PROVIDING TARGETED RESPONSIBLE FOREIGN AID

With

Prof. Kyungsik Han
Department of Data Science

Hanyang University

&

ByungJoo Choi

Data Scientist, Coupang

Contributors:

Jimin Woo – MS in Data Science, Hanyang University

Seunggil Yu – BS in Data Science, Hanyang University

Kushal Navghare – MS in Data Science, DePaul University

Ronaldlee Ejalu – MS in Data Science, DePaul University

DATA SCIENCE SOUTH KOREA/US GLOBAL LAB

Contents

Introduction & Literature Review

What is Foreign Aid?

Although the term 'Aid' is simply defined as a noun: *help, typically of a practical nature* in the dictionary and its meaning is easy to catch, The Foreign Aid is the field which requires a complex knowledge to fully understand its nature. Many articles define foreign aid as where the rich nations transfer capital, goods or services to the poor ones for the benefit of both the country and its population [2] But because a country consists of many elements, like military service, agriculture, medical services, and tons of other things, transferring occurs in various aspects. Although there are many points of views on types of Foreign aid, the author of [2] has divided Foreign Aid into 6 types.

1. Humanitarian foreign aid,
2. Subsistence foreign aid,
3. Military foreign aid,
4. Bribery foreign aid
5. Prestige foreign aid
6. Foreign aid for economic development

In this project, we will focus on Humanitarian foreign aid as other foreign aids can have political meaning which is hard to handle in this data science project.

but even in the Humanitarian foreign aid, this part can be divided into subsections.

1. Disbursements: represent the actual funds that have been paid out or distributed to the recipient countries or organizations as part of the foreign assistance program.
2. Obligations: refer to the commitments made by the U.S. government to provide financial assistance to specific countries or organizations.

in the project we did, we used both of those sections into the modelling.

Why Foreign Aid: Example of South Korea

There might be a question whether Foreign Aid is helping a country's economic development. South Korea is the prime example of a country that has benefited most from

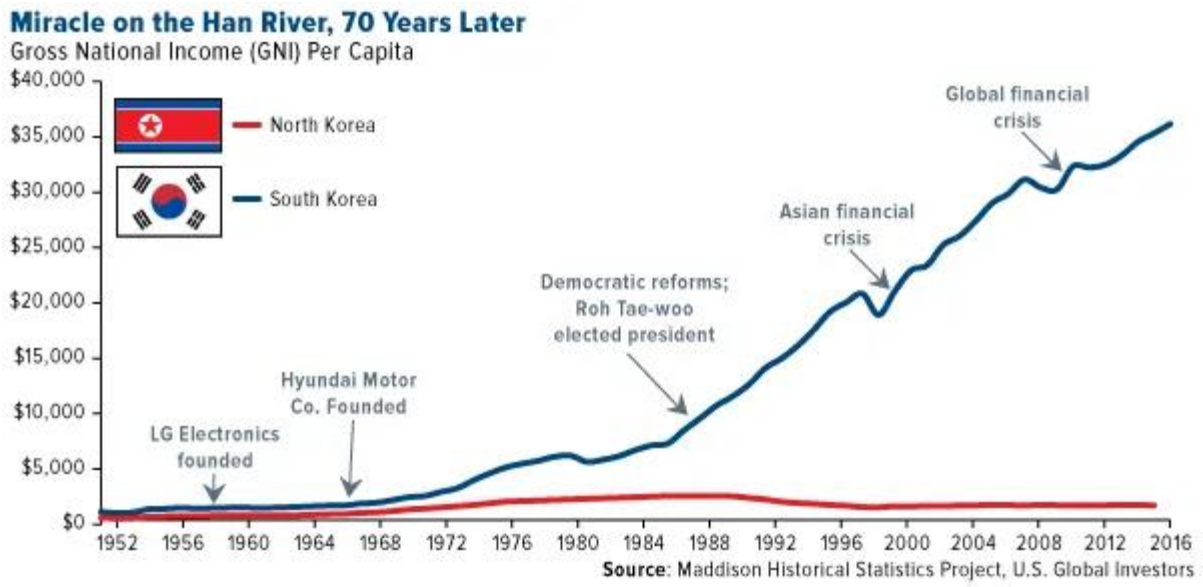
foreign aid. After the armistice of the Korean War, which happened in 1953, South Korea was one of the poorest countries in the world. South Korea's GDP per capita at the year of 1953, was just 67$ which could be calculated as the current currency of 370$ and indicated that this number was the lowest GDP per capita in the world at that time.

The condition for economic development of South Korea was very bad, South Korea was a very small country without any natural resources like oil or gas. Most of the natural resources in the Korean Peninsula, like coal, were distributed in North Korea. Edward Willett Wagner, one of the American professor of Korean studies at Harvard University, stated situation of South Korea at the year of 1961 as:

"Regrettably, the present prospect in Korea is no less bleak than the record of the past. The economic problem is the most glaring. Unemployment is estimated at as high as 25 percent of the labor force. Gross national product in 1960 was less than $2 billion and per capita income well under $100. Electric generating capacity is only one-sixth that of, for example. Mexico and annual output is less than 70 kilowatt hours per capita. The only other source of energy is coal, there being no oil or natural gas. Mineral resources are deficient in several other vital categories. As much as three-fourths of the forest area is either denuded or covered with scrub growth. Exports have averaged a scant $20,000,000 annually as opposed to a volume of imports (exclusive of military items) amounting to $200,000,000 per year. Thus it can easily be seen that there is no possibility of an economic miracle being wrought in South Korea. Judging from the record to date, it will be ***miracle*** *enough if the economy can be made to grow even a little faster than the burgeoning population. All this is not to imply that the U.S. economic aid program in South Korea has been a dismal failure; indeed, it has a number of very real achievements to its credit. As will be mentioned in a moment, the conditions for economic growth are more favorable in the north than in the south. In the long run South Koreans will not choose between Washington and Moscow but between Seoul and Pyongyang." [1]*

The Miracle on the Han River from US aid

Even with the pessimistic perspective about the development of South Korea, US government and private capital kept support South Korea economic development, with continuous aid given from US, The economy of South Korea showed improvement little by little over 20 years and at the beginning of 1970s, South Korea's economic development accelerated its speed, reaching total amount of exports of 10 billion dollars in 1977.



