

**Forecasting GDP Growth: How Can Machine Learning Improve Predictions in Economics?**

**BC2406 Analytics 1 - Visual & Predictive Techniques**

Team 6 S09

AVERINA QUAK XUE MIN (U1810311G)

HOU CHEN YING @HOU JENNY

NATALIE CHRISTYFERA YU

SNG YI XUAN (U2021009B)

SUN CHANG

NOV, 2021

**Executive Summary**

There lies a business opportunity for EIU to improve their value proposition to their customers through the improvement of economic forecasting. Two important ways have been identified to improve EIU’s current forecasting–methods used and variables used specifically in the prediction of real GDP growth. EIU currently uses an analytical method of regression for their prediction which could be replaced with Machine Learning methods which aim to increase both efficiency and accuracy. In addition, the EIU uses a small pool of indicators which are mostly economic indicators for the prediction of real GDP growth. However, studies have shown that there exist variables outside of the current pool that are related to a country’s economy and could be used to obtain a better prediction of real GDP growth.

The report hence aims to explore the use of Machine Learning techniques, particularly Linear Regression and CART, with the possible inclusion of other variables to provide a more optimal combination of indicators to predict real GDP growth.

Our final dataset includes a total of 21 independent variables (inclusive of EIU’s variables) and countries in the Asia continent across 10 years. The final dataset additionally went through rounds of data cleaning and feature scaling to obtain the most optimal and reliable set of data that will be used by our machine learning models.

In the training of our LR model, we took a systematic approach to train our model to obtain an optimal LR model with statistically significant variables, taking into account multicollinearity and significance. A similar approach was taken for CART for training of the model to obtain an optimal tree and identified the important indicators. In both models, comparisons are made with the baseline EIU model, where we conclude that the inclusion of other specific variables not present in the original EIU model, will improve real GDP growth predictions.

However, the presence of certain limitations to our approach have hindered us in achieving better results for our machine learning models. We note that a model could have been developed specifically for a country to obtain better results but due to the insufficient data available, we were unable to explore this approach. Nonetheless, we strongly believe that the benefits and insights gained from our findings are highly relevant to EIU, and that new directions and pilot studies could be initiated by EIU, paving the way for exciting opportunities.

# 

# Contents

[**Contents**](#_noyqwqdzp4vs) **3**

[**1. Business Problem**](#_sjes5ecjznfv) **5**

[1.1 EIU’s current method](#_i81ho92hav40) 5

[1.2 Business Opportunity](#_2xy12h2c50u1) 6

[**2. How Machine Learning is Used**](#_vx6pxo3tac8w) **6**

[2.1 How Machine Learning differs from Analytics Models](#_jiebvms1qxaq) 6

[2.2 Success stories of adopting Machine Learning](#_b62yqeqv48wz) 6

[2.3 Value of Machine Learning to EIU](#_c0z0weos3r7m) 8

[2.3.1 Accuracy and Efficiency](#_liryn4thg98q) 8

[2.3.2 More Comprehensive Report](#_rkqhpgkyoxht) 8

[**3. Proposed Approach**](#_15zyi9k83o) **8**

[**4. Data**](#_jb3vnf8kdxo0) **9**

[4.1 New Set of Independent Variables](#_ldt5g35s8y2a) 9

[**5. Data Extraction**](#_iabdtqkiunp2) **10**

[**6. Data Cleaning**](#_wrtn0lsdtaij) **10**

[6.1 Dealing with NAs](#_yt6vt2qmtjzp) 10

[6.1.1 Removing Variables](#_xvlrg161qbib) 10

[6.1.2 Removing Countries](#_dg48kh876qir) 11

[6.2 Feature Scaling](#_iocj82ctzfdz) 11

[**7. Linear Regression**](#_57drsgb6pxjg) **12**

[7.1 Linear Regression Process](#_fge8lflrgg9m) 12

[7.2 Comparison of results with EIU baseline model](#_t81s8qtertiu) 14

[7.2.1 Comparison of Result](#_7inaf1uosnap) 14

[7.2.2 How EIU can Benefit from Adopting Some Variables](#_qfw2n9j4mesi) 14

[**8. Classification and Regression Tree (CART)**](#_dgrrwsgvifr1) **15**

[8.1 Dataset used in CART Model](#_xoa2sqnqwkf0) 15

[8.1.1 Original Cleaned Dataset](#_kdk18eeuf0fd) 15

[8.1.2 Dataset with less than 11 NAs](#_y1lh8b6gfatf) 15

[8.1.3 Dataset with no NAs](#_36u0c82u4g1i) 15

[8.1.4 Dataset with Mean in replacement of NAs](#_b0ggqthojm4w) 16

[8.1.5 Final choice of dataset used in CART model](#_7vtd2bc1k6js) 16

[8.2 CART Process](#_puex51vbkckd) 16

[8.3 Comparison of results with EIU baseline model](#_qb7z3m5hts0i) 19

[**9. Insights gained from Linear Regression and CART**](#_e124bpkp7o7g) **20**

[**10. Limitations & Future directions**](#_ggvave5l08ta) **21**

[10.1 Limitations](#_acni7s9tbi9m) 21

[10.2 Future Directions](#_wmjbmuq9u9qe) 22

[10.3 Pilot Studies that can be initiated at EIU to evaluate Machine Learning](#_yr0djald0hi8) 23

[10.3.1 Trade-offs in fairness in predictive algorithms](#_okm28v8vji9k) 23

[10.3.2 Interpretability of models](#_bjshjkrlhxmr) 23

[10.3.3 Identification and evaluation of inconsistencies in data collection](#_zfu1cjsttmwy) 24

[**References**](#_dyiw4l9788cm) **25**

[**Appendix**](#_2c0nz2df9fmk) **28**

[Appendix A: Data Dictionary](#_fn810wdrxphj) 28

[Appendix B: Data Extraction and Cleaning](#_subi0oirekdq) 33

[4. Data](#_9w8vjvwhecib) 33

[5. Data Extraction](#_6ps4ra6w5up4) 34

[6.1 Data Cleaning](#_1v400knbzvct) 34

[6.2 Feature Scaling](#_h1k5xppeesch) 35

[Appendix C: Linear Regression](#_y3rrrsv2h1bi) 36

[Appendix D: Classification & Regression Tree](#_1pdlirw0fm4d) 37

[8.1 Dataset used in CART Model](#_lro8ewmhzrpw) 37

[8.1.2 Dataset with < 11 NAs](#_gdepzvkwck9o) 37

[8.1.3 Dataset with no NAs](#_ks1qf9aid6tn) 38

[8.1.4 Dataset with Mean in place of NAs](#_ctz4gr424u84) 39

# 

# 1. Business Problem

The Economist Intelligence Unit (EIU) with their flagship country reports aims to provide forecasting and advisory services to assist entrepreneurs, financiers and government officials in their decision making. In the country report, it contains in-depth and ongoing analysis and forecasts of political, policy and economic conditions to help interested stakeholders better understand the current and future positions of a country.

For example, through the use of forecasting methodologies, EIU is able to make future estimates of important economic indicators such as Gross Domestic Product growth, exchange rates, consumer price inflation, unemployment rate, etc, providing decision-makers a future outlook of the country.

GDP and Real GDP growth are a key macroeconomic indicator included in every economic forecast agency’s forecast. Gross Domestic Product (GDP) measures the economic state of a country by combining the monetary value of all final goods and services produced in a country at a given time (Callen, 2020).

Real GDP growth measures the year-on-year changes in GDP, accounting for inflation. In fact, Real GDP growth is the most widely used indicator of a country’s economy (Dynan & Sheiner, 2007), as it allows for easy tracking of growth without the distorting influence of inflation.

## 1.1 EIU’s current method

For the EIU, to estimate real GDP growth per capita, it conducts a growth *regression* that links real GDP growth per capita to a set of growth determinants (independent variables). The set includes (but is not limited to):

1. The availability of an educated workforce
2. GDP per worker
3. Life Expectancy
4. The quality of institutions (including the legal framework and the quality of the bureaucracy)
5. Fiscal policy
6. The degree of government regulation
7. Movements in the population of working age relative to the overall population
8. The development of information and communication technology infrastructure

However, despite GDP’s value as a macroeconomic indicator, there are limitations to this measure. GDP fails to account for numerous factors that are crucial in determining social and environmental issues which are important in measuring economic growth (European Commision, 2007). As such, we believe this has led EIU’s current method of estimating real GDP growth to be incomprehensive.

## 1.2 Business Opportunity

Due to the limitations present in EIU’s current forecasting methodology for real GDP growth, there is an opportunity to improve on it. In particular, the current set of variables used in the prediction of real GDP growth may not be all-encompassing, research has also shown that there are variables outside of the current set that affects GDP (Ilter, 2017), thus there could exist a more optimal combination of variables that can better predict real GDP growth. Hence, there lies an opportunity for EIU to improve on the existing methodology in terms of accuracy and comprehensiveness.

For EIU to continue maintaining its top spot as an economic forecast agency, it is critical for EIU to continuously enhance its methodology to improve its forecast. Our group believes that this can be done via performing predictive analysis using machine learning techniques and the possible addition of several other variables into the current model.

# 

# 2. How Machine Learning is Used

## 2.1 How Machine Learning differs from Analytics Models

Machine learning focuses on exploring the algorithm learnt from data automatically, aiming to construct and improve the model without any human intervention. These processes are built so as to make better and appropriate predictions for the future. However, data analytics refers to the process of obtaining insights from currently available data through some measures such as cleaning, inspecting, transforming, and modelling. From insights and conclusions that have been made, businesses can have the chance to enhance their decision-making system.

As a whole, data analytics focuses on extracting any valuable information from currently available data, whereas machine learning focuses on computing and training algorithms with the data for automatically functioning in the future.

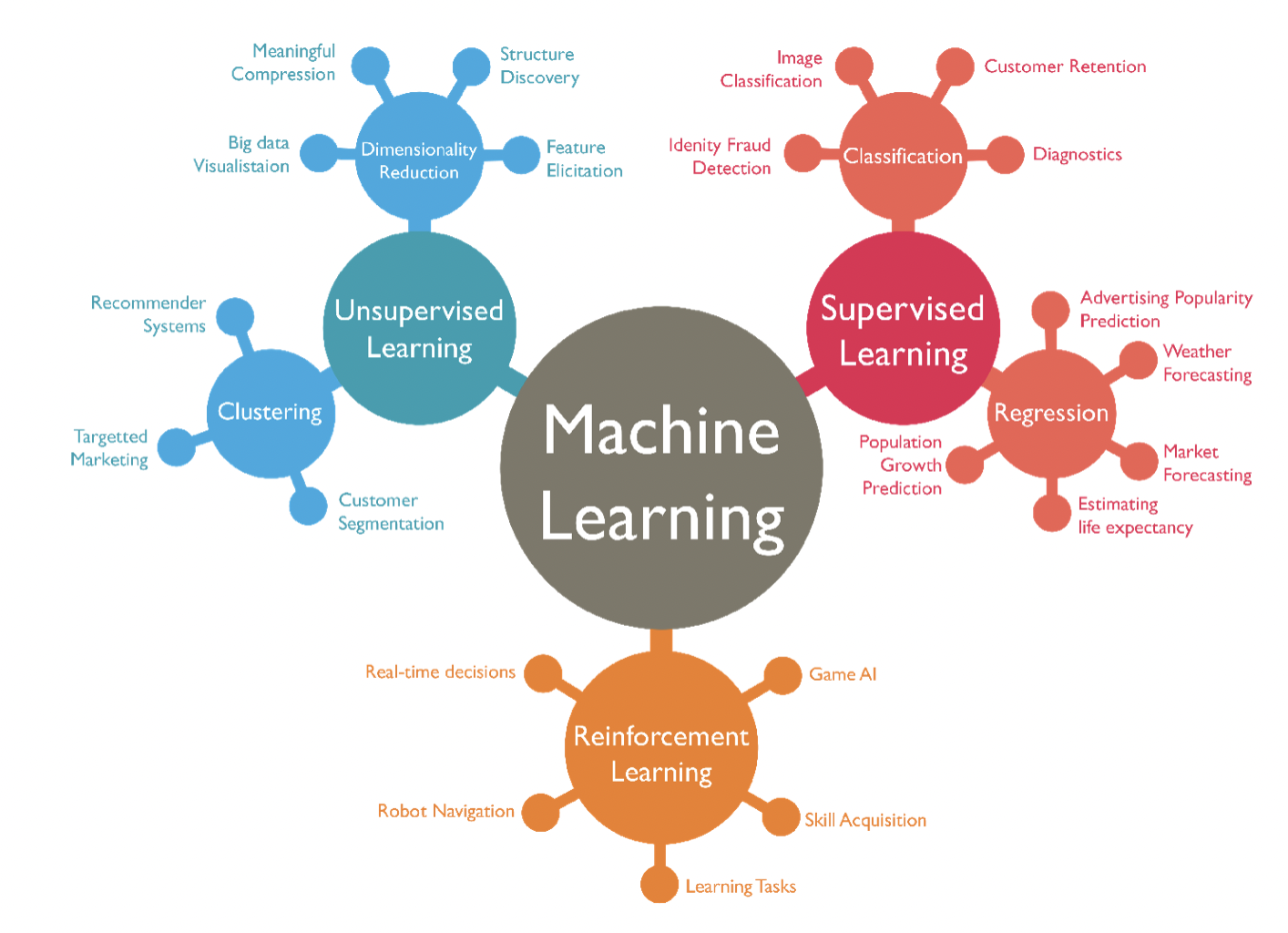
## 2.2 Success stories of adopting Machine Learning

In business forecasting, mispredictions happen when assumptions are coupled with unforeseen occurrences. Machine Learning and AI ensure that these mistakes are minimised. Business forecasts are utilized to construct strategies based on those precise predictions. Historical data are collected and assessed using quantitative or qualitative models. This information is essential for the identification of trends, which are then applied to influence decision-making. These activities might include operations, marketing, and finance. There are several ways businesses use forecasting; such as demand forecasting, predictive analytics, and demand sensing.

