson, J. (2021, January 8). *Machine learning is a helpful forecasting tool, but challenges remain*. Retrieved from Oxford Economics: https://blog.oxfordeconomics.com/world-post-covid/machine-learning-is-a-helpful-forecasting-tool-but-challenges-remain

Tofallis, C. (2015, August). *A better measure of relative prediction accuracy for model selection and model estimation*. Retrieved from Research Gate: https://www.researchgate.net/publication/280222125_A_better_measure_of_relative_prediction_accuracy_for_model_selection_and_model_estimation

Rpubs (n.d.) TuneGrid and TuneLength in Caret. Accessed 28 October 2021. Retrieved from: https://rpubs.com/Mentors_Ubiqum/tunegrid_tunelength

Steinberg D. (2013). Introduction to CART Decision Trees. Accessed 28 October 2021. Retrieved from: https://cdn2.hubspot.net/hub/160602/file-249977783-pdf/docs/JSM

StatisticsHowTo (n.d). RMSE: Root Mean Square Error. Accessed 29 October 2021. Retrieved from: https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/

Springer. (n.d). Mean Absolute Percentage Error (MAPE). Accessed 29 October 2021. Retrieved from: https://link.springer.com/referenceworkentry/10.1007%2F1-4020-0612-8_580#howtocite

Gallo A. (16 February 2016). A Refresher on Statistical Significance. Accessed 29 October 2021. Retrieved from: https://hbr.org/2016/02/a-refresher-on-statistical-significance

Team, I. E. (2021). How to Calculate Sample Mean (with Examples). Retrieved from:

https://www.indeed.com/career-advice/career-development/how-to-calculate-sample-mean

EliteDataScience. (n.d). Overfitting in Machine Learning: What it is and How to prevent it. Accessed 29 October 2021. Retrieved from: https://elitedatascience.com/overfitting-in-machine-learning

C2ES. (n.d). Regulating power Sector Carbon Emissions. Accessed 29 October 2021. Retrieved from: https://www.c2es.org/content/regulating-power-sector-carbon-emissions/

Lewinson E. (1 November 2020). Choosing the correct error metric: MAPE vs sMAPE. Accessed 29 October 2021. Retrieved from: https://towardsdatascience.com/choosing-the-correct-error-metric-mape-vs-smape-5328dec53fac

Molnar, C. (2021). Interpretable Machine Learning. Retrieved from: https://christophm.github.io/interpretable-ml-book/index.html

Huber J. (1977). Local Smoothing: a method of controlling error and estimating relationships in consumer research. Accessed 29 October 2021. Retrieved from: https://www.acrwebsite.org/volumes/9323/volumes/v04/NA-04

Grossfield, B. (2020). Deep learning vs machine learning: a simple way to learn the difference. Retrieved from: https://www.zendesk.com/blog/machine-learning-and-deep-learning/

Cheng Q., Zhu, J., Luo, J, Xu Z., Zhu L. (2018). Directional Modulation Aided Secure Spatial Modulation. Accessed 29 October 2021. Retrieved from: https://www.researchgate.net/figure/Complexity-Comparison-for-the-ML-and-NML-Detectors-2-a-N_fig2_326661012

GitHub. (n.d.). Variable Importance. Accessed 30 October 2021. Retrieved from: https://topepo.github.io/caret/variable-importance.html

JJ. (2016). MAE and RMSE – Which Metric is Better? Retrieved from: https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d

Kenway N. (30 September 2021). ESG data spending is growing 20% per year. Accessed 30 October 2021. Retrieved from: https://esgclarity.com/esg-data-spending-is-growing-20-per-year/

EIU. (n.d) Metholodology: Overview. Accessed 30 October 2021, Retrieved from: http://graphics.eiu.com/data_services/contentguide/methodol.htm

IBM. (n.d.). IBM SPSS Predictive Analytics Enterprise. Accessed 30 October 2021, Retrieved from: https://www.ibm.com/sg-en/products/spss-predictive-analytics-enterprise

8. Appendices

Appendix A: Data dictionary for ESG variables

Factor	Data dictionary	Type of factor
Voice and Accountability	Reflects perceptions of the extent to which a country's citizens can participate in selecting their government, as well as freedom of expression, freedom of association, and a free media.	Governance
Political Stability and Absence of Violence/Terrorism	Political Stability and Absence of Violence/Terrorism measures perceptions of the likelihood of political instability and/or politically motivated violence, including terrorism.	Governance
Government Effectiveness	Reflects perceptions of the quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies.	Governance
Regulatory Quality	Reflects perceptions of the ability of the government to formulate and implement sound policies and regulations that permit and promote private sector development.	Governance
Rule of Law	Reflects perceptions of the extent to which agents have confidence in and abide by the rules of society, and in particular the quality of contract enforcement, property rights, the police, and the courts, as well as the likelihood of crime and violence.	Governance
Control of Corruption	Reflects perceptions of the extent to which public power is exercised for private gain, including both petty and grand forms of corruption, as well as "capture" of the state by elites and private interests.	Social
Life expectancy at birth, total(years)	Life expectancy at birth indicates the number of years a new-born infant would live if prevailing patterns of mortality at the time of its birth were to	Social

	stay the same throughout its life.	
Fertility rate, total (births per woman)	Total fertility rate represents the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with age-specific fertility rates of the specified year.	Social
CO2 emissions (metric tons per capita)	Carbon dioxide emissions are those stemming from the burning of fossil fuels and the manufacture of cement. They include carbon dioxide produced during consumption of solid, liquid, and gas fuels and gas flaring	Environment
Nitrous oxide emissions (Thousand metric tons of CO2 equivalent per capita	Nitrous oxide emissions are emissions from agricultural biomass burning, industrial activities, and livestock management.	Environment

Appendix B: Renamed variables

Original name	Renamed
Voice and Accountability	Voice
Political Stability and Absence of Violence/Terrorism	Political
Government Effectiveness	GovtEff
Regulatory Quality	RegQual
Rule of Law	Law
Control of Corruption	Corruption
Life expectancy at birth, total(years)	LifeExp
Fertility rate, total (births per woman)	FertRate
CO2 emissions (metric tons per capita)	Carbon
Nitrous oxide emissions (Thousand metric tons of CO2 equivalent per capita)	Nitrous

Appendix C: Descriptive analysis (Mean, median & mode)