<Figure 1. comparison of GNI per capita growth between North and South Korea>

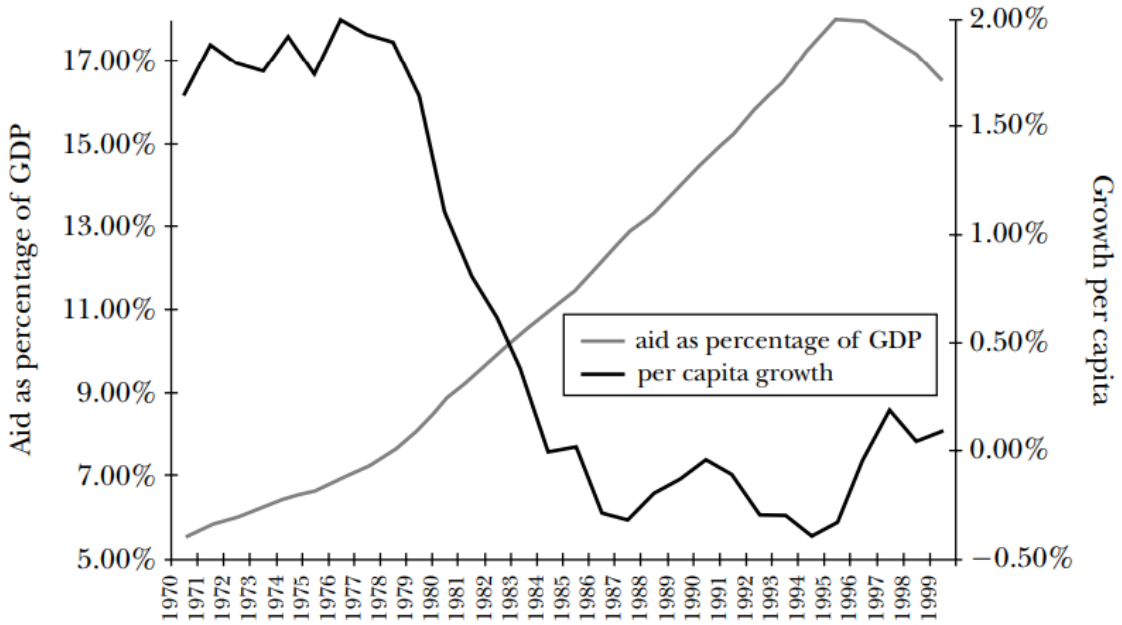
With continuing economic development, South Korea had been designated as the advanced country by UN in 2022. South Korea now has become one of the strong allies of the USA, South Korea sent troops to Iraq in 2003, and South Korea and U.S forces conduct joint exercises every year. and for the private sector, many Korean companies and U.S companies have a beneficial relationship with each other. like Samsung and LG are providing displays for Apple's iphone. the big difference between North Korea and South Korea after Korean war was their ally, USA for South Korea and the Communist party for North Korea, and this difference made a big gap between two Korea.

<Figure 2. The President of USA and Korea >

<Figure 3. U.S. and Korean Force>

Skepticism about Foreign Aid

Although there were marvellous economic development of South Korea from most poorest to the one of the advanced country, many country just stay as poor, even with the lots of Aid given by U.S or other countries. Especially after the 1990s, some country's economic indicators showed that it was worse than it had been in the 1960s.<Fig. 4> shows that growing amount of aid doesn't affect.



<Figure 4. Aid and Growth in Africa [3]>

With many donors of financial aid defining their final objective as "poverty reduction", many donors like U.S. found no positive result out of aid. The author of [3] pointed out that the governments of the recipient countries found little incentive to raise the productive potential of the poor, because it might engender political activism that threatens the current political elite.

Also, there are limitations for observing positive relationships with foreign aid and its economic growth because there are many factors affecting economic growth. Those are the reasons why a pessimistic point of view on Foreign Aid exists.

Providing "Targeted", "Responsible" Foreign Aid

Inspired by the Difficulty that the US faced, our goal is to provide the US embassy useful information of what will be Targeted, Responsible Aid to the recipient country. here, we'd like to define each term as

1. **Targeted**: For which country does aid be helpful?

   As mentioned above, the government of some poor country doesn't want aid to be helpful for their citizens suffering poverty. For those countries, there are other ways to help them, like through NGO or direct private aid. so, we will figure out for which country that the US embassy can help.
2. **Responsible**: The aid should be given in a way that maximizes outcome

   After specifying some countries as a recipient, we will find out what will be the most effective way to maximize Foreign Aid's objective. Here, we will focus on how much aid should be given to the country and forecast its result as an outcome.

We have combined all the results and combined all the information for the US government so that the actual user U.S. embassy is easy to catch the result of our project.

Data Description

Our analysis is divided into supervised learning (GDP prediction using world development indicators) and unsupervised learning (clustering of countries using various features). For each data analysis, we looked for a suitable dataset for each pipeline.

To advance our data analysis on targeted foreign aids, we searched datasets available on the Internet. As a result, we were able to obtain data from the World Bank data and OECD data sites [4]. This data included the World Development Indicators (WDI), which also included country codes and names. For supervised learning, we used this dataset. The data was downloaded in the raw state without filtering for countries and indicators, and filtering was performed at the pre-processing stage before proceeding with modeling after downloading.

We also used the corruption perception index(CPI) and foreign aid data in raw format for topic modelling and clustering [5]

Pre-processing

Year Filtering

The World Data Index and other data commonly had a large number of missing values in the data before 2000. In addition, because cultural and social aspects before 2000 were very different from those after 2000, we used only data after 2000 for our analysis, and filtered all data before that.

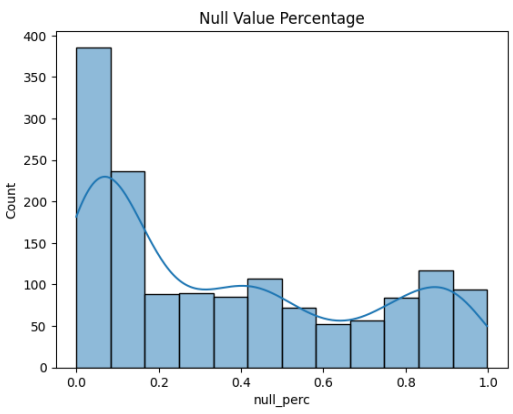
Country Filtering

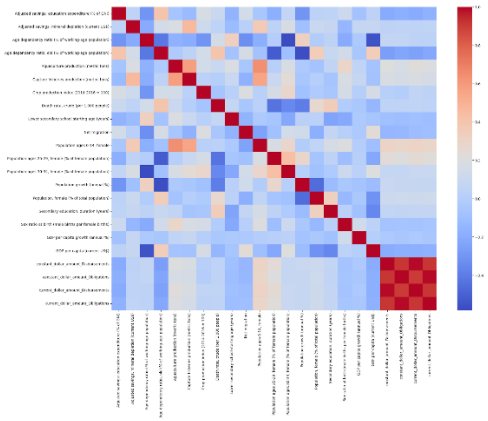
We sought to analyze targeted foreign aid for Latin America and Southeast Asia. The reason for this is that a large amount of foreign aid is already occurring in Africa, where the most foreign aid is currently practiced, and it cannot be well-performed due to corruption. Therefore, only countries belonging to Latin America or Southeast Asia were filtered. In addition, countries with too few populations or very few data collected could be outliers and could have a negative impact on the analysis, so they were also dropped. As a result, the remaining countries were 24 Latin American countries and 10 Southeast Asia countries. Therefore, a total of 34 countries were used for data analysis.