**Demand forecasting** is the projection of the expected future demand. Operations, such as warehousing, shipping, and pricing, come under forecasting because these jobs require information on clients’ demands.

**Predictive analytics** is a popular type of machine learning that many businesses use. Statistical methods allow not only demand estimates, but also an understanding of what drove the sales for the firms’ products or services. It also offers insights into how customers might act in specific conditions. Firms can integrate company data with other indicators with the use of predictive analytics tools. This ranges from anything between economic indicators to even changes in weather as these factors affect consumer bargaining power and simplify the discovery of potential markets.

**Demand sensing** employs machine learning to capture changes in spending activity in real time. Solutions in demand sensing take daily data and then compares it with prior trends to identify an increase or decrease in sales. The system evaluates and looks at the reasons for sales outcomes and proposes adjustments.



*Figure 2a: The different types of machine learning used by businesses*

For instance, in 2015, a company called Kroger acquired, transported and renamed a UK analytics business '84.51°'. The acquisition led to a scheme known as "Embedded Machine Learning," that routinely frames, constructs and deploys machine learning solutions in its entirety. Managing the supply chain is a major application of this embedded machine learning for the retailer. In 2018, its application for AI-powered sales projections could anticipate sales statistics on a daily basis on each of the goods in 2500 outlets. Kroger is one of few non-tech companies that has made the required funding available for bringing a full AI and ML in-house team. This team designs, builds and deploys all sorts of solutions across all sections of the corporation. By integrating AI, Kroger aims to minimise expenses and to maximise every opportunity through reduction of inefficiency.

Regarding another example in the same industry as EIU, Oxford Economics provides full economic solutions involving databases, publications, and analytical tools. They assist their clients in tracking and analysing global trends in countries, industries, cities, and tourism. Other than that, they also help clients determine top locations that would give the most potential for their firms to flourish by using a “new market evaluation” technique. The results of these aid in the development of location strategy plans and investment strategies. Oxford Economics also helps to recognise the implications of economic risks on the investing strategy via scenario planning, where made-up scenarios are tailored to the client's business, informing them of the likely effects on their national, municipal, and regional markets. These scenarios can include random events such as Brexit, policy changes, evolving consumer habits, or a recession. Consequently, their economic forecasts also help clients navigate through global crises by discovering challenges and opportunities for their firms.

## 2.3 Value of Machine Learning to EIU

Machine Learning provides EIU the opportunity to enhance their product, through automation and continual learning of the Machine Learning model.

## 2.3.1 Accuracy and Efficiency

Using machine learning to generate predictions is more efficient, with the process being automated. As such, there is less chance for human error and the system can offer more accurate predictions, especially with properly trained machine learning models. Due to its nature of assessing historical data, the system will also increase in accuracy over time as more data is collected and analysed. By improving on the quality and accuracy of its reports through application of Machine Learning on predictive analysis, EIU can improve its value proposition to its customers.

## 2.3.2 More Comprehensive Report

Machine learning models can determine significant factors that affect desired output variables (*e.g. real GDP Growth)*. Knowing these factors allow clients to better understand the strengths and weaknesses of a country or region, depending on the applied model. It provides further understanding of the data and allows clients to make better, more informed decisions (i.e policymakers can formulate policies that can enhance these factors which will in turn improve GDP and real GDP growth for the country). EIU’s product will also be more comprehensive, and they can stay ahead of its competitors and remain the preferred choice for economic insights.

# 3. Proposed Approach

Our proposed approach starts with extracting previous 10 years' data from approximately 60 countries in EIU’s list of countries they currently provide a report of. The countries were chosen randomly and evenly across the regions set by EIU. While the years would affect prediction due to world phenomena, it would not be feasible to only take one year’s data since we cannot predict the appearance of random phenomena (e.g. COVID-19) nor its effects on GDP. Since different countries achieve their GDP differently, we had to find a subset that would make the importance of the variables more uniform.

We chose a region, specifically 20 countries in Asia. Our reasoning was that a region could have more similarities in the variables that will affect real GDP growth more significantly.

Additionally, it may not be the best approach to directly compare the results from EIU’s analytical model to our machine learning model as we were unable to obtain the exact datasets for some EIU’s variables. A basis of comparison should be established to ensure a fair analysis of the results. Therefore, to test the hypothesis of whether other indicators (those not used by EIU) are significant in the prediction of real GDP growth, iterations of Linear Regression and CART using only EIU variables were run and compared to our model, which comprises several other variables. Ultimately, we hope to showcase a sample of how Machine Learning Models can help achieve better results with addition of new variables.

# 4. Data

As mentioned previously, EIU uses a series of indicators, mostly economic indicators in their forecasting of real GDP Growth. These variables are widely used and are appropriate indicators for the prediction of real GDP Growth. However, this current set of variables used in the prediction may not be all-encompassing.

Nonetheless, to establish a baseline that models after these current variables that EIU uses, we have included and used variables that EIU had considered under their prediction model into our dataset. However, it should be noted that the data used may not match exactly to that of EIU as some of the data were not available. In such cases, we made great effort to obtain variables that are similar and of good replacement.

In our approach in developing our machine learning model, we aimed for our model to be able to predict the real GDP Growth of countries in Asia. As a result, in the data extraction and mining process [(Appendix B)](#1lo74gr6kgp3), we considered 20 countries, all of which are relevant and of interest to the EIU. We then proceeded to extract 10 years’ worth of data, spanning from 2010 to 2019.

As seen in [Appendix B](#wq5j5h1ct4pn):

|  |  |
| --- | --- |
| Dependent Variables | Independent Variables considered under EIU |
| Real GDP Growth | Market Size, Life Expectancy, Mean Years of Schooling, Government Indicator, Internet Penetration, Trade Openness, Labour Force, Domestic Credit to Private Sector |

## 4.1 New Set of Independent Variables

In our approach to our business opportunity statement - finding a more optimal set of variables that can better predict real GDP Growth. We researched and included key variables, both economic and non-economic indicators that have proven to have some relation to real GDP growth.

We aim to include a diverse and large set of variables that are, importantly, related to a country’s growth potential. We do not aim to conduct a too extensive filtering of these variables at this point, as we feel that all these variables are potentially relevant in real GDP growth prediction. We aim to include these variables into our dataset and when we develop our machine learning models (Linear Regression and CART), the process of which will filter out the statistically significant and important variables.

The list of the new independent variables with their respective definitions, justifications and their data source can be found in our data dictionary document which is also available in [Appendix A](#_fn810wdrxphj).

# 5. Data Extraction

Data extraction is the process of extracting data from primary studies into a standardised table (Mathes et al., 2017). In our case, the group extracted data from credible data sources such as World Bank and International Monetary Fund (IMF) where CSV versions of the datasets for the specific variables were available for most countries, before combining them into a final spreadsheet which will undergo data cleaning.

Original datasets of the variables needed were downloaded from the respective sites where it went through Data Extraction ([Appendix B](#_6ps4ra6w5up4)) to extract only data for the countries and years we have in our final spreadsheet. In this process, differences in naming convention for countries between the original dataset and our final spreadsheet were identified to allow us to standardise them. The data extracted were then combined in the final spreadsheet.

# 

# 6. Data Cleaning

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset (Paterson, n.d.). It is especially important for our dataset to be cleaned, as our dataset comprises data combined from multiple data sources, for different countries across the span of 10 years. There is a need to handle inconsistencies in data - missing, incorrect or inconsistent data values, to prevent false conclusions when developing our machine learning model as it would cause our outcomes and algorithms to be unreliable.

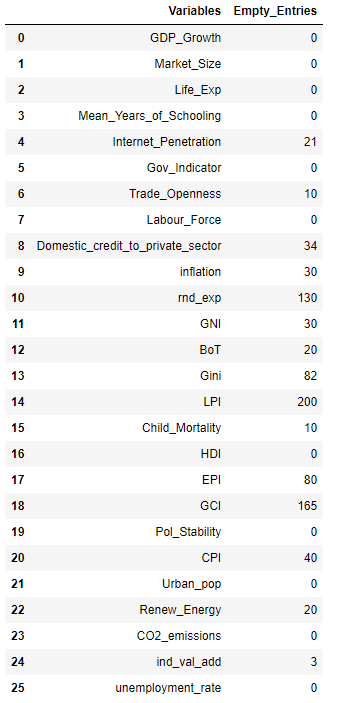
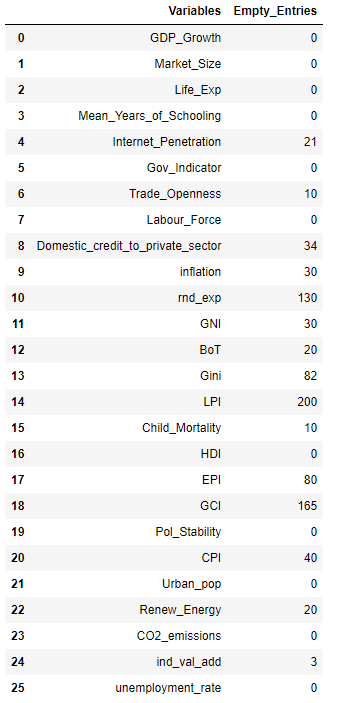
In order for our machine learning models to interpret features on the same scale, we performed feature scaling.

## 6.1 Dealing with NAs

It is important that the independent variables, countries, and years we use in our machine learning model should not have too many missing values as it can affect the accuracy and training of our models. As such, we performed data cleaning to remove the variables and countries that contain too many NAs. [(Appendix B)](#onqfnf9tnijy)

### 6.1.1 Removing Variables

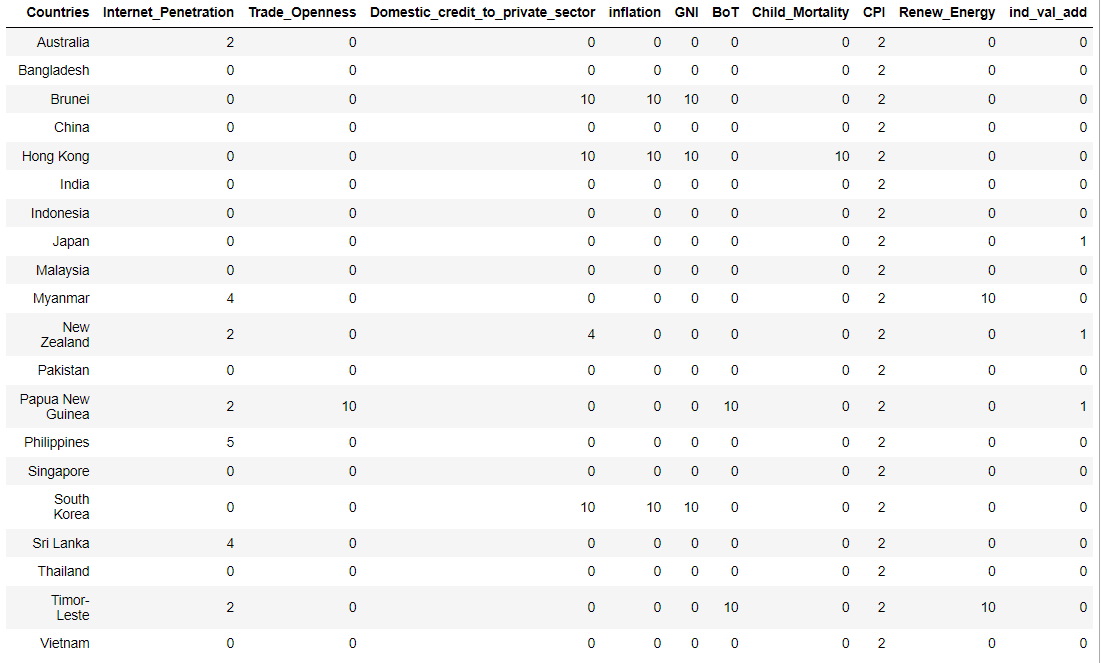
We count the number of missing values corresponding to each independent variable and we will remove those variables that have more than 20% of their data missing.