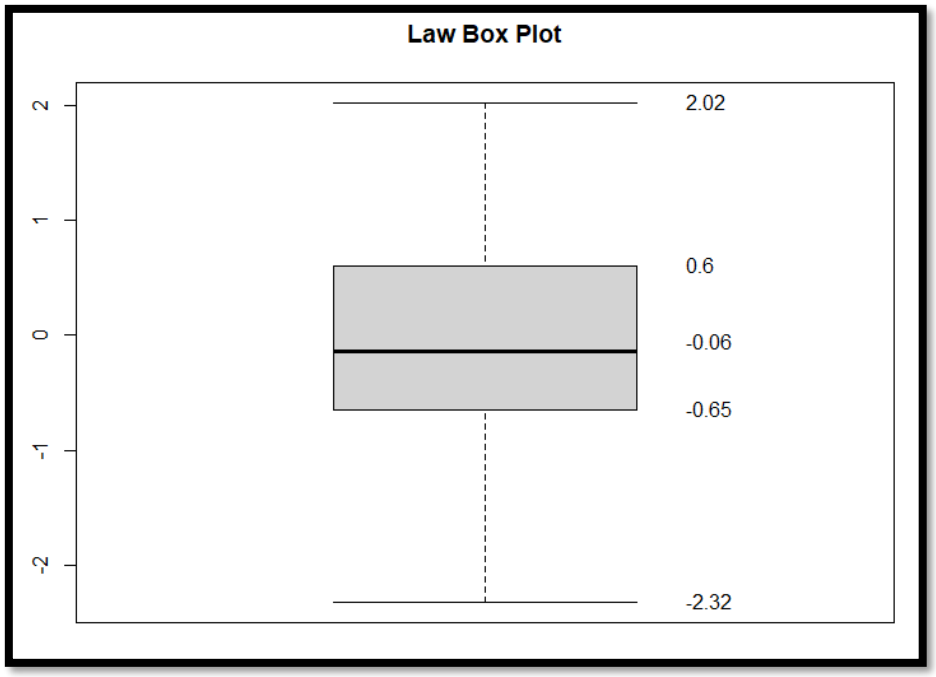


Figure 4.1.1.1 Law Box Plot

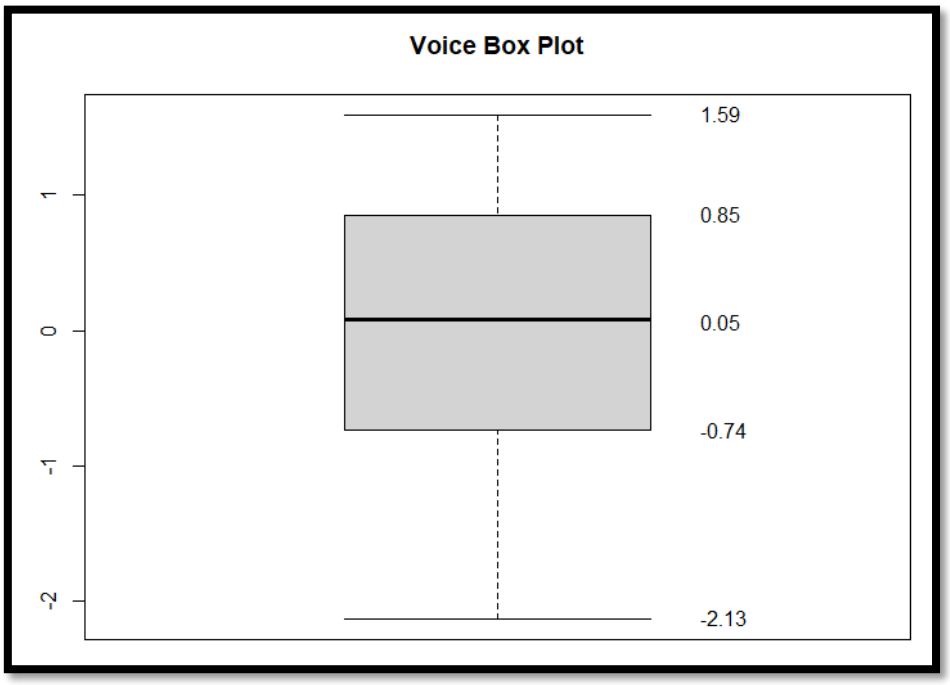


Figure 4.1.1.2 Voice Box Plot

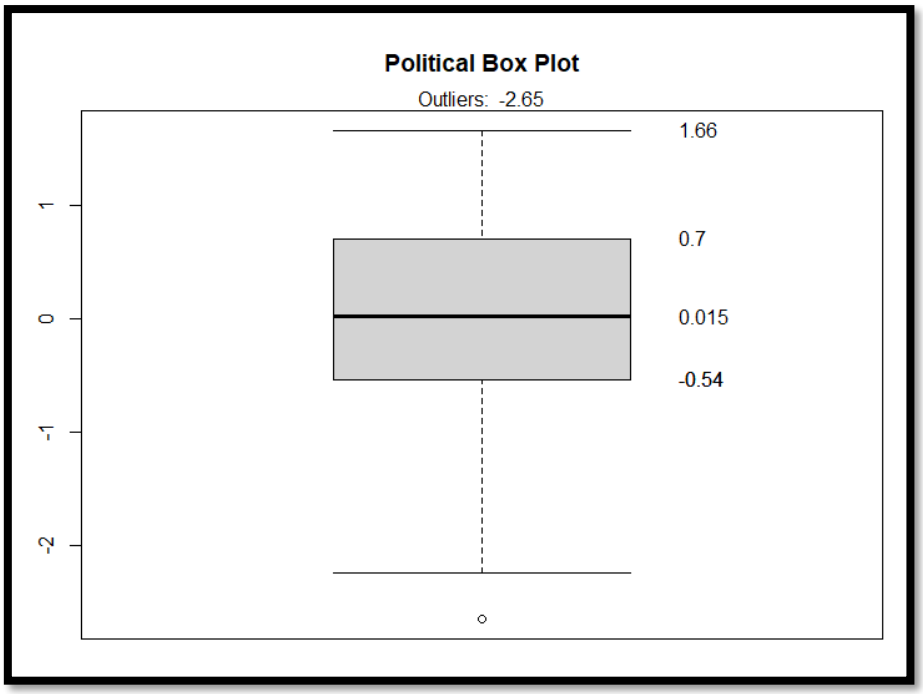


Figure 4.1.1.3 Political Box Plot

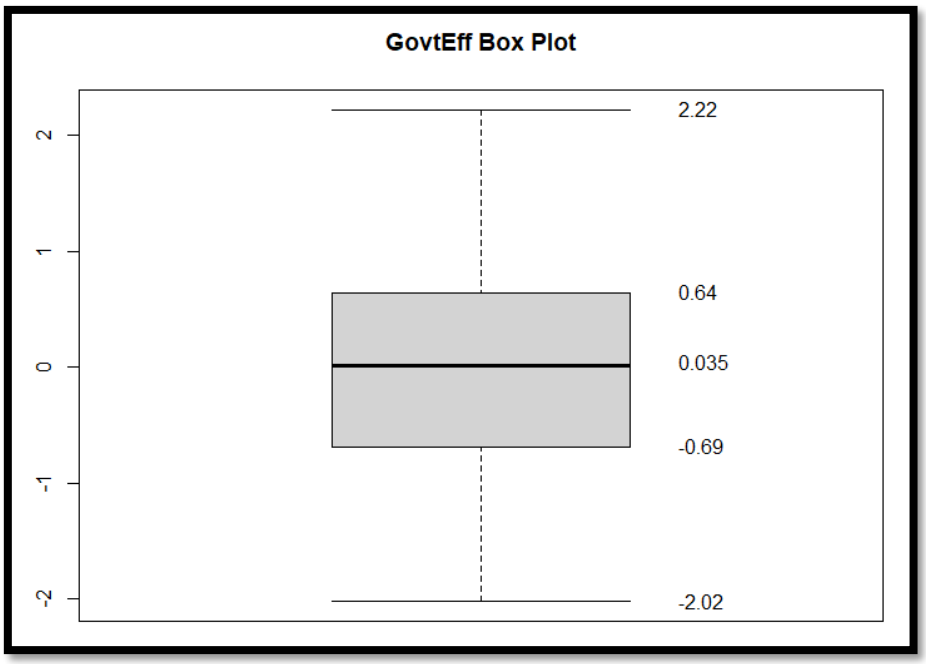


Figure 4.1.1.4 GovtEff Box Plot

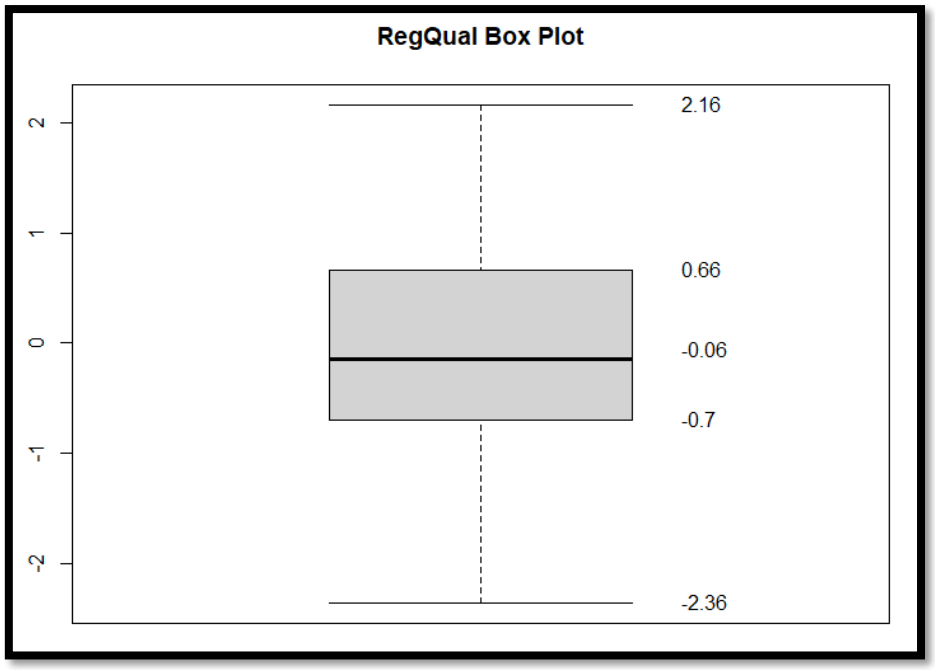


Figure 4.1.1.5 RegQual Box Plot

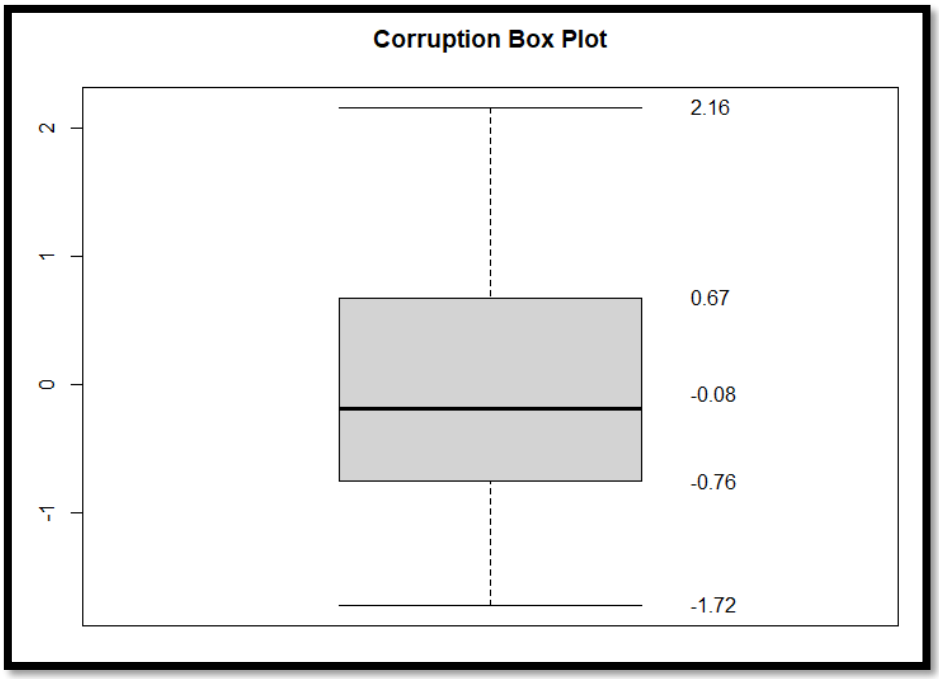


Figure 4.1.1.6 RegQual Box Plot

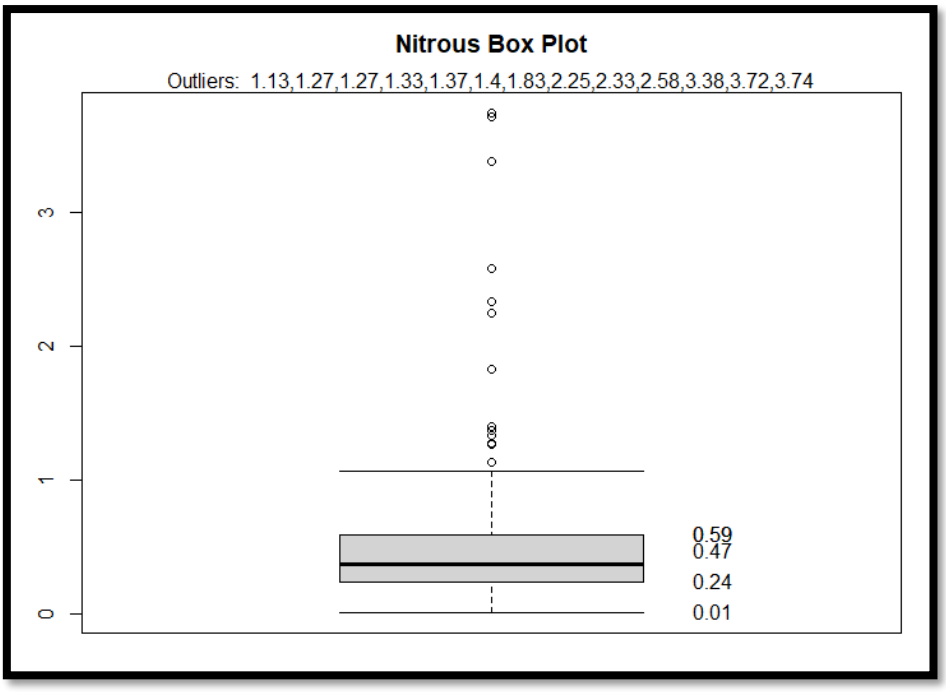


Figure 4.1.1.7 Nitrous Box Plot

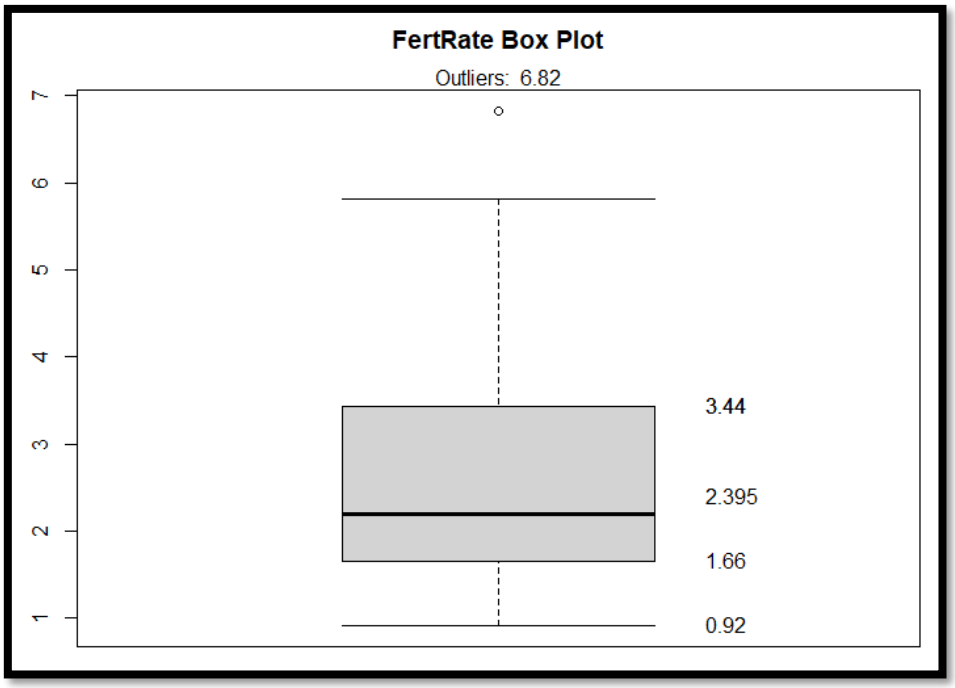


Figure 4.1.1.8 FertRate Box Plot