Supervised Learning

World Development Indicators

Since we downloaded the data in raw format, all indicators were downloaded without filtering. There were 1478 indicators in total before filtering. When used in our analysis, too many indicators can negatively affect predictions and clustering, so we proceeded with several strategies to get the right number of features. The first is to remove features with missing value. There have been active discussions about how to fill missing values in some features. It was able to use the interpolation technique and replace missing values with the feature's feature, but due to the characteristics of WDI that can change greatly due to various external factors, the replacement of missing values was determined not to do. For this reason, only features without null values were used. The number of features with and without missing values is indicated in the features below <Figure 5>.





<Figure 5. Null value percentage chart> <Figure 6. Correlation plot for selected features>

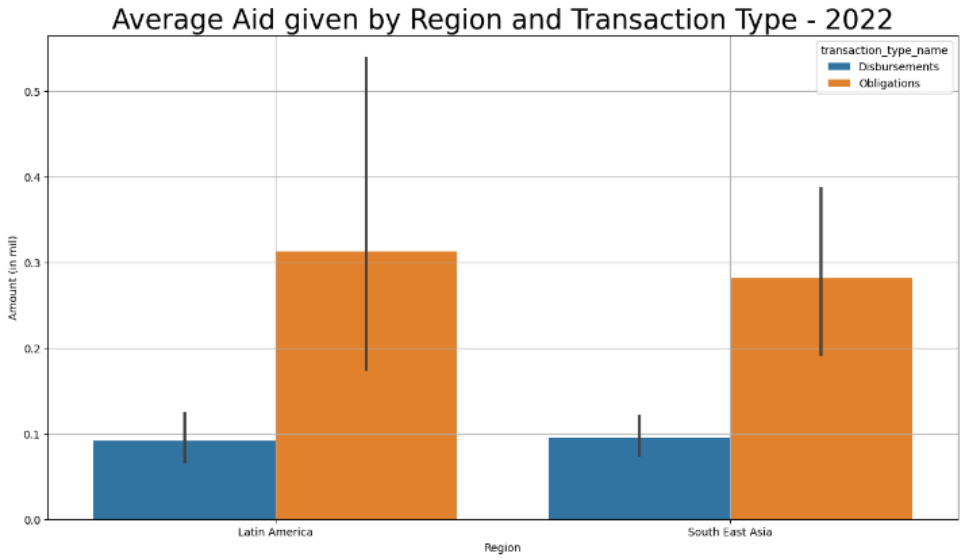
Even after removing all the features that have missing values, there were still too many features left. Multicollinearity can be a major obstacle in building a regression model, so we

proceeded with correlation analysis and removed features with too high correlation. All features with more than 70% correlation were dropped.

The number of remaining features after all handling and filtering was 19. We used these features for supervised learning. Correlation plots of the remaining features are in the following <Figure 6>.

US Foreign Aid Data

There are 2 types of foreign aid given to countries we downloaded. First is Disbursements. It represents the actual funds that have been paid out or distributed to the recipient countries or organizations as part of the foreign assistance program. Another one is Obligations. It is, on the other hand, refer to the commitments made by the U.S. government to provide financial assistance to specific countries or organizations. <Figure 7> is the average aid given by region and transaction type. In both countries, average transactions show similarly.



<Figure 7. Average Aid given by Region and Transaction types >

There were foreign aid data values less than 0, and all such data were dropped. After preprocessing and handling missing data, foreign aid data and filtered world development indicators were merged into a comprehensive dataset for supervised learning and used for modeling. Therefore, the total number of features is 21. The <Table1> below is the Feature used when we did supervised learning.

Population ages 30-34, female (% of female population)	Death rate, crude (per 1,000 people)
Sex ratio at birth (male births per female births)	Crop production index (2014-2016 = 100)
Secondary education, duration (years)	Aquaculture production (metric tons)
Population, female (% of total population)	Aquaculture production (metric tons)
Population growth (annual %)	Adjusted savings: mineral depletion (current US$)
Population ages 30-34, female (% of female population)	Adjusted savings: education expenditure (% of GNI)
Population ages 25-29, female (% of female population)	current_dollar_amount_Disbursements **(Foreign Aid)**
Population ages 0-14, female	constant_dollar_amount_Obligations **(Foreign Aid)**
Net migration	constant_dollar_amount_Disbursements **(Foreign Aid)**
Lower secondary school starting age (years)	current_dollar_amount_Obligations**(Foreign Aid)**
Death rate, crude (per 1,000 people)	

<Table 1. Selected Features in Supervised learning>

Unsupervised Learning

Foreign Aid Description Text Data

For unsupervised learning, we proceeded with topic modelling together with the foreign aid column above. Only data from 2021–2022 were used, and data on foreign aid activity descriptions were used for topic modelling and clustering. All this data consisted of long sentences, and we tried to cluster by extracting key words. In all text analysis, all stop words were removed. There were 53,778 data for Latin America and 17,930 for Southeast Asia.