*Figure 6a: Number of missing values per variable*

From Figure 6a, rnd\_exp, Gini, LPI, EPI, and GCI have more than 40 entries that are not filled, thus we removed these variables from our dataset, bringing our total number of variables used down from 26 to 21.

### 6.1.2 Removing Countries



*Figure 6b: Number of NA values per country per variable*

For each country, there are 10 years worth of data. To ensure that we do not have too many consecutive missing values, we aim to remove those countries that have more than 2 variables with 10 years worth of missing data.

Referring to Figure 6b, we removed Brunei, Hong Kong, Papua New Guinea, South Korea, and Timor-Leste, bringing our total countries down from 20 to 15.

## 6.2 Feature Scaling

Our dataset deals with several different indicators, all of which have varying degrees of magnitude, range and units. In order for our machine learning models to interpret these features on the same scale, we need to perform some sort of feature scaling. This is especially so for machine learning algorithms like linear regression, neural network that use gradient descent as an optimization technique (Bhandari, 2021)

Here, we perform two methods of feature scaling, Normalization and Standardization ([Appendix B](#_h1k5xppeesch)). The goal of feature scaling is to obtain a set of independent variables that are scaled bringing every feature in the same footing without any upfront importance. After which, we will fit our models to raw, normalized and standardized data and compare the performance for best results.

We do note that algorithms such as CART which rely on rules and decision split do not require normalization or standardization. As such, the normalized and standardized dataset will only be used by our Linear Regression Model.

# 

# 7. Linear Regression

## 7.1 Linear Regression Process

For continuous variable Y which is our real GDP Growth, we require Root Mean Square Error (RMSE) to determine prediction errors and hence accuracy of the model. As explained in the proposed approach, for a more fair baseline of comparison, we will perform two iterations of Linear Regression, one with just EIU variables and the other with EIU variables and the additional variables we aim to explore.

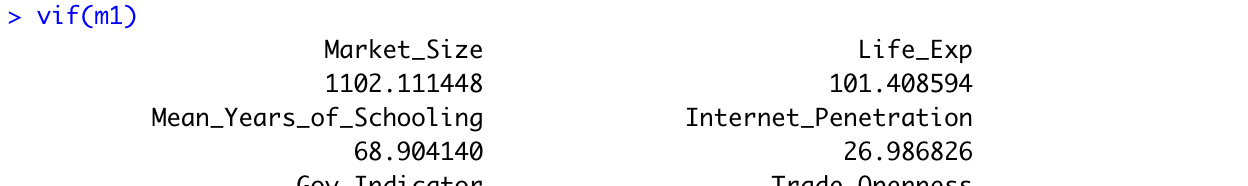
For our model, we ended up using standardised data for consistency (Jaadi, Z., 2019). ​​This was important for our dataset since our x-variables in the unstandardised dataset had very different ranges of values. In machine learning, the model will place more importance on the variable with larger range, and we want to avoid that bias.

To illustrate the process, we will train our linear regression model using the standardised dataset as it is more comprehensive and consistent. *This section goes through the more advanced parts and explains our thought process. For more detailed explanations, please refer to the R script for Linear Regression.*

**Step 1: Cleaning Dataset**

We made the choice to remove columns that we did not need for the model and any rows with NULL values, by using na.omit().We were thus able to obtain an optimal, cleaned data with no missing values. Additionally, we did not use mean as a replacement of the missing values, as it could have skewed a time-series pattern across the x variables if it had been present.

**Step 2: Addressing Multicollinearity**



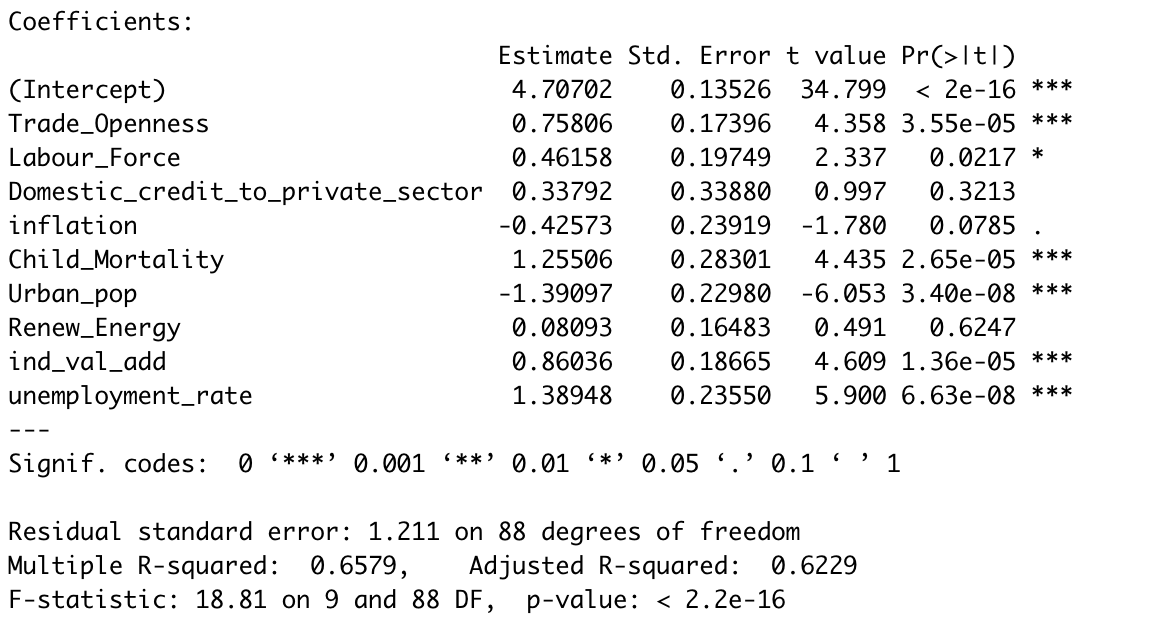
*Figure 7a*

Multicollinearity is when two or more independent variables are highly intercorrelated. Estimated coefficients end up being able to affect one another, affecting stability of results. It will also result in an overfitting problem.

We used VIF (variance inflation factor) to eliminate multicollinearity. It measures how multicollinear variables could affect the model. The cut-off we used was VIF <10. From our research, credible sources cite​​d (Hair et al., 1995) this as an appropriate benchmark. After running vif on our linear model, the variable with the highest vif is removed.

The linear model is then updated, and the process is repeated until all VIF values are below 10. The entire set is then updated to remove factors causing multicollinearity.

**Step 3: Removing Non-significant x-variables**



*Figure 7b*

Running a summary() on a linear model gives us the above output. We refer to Pr(>|t|) [p-value] as well as Adjusted R-squared. The variable with the highest Pr(>|t|) value is the least significant in determining y-variable (the more \*, the more significant). This gets removed and we check for changes in Adjusted R-squared. An increase indicates that there is an improvement in the model. Non-significant factors will be removed one-by-one, as the Pr(>|t|) changes. When the Adjusted R-squared drops back down, we trace back to the previous step, and stop the removal there. Afterwhich, we will update the dataset once again.

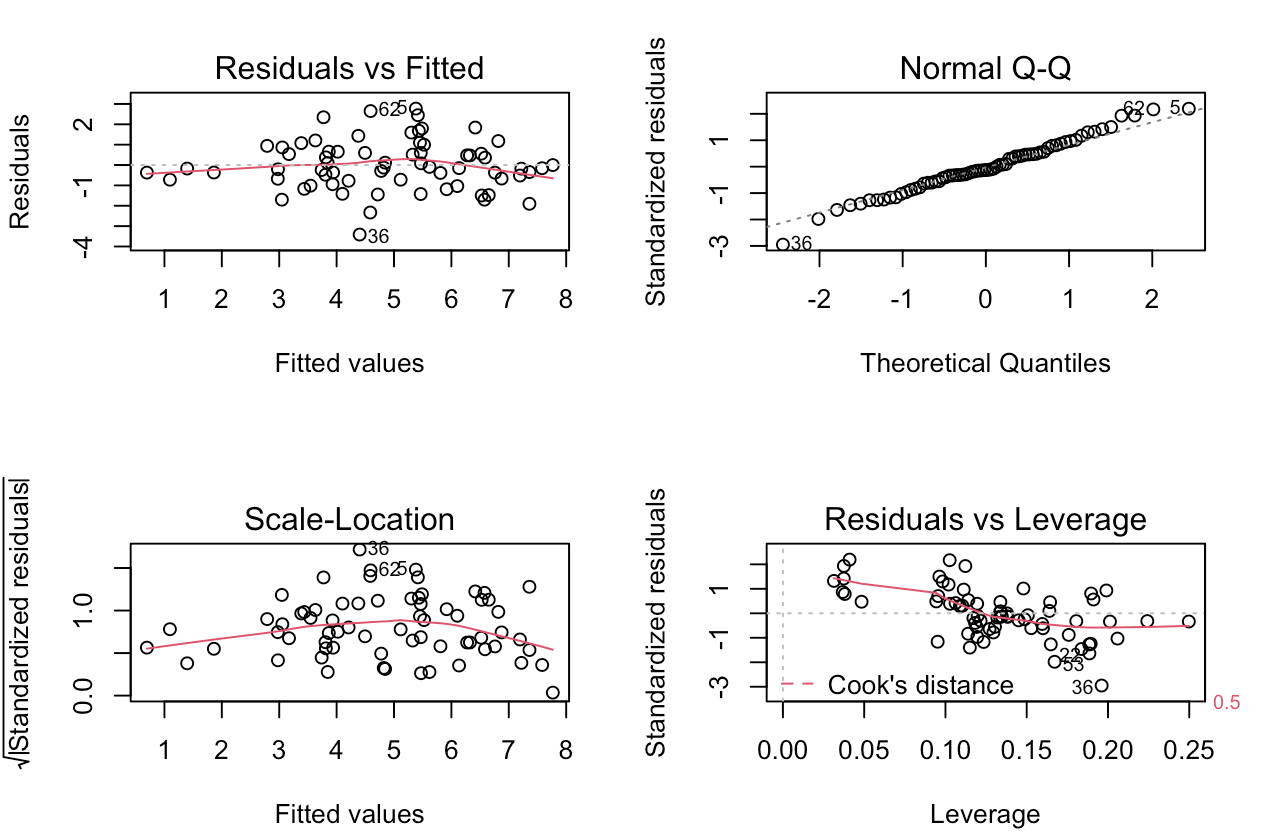
**Step 4: Create train-test split**

Following our proposed approach, we used a 70-30 split for train-test remembering to compare the two sets for even distribution.

**Step 5: Create model on train set**

We created a linear model for the train set using lm().

**Step 6: Plot and check if train set is suited for linear regression**



*Figure 7c: Diagnostic checks with plot() function*

The model diagnostic plots tell us that our model is good to be used for linear regression, by being being able to show:

1. Linear Association between Y and Xs.
2. Errors have a normal distribution with mean 0.
3. Errors are independent of X and have constant standard deviation.
4. Presence of influential outliers

Things we look for to show they prove the above:

Residuals vs Filled: horizontal straight line not far from the dotted line

Normal Q-Q: points do not deviate from diagonal dotted line

Scale-Location: red line is almost horizontal, points have roughly same variability for each fitted value

Residuals-Leverage: lack of influential outliers which sit outside of the dotted line

**Step 7: Apply model from train set to predict on test set**

We use predict() to apply our train set model to the test set data. From there, we can derive a test set error for RMSE calculation.

## 7.2 Comparison of results with EIU baseline model

### 7.2.1 Comparison of Result

RMSE-train: 1.201735 | RMSE-test: 1.065562

There are two things we look out for when comparing RMSE, namely the similarity of value between train and test set RMSE, and the value itself. A smaller value would indicate higher accuracy. Our RMSE for train and test set were fairly similar.

Following the same model creation process for the EIU baseline model, the resulting RMSE was

RMSE-train (EIU baseline model): 1.830851 | RMSE-test (EIU baseline model): 1.991584

Since our RMSE value for the test set is smaller than that of the EIU baseline model we concluded that that our RMSE value was acceptable since it showed improvement in accuracy.

### 7.2.2 How EIU can Benefit from Adopting Some Variables

After all the iterations and eliminations, the outcomes of the indicators were as follows:

* ​​y-variable: GDP\_Growth
* Variables present in EIU baseline model: Domestic Credit to Private Sector, Trade Openness, Labour Force ([Figure C3 Appendix C](#dc91tbwe4s3y))
* New Variables obtained: Inflation, Child Mortality, Urban Population, Industry Value Add, Unemployment Rate ([Figure C2 Appendix C](#nfh9fuwlvouw))

The remaining variables are all statistically significant in predicting real GDP Growth. Other than the variables that EIU had used, there were additional variables, namely Inflation, Child Mortality, Urban Population, Industry Value-Add and Unemployment Rate which are new variables not previously present in the original EIU baseline model. The better RMSE of our model indicates that these variables helped improve upon the accuracy of predictions.