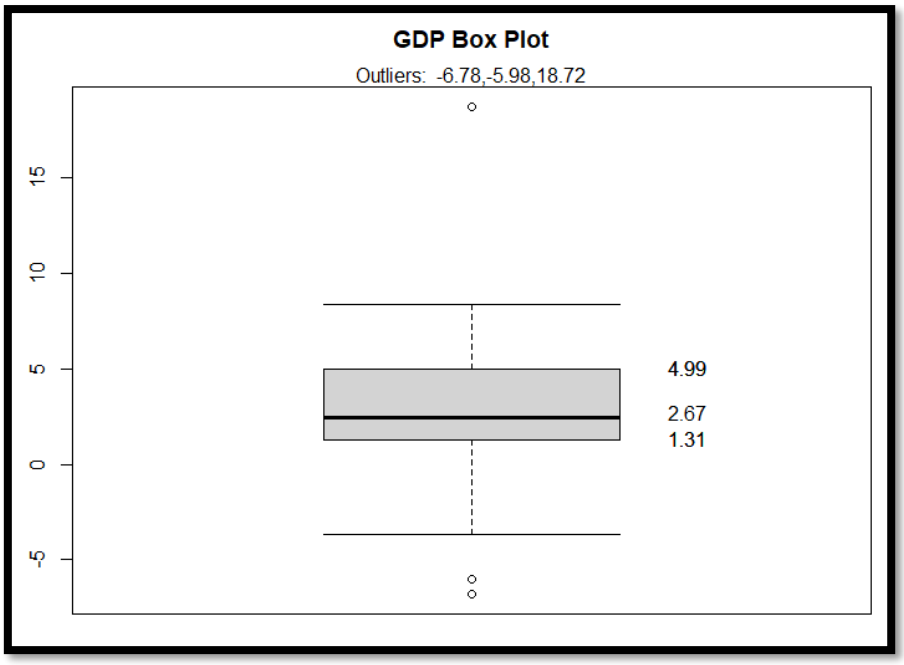


Figure 4.1.1.9 GDP Box Plot

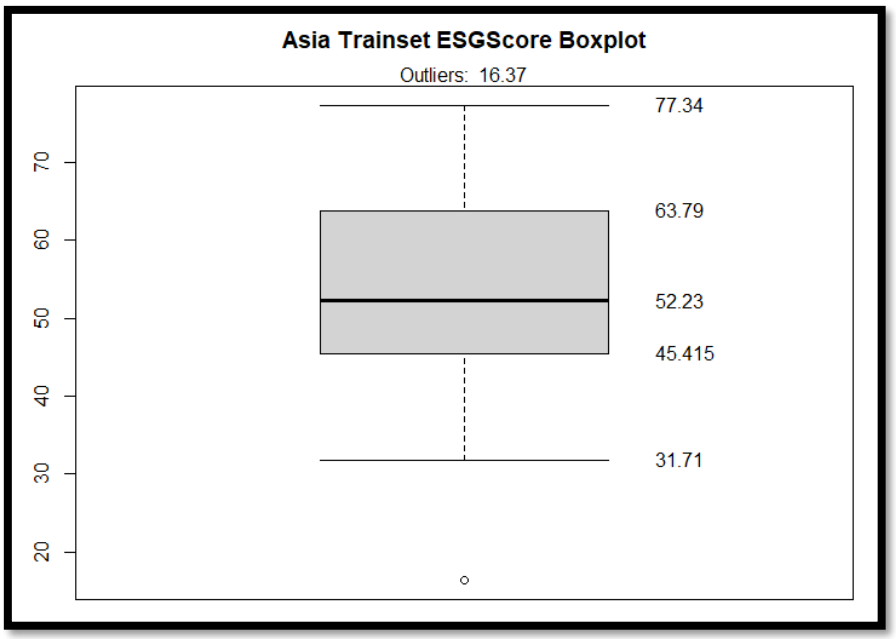


Figure 4.2.1.1 Asia Box Plot

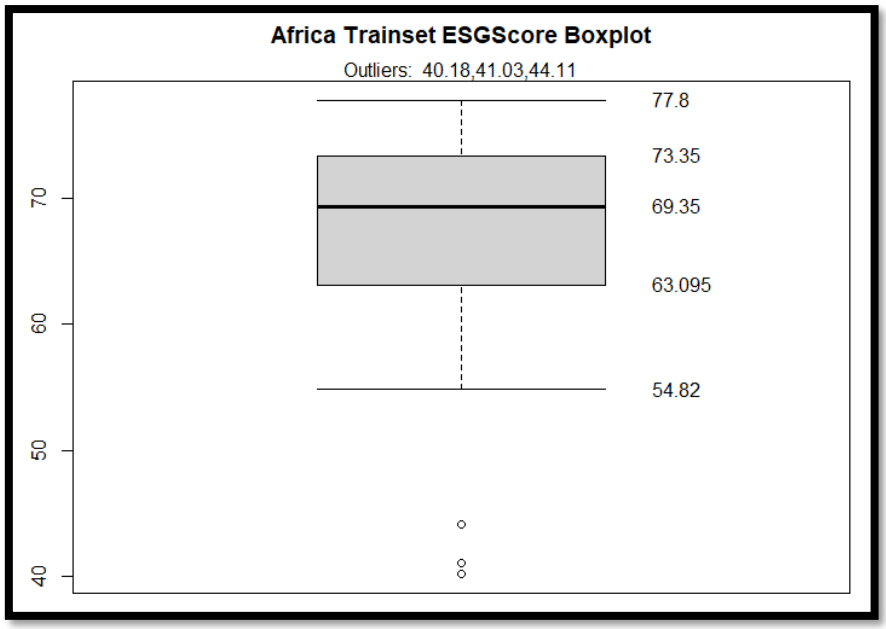


Figure 4.2.1.2 Africa Box Plot

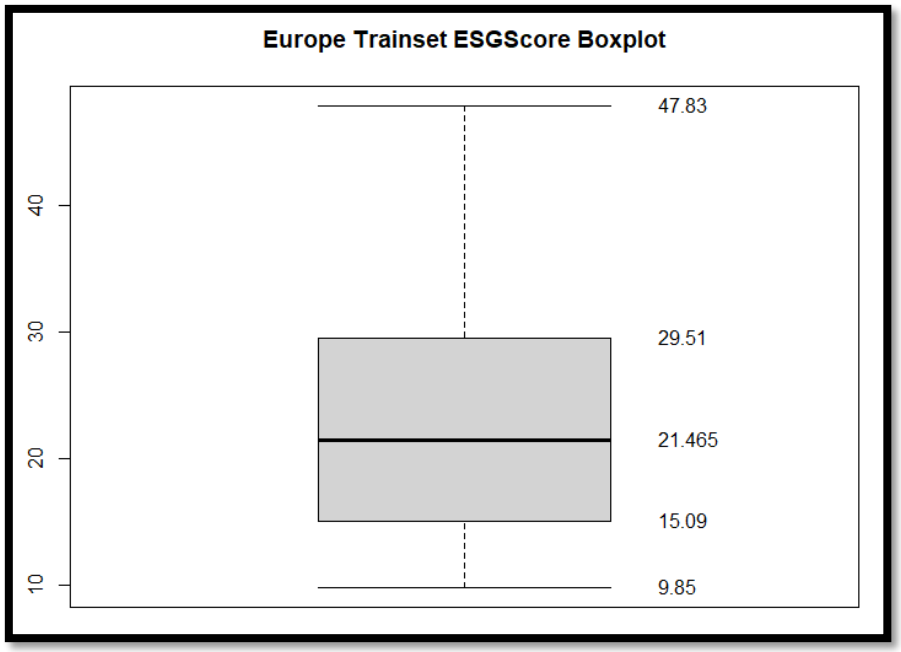


Figure 4.2.1.3 Europe Box Plot

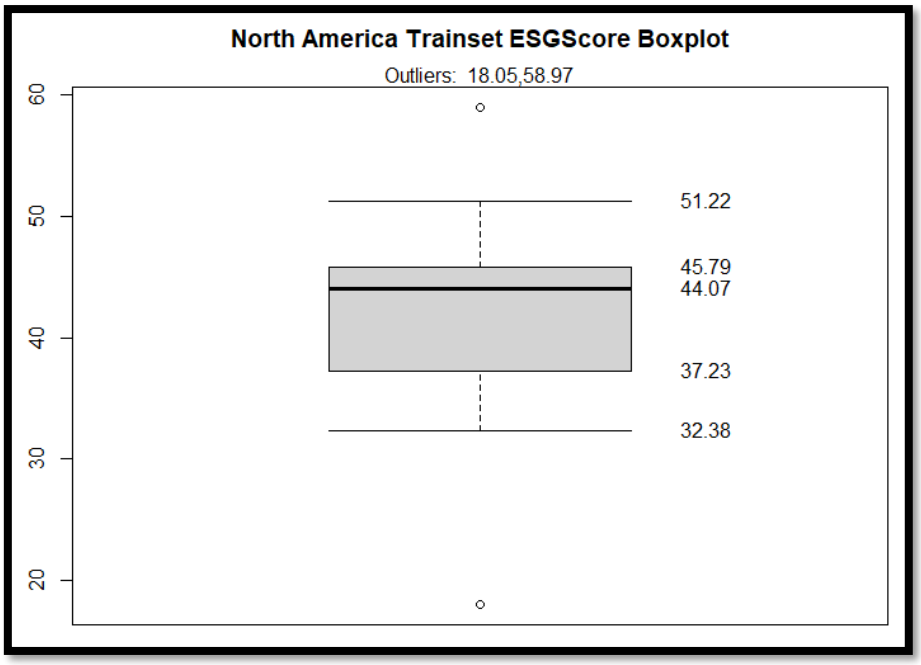


Figure 4.2.1.4 North America Box Plot

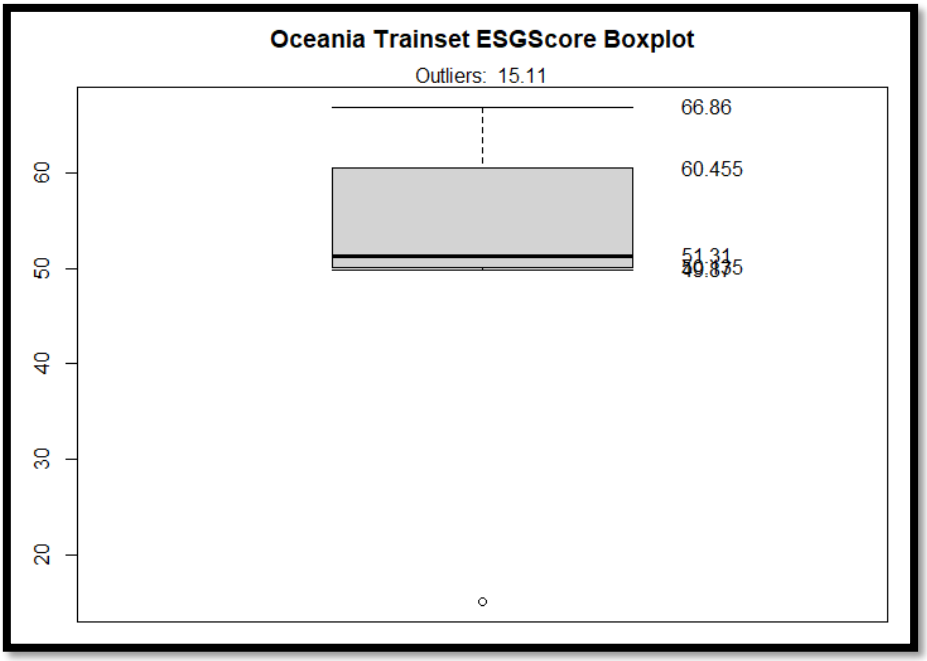


Figure 4.2.1.5 Oceania Box Plot

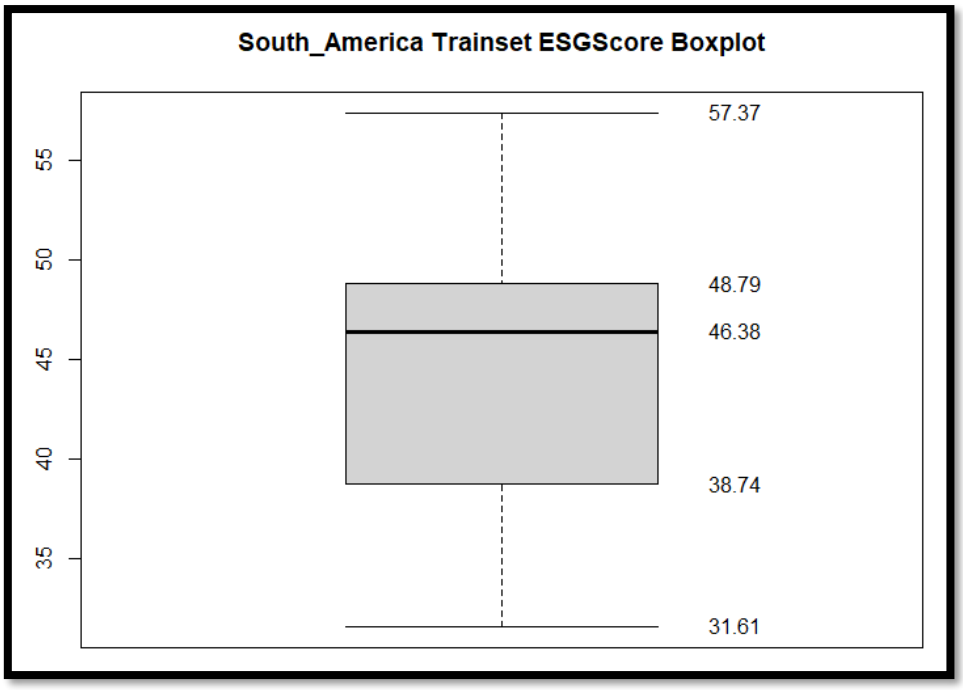


Figure 4.2.1.6 Oceania Box Plot

Appendix D: Machine learning (Linear equations)