Corruption Perception Index

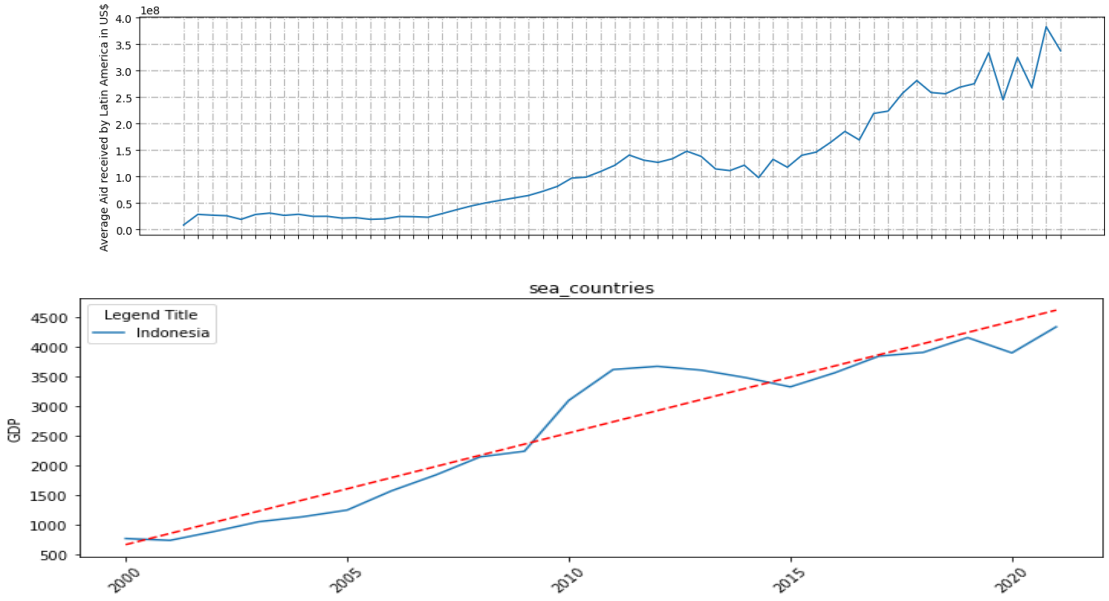
We also tried to advance clustering for corruption index. This also used the GDP equivalent data from the World Development Indicator, and only the data for 2021 were used. Data corresponding to the Corruption Perception index were downloaded from the Transparency International site [6].

Exploratory Data Analysis

In order to gain sufficient insight into the data, we proceeded with a number of EDAs. However, we will introduce some of them in this section that strongly supports our hypothesis.

Trends and Correlation of Foreign Aid and GDP

<Figure 6> below shows the average foreign aid received by Latin America from the United States, and below that is the GDP trend line and its actual value in Southeast Asia. All countries in Latin America and Southeast Asia show that average aid increases over time, and so does GDP. Not all correlations are causations, and it is possible to think that an increase in GDP simply derived from an increase in living standards. But it is certain that foreign aid has improved people's living standards in multiple ways.

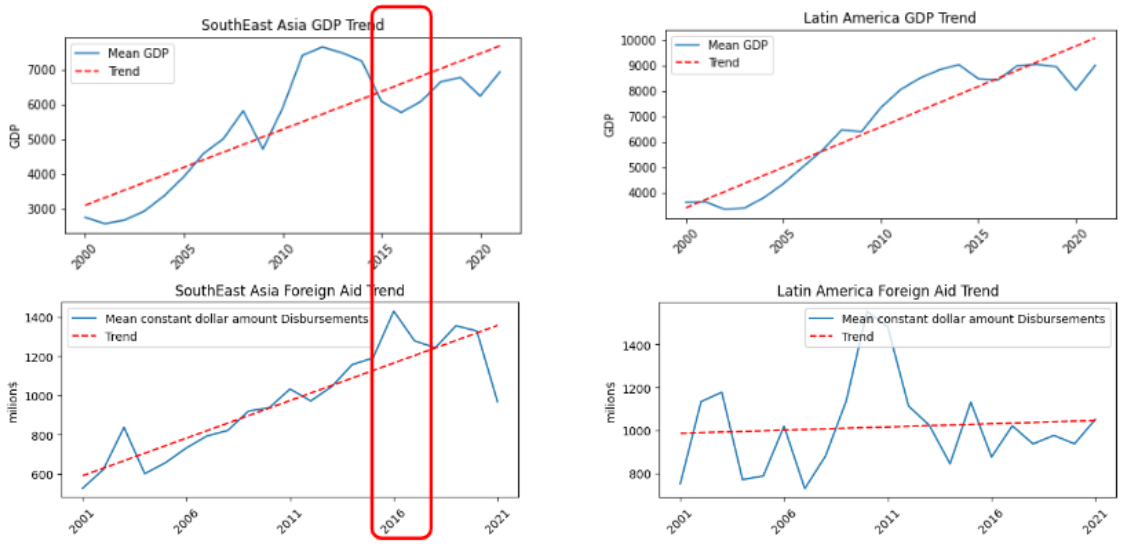


<Figure 8. GDP and Average aid given by USA>

Therefore, we can see that there is a correlation between GDP and foreign countries.

Current status of Foreign Aid and GDP

<Figure 9> below shows GDP and foreign countries from the 2000s to the present. In Southeast Asia, both Foreign aids and GDP are clearly increasing, but in Latin America, the increase in Foreign aids does not show a clear trend compared to the increase in GDP. This point suggests that even if the United States makes a large investment in foreign aid, it does not always provide a large benefit. Sufficient benefits are seen even with small investments and other policies.



<Figure 9. GDP Trend>

Also, looking at the red marked part in <Figure 7>, it seems that the total amount of Foreign Aid increased when the GDP of Southeast Asia decreased. This shows that Southeast Asia was able to effectively emerge from the economic slump through USAid. As such, determining and budgeting current policy relies exclusively on previous data. Therefore, it should be possible to predict next year's budgeting through a more efficient and statistical basis. Our solution is to proceed with predictive modeling on historical data and other diverse data. Ultimately, our approach to Foreign aid is based on two sets of data.

1. **Supervised Learning** - Prediction of GDP using World Development index and US foreign aid data.
2. **Unsupervised Learning** - Clustering countries by using Foreign aid description of US AID dataset and Corruption Perception Index.

Modelling

In this project, we adopted a comprehensive modelling approach that integrated both supervised and unsupervised machine learning techniques. The unsupervised modelling involved clustering countries based on their CPI and GDP using K-means, providing a deeper understanding of the underlying patterns and groupings in the data. For the supervised modelling, we employed a range of techniques, starting with a linear regression model as a baseline to establish hypotheses and then progressing to more advanced models such as decision trees and Random Forest. This iterative approach allowed us to explore the relationships between the input features and the target variable, refine our models, and ultimately develop a predictive model that offers valuable insights and actionable recommendations for decision-makers.