It affirms our hypothesis that there are other factors that EIU can integrate into their prediction model to improve upon it. Through adopting these extra variables, EIU may be able to increase the accuracy of their report improving upon their value proposition.

# 8. Classification and Regression Tree (CART)

A Classification and Regression Tree (CART), is a predictive model, which explains how an outcome variable's values can be predicted based on other values. In our case, we aim to explain how real GDP growth can be predicted based on several economic and non-economic variables. Since real GDP growth is a continuous variable, the tree formed will be a regression tree.

## 8.1 Dataset used in CART Model

### 8.1.1 Original Cleaned Dataset

When running the CART model on the final dataset, our optimal model included 5 branches [(Figure 8b in 8.2 CART Process)](#51tjec32qplh) with a RMSE of 1.839 for the test set and a RMSE of 1.131 for the train set. The group wanted to explore ways to improve the accuracy and possibly the inclusion of more variables in the optimal model. As such, the dataset went through additional sequential cleaning, producing a few iterations of cleaned datasets.

### 8.1.2 Dataset with less than 11 NAs

One iteration which the group had explored is to remove NAs by looking at independent variables with more than 10 NAs ([Appendix D](#_gdepzvkwck9o)), which are Internet Penetration and Corruption Perception Index (CPI).

The CPI variable had missing values across 12 countries ([Appendix D](#6zfmvttepn7x)), hence we chose to drop the entire column since it would not be able to showcase an accurate and fair depiction of CPI’s influence for real GDP growth for all countries in our dataset. The Internet Penetration variable had missing values for 5 countries. We chose to drop the rows of missing values as the Internet has an influence on economic growth particularly for large and developed economies (Mayinka, 2011). Hence, we believe that preserving this variable is important.

With this dataset, we ran a CART model to obtain the optimal model with 6 branches ([Appendix D](#nlnvllcmkpz7)) and RMSE of 2.589 (3dp) for the test set and RMSE of 1.068 (3dp) for the train set.

### 8.1.3 Dataset with no NAs

In this iteration, we remove all NAs in the dataset ([Appendix D](#_ks1qf9aid6tn)). Similar to 7.1.1, we removed the CPI column and the rows containing NAs for Internet Penetration.

For the rest of the variables containing NAs, namely Renewable Energy, Domestic Credit to Private Sector and Industry Value Add, we remove the rows instead of the columns as the NAs are consistent to only a country respectively. As such, we are still able to explore the influence of the variables with respect to real GDP growth in other countries.

With this dataset with no NAs, we obtain an optimal model with 4 branches with a RMSE of 1.803 (3dp) for the test set and RMSE of 1.237 (3dp) for the train set.

### 8.1.4 Dataset with Mean in replacement of NAs

In this iteration, the group replaced NAs present in independent variables with mean values across available years specific to the countries. For Myanmar, since data for Renewable Energy was missing throughout all 10 years, the rows were removed instead.

With this dataset, an optimal model of 2 branches was obtained ([Appendix D](#_ctz4gr424u84)), with RMSE of 2.157 (3dp) for the test set and RMSE of 1.400 (3dp) for the train set.

### 8.1.5 Final choice of dataset used in CART model

After comparing the 4 CART Models obtained from the different iterations of dataset, the group decided to utilise the **original dataset** for the CART Model as it had

1. Greatest number of branches
2. Relatively low test set RMSE
3. Relatively low overfitting based on the difference in train set and test set’s RMSE

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | No.of Branches | Trainset RMSE | Testset RMSE | Difference in RMSE |
| **Original** | **5** | **1.131** | **1.839** | **0.708** |
| < 10 NAs | 6 | 1.068 | 2.589 | 1.521 |
| No NAs | 4 | 1.839 | 1.803 | 0.566 |
| Using mean | 2 | 2.236214 | 2.157 | 0.757 |

*Table 8a: Comparing across datasets*

## 8.2 CART Process

As discussed above, the team proceeded with the analysis of the CART model using the original cleaned dataset. In this section, we explain the steps taken to obtain the results as shown above.

The goal of our CART machine learning model is to explain how real GDP Growth can be predicted based on several economic and non-economic variables and to identify important variables. The output of CART is a decision tree where each fork is a split in a predictor variable and each leaf node contains a prediction of real GDP growth based on the mean of the values for the real GDP growth at such leaf nodes.

**Step 1: Splitting the Data**

The group decided to use 10 cross validation folds with 70-30 split ratio as it is the optimal for CART (Vrigazova, 2021).

**Step 2: Growing the Tree**

Next, growing the tree involves deciding on which features to choose and what conditions to use for splitting, along with knowing when to stop. At each node, CART will consider and test all independent variables and all possible values of the variable to determine the best possible split. To identify the variable and the specific split or cut-point used, we calculate how much accuracy each split will cost us, using a function: sum of squared estimate of errors (SSE). Tree construction will then end using a predefined stopping criterion, which is a minimum of 2 training instances assigned to each leaf node of the tree.

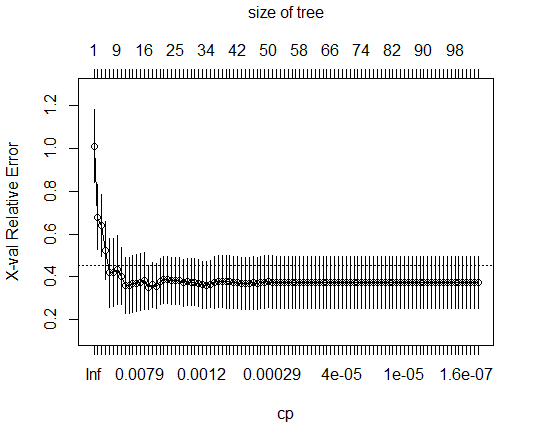
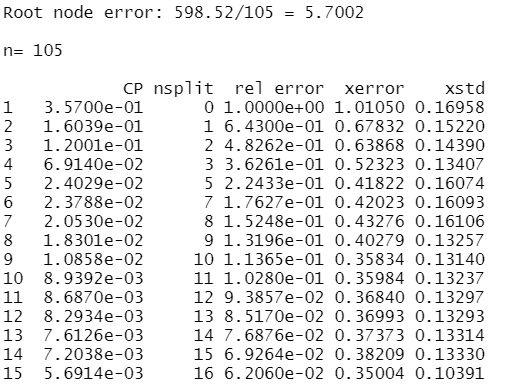
However, growing the tree to its maximum is not ideal as such a tree is complex and leads to overfitting. The complexity of a decision tree is the number of splits in the tree. Hence, simpler trees are preferred. They are easy to understand, and less likely to suffer from overfitting.

**Step 3: Pruning the Tree**

The performance of a tree can be further increased by pruning through the removal of branches that make use of features with low importance. This way, we reduce the complexity of the tree, and increase its predictive power by reducing overfitting. We used the Minimal Cost-Complexity Pruning method to prune our tree.

This algorithm is parameterized by α(≥0) known as the complexity parameter. The complexity parameter is used to define the cost-complexity measure, Rα(T) of a given tree T: Rα(T)=R(T)+α|T| where |T| is the number of terminal nodes in T and R(T) is defined as the total misclassification rate of the terminal nodes.

**Step 4: Finding CP Optimal**

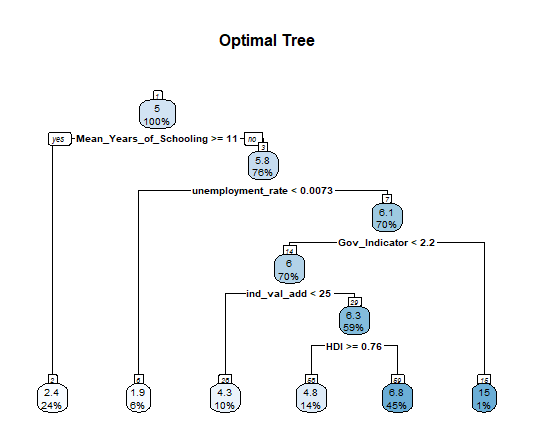


*Figure 8a: printcp() and plotcp()*

To find out how the tree performs, we ran the printcp() function to find the one with the least cross-validated error, afterwhich, add its 1 standard error to obtain the cv error cap, which will be used to find the cp optimal. Trees whose cv error is below the cap are statistically equivalent in terms of error. As a result, the optimal cp region is the one whose cp error is just below the cv error cap. From our calculations, the optimal cp is 0.04075956.

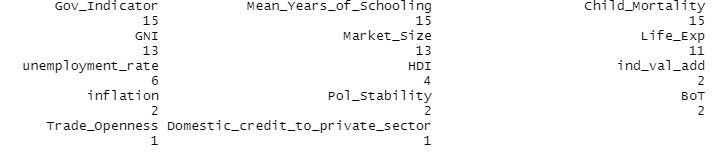
**Step 5: Optimal Tree**

Now, with the cp optimal obtained, we are able to prune the tree to its optimal size. Figure 8b shows the final optimal tree achieved for our CART model. Referring to the figure, we observed that the indicators used as the decision splits are Mean Years of Schooling, Unemployment Rate, Government Indicator, Industry Value Add, and Human Development Index.



*Figure 8b: Optimal tree obtained after pruning*

**Step 6: Variable Importance**



*Figure 8c: Variable Importance of Original Dataset*

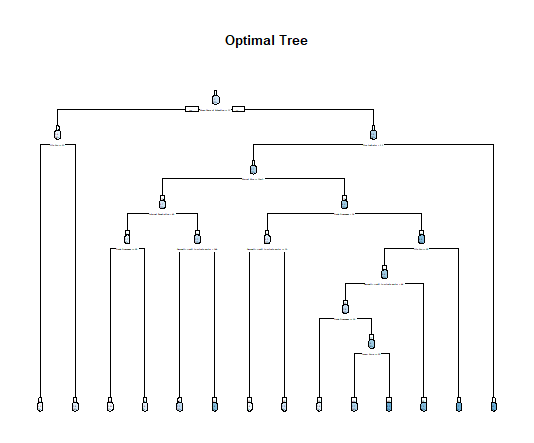
Variable importance is the sum of the improvement in all nodes where the variable appears as a decision rule (Steinberg, 2009). Being a surrogate also increases a variable’s importance. The relative importance of variables were obtained (Figure 8c) for the optimal CART model.

Variables that appear as nodes on the tree are important variables as feature selection is completed automatically (Kotu & Deshpande, 2019). Our threshold to determine if a variable is important will be based on the least important variable that is present in the optimal tree, in this case, it is Industry Value Add. Hence, any variables that are ranked higher than Industry Value Add are important. Certain variables such as Child Mortality, GNI, Market Size and Life Expectancy were not variables present in the optimal tree, but are shown to have high variable importance, suggesting that they could’ve been used as a surrogate thus increasing their importance.

Thus, the important variables obtained from our CART model are Government Indicator, Mean Years of Schooling, GNI, Market Size, Life Expectancy, Unemployment Rate, HDI and Industry Value Add.

## 8.3 Comparison of results with EIU baseline model

As mentioned in Section 2.4: Proposed Approach, we will perform an iteration of the CART model using just EIU’s variables. The following figure obtained is the optimal tree after pruning.



*Figure 8d: Optimal tree obtained using only EIU Variables*

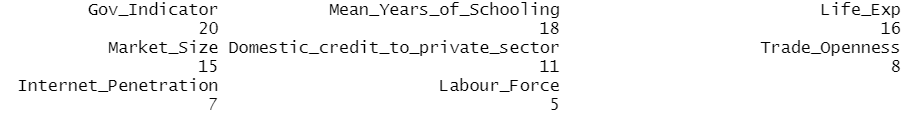
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | No.of Branches | Trainset RMSE | Testset RMSE | Difference in RMSE |
| **Original** | **5** | **1.131** | **1.839** | **0.708** |
| EIU Variables | 13 | 0.8067812 | 2.236214 | 1.430 |

*Table 8b: Comparison of the original dataset and EIU baseline model*

From Table 8b, we note that the testset RMSE for EIU baseline model is 2.24, the difference in RMSE with the train set is 1.430 which is 2 times more than that obtained from our model. This suggests that the EIU baseline model is more prone to overfitting as compared to our model.

Additionally, the testset RMSE value for the EIU baseline model is also higher than that obtained from our model. This suggests that our model is also more accurate in the prediction of real GDP growth, which further affirms the hypothesis that the inclusion of other indicators besides those used by EIU, are significant and adds value to the prediction of real GDP Growth.