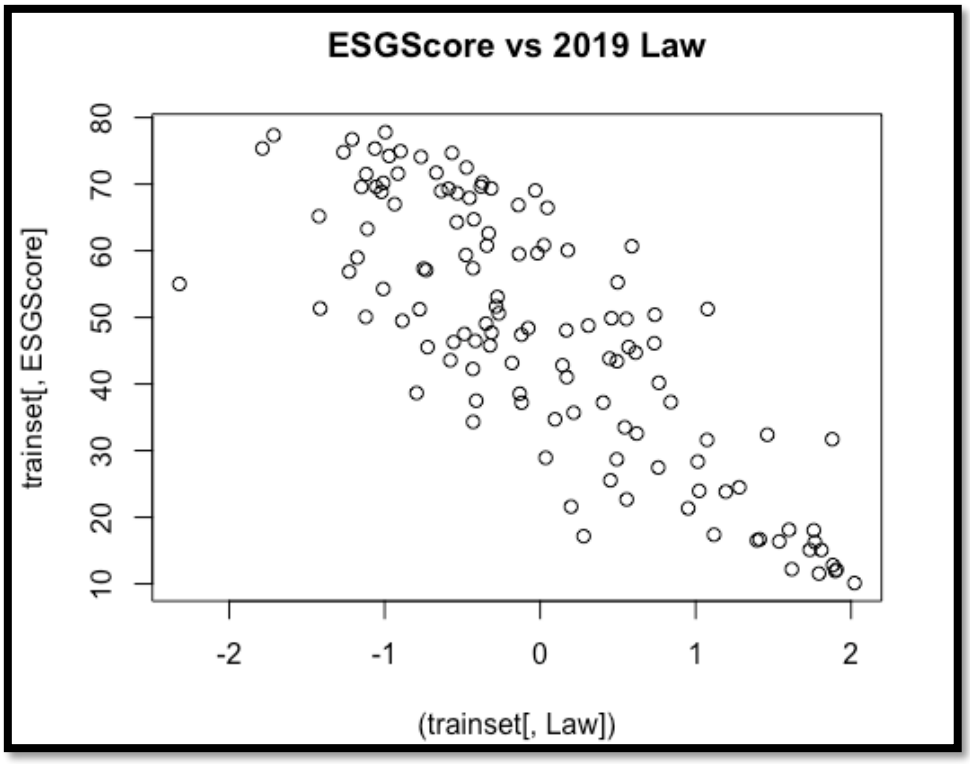


Figure 5.1.1.1: ESGScore vs 2019 Law

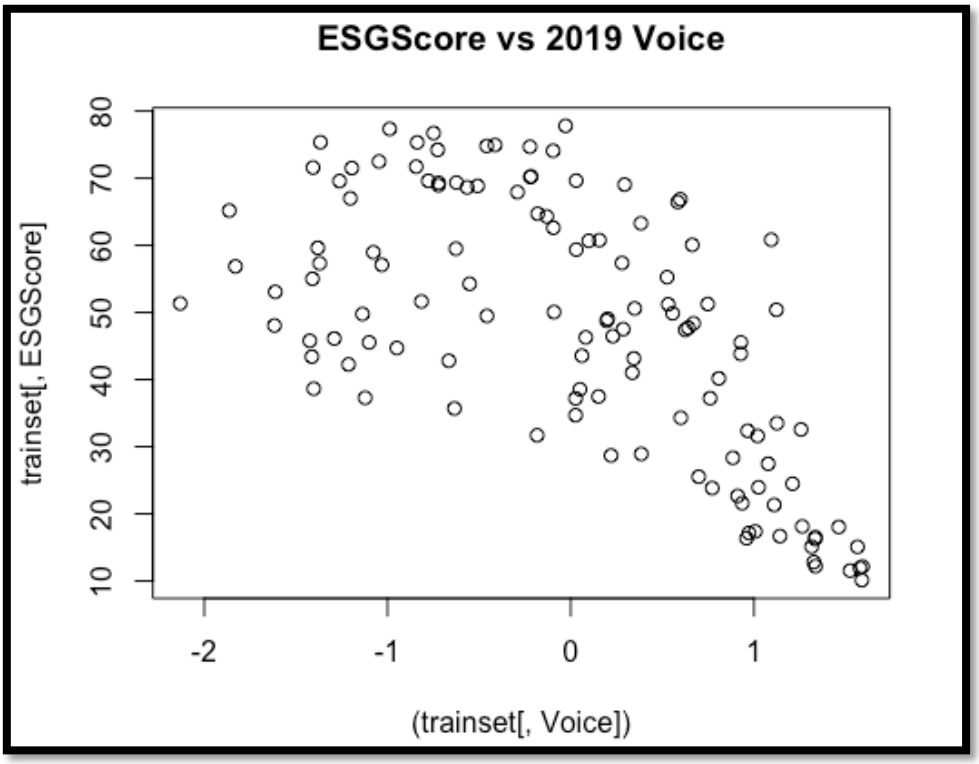


Figure 5.1.1.2: ESGScore vs 2019 Voice

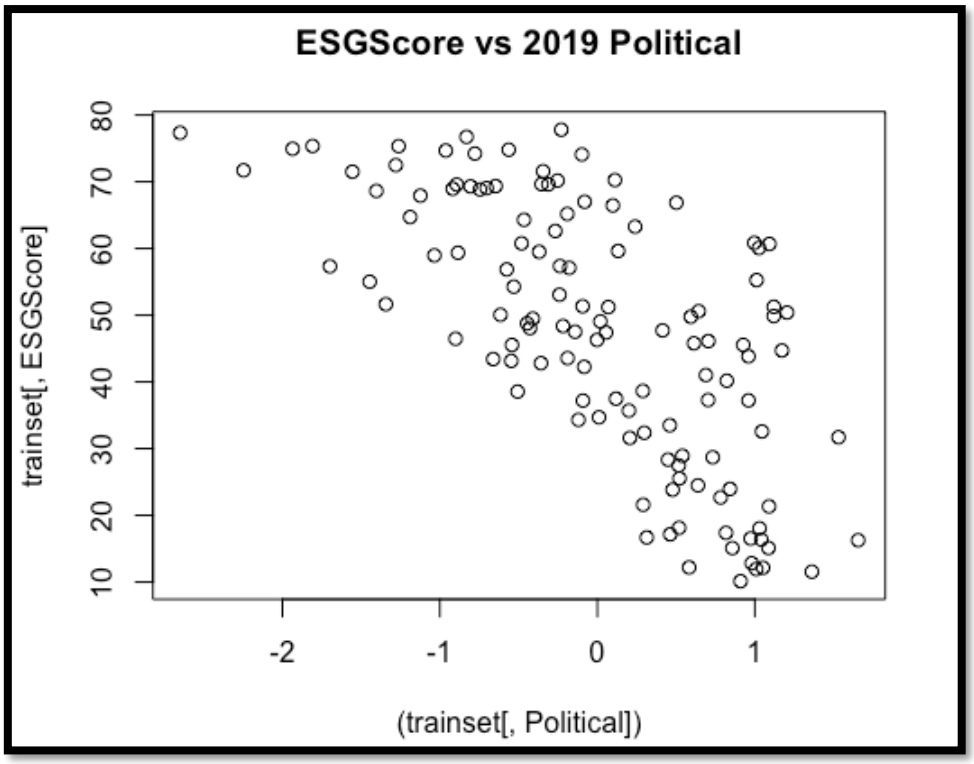


Figure 5.1.1.3: ESGScore vs 2019 Political

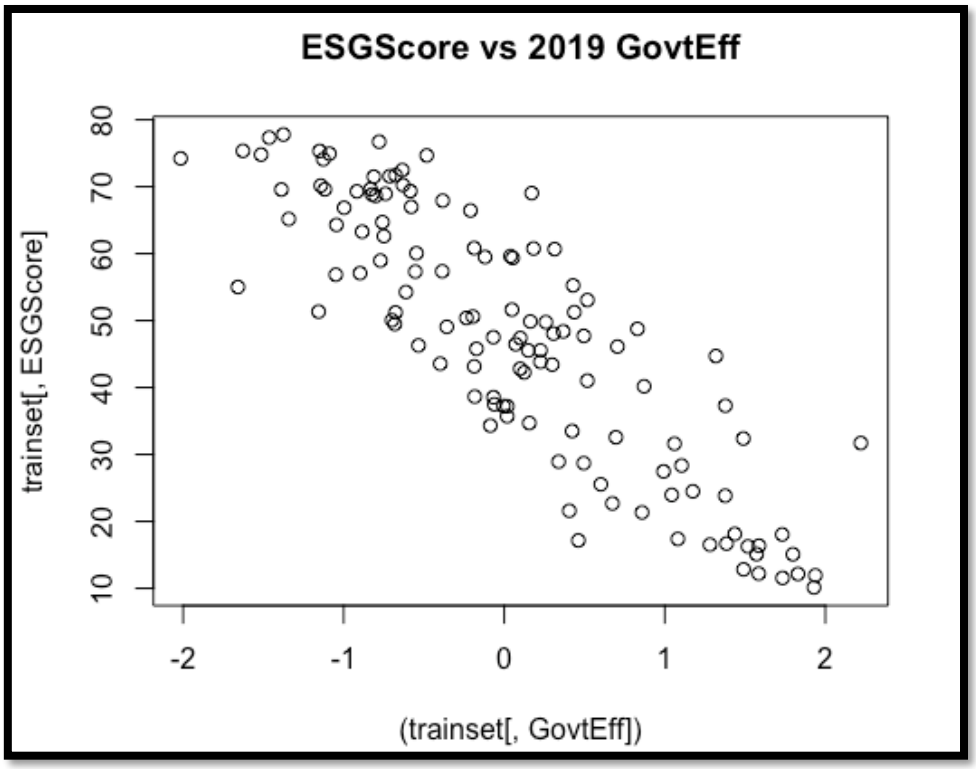


Figure 5.1.1.4: ESGScore vs 2019 GovtEff

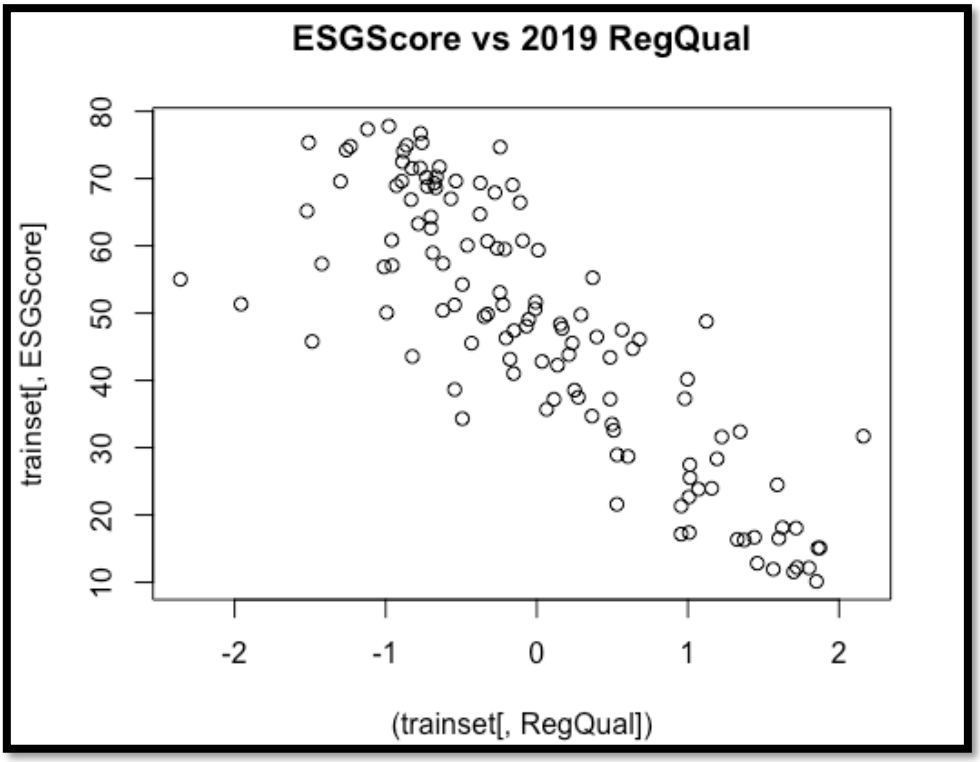


Figure 5.1.1.5: ESGScore vs 2019 RegQual

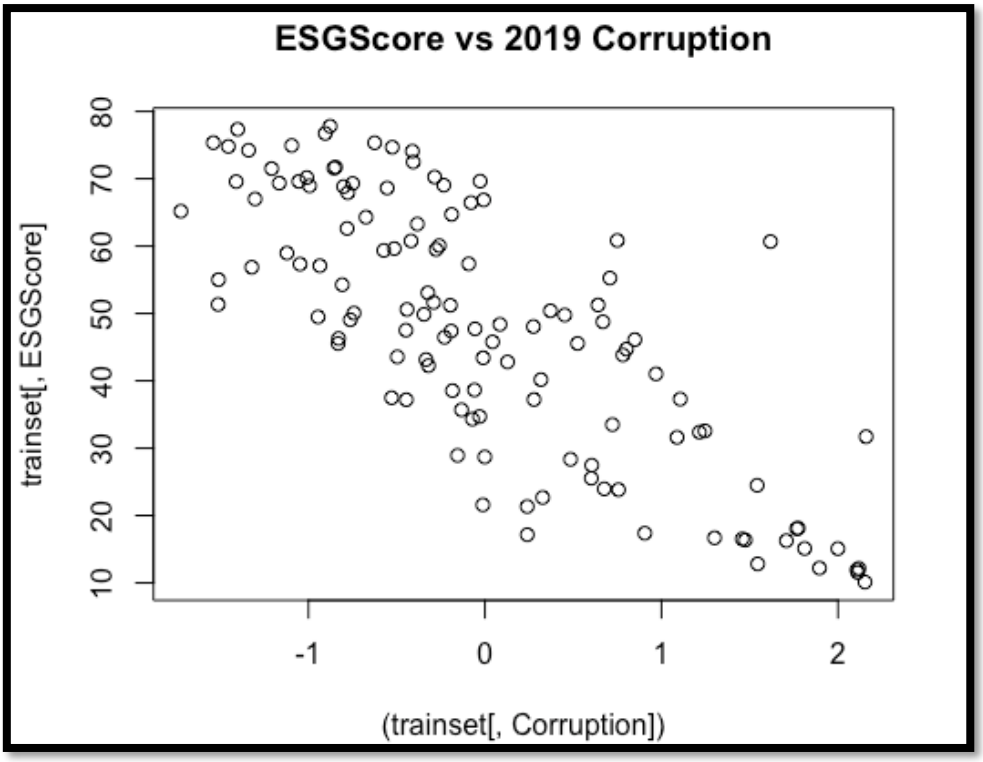


Figure 5.1.1.6: ESGScore vs 2019 Corruption

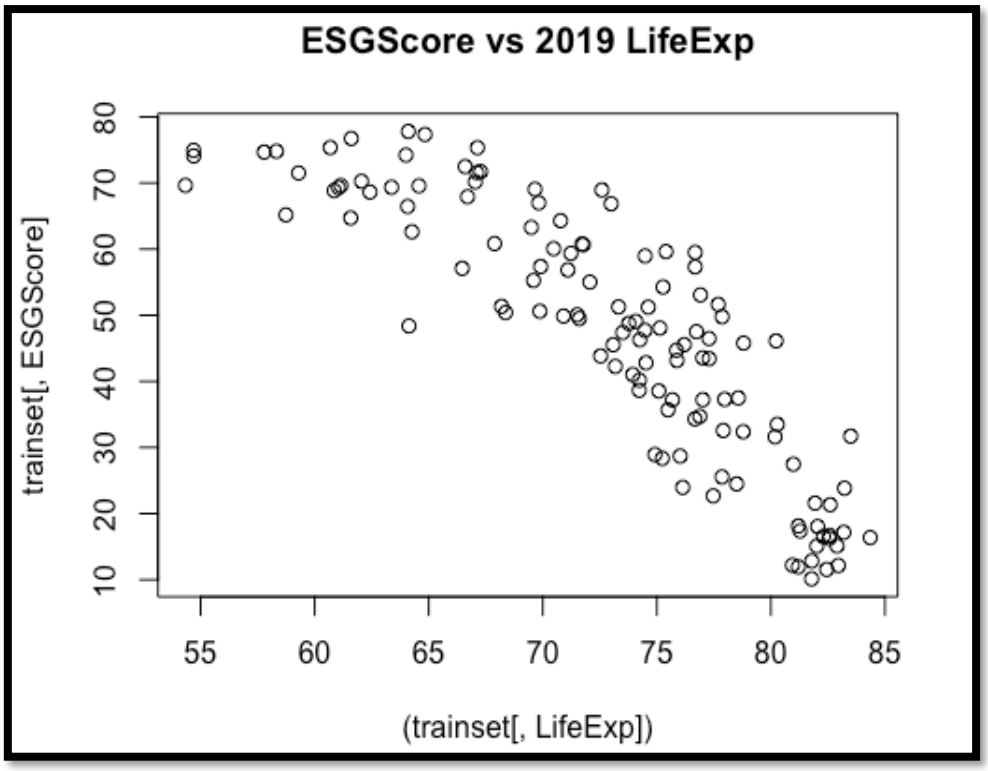


Figure 5.1.1.7: ESGScore vs 2019 LIfeExp

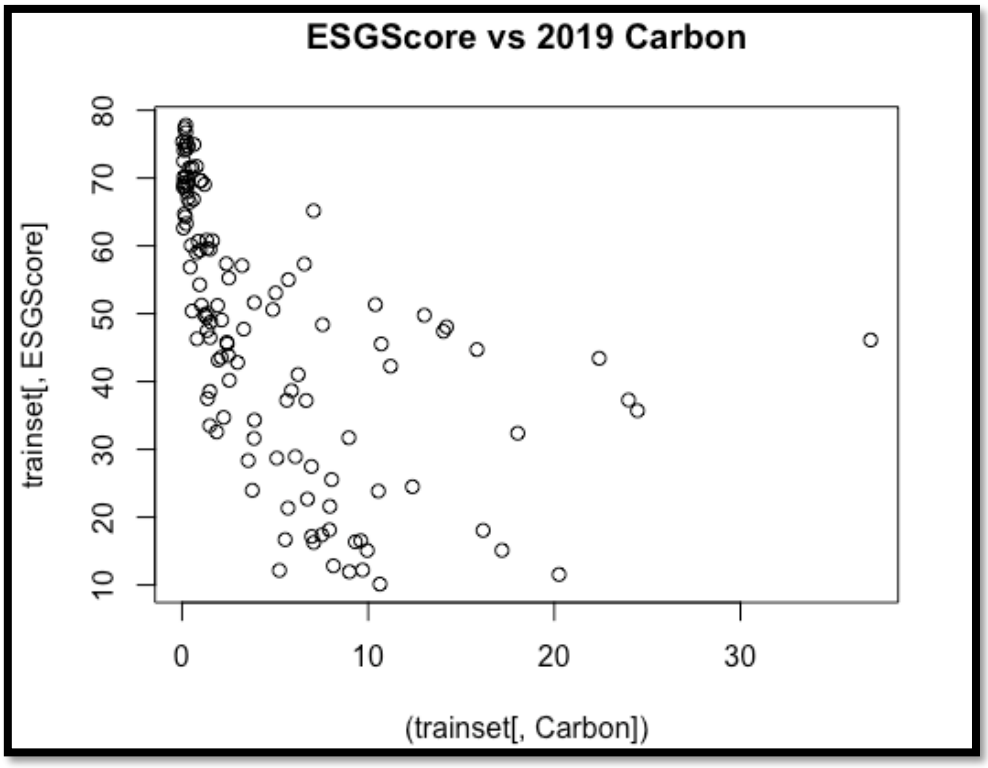


Figure 5.1.1.8: ESGScore vs 2019 Carbon

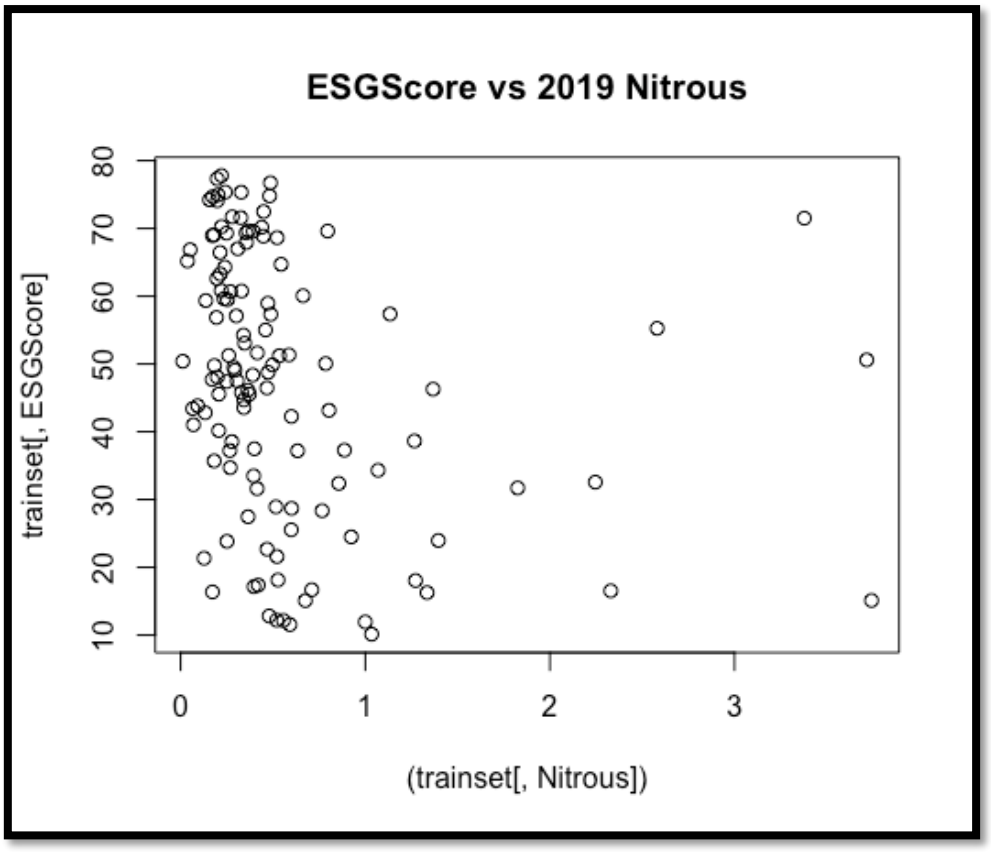


Figure 5.1.1.9: ESGScore vs 2019 Nitrous

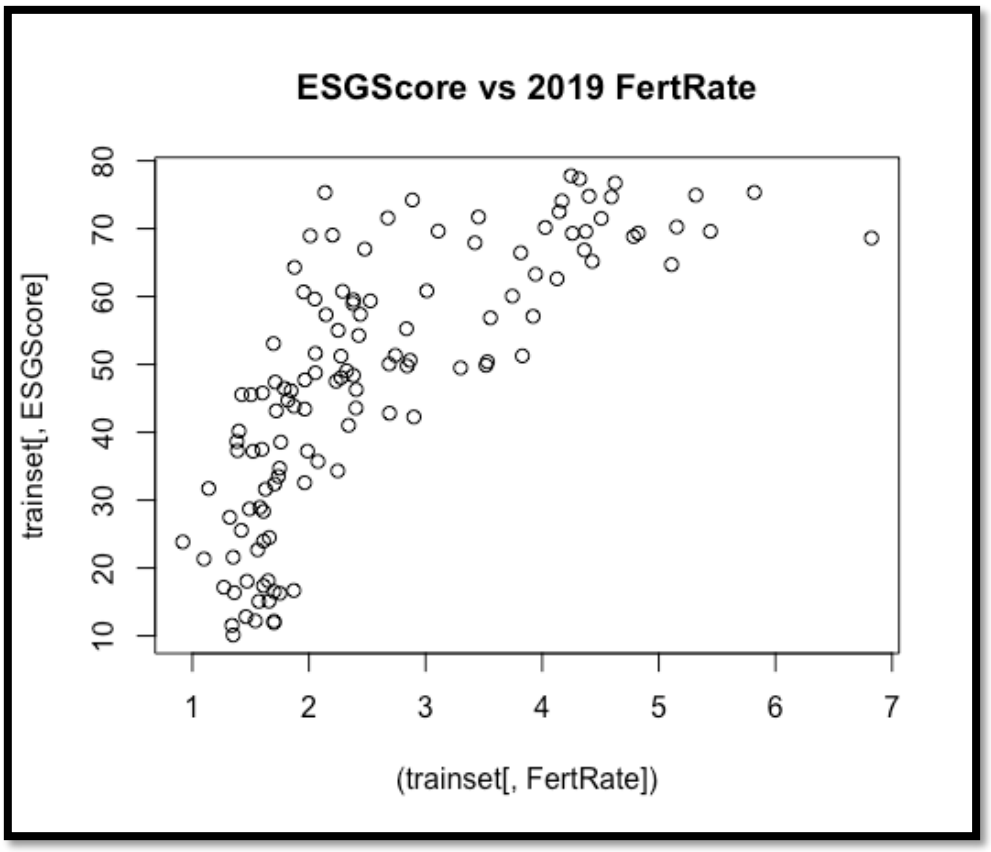