Unsupervised Modeling: Clustering

To gain insights into the relationships between countries based on CPI and GDP, a clustering analysis was performed. K-means clustering was selected due to its ability to group similar data points into clusters based on their feature similarity. The number of clusters was determined using the elbow method, which identified 4 clusters as the optimal choice. Pre-processing techniques such as feature scaling were applied to ensure equal importance to both CPI and GDP during clustering. The results of the clustering analysis revealed distinct clusters, each representing a group of countries with similar CPI and GDP characteristics. The clusters were further analyzed to understand the unique attributes and characteristics of each group.

Clustering Algorithm

K-means clustering was selected as the clustering algorithm due to its simplicity and efficiency in grouping similar data points into clusters based on their feature similarity. The number of clusters was determined using the elbow method, which involved evaluating the within-cluster sum of squares (WCSS) for different values of k and selecting the value of k that resulted in the "elbow" point in the WCSS curve. The optimal number of clusters was found to be 4. The process will use the results of clustering to understand which group of countries should require foreign aid and yield towards a meaningful impact on key indicators.

Cluster Analysis

After clustering, an in-depth analysis of each cluster was conducted to understand the unique attributes and characteristics of the countries within each group. This analysis included examining the average values of CPI and GDP, identifying the countries with the highest and lowest values within each cluster, and exploring additional features to gain insights into the distinguishing factors of each cluster. Additionally, the clustering results were incorporated into the other part of modelling approach to understand and decide which group of countries should be focused on to provide foreign aid based on the performance of the model in these clusters. This will serve as providing a ‘targeted foreign aid’ phase of the study.

Supervised Learning: Regression Modelling

To predict GDP per capita (current $US), a supervised learning approach was employed using various regression models. Initially, a linear regression model was utilized as a baseline model to establish a hypothesis. The model was trained using 16 different world indicators which, in a way, contribute towards the economic development of developing countries. Additionally, the foreign aid was also considered as explanatory variables. The idea here is to have foreign aid as a controllable parameter, which can be controlled by the users to predict GDP. This will help in estimating the budget to be considered towards foreign aid.

The coefficients of the linear regression model were interpreted to understand the relationship between the predictors and the target variable. The model showed promising results as per building the hypothesis that the relationship between GDP, foreign aid and world development indicators can be modelled using the available data. The results were showing most of the explanatory features used were statistically significant to consider the relationship between dependent and independent features. Also, the F-test was a success, giving a

significant p-value for rejecting the null hypothesis. The model's performance served as a benchmark for subsequent advanced modeling techniques.

Advanced Modelling Techniques

To improve upon the baseline model's performance, advanced modeling techniques were explored, including decision trees and random forests. While there was a relationship between the dependent and independent features, we still considered multiple other models to reduce the error rate from the linear regression model. Decision trees provide a non-linear approach to capture complex relationships, while random forests leverage ensemble learning to enhance predictive accuracy. Hyperparameter tuning was conducted to optimize the models, and model evaluation metrics, such as Root Mean Squared Error, were employed to assess their performance. The results were compared against the baseline linear regression model to determine if the advanced models outperformed it in terms of predicting the GDP with lower error rate.

Regression Algorithm: Random Forest

Random Forest is a powerful ensemble learning method widely used in predictive modeling tasks. It combines the strength of multiple decision trees to make accurate predictions. In this project, Random Forest was employed to predict [specific outcome] based on the available features.

Random Forest works by creating a collection of decision trees, where each tree is built using a random subset of the data and a random subset of the features. The randomness in feature selection and data sampling helps to reduce overfitting and increase the model's robustness.

The modeling process using Random Forest consisted of the following steps:

Dataset Split

The dataset was split into a training set and a validation/test set. The training set was used to train the Random Forest model, while the validation/test set was used to evaluate its performance.

Hyperparameter Tuning

Random Forest has several hyperparameters that can be tuned to optimize model performance. These include the number of trees in the forest, the maximum depth of each tree, and the number of features considered for splitting at each node. Techniques such as grid search or random search were employed to find the optimal combination of hyperparameters.

Model Training

The Random Forest model was trained on the training set using the chosen hyperparameters. Each decision tree in the forest was constructed by recursively splitting the data based on the selected features and their respective thresholds. The splitting process continued until a stopping criterion, such as reaching the maximum tree depth or minimum number of samples at a node, was met.

Model Evaluation

The trained Random Forest model was evaluated using the validation/test set. Predictions were made on the validation/test set, and evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R2 Score, etc. were calculated to assess the model's performance. Additionally, metrics like feature importance were examined to identify the most influential features in the prediction process.

Model Interpretation

Random Forest provides a measure of feature importance, which indicates the relative contribution of each feature in the prediction process. This information can be used to gain insights into the factors driving the predicted outcome. Additionally, visualization techniques, such as decision tree visualization or partial dependence plots, can be employed to interpret the individual decision trees' behaviour and understand the model's decision-making process.

Advantages of Random Forest

The Random Forest model offers several advantages, including:

- Robustness to overfitting: The ensemble nature of Random Forest helps to mitigate overfitting by averaging the predictions of multiple trees.

- Non-linearity handling: Random Forest can capture nonlinear relationships between features and the target variable, allowing for more accurate predictions in complex datasets.
- Feature importance: The feature importance provided by Random Forest aids in identifying the most influential features, contributing to better understanding and interpretation of the predictive model.

It is important to note that model selection should be based on the specific characteristics of the dataset, the nature of the problem, and the desired performance metrics. Random Forest was chosen in this project due to its ability to handle complex relationships and provide reliable predictions.

Model Performance

The performance of the Random Forest model was assessed using various evaluation metrics, including Mean Squared Error, Root Mean Squared Error, R2 score, etc. These metrics provide insights into the model's predictive power and its ability to predict the GDP per capita. The performance of the Random Forest model was compared to the baseline linear regression model to gauge its effectiveness in improving prediction.

Below are the results of different models and their performances:

Model Name	**Error (RMSE)**
Linear Regression	4854.85
K-Nearest Neighbours	8464.37
Decision Tree	2404.05
Random Forest	1120.86

<Table 2> Result of different models

Model Interpretability

Model interpretability is a crucial aspect of the modelling process, as it enables us to understand and explain the factors driving the model's predictions. It helps build trust and

confidence in the model's outcomes and allows stakeholders to make informed decisions based on the insights gained. One approach to enhancing model interpretability is through the use of SHAP (SHapley Additive exPlanations) values.