Next, we compare the indicators obtained from both CART models. For the EIU baseline model, the figure below shows the variables importance of the indicators used.



*Figure 8e: Variable Importance for EIU Variables*

Referring to Figures 8c and e, we observe that in both models, Government Indicator and Mean Years of Schooling (EIU variables) are both the top indicators in terms of variable importance.

However, in our model, GNI is the next most important variable which is an indicator not present in the EIU baseline model. In particular, Unemployment Rate, Industry Value Add and HDI which were also not part of the original variables used by EIU, are variables used and are important in our optimal CART model. With the inclusion of these variables, our model achieved a lower test set RMSE, implying a better prediction of real GDP growth. As such, EIU could consider the inclusion of these variables to obtain a more robust model that can better predict real GDP growth.

# 9. Insights gained from Linear Regression and CART

After developing our machine learning models for Linear Regression and CART, we will now compare the results between the models and evaluate the outcomes.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Trainset RMSE | Testset RMSE | Important Variables |
| LR Optimal | 1.201735 | 1.065562 | DomestPrivSec, TradeOpen, LabForce,  **Inflation**, **ChildMort**, **UrbanPop**, **IndValAdd**, **UnempRate** |
| LR EIU Baseline | 1.830851 | 1.991584 | MarketSize, TradeOpen, LabForce, DomestPrivSec, MeanYearsofSchooling |
| CART Optimal | 1.131 | 1.839 | GovInd, MeanYearsOfSchooling, **ChildMort**, **GNI**, MarketSize, LifeExp, **UnempRate**, **HDI**, **IndValAdd** |
| CART EIU Baseline | 0.8067812 | 2.236214 | GovInd, MeanYearsOfSchooling, MarketSize, LifeExp, DomestPrivSec, TradeOpen, IntPen, LabForce |

*Table 9: Comparison of RMSE across the LR and CART models*

Referring to the table above, we observed that both optimal models for LR and CART achieved lower RMSE values as compared to its EIU baseline model counterparts. This suggests that regardless of whichever machine learning model is used, the outcome is that with the inclusion of other indicators that were originally not used in the EIU baseline model, it will lead to a better prediction of real GDP growth. Further affirming the hypothesis that other indicators are important to a country's economy and may not necessarily need to be economic indicators to better predict real GDP growth.

From our results, we conclude that Inflation, Child Mortality, Urban Population, Industry Value Add, Unemployment Rate, GNI and HDI are such indicators that EIU could consider to include into their model, with greater emphasis placed on **Child Mortality**, **Unemployment Rate**, **Industry Value Add**.

Next, as we compare between LR and CART, we note that the model for LR was able to better predict real GDP Growth as seen with its lower RMSE value. This could potentially be due to 2 reasons, firstly, LR model was trained on data with all NAs removed thus removing all noise from the dataset whereas the CART model was trained on a dataset that had several NAs. Additionally, from the diagnostic plot of LR, we noted the model is a good fit for Linear Regression as it follows the assumptions behind a linear model.

Next, comparing to EIU’s actual analytical model, we extracted EIU’s estimated values for real GDP growth. However only 14 datapoints could be found across 3 countries (i.e Australia, China and Japan) and the resulting RMSE obtained was 0.738. Comparing the accuracy, EIU’s analytical model is slightly more accurate compared to the optimal LR model’s of 1.066. However, this could be due to having a larger dataset and specific analytical models for the respective countries. However, our insights suggest that EIU should still adopt machine learning in place of their analytical models as with the resources and dataset they have, a much better result could definitely be obtained should they adopt machine learning.

# 10. Limitations & Future directions

## 10.1 Limitations

Regarding our expectation to include more appropriate and well-performed economic and non-economic predictor variables in explaining the response variable, i.e. real GDP growth, we did consider and selected quite a few relevant variables and designed the procedure to refine. However, since the data we included in our model are from multiple countries and there are, indeed, different economic conditions and developments among countries depending on their national backgrounds, and motivations in promoting economic growth, etc, it would be quite arbitrary to include different countries as a whole to predict a final real GDP growth. Even though we took this problem into consideration and included only the countries in Asia to limit the regional differences, there could still exist differences between these countries, affecting our results obtained from the trained model.

In the section of linear regression model, we eliminated variables with too many missing values as it may disturb the final result. However, it would be quite arbitrary for us to ignore them entirely rather than taking them into account. Therefore, the model would possibly be more comprehensive and robust if we had the dataset for those deleted variables included into the training.

Apart from the data preprocessing, we also encountered some obstacles while collecting the valid and accurate data for those variables we selected to include into our adjusted model over a long term. For instance, there was insufficient data over 20 years or above in record for many countries. As such, we compromised to use only 10 years of records. We could have possibly obtained a more optimised and desirable model accuracy if more data was available.

Lastly, the newly added predictors are selected based on our observations, discussions and certain research reports. There may exist other variables outside of what we have explored, which could better predict real GDP growth. This will have to be further explored in future works due to the time constraints we had with the project.

## 10.2 Future Directions

Despite the limitations, we strongly believe that the benefits and insights gained from our findings are highly relevant to EIU. New directions and pilot studies could be initiated by EIU, paving the way for new exciting opportunities.

From our findings of the report, it is evident that EIU should consider adding other indicators such as Child Mortality, Unemployment Rate and Industry Value Add into their preexisting models for their prediction of real GDP growth. EIU could consider integrating these indicators into their flagship country report, so as to obtain a more comprehensive, holistic as well as indicative estimate to real GDP growth, providing stakeholders a better understanding of the economic and political conditions of the country. In addition, as mentioned in our limitations, due to our time constraints, we could only explore some variables, thus, EIU could consider further exploring more variables that could have an influence on a country’s economy to include in their model.

In addition, machine learning techniques could provide greater insights and benefits to EIU. EIU should thus consider exploring the adoption of machine learning to forecast, and pilot studies to further evaluate the use of machine learning. Coupled with their extensive resources and expertise, this could yield much better results than the team’s current exploration.

In terms of future direction for EIU to expand its offerings, we have a few suggestions.

EIU can expand their services by providing data for global industry services such as forecasting commodity prices, specific-industry forecasts, as well as regional industry services. Within a comprehensive and internationally integrated modelling approach, EIU can provide short-term to long-term forecasts and analysis of production and investment patterns throughout industries. They can also evaluate historical data and forecasted data, acquire insights in the influence of economic trends, and assess the performance of industrial sectors under various economic circumstances using reports, databases, and econometric models. On top of that, EIU can also consider posting daily publications, research & analyses, and articles. For instance, industry-related indexes like Coal index for daily tracking.

Regarding education, EIU can consider having a Teaching Hub, where students, analysts and academics can look into powerpoints, interactive notes, dashboards and other learning materials. EIU can also provide new on-demand consultation services to provide solutions by having a global team of experts ready to give in-depth knowledge on the economic themes such as trade, finance, sustainability, policies, shipping, manufacturing and more. The experts are to keep a close eye on global and local markets and provide a balanced view of the dynamics that affect a client’s business.

## 10.3 Pilot Studies that can be initiated at EIU to evaluate Machine Learning

### 10.3.1 Trade-offs in fairness in predictive algorithms

One topic that has gotten a lot of attention is the issue of fairness and discrimination, such as whether algorithms will encourage discrimination in specific situations (Eg. Gender or ethnicity-based discrimination in situations such as hiring, court rulings, and bank lending). As a result, economic issues such as unemployment can arise.

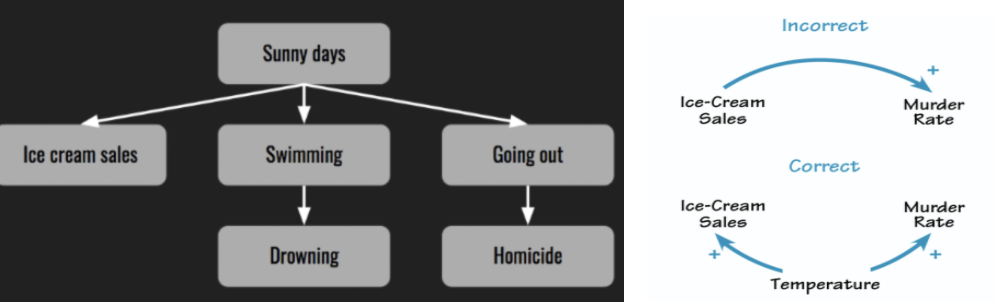
Womens' attempts for finding jobs are hampered by employers' bias towards married and unmarried women, lack of job replacement assistance due to occupational segregation by gender, salary discrimination, and a lack of a social network.

**Case study:** Amazon used to have a hiring algorithm that was found to discriminate against women. The program was designed to scan the internet for plausible applicants and rate them on a scale of 1 to 5. The algorithm, on the other hand, learned to demote women's CVs for STEM job applications, like software developer. Amazon's algorithm learned that male applicants were favoured. CVs which included the term "Women" such as "Women's Rugby Union" were penalised. It also immediately rejected female applicants from two specific all-girls' colleges. To assist forecast outcomes, AI systems are trained to look for patterns in massive datasets. Amazon's algorithm learned how to discover the best candidates by analysing all CVs submitted to the business over 10 years. Given the company's low female employee ratio, which is typical of most technological firms, the algorithm rapidly identified male hegemony and assumed it was a factor in success. Thus, the algorithm became locked in a cycle of bias against female candidates in the process of increasing accuracy. Also, because the data used in the train set was made by biased, flawed humans, the algorithm gained negative human characteristics such as prejudice and discrimination that have long been an issue in hiring.

Due to pre-existing prejudice in some machine learning models, it can affect the unemployment rate. When women aren’t hired due to prejudiced predictive recruiting tools, structural and frictional unemployment can hinder a country’s economic growth. EIU could develop pilot studies to evaluate machine learning algorithms which display such cases of discrimination, especially so if these algorithms are related to economics and efforts by EIU. These include finding out how to establish fairness constraints, how to ensure that algorithms promote equity, and what kind of fairness is favourable. For instance, when a predictive model is used to arrange job interviews based on resumes, evaluations can be done on how both type I and type II errors may be equalised among two separate groups of people (eg. males and females).

### 10.3.2 Interpretability of models

There are debates over what interpretability is and if simpler models provide more benefit. Economists recognise that simple models may be deceiving. For example, two attributes can be positively correlated, but as a matter of fact it only seems that way because of coincidence and/or underlying factors. For example, sales of ice cream and murders in the US have a positive correlation. However, one does not cause the other. Correlation doesn't imply causation.



*Figure 10a showing how ice cream sales are related to murders*

However, this may not be as obvious in more complicated topics, and thus interpretability may sometimes be a problem. EIU could initiate studies to better understand correlation and causation between independent variables and response variables. If such studies can be done, more accurate predictions can be made as these variables may seem correlated, but are actually in fact cause and effect of each other. If this is so, it gives stakeholders a much clearer understanding and direction to take when making decisions. Thus, EIU can consider questions such as how to determine better correlation between variables, especially in cases where they seem correlated but are only so because of underlying reasons.

### 10.3.3 Identification and evaluation of inconsistencies in data collection

EIU could identify and evaluate inconsistencies in data collection from different countries. The World Bank is considered a credible source for datasets, but there still exists some limitations in the accuracy of the data collected versus actual cases, due to differences in the data collection processes. There should be consideration of the potential existence of biases between the real cases and the information indicated by the data collected. Furthermore, because different countries have different methods, procedures, and motivation for collecting raw data, and different criterias are followed to calculate or measure the predictors across different countries, this may lower the accuracy of the forecasts due to the inconsistency in data. Thus, pilot studies can be aimed at spotting these inconsistencies and an evaluation can then be made to study their impact on machine learning algorithms.

# 

# References

Amiri, Arshia, et al. (2013). *Impact of Maternal and Child Health on Economic Growth: New Evidence Based Granger Causality and DEA Analysis*. Lund University, <https://www.who.int/pmnch/topics/part_publications/201303_Econ_benefits_econometric_study.pdf>.

Bhandari, A. (2021, August 26). Feature Scaling | Standardization Vs Normalization. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/>

Callen, T. (2020). *Finance & Development*. Finance & Development | F&D. Retrieved October 29, 2021, from <https://www.imf.org/external/pubs/ft/fandd/basics/gdp.htm>.

Cederborg, Jenny, and Sara Snöbohm. (2016). *Is There a Relationship between Economic Growth and Carbon Dioxide Emissions?* . Södertörns University, <https://www.diva-portal.org/smash/get/diva2:1076315/FULLTEXT01.pdf>.

Dynan, K., & Sheiner, L. (2018). GDP as a measure of economic well-being. Work. pap, 43.

Elistia, E., & Syahzuni, B. A. (n.d.). *The correlation of the human development index (HDI) towards economic growth (GDP per capita) in 10 ASEAN member countries*. JHSS (JOURNAL OF HUMANITIES AND SOCIAL STUDIES). Retrieved November 3, 2021, from https://journal.unpak.ac.id/index.php/jhss/article/view/949.