Figure 5.1.1.10: ESGScore vs 2019 Nitrous

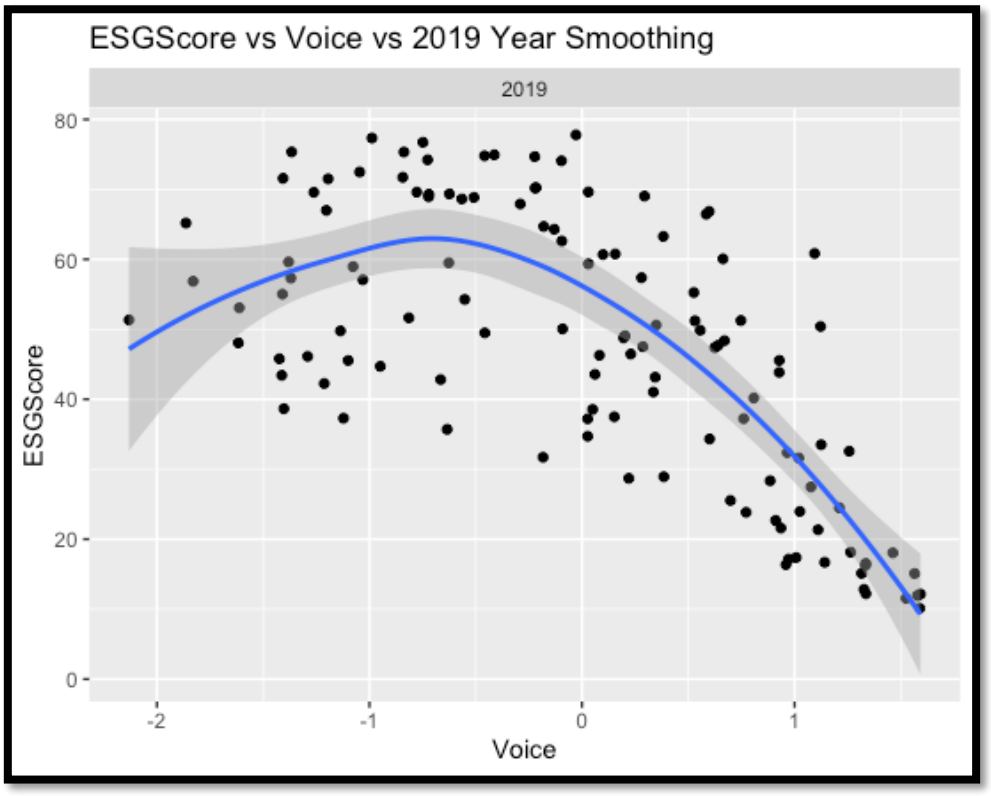


Figure 5.1.1.11: ESGScore vs Voice 2019 weighted smoothing

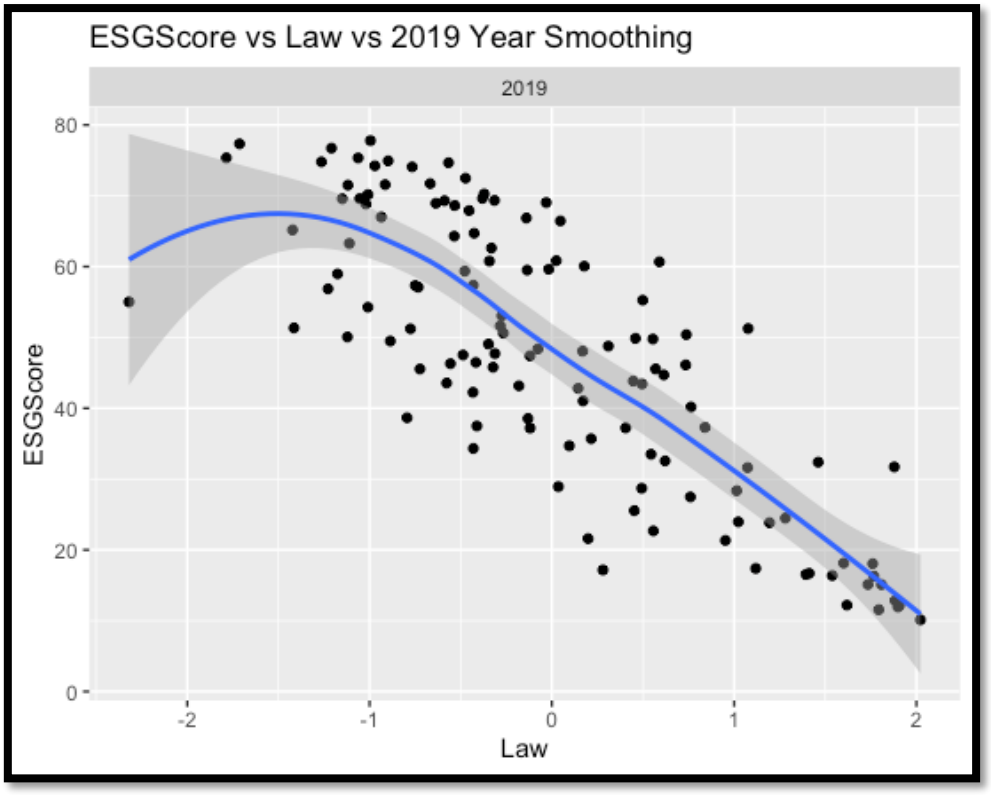


Figure 5.1.1.12: ESGScore vs Law 2019 weighted smoothing

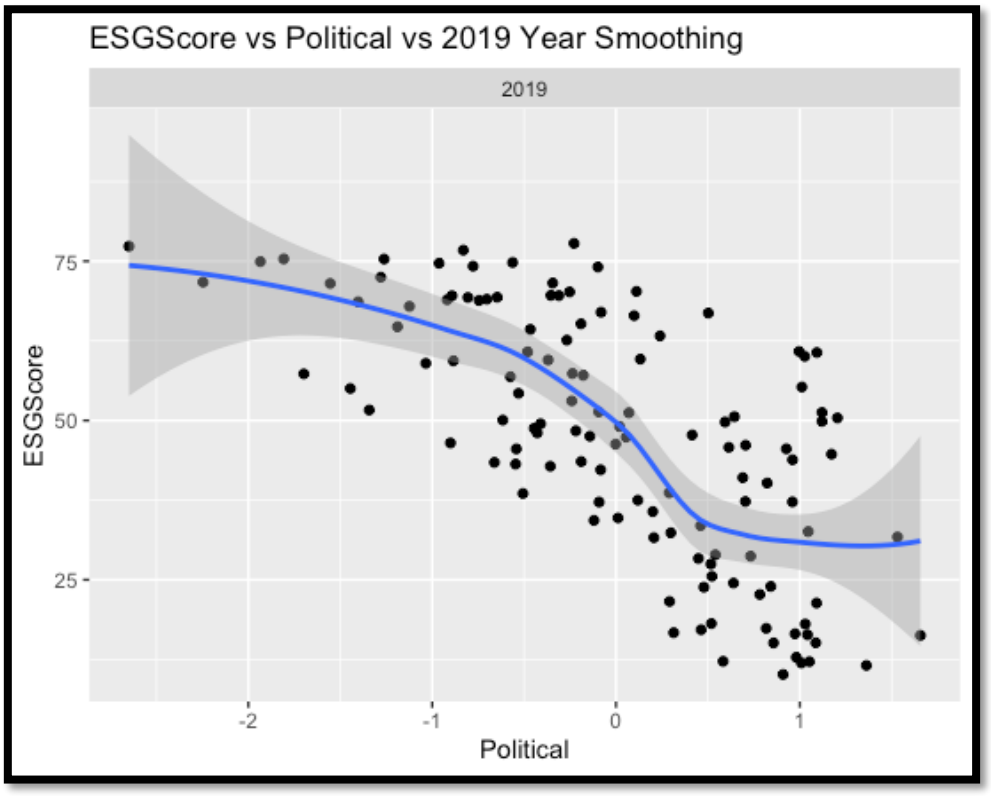


Figure 5.1.1.13: ESGScore vs Political 2019 weighted smoothing

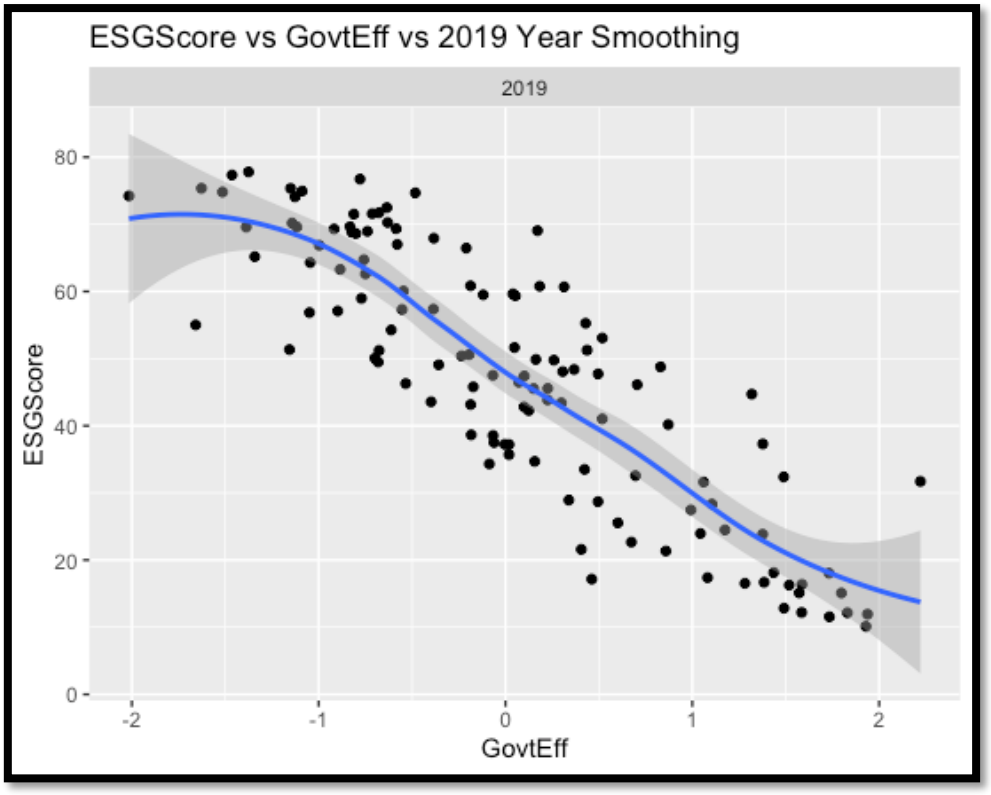


Figure 5.1.1.14: ESGScore vs Political 2019 weighted smoothing

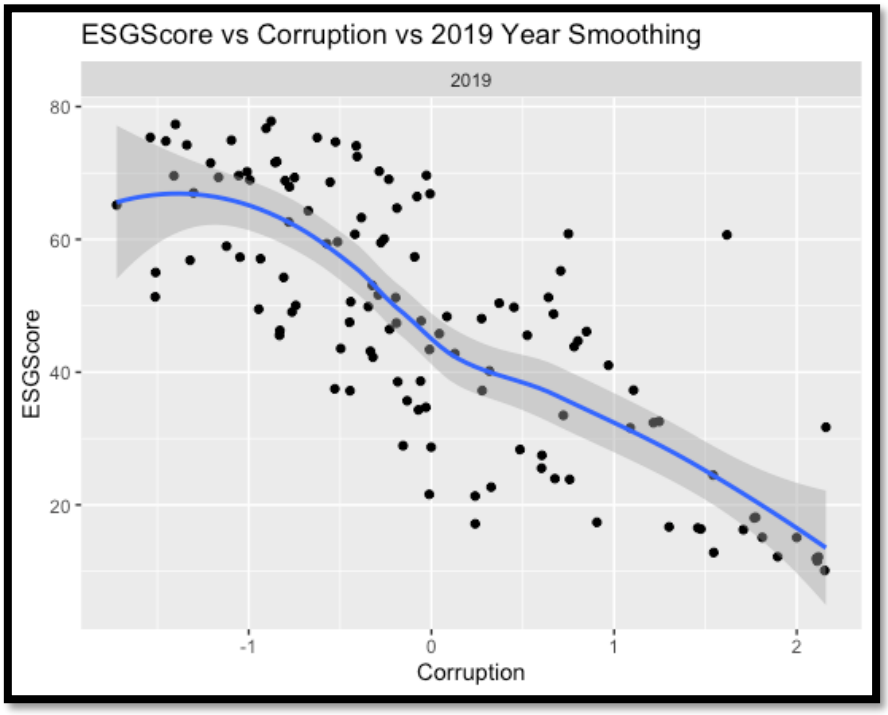


Figure 5.1.1.15: ESGScore vs Corruption 2019 weighted smoothing

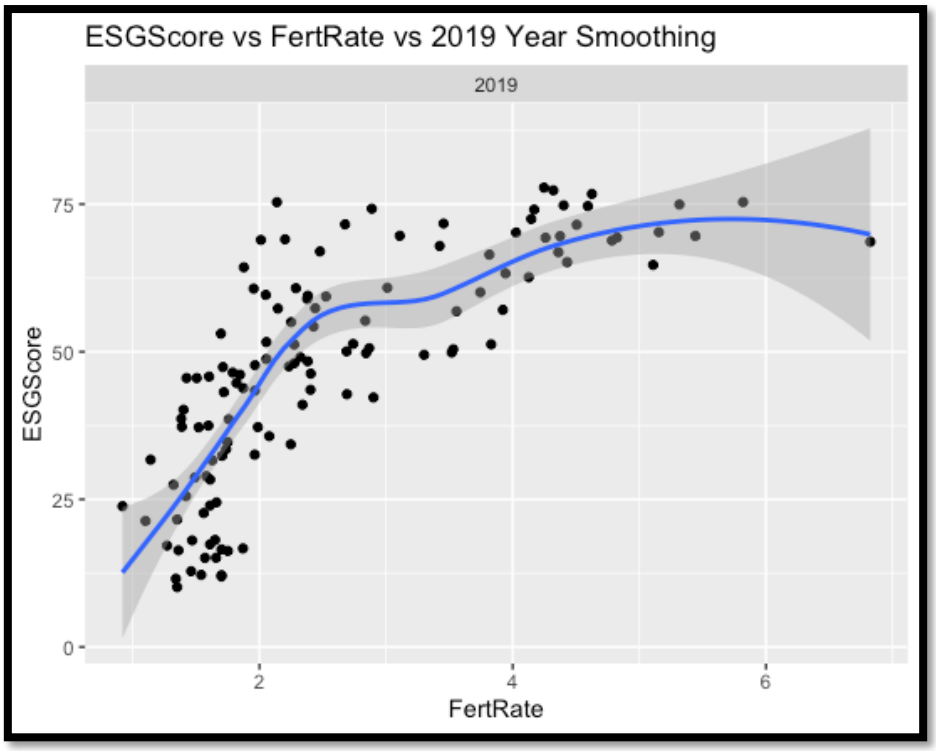


Figure 5.1.1.16: ESGScore vs Fertrate 2019 weighted smoothing

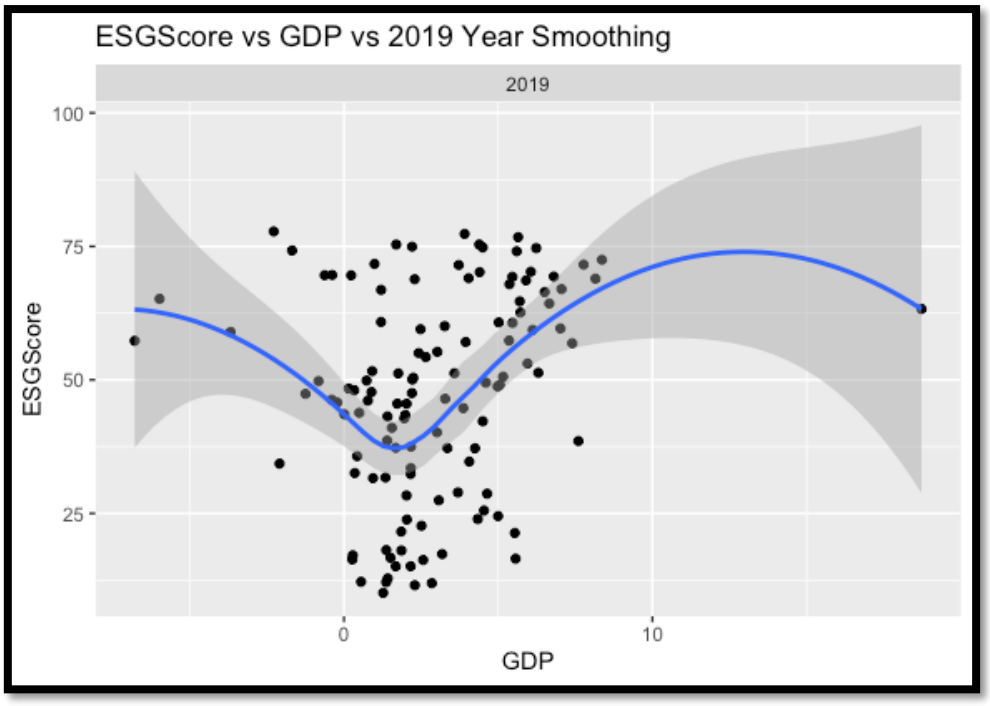


Figure 5.1.1.17: ESGScore vs GDP 2019 weighted smoothing

```
Call:
lm(formula = ESGScore ~ Law + Voice + Political + GovtEff + RegQual +