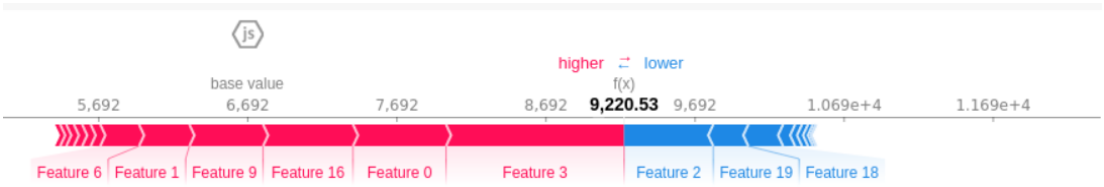
SHAP (SHapley Additive exPlanations) values

SHAP values provide a unified framework for interpreting the impact of individual features on the model's predictions. They are based on Shapley values from cooperative game theory, which assign values to each feature based on its contribution to the prediction. SHAP values provide both global and local interpretability, allowing us to understand the importance of features across the entire dataset and within individual instances.

In the context of the foreign aid, SHAP values can be used to analyze the influence of different features on the predicted GDP per capita. By calculating SHAP values for each feature in the model, we can gain insights into the relative importance and directionality of the features' effects on the predictions.

Local Interpretability

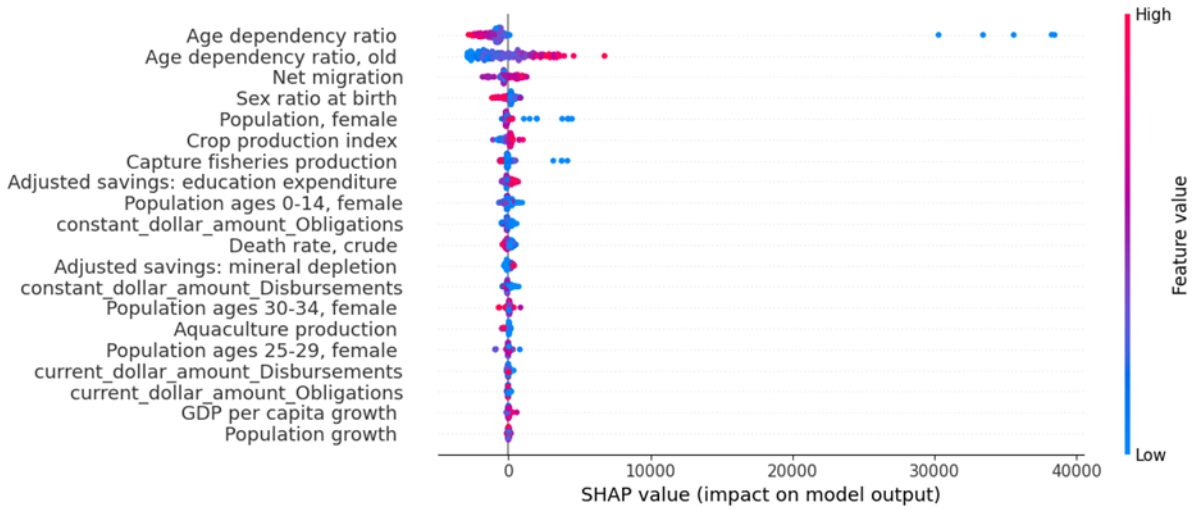
SHAP values also provide insights into the impact of features on individual predictions. By analyzing the SHAP values for a specific instance, we can understand why the model made a particular prediction. Positive SHAP values (shown in red color below) indicate that a feature positively contributes to the prediction, while negative values (shown in blue color below) indicate a negative contribution. By examining the SHAP values for multiple instances, we can identify patterns and understand the decision-making process of the model.



<Figure 10. SHAP value analysis>

Global Interpretability

By examining the average SHAP values across the entire dataset, we can determine the global importance of each feature. Features with higher absolute SHAP values have a stronger influence on the predictions. This information helps us understand which variables are the most critical in determining the GDP per capita.



<Figure 11. SHAP features values analysis>

SHAP conclusion

During the analysis of the Random Forest model, we observed that the Age dependency ratio of the working-age population has a significant impact on the model's predictions. We found that a higher age dependency ratio draws the model predictions away from the expected value of $6692.27 for GDP. This suggests that countries with a higher proportion of dependents relative to the working-age population may experience lower GDP values. This insight highlights the importance of promoting policies and initiatives that address the challenges associated with an aging population to foster economic growth.

Additionally, we found that an increase in the Crop production index has a higher impact on the model's predictions. This indicates that improvements in agriculture production can positively influence the economic condition of countries. Enhancing agricultural practices, investing in technology, and supporting farmers can lead to increased crop production, thereby contributing to overall economic development.

Text Analysis

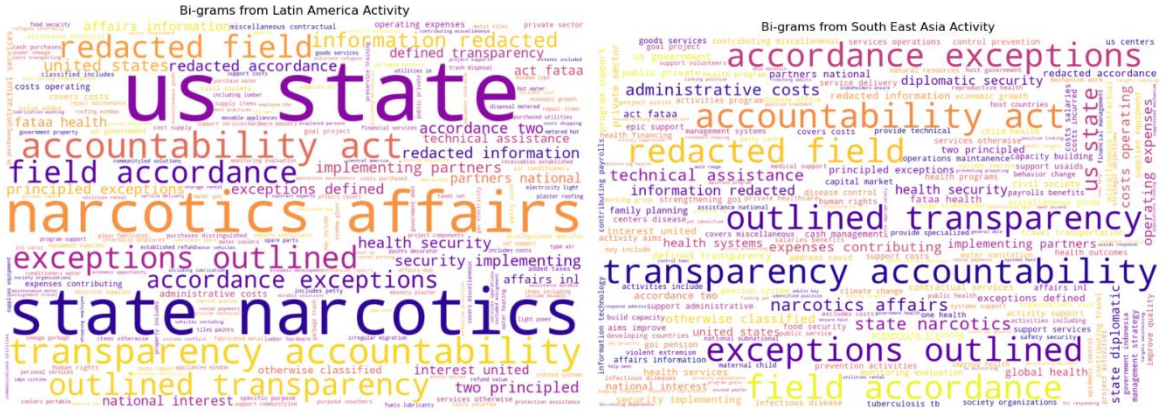
Word cloud analysis and topic modeling are powerful techniques used to gain insights from textual data. In this report, we present the findings and methodology of these approaches as part of our comprehensive analysis.

We found that in the dataset, a field named ‘activity’ was present which had a lot of text information involving various rules and regulations associated with the respective country, aid amount, type of aid, etc. We thought of exploring this variable for different regions using natural language processing techniques.