European Commission. (2007). *Press corner*. European Commission - European Commission. Retrieved October 29, 2021, from <https://ec.europa.eu/commission/presscorner/detail/en/MEMO_07_472>.

Hair, J. F. Jr., Anderson, R. E., Tatham, R. L. & Black, W. C. (1995). Multivariate Data Analysis (3rd ed). New York: Macmillan

Ilter, C. (2016). What economic and social factors affect GDP per capita? A study on 40 countries. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2914765>

Jaadi, Z. (2019, September 4). When and Why to Standardize Your Data? *built in*. <https://builtin.com/data-science/when-and-why-standardize-your-data>

Kleinberg, Jon, et al. (2016). “Inherent Trade-Offs in the Fair Determination of Risk Scores.” *ArXiv:1609.05807 [Cs, Stat]*, *arXiv.org*, http://arxiv.org/abs/1609.05807.

Kotu, V., & Deshpande, B. (2019). *Data Science: Concepts and Practice*. Elsevier/Morgan Kaufmann Publishers.

Mathes, T., Klaßen, P., & Pieper, D. (2017). Frequency of data extraction errors and methods to increase data extraction quality: a methodological review. *BMC medical research methodology*, *17*(1), 152. https://doi.org/10.1186/s12874-017-0431-4

Mayinka. (2011). *Impact of internet on economic growth - mckinsey & company*. Retrieved October 29, 2021, from <https://www.mckinsey.com/~/media/McKinsey/Industries/Technology%20Media%20and%20Telecommunications/High%20Tech/Our%20Insights/The%20great%20transformer/MGI_Impact_of_Internet_on_economic_growth.pdf>.

Paterson, K. (n.d.). *Data Cleansing: What Is It and Why Is it Important?* Blue-Pencil. <https://www.blue-pencil.ca/data-cleansing-what-is-it-and-why-is-it-important/>

Patidar, P., & Tiwari, A. (2013). Handling missing value in decision tree algorithm. *International Journal of Computer Applications*, *70*(13), 31–36. <https://doi.org/10.5120/12023-8063>

Singh, Seema. (2019). “Why Correlation Does Not Imply Causation?” *Medium*, https://medium.com/@seema.singh/why-correlation-does-not-imply-causation-5b99790df07e.

Steinberg, D. (2009). Chapter 10 CART : Classification and Regression Trees.

Vrigazova, B. (2021). The proportion for splitting data into training and test set for the bootstrap in Classification problems. *Business Systems Research Journal*, *12*(1), 228–242. <https://doi.org/10.2478/bsrj-2021-0015>

Wu, S. (2020, May 19). Multicollinearity in Regression. *towards data science*. https://towardsdatascience.com/multi-collinearity-in-regression-fe7a2c1467ea

*10 companies using machine learning in cool ways*. (n.d.). Retrieved November 4, 2021, from https://www.wordstream.com/blog/ws/2017/07/28/machine-learning-applications

*50 ai examples from 50 different companies | manceps*. (n.d.). Retrieved November 4, 2021, from https://www.manceps.com/ai-examples

*Bespoke business forecasting*. (n.d.). Oxford Economics. Retrieved November 4, 2021, from https://www.oxfordeconomics.com/consulting/bespoke-business-forecasting

*Gross national income(Gni)*. (n.d.). Investopedia. Retrieved November 4, 2021, from https://www.investopedia.com/terms/g/gross-national-income-gni.asp

*How is machine learning used in business forecasting?* (n.d.). Knowledgenile. Retrieved November 4, 2021, from https://www.knowledgenile.com/blogs/how-is-machine-learning-used-in-business-forecasting/

*Increasing world’s share of renewable energy would boost global gdp up to $1. 3 trillion*. (n.d.). /Newsroom/Pressreleases/2016/Jan/Increasing-Worlds-Share-of-Renewable-Energy-Would-Boost-Global-GDP-up-to-13-Trillion. Retrieved November 4, 2021, from https://www.irena.org/newsroom/pressreleases/2016/Jan/Increasing-Worlds-Share-of-Renewable-Energy-Would-Boost-Global-GDP-up-to-13-Trillion

*Why is the consumer price index controversial?* (n.d.). Investopedia. Retrieved November 4, 2021, from https://www.investopedia.com/articles/07/consumerpriceindex.asp

# 

# Appendix

## Appendix A: Data Dictionary

|  |  |  |  |
| --- | --- | --- | --- |
| **Variables** | **Calculation/Units** | **Description** | **Justification** |
| **Dependent Variable** |  | | |
| GDP Growth  (GDP\_Growth) | % of GDP | Real GDP growth measures economic growth, as expressed by gross domestic product (GDP), from one period to another, adjusted for inflation or deflation. | Dependent Variable |
| **Independent Variables used by EIU** |  | | |
| Market Size  (Market\_Size) | Billions | GDP (Gross domestic product (GDP) valued at purchasing power parity in billions of international dollars) | Used in EIU model |
| Life Expectancy  (Life\_Exp) | Number of Years | Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life. |
| Mean\_Years\_of\_Schooling | Number of Years | Average number of completed years of education of a country’s population aged 25 years and older, excluding years spent repeating individual grades. |
| Internet Penetration  (Internet\_Penetration) | % of population | Internet users are individuals who have used the Internet (from any location) in the last 3 months. The Internet can be used via a computer, mobile phone, personal digital assistant, games machine, digital TV etc. |
| Government Indicator  (Gov\_Indicator) | -2.5 to 2.5 | Control of Corruption captures 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. Estimate gives the country's score on the aggregate indicator, in units of a standard normal distribution, i.e. ranging from approximately -2.5 to 2.5. |
| Trade\_Openness | Sum of import and export divided by total GDP | Trade Openness is a measure of the extent to which a country is engaged in the global trading system. Trade openness is usually measured by the ratio between the sum of exports and imports and gross domestic product (GDP). |
| Labour Force Participation Rate  (Labour\_Force) | % of population ages 15+ | The labour force participation rates is calculated as the labour force divided by the total working-age population. The working age population refers to people aged 15 to 64 |
| Domestic\_credit\_to\_private\_sector | % of GDP | It refers to financial resources provided to the private sectorby financial corporations, such as through loans, purchases of non-equity securities, and trade credits and other accounts receivable, that establish a claim for repayment. |
| **New Independent Variables** |  | | |
| Inflation | (new CPI - old CPI)/old CPI | Inflation is the rate of increase in prices over a given period of time. Inflation is typically a broad measure, such as the overall increase in prices or the increase in the cost of living in a country. | Inflation occurs when there is a rise either in consumption, investments, and government spending or net exports in an economy. GDP is the sum of all these factors, and hence an increase in any of them translates to an increase in GDP, vice versa. |
|
|
|
|
|
| Research & Development Expenditure  (Rnd\_exp) | Number of people who engaged in R&D. | R&D personnel in a statistical unit include all persons engaged directly in R&D, whether employed by the statistical unit or external contributors fully integrated into the statistical unit’s R&D activities, as well as those providing direct services for the R&D activities (such as R&D managers, administrators, technicians and clerical staff). | Increasing R&D expenditure means that there would be higher efficiency in the production, which in turn would cause an increase of real GDP growth. |
| Gross National Income  (GNI) | GDP + Money flowing from foreign countries - money flowing to foreign countries | GNI is the sum of value added by all resident producers plus any product taxes (less subsidies) not included in the valuation of output plus net receipts of primary income (compensation of employees and property income) from abroad | GNI includes a nation’s GDP and the income received from its citizens overseas. |
| Balance of Trade  (BoT) | Exports - Imports (% of GDP) | The difference in value between a country's imports and exports. In economics, this difference is an aspect of GDP | An increase in net exports (trade surplus) leads to a higher GDP, and vice versa. |
| Gini Coefficient  (Gini) | 1 - (sum of (Fraction of Income \* (Fraction of Population + 2 \* % of Population that is richer))) | A measure of statistical dispersion intended to represent the income inequality or the wealth inequality within a nation or a social group. | Higher GDP per capita correlates to a higher Gini coefficient. |
| Logistics Performance Index  (LPI) | Index:  Scale of 1 to 5 | The Logistics Performance Index is an interactive benchmarking tool created by the World Bank to help countries identify the challenges and opportunities they face in their performance on trade logistics and what they can do to improve their performance. | An efficient, performing logistic system is a key factor of sustainable economic growth (Havenga, 2010). |
| Child Mortality Rate  (Child\_Mortality) | Number of deaths of children below 5 years of age per 1000 births | Child mortality is the mortality of children under the age of five. The child mortality rate, also under-five mortality rate, refers to the probability of dying between birth and exactly five years of age expressed per 1,000 live births | GDP and Child Mortality are highly negatively correlated. (Amiri et al., 2013) |
| Human Development Index  (HDI) | (Health index x Education Index x Living Standard Index)^1/3 | The HDI is a summary composite measure of a country's average achievements in three basic aspects of human development: health, knowledge and standard of living. | A high HDI shows good economic welfare of a country, which also means that there is higher human power resources, and thus leading to an increase in GDP, vice versa. |
| Environmental Performance Index  (EPI) | Index:  Scale of 0 to 100 | The Environmental Performance Index is a method of quantifying and numerically marking the environmental performance of a state's policies. | There is a strong inverse relationship between  EPI and GDP |
| Global Competitiveness Index  (GCI) | Index:  Scale of 0 to 100 | The Global Competitiveness Index (GCI), a highly comprehensive index, which captures the microeconomic and macroeconomic foundations of national competitiveness. Competitiveness as the set of institutions, policies, and factors that determine the level of productivity of a country. | GCI is one of the key drivers for economic growth, which in turn increases GDP and vice versa. |
| Political Stability  (Pol\_Stability) | Index:  Scale of -2.5 to 2.5 | Propensity of no government collapse either because of conflicts or rampant competition between various political parties. | Political stability may increase investment, which will increase economic growth, and thus leading to an increase in GDP, vice versa. |
| Corruption Perceptions Index  (CPI) | Index:  Scale of 1 to 100 | Corruption Perceptions Index is an indicator of corruption worldwide calculated that classifies countries according to how corrupt the public sector of the country is. | Corruption in a country can deter investments into an economy and thus less corruption can mean more expenditure in the area of investments, boosting the economic health. |
| Urban Population  (Urban\_pop) | Expressed as a percentage ln(UPt/UP0)/n where n is the length of the period and UP is the urban population | Urbanisation refers to the decrease in the proportion of people living in rural areas and shifting to urban areas | At least 80% of global GDP is generated in cities. Hence, urbanisation affects economic growth, which will in turn increase GDP. |
| Electricity generated from Renewable Energy sources  (Renew\_Energy) | TWH | This refers to the sum of absolute electricity production by renewable sources (wind, hydro, solar, other renewables) in TWH | Renewable energy not only has the potential to reduce the negative effects of greenhouse gases, it can also increase economic growth. Increasing the world's shares of renewable energy would greatly increase GDP by $1.3 trillion. |
| CO2 Emissions  (CO2\_emissions) | tonnes | CO2 emissions refers to how much CO2 a country emits each year | At the highest level of significance, the association between GDP and CO2 emissions is statistically supported. (Cederborg et al. 2016) |
| Industry Value Add  (Ind\_val\_add) | % of GDP | Industry corresponds to ISIC divisions 10-45 and includes manufacturing (ISIC divisions 15-37). It comprises value added in mining, manufacturing (also reported as a separate subgroup), construction, electricity, water, and gas. Value added is the net output of a sector after adding up all outputs and subtracting intermediate inputs. It is calculated without making deductions for depreciation of fabricated assets or depletion and degradation of natural resources. The origin of value added is determined by the International Standard Industrial Classification (ISIC), revision 3. Note: For VAB countries, gross value added at factor cost is used as the denominator. | This refers to how much a private or public sector contributes to the overall GDP. Hence, GDP is directly affected by industry value added. |
| Unemployment\_rate | (Unemployed/Total labour force) x 100 | The share of workers in the labor force who do not currently have a job but are actively looking for work | High levels of unemployment usually lead to a decrease in consumption, investment, and net exports in an economy. Since GDP is a sum of these factors, a decrease in any of them leads to a decrease in GDP. |

## 

## Appendix B: Data Extraction and Cleaning

For more detailed explanation and steps taken, refer to the Jupyter Notebook titled:

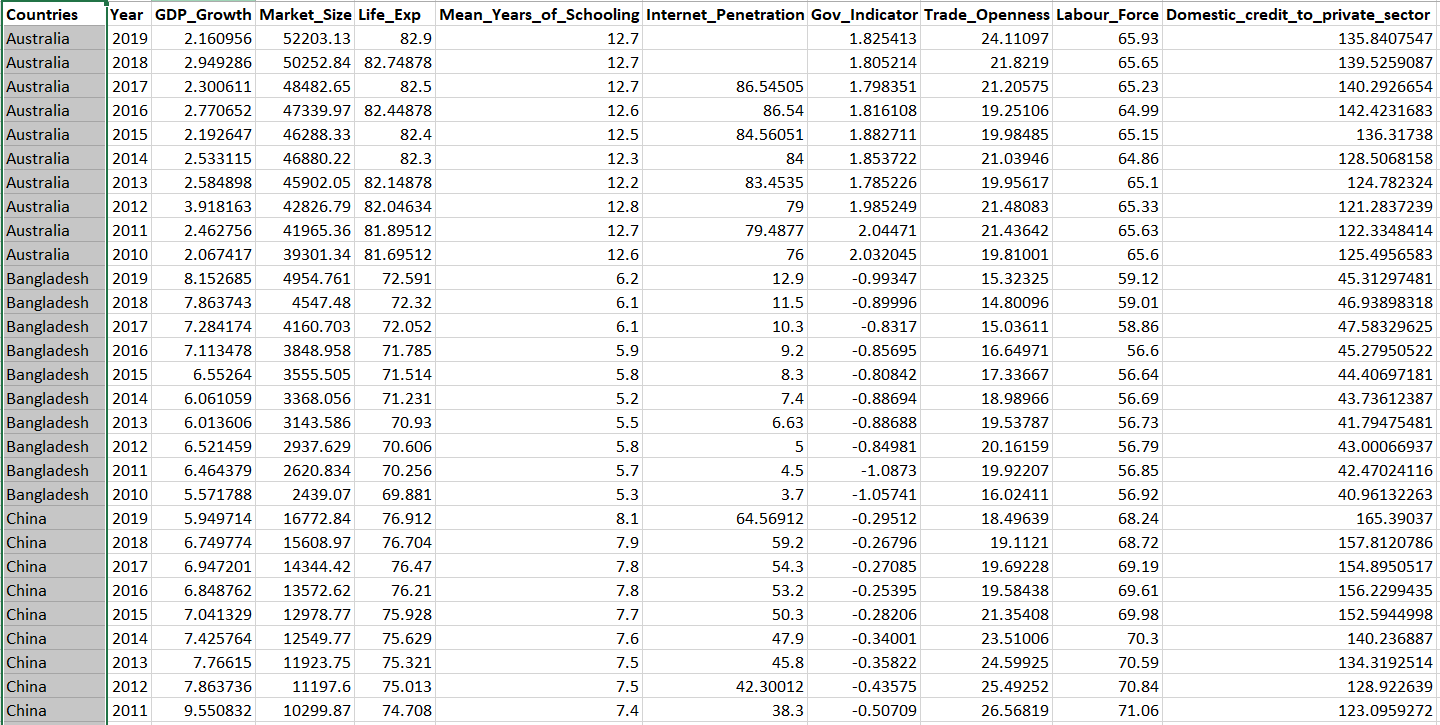
1. Final Data Extraction
2. Final Data Cleaning
3. Feature Scaling

### 4. Data

We chose 20 countries (Figure B1) from EIU list of countries in the Asia continent which are relevant and of interest to EIU to perform predictive analysis with Machine Learning Models. Figure B2 



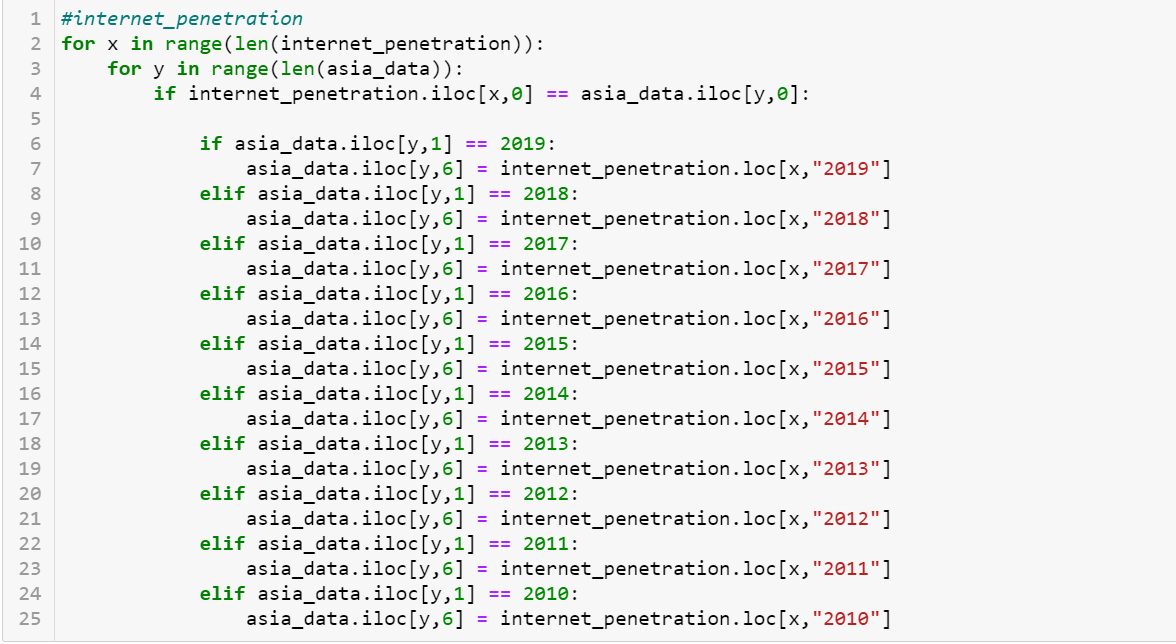
*Figure B1: 20 countries from EIU list*



*Figure B2: Dependent variable and the independent variables considered by EIU*

### 5. Data Extraction

Using pandas dataframe, we read the downloaded CSV files for the various variables from their respective websites and the final spreadsheet into jupyter notebook.

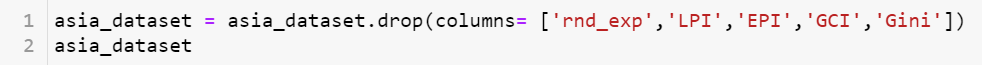


*Figure B3: Extraction process for variables*

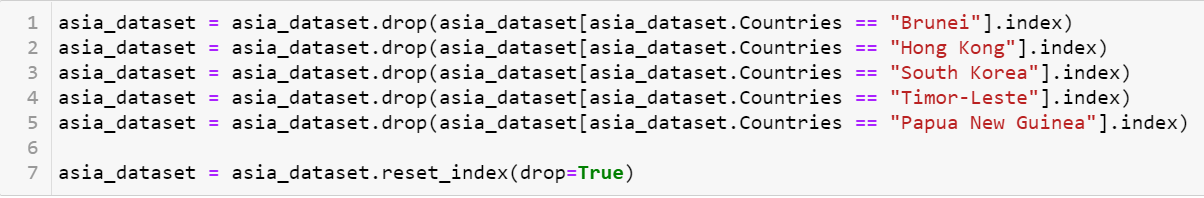
Figure B3 showcases an example of how we extracted data into our final spreadsheet using integer location. We extracted only data that has countries and years equivalent to our final spreadsheet. This process is repeated for all variables.

### 6.1 Data Cleaning

In the data cleaning process, we attempt to remove variables with less than 80% filled (Figure B4) and countries with more than 2 variables with 10 years’ worth of data missing (Figure B5). We also converted datatypes of all our variables to numeric to run our Linear Regression and CART Model (Figure B6).



*Figure B4: Dropping variables with < 80% filled*



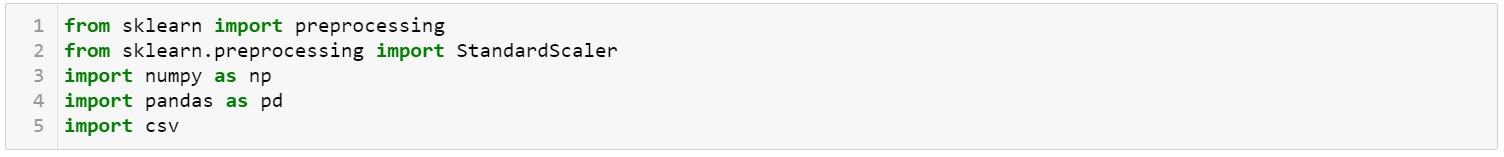
*Figure B5: Dropping countries with more than 2 variables with 10 years’ worth of data missing*



*Figure B6: Converting data types for usage*

### 6.2 Feature Scaling

We perform two feature scaling. Normalization and Standardization using sklearn package for independent variables in our dataset

****

*Figure B7: Packages used for Feature Scaling*

## 

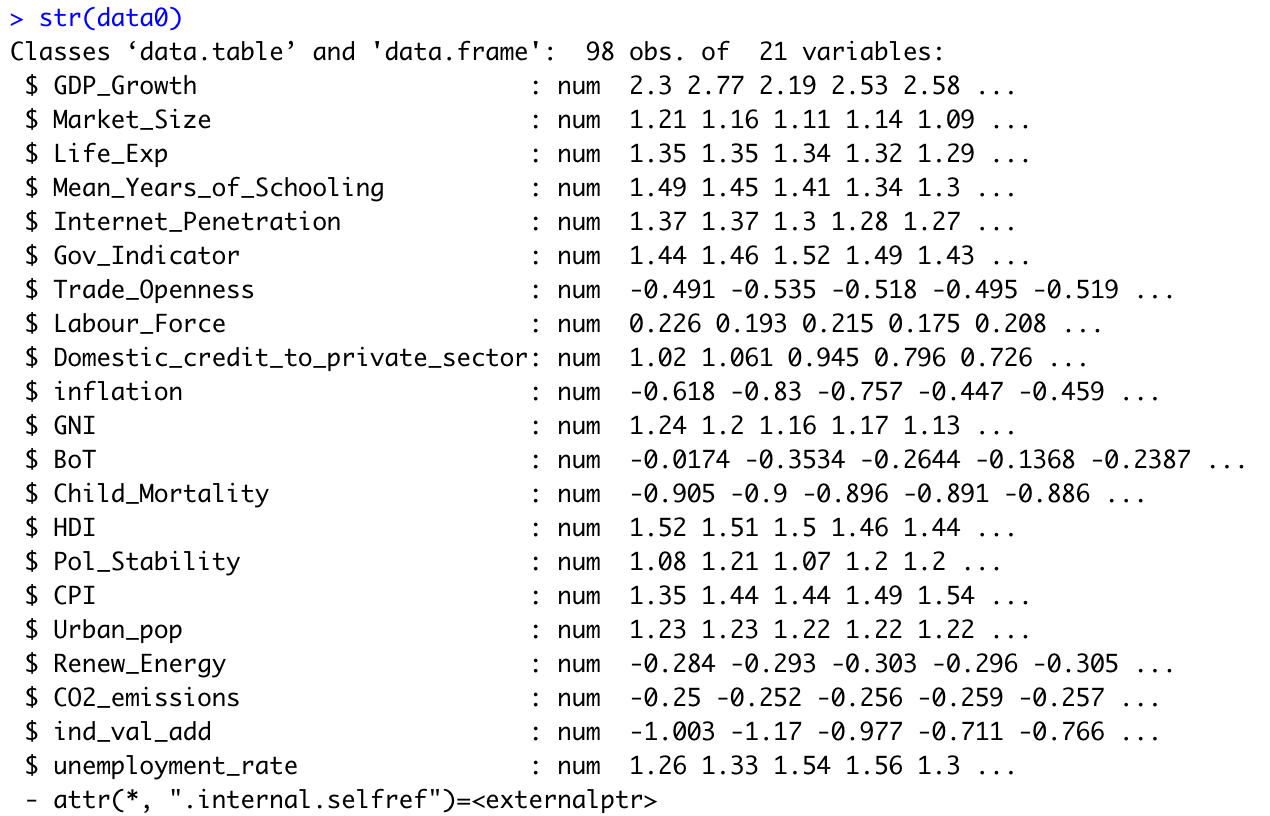
## 

## 

## Appendix C: Linear Regression

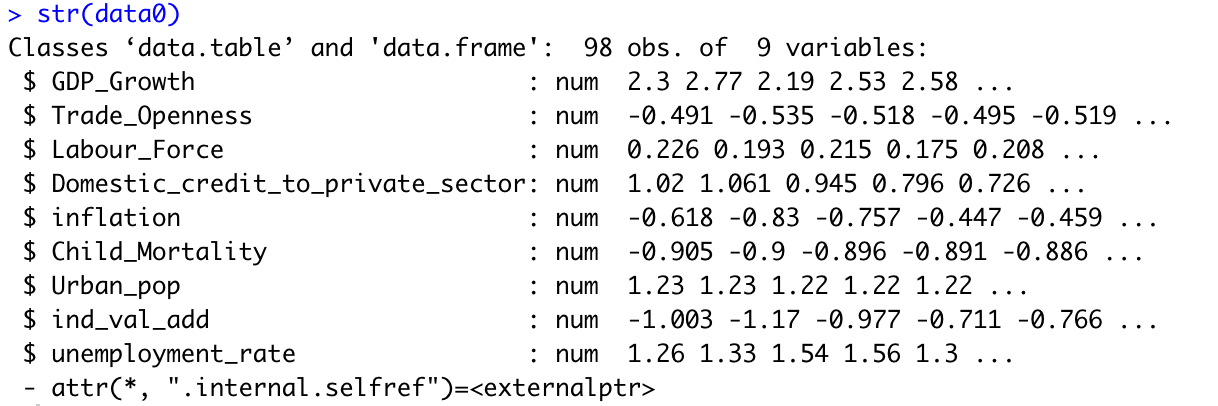
For a more detailed look into steps taken, please refer the R-Script tilted “Linear Regression Model.R”