    Corruption + LifeExp + Carbon + Nitrous + FertRate, data = trainset)

Residuals:
    Min      1Q  Median      3Q     Max
-25.877  -5.851  -0.191   5.242  22.948

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 71.95897    2.38093  30.223  < 2e-16 ***
Law          0.50107    0.59701   0.839    0.401
Voice       -3.70967    0.26648 -13.921  < 2e-16 ***
Political   -1.22374    0.23096  -5.299 1.25e-07 ***
GovtEff     -3.01566    0.58479  -5.157 2.68e-07 ***
RegQual     -3.20658    0.43200  -7.423 1.49e-13 ***
Corruption  -0.01605    0.44986  -0.036    0.972
LifeExp     -0.45347    0.02890 -15.689  < 2e-16 ***
Carbon      -0.35267    0.03193 -11.047  < 2e-16 ***
Nitrous     -1.33539    0.17802  -7.501 8.29e-14 ***
FertRate     3.07652    0.16251  18.932  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.197 on 2981 degrees of freedom
Multiple R-squared:  0.858,     Adjusted R-squared:  0.8576
F-statistic:  1802 on 10 and 2981 DF,  p-value: < 2.2e-16
```

Figure 5.1.2.1: Linear Regression for all variables

```
Call:
lm(formula = ESGScore ~ Law + Voice + Political + GovtEff + RegQual +
    Corruption + LifeExp + Carbon + Nitrous + FertRate, data = trainset)

Coefficients:
(Intercept)          Law        Voice    Political      GovtEff      RegQual   Corruption      LifeExp       Carbon      Nitrous     FertRate
   71.95897      0.50107     -3.70967     -1.22374     -3.01566     -3.20658     -0.01605     -0.45347     -0.35267     -1.33539      3.07652
```

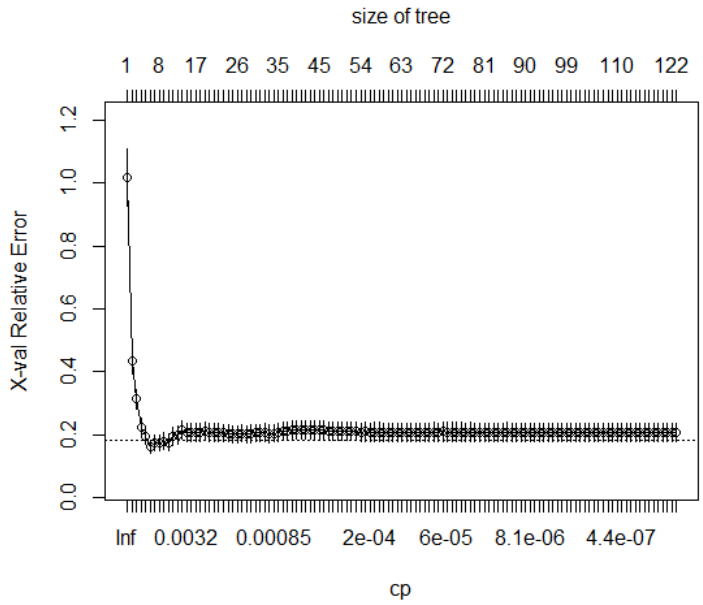
Appendix E: Mean value imputation - Optimal CART Model (Print)