WordCloud

Word cloud analysis involves visualizing the most frequent words in a corpus. By creating word clouds, we can quickly identify the key themes and patterns present in the text. The size and prominence of each word in the cloud indicate its frequency and importance within the dataset. This technique allows us to identify prominent topics and keywords that can aid in understanding the underlying narrative.

The methodology for word cloud analysis involved pre-processing the text data by removing stopwords, punctuation, and performing lemmatization. We then calculated the term frequencies and visualized the results using a word cloud library. The generated word clouds offered a concise summary of the most important terms and concepts within the corpus.



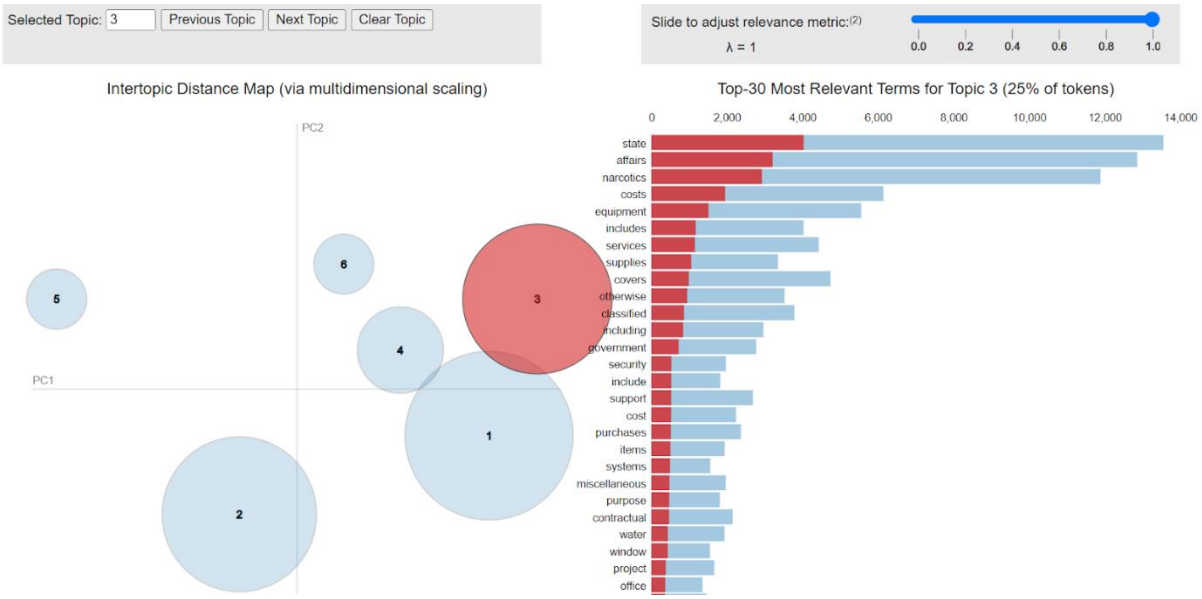
<Figure 13. The result of WordCloud>

<Figure 13> represents the 2 word clouds created for two different regions in focus of this study.

Topic Modelling

Topic modelling, on the other hand, is a statistical modelling technique used to discover latent topics within a collection of documents. It automatically clusters the documents based on the distribution of words and identifies the most prominent topics. In our analysis, we utilized the Latent Dirichlet Allocation (LDA) algorithm for topic modelling. LDA assigns probabilities to each word for each topic, allowing us to identify the most relevant topics in the corpus.

For topic modeling, we utilized the Gensim library in Python. We prepared the text data by tokenizing, creating a bag-of-words representation, and fitting the LDA model. We determined the optimal number of topics using coherence scores and visualized the results using topic-word and document-topic distributions.



<Figure 14. Topic modelling highlight>

Above is the example of topic modelling highlighting the results of topic #3 from 6 topics created using LDA model.

- Topic size: The size of the bubble represents the prevalence of the topic in the corpus. Larger bubbles indicate more dominant topics.
- Term relevance: The relevance of terms to each topic is displayed in the topic-term distribution. Focus on the top-ranked terms, as they provide insights into the main themes of each topic.
- Overlapping bubbles: If bubbles overlap, it suggests that the topics are closely related and share common terms. Explore the overlapping bubbles to understand the connections between topics.
- Coherence score: Some visualizations include a coherence score, which indicates the coherence or interpretability of the topics. Higher coherence scores indicate more coherent topics.

Text Analysis Conclusion

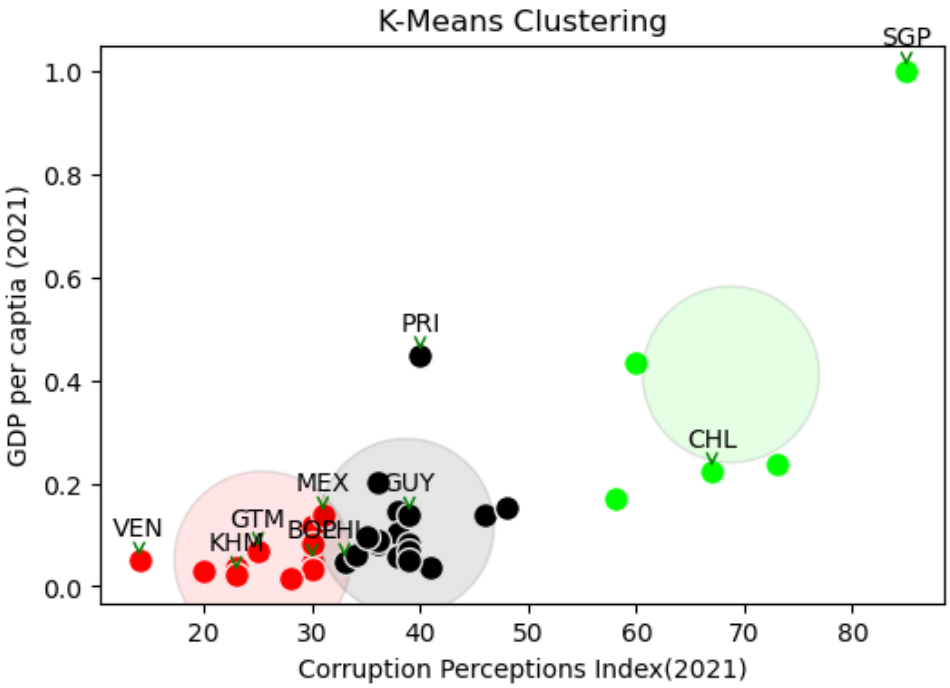
In conclusion, word cloud analysis and topic modelling are effective approaches for gaining insights from textual data. By visualizing word frequencies and identifying latent topics, we can extract meaningful information and enhance our understanding of the dataset. The findings from these techniques contribute to a more comprehensive analysis and pave the way for further investigation in text analysis and related domains. Also, this can serve as part of future work where we can utilize this text data and develop further solutions:

- Group of countries (clusters) which has similar set of activities. This exercise also includes using more complex transformer-based models which understands the context behind words rather that rudimentary bag of words model
- Creating a set of policies for countries, based on the activities carried out previously. This will provide insights on budgeting foreign aid for upcoming year and targeting particular aspect of development in countries.