In Figure C1, the original variables used in the Linear Regression is as shown,



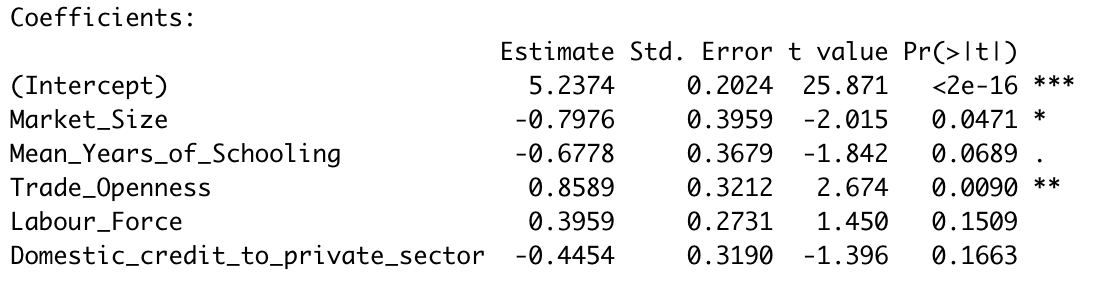
*Figure C1: Original variables in Linear Regression*

In Figure C2, we obtain the following variables in our optimal Linear Regression Model after removing variables that are high in multicollinearity and insignificant.



*Figure C2: Variables in Optimal Linear Regression Model*

In Figure C3, the remaining variables for Linear Regression on EIU baseline Model after removing variables causing multicollinearity and less statistically significant variables are shown:



*Figure C3: Variables in Optimal EIU Baseline Model*

## 

## Appendix D: Classification & Regression Tree

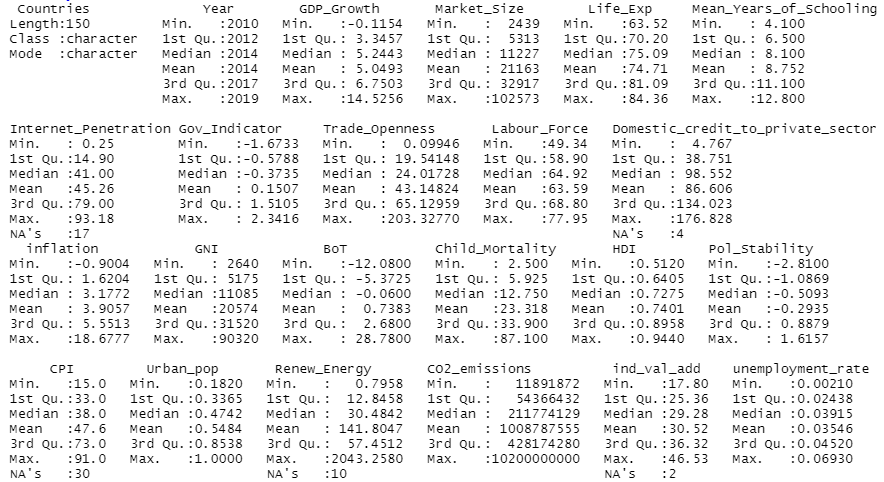
For a more detailed look into the steps taken, please refer to the R-Script tilted “CART model”

### 8.1 Dataset used in CART Model

#### **8.1.2 Dataset with < 11 NAs**

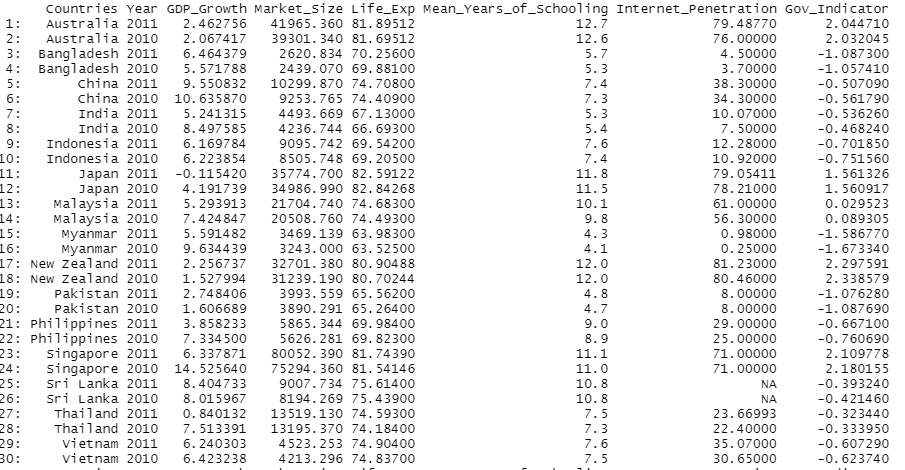
In this dataset, we attempt to reduce the number of NAs per variable to 10 and below.

In Figure D1, using the summary() function, we get an overview of the number of NAs per variable.



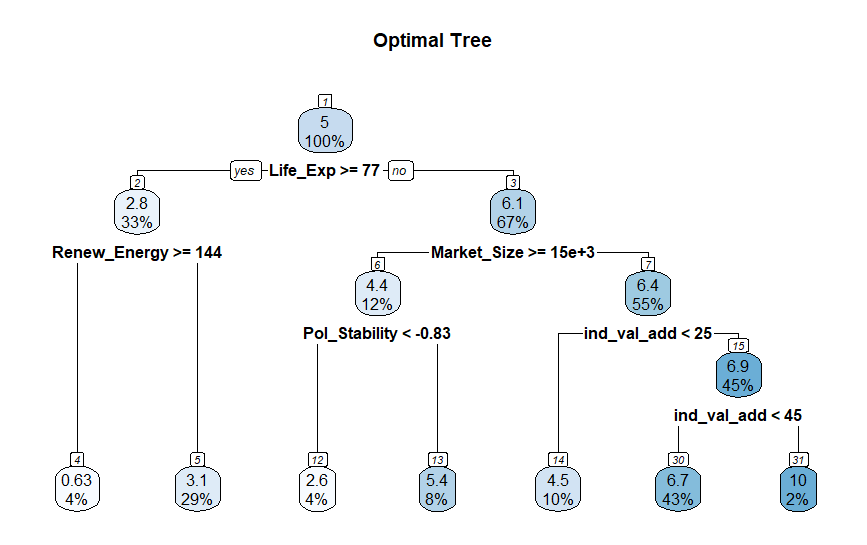
*Figure D1: Summary of variables*

After which, we look into the NAs of variables with more than 10 NAs. For example, in Figure D2, it displays rows with NA values in the CPI column. We will then remove the NAs by column or by rows logically as mentioned in our report. This process is repeated for other variables as well.

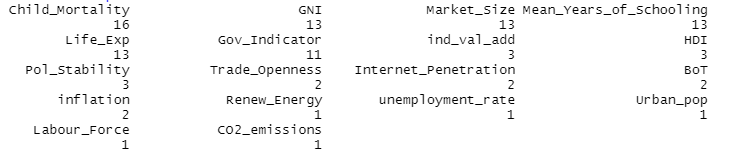


*Figure D2: Data with NA values in CPI column*

The optimal model (Figure D3) obtained from the dataset with less than 11 NAs has 6 nodes, with the variable importance as seen in Figure C4.



*Figure D3. Optimal Tree with 6 branches for dataset containing variables with less than 11 NAs*

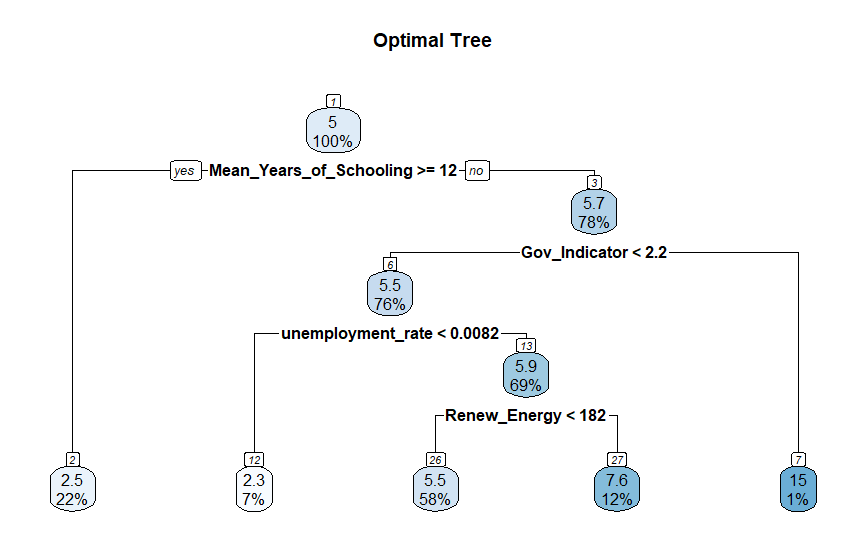


*Figure D4. Variable Importance for dataset containing variables with less than 11 NAs*

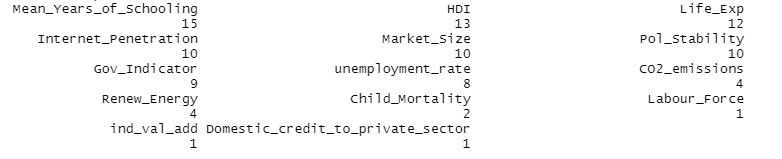
#### **8.1.3 Dataset with no NAs**

In this section, we removed variables with NAs by column and rows accordingly, as mentioned in our report. The summary of variables are as seen in Figure D1.

The optimal model (Figure D5) obtained from the dataset with no NAs has 4 nodes, with the variable importance as seen in Figure D6.



*Figure D5. Optimal Model with 4 branches for dataset with no NAs*

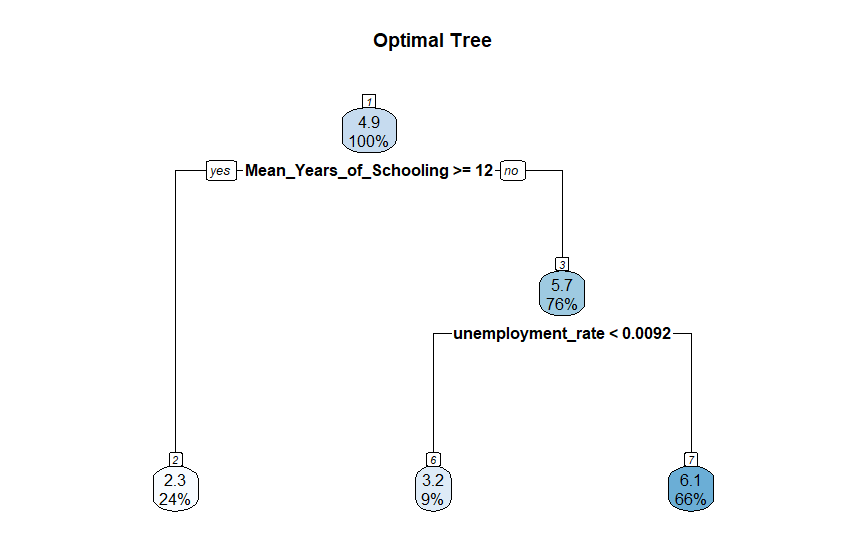


*Figure D6. Variable Importance for dataset with no NAs*

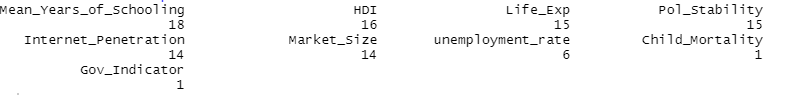
#### **8.1.4 Dataset with Mean in place of NAs**

In this iteration, we replace all NA values with the mean of the variable value specific to the country by using the =AVERAGE() function in excel.

The optimal model (Figure D7) obtained from the dataset replacing NAs with mean values has 2 nodes, with the variable importance as seen in Figure D8.



*Figure D7. Optimal tree with 2 branches for dataset replacing NAs with mean values*



*Figure D8. Variable Importance for dataset replacing NAs with mean values*