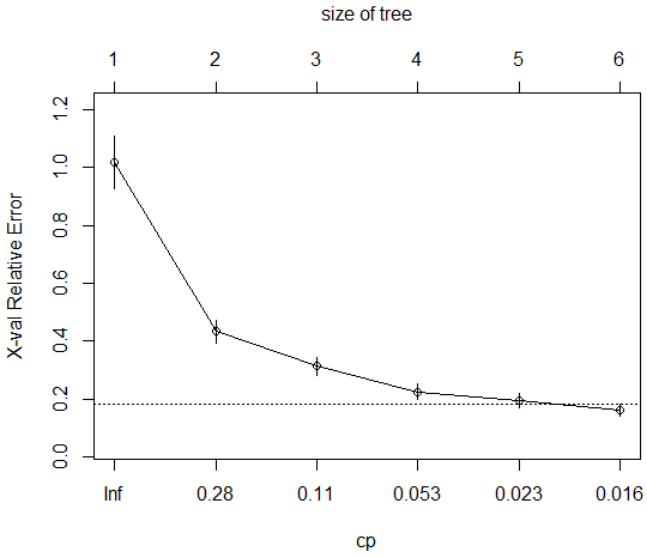
```
n= 123

node), split, n, deviance, yval
      * denotes terminal node

 1) root 123 45495.5200 47.40724
   2) RegQual>=0.0219588 54  8601.4410 30.13296
     4) LifeExp>=80.61023 21   588.1346 17.31095 *
     5) LifeExp< 80.61023 33  2363.7890 38.29242
      10) FertRate< 1.753 16   628.7698 32.03438 *
      11) FertRate>=1.753 17   518.6567 44.18235 *
   3) RegQual< 0.0219588 69  8169.7830 60.92623
     6) LifeExp>=67.578 40  2721.4960 53.79800
      12) Carbon>=1.771566 18   669.9060 48.00889 *
      13) Carbon< 1.771566 22   954.7747 58.53455 *
     7) LifeExp< 67.578 29   612.4152 70.75828 *
```

Appendix F: Surrogate values – Optimal CART Model (Print)

```
n= 123

node), split, n, deviance, yval
      * denotes terminal node

 1) root 123 45495.5200 47.40724
   2) RegQual>=0.0219588 54  8601.4410 30.13296
     4) LifeExp>=80.61023 21   588.1346 17.31095 *
     5) LifeExp< 80.61023 33  2363.7890 38.29242
      10) FertRate< 1.753 16   628.7698 32.03438 *
      11) FertRate>=1.753 17   518.6567 44.18235 *
   3) RegQual< 0.0219588 69  8169.7830 60.92623
     6) LifeExp>=67.578 40  2721.4960 53.79800
      12) LifeExp>=73.03995 20   936.1238 48.98750 *
      13) LifeExp< 73.03995 20   859.7361 58.60850 *
     7) LifeExp< 67.578 29   612.4152 70.75828 *
```

Appendix G: Surrogate values – unpruned & pruned tree X-Val Relative Error (plotcp)

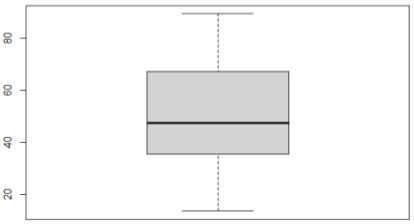




Appendix H: SVM Model outputs

Appendix H.1: SVM predicted ESG Score results without scenario-based analysis

```
> test_pred
       1        2        3        4        5        6        7        8        9       10       11       12       13       14       15       16       17       18       19       20
75.60415 31.44830 46.11272 36.70353 41.87730 59.27555 43.09962 89.52982 76.45518 85.60898 46.63269 22.84486 18.08231 20.22071 20.43325 13.60867 68.00015 37.39797 70.44706 81.28433
      21       22       23       24       25       26       27       28       29       30       31       32       33       34       35       36       37       38       39       40
26.13396 51.57704 55.12003 53.42240 47.51648 20.05524 40.48269 18.79716 66.58530 69.80140 27.47352 39.45087 35.67084 56.47512 75.76407 57.55498 14.01904 35.45714 38.31217 60.43055
      41       42       43       44       45       46       47       48       49       50       51
28.21304 49.41818 69.10725 74.49861 61.77018 68.90624 64.53808 35.94639 64.05714 74.86424 47.57544
```




```
> summary(test_pred)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13.61   35.56   47.58   49.29   67.29   89.53
```

Appendix H.2: SVM results accuracy and analysis without scenario-based analysis

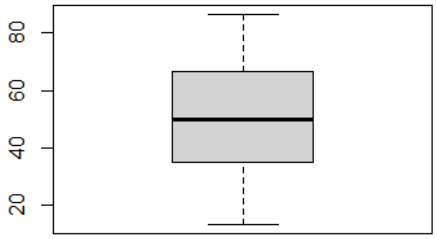
```
> print(model)
Support Vector Machines with Linear Kernel

125 samples
 10 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 113, 112, 113, 113, 112, 112, ...
Resampling results:

  RMSE      Rsquared  MAE
  7.309046  0.866045  6.111154

Tuning parameter 'C' was held constant at a value of 1
>
```

Appendix H.3: Output Summary and MAPE of non-scenario based SVM

```
> summary(table(test_pred, testing$ESGScore))
Number of cases in table: 51
Number of factors: 2
Test for independence of all factors:
        Chisq = 2550, df = 2500, p-value = 0.2383
        Chi-squared approximation may be incorrect
> MAPE(testing$ESGScore, test_pred) #MAPE for non-scenario-based analysis is 0.1429
[1] 0.1429487
>
```

Appendix H.4: SVM predicted ESG Score results with Scenario-based analysis

```
> test_pred2
       1        2        3        4        5        6        7        8        9       10       11       12       13       14       15       16       17       18       19       20
75.61782 34.34022 45.58831 42.42815 43.43649 58.41978 43.19118 86.63893 75.27072 85.09118 43.31703 24.62043 17.02576 21.74612 22.34825 13.57342 67.86598 35.71638 68.96005 83.40936
      21       22       23       24       25       26       27       28       29       30       31       32       33       34       35       36       37       38       39       40
29.33967 52.13080 53.73743 52.75317 47.70544 18.87182 40.98918 21.06122 64.62650 68.53094 28.55188 39.08097 36.26777 54.57777 74.71513 56.75207 13.26167 37.87092 39.65017 61.91883
      41       42       43       44       45       46       47       48       49       50       51
29.86351 51.73399 67.49349 75.23095 60.68445 67.10050 65.96716 32.37955 63.04466 73.84942 49.89203
```




```
> summary(test_pred2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13.26   35.03   49.89   49.38   66.53   86.64
```

Appendix H.5: RMSE of scenario based SVM

```
> RMSE.svm.test2 <- sqrt(mean(testset4.error^2))
> RMSE.svm.test2 #8.231792
[1] 8.231792
```

Appendix H.6: MAPE of scenario based SVM

```
> MAPE(testing$ESGScore, test_pred) #MAPE for non-scenario-based analysis is 0.1429
[1] 0.1429487
```

Appendix I: Variable importance of CART Models

Appendix I.1: Variable importance from CART with mean values replacing NA

```
>
>
> CART2$variable.importance
   RegQual    GovtEff        Law    LifeExp      Voice Corruption   FertRate     Carbon    Nitrous  Political
 748037.79  629234.52  589652.04  553647.64  521481.20  495305.01  171764.54  167849.78   42390.98   35777.67
>
>
```

Appendix I.2: Variable importance from CART with surrogates including Carbon and Nitrous Null Values

```
>
>
> esgcart3$variable.importance
   RegQual    LifeExp    GovtEff        Law   FertRate Corruption      Voice  Political
 34867.590  30028.622  28467.454  25544.259  25366.500  23447.804   5807.958   2123.678
>
```

Appendix I.3: Variable importance from CART with surrogates excluding Carbon and Nitrous Null Columns

```
>
>
> esgcart5$variable.importance
   RegQual    LifeExp    GovtEff        Law   FertRate Corruption      Voice  Political
 34867.590  30028.622  28467.454  25544.259  25366.500  23447.804   5807.958   2123.678
>
>
>
```

Appendix J: Variable importance of SVM Models

Appendix J.1: Variable importance from SVM without scenario-based analysis

```
>
>
>
> varImp(svm_Linear, scale = FALSE) #Shows variable importance
loess r-squared variable importance

           Overall
RegQual     0.7977
LifeExp     0.7973
GovtEff     0.7584
Carbon      0.7193
Law         0.7018
FertRate    0.6667
Corruption  0.6090
Voice       0.5973
Political   0.4805
Nitrous     0.1254
>
>
>
```

Appendix J.2: Variable importance from SVM with scenario-based analysis

```
>
>
>
> varImp(svm_Linear_Grid, scale = FALSE) #Shows variable importance
loess r-squared variable importance

           Overall
RegQual     0.7977
LifeExp     0.7973
GovtEff     0.7584
Law         0.7018
FertRate    0.6667
Corruption  0.6090
Voice       0.5973
Political   0.4805
>
>
>
```

Appendix K

CART Model (optimal print) for CART with Mean values

```
> print(CART2)
n= 123

node), split, n, deviance, yval
      * denotes terminal node

 1) root 123 45495.5200 47.40724
   2) RegQual>=0.0219588 54  8601.4410 30.13296
     4) LifeExp>=80.61023 21   588.1346 17.31095 *
     5) LifeExp< 80.61023 33  2363.7890 38.29242
      10) FertRate< 1.753 16   628.7698 32.03438 *
      11) FertRate>=1.753 17   518.6567 44.18235 *
   3) RegQual< 0.0219588 69  8169.7830 60.92623
     6) LifeExp>=67.578 40  2721.4960 53.79800
      12) Carbon>=1.771566 18   669.9060 48.00889 *
      13) Carbon< 1.771566 22   954.7747 58.53455 *
     7) LifeExp< 67.578 29   612.4152 70.75828 *
```

Appendix L

Variables used in plotting CART with mean values regression tree and node split (printcp)

```
> printcp(CART2, digits = 3)

Regression tree:
rpart(formula = ESGScore ~ Law + Voice + Political + GovtEff +
    RegQual + Corruption + LifeExp + Carbon + Nitrous + FertRate,
    data = trainset, method = "anova", control = rpart.control(minsplit = 2,
        cp = 0))

Variables actually used in tree construction:
[1] Carbon   FertRate LifeExp  RegQual

Root node error: 45496/123 = 370

n= 123

      CP nsplit rel error xerror   xstd
1 0.6314      0    1.0000  1.007 0.0905
2 0.1242      1    0.3686  0.443 0.0383
3 0.1063      2    0.2445  0.321 0.0333
4 0.0267      3    0.1382  0.228 0.0272
5 0.0241      4    0.1114  0.201 0.0239
6 0.0130      5    0.0873  0.165 0.0229
```

Appendix M

CART Model (optimal print) for CART with Surrogate values (with carbon and nitrous)

```
> print(esgcart3)
n= 123

node), split, n, deviance, yval
      * denotes terminal node

 1) root 123 45495.5200 47.40724
   2) RegQual>=0.0219588 54  8601.4410 30.13296
     4) LifeExp>=80.61023 21   588.1346 17.31095 *
     5) LifeExp< 80.61023 33  2363.7890 38.29242
      10) FertRate< 1.753 16   628.7698 32.03438 *
      11) FertRate>=1.753 17   518.6567 44.18235 *
   3) RegQual< 0.0219588 69  8169.7830 60.92623
     6) LifeExp>=67.578 40  2721.4960 53.79800
      12) LifeExp>=73.03995 20   936.1238 48.98750 *
      13) LifeExp< 73.03995 20   859.7361 58.60850 *
     7) LifeExp< 67.578 29   612.4152 70.75828 *
```

Appendix N

Variables used in plotting CART with Surrogate values regression tree and node split (printcp) (with carbon and nitrous)

```
> printcp(esgcart3, digits = 3)

Regression tree:
rpart(formula = ESGScore ~ Law + Voice + Political + GovtEff +
    RegQual + Corruption + LifeExp + Carbon + Nitrous + FertRate,
    data = trainset2, method = "anova", control = rpart.control(minsplit = 2,
        cp = 0))

Variables actually used in tree construction:
[1] FertRate LifeExp  RegQual

Root node error: 45496/123 = 370

n= 123

      CP nsplit rel error xerror   xstd
1 0.6314      0    1.0000  1.019 0.0918
2 0.1242      1    0.3686  0.434 0.0392
3 0.1063      2    0.2445  0.313 0.0326
4 0.0267      3    0.1382  0.225 0.0275
5 0.0203      4    0.1114  0.195 0.0236
6 0.0120      5    0.0911  0.162 0.0219
```