Results and discussion

The results of this empirical analysis start with figure 10 as shown below:

First, K-means clustering provided the ability to group similar data points into clusters based on their feature similarity. The elbow method identified four clusters revealing distinct clusters where each represented a group of countries with similar CPI and GDP characteristics. Clustering identified countries with the highest and lowest values of both CPI and GDP.



<Figure 15. Result of K-Means Clustering in chart>

Group 0(Red)	Group 1(Black)	Group 2(Green)
'Bolivia', 'Dominican Republic', 'Guatemala', 'Honduras', 'Haiti', 'Cambodia', 'Lao PDR', 'México', 'Myanmar', 'Nicaragua', 'Paraguay', 'Venezuela, RB'	'Argentina', 'Brazil', 'Colombia', 'Cuba', 'Ecuador', 'Guyana', 'Indonesia', 'Malaysia', 'Panama', 'Peru', 'Philippines', 'Puerto Rico', 'El Salvador', 'Suriname', 'Thailand', 'Timor-Leste', 'Vietnam'	'Brunei Darussalam', 'Chile', 'Costa Rica', 'Singapore', 'Uruguay'

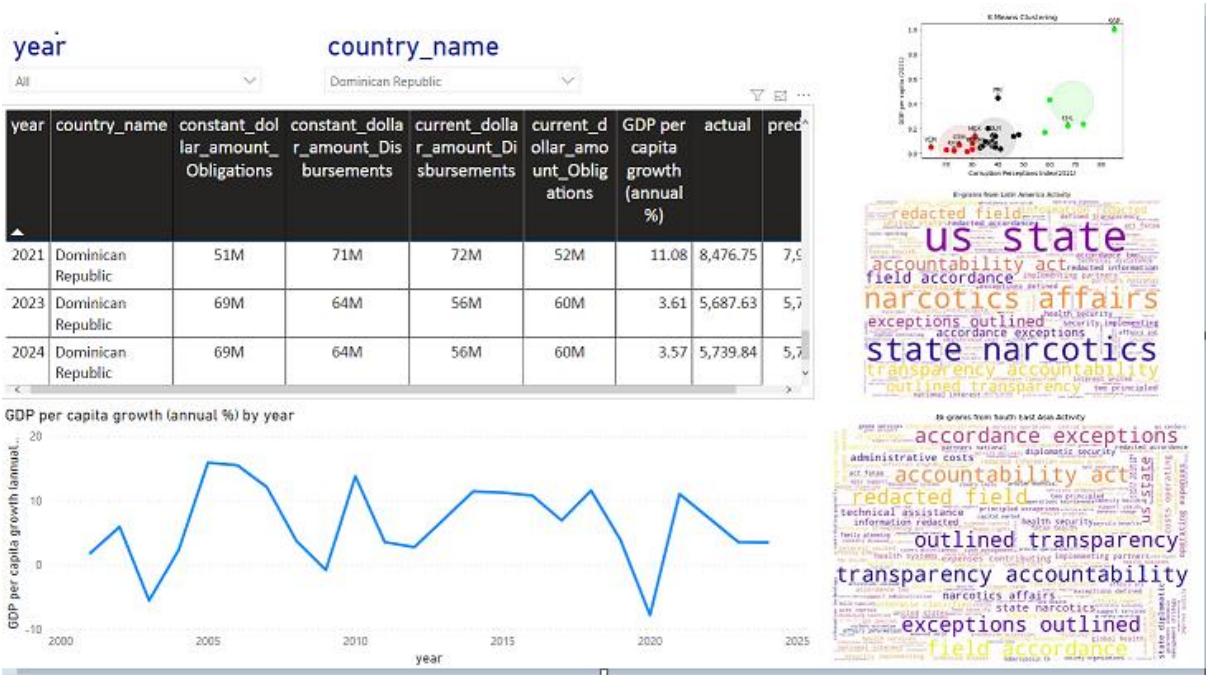
<Table 3, The result of K-mean Clustering>

From the result of clustering, we can intuitively think that there is a linear relationship between GDP per capita and Corruption Perception Index. We can think of this as we have discussed in the Introduction, governments of poor countries don't want their nation to be developed. So using the results of clustering, we can prioritize Group 1 as the first target because although Group 2 has a high CPI, they also have relatively high GDP per capita which means they don’t need aid desperately.

Second, using 16 World indicators a linear regression model was used to establish the hypothesis: whether foreign aid increases GDP per capita and the relationship between the 16 world indicators including foreign aid and GDP. The independent variables used to predict the model were statistically significant. The adjusted R squared value was 0.70 with a high RMSE value of 4854.85 and can be comparable to those attained in other studies. This value means that a set of the 16 worlds indicator variables can explain the variation in the response variable. Some of estimated coefficients are negative meaning that GDP per capita tends to have adverse effects on some of the world development indicators. Our linear regression base model

establishes the relationship where GDP can be modeled as a function of multiple World indicators including Foreign Aid. Third, as shown in <Table 2>.advanced modeling techniques were explored to improve the baseline model's performance. Between the Linear regression, k-Nearest Neighbours, Decision Trees and Random Forest, Random Forest produced the lowest Root Mean Squared Error of 1120.85.

Fourth, SHAP (Shapley Additive explanations) values were used to enable model interpretability by analyzing the influence of the different features on the predicted GDP per capita and the age dependency ratio of the working age population and Crop production index were found to be the most critical variable in determining the GDP per capita. Fifth, through NLP, word cloud analysis and topic modeling produced effective approaches for gaining insights from the textual data where we extract useful information through the visualized word frequencies. Last but not least, a parameterized Power BI dashboard was implemented to allow the different stakeholders at the U.S embassy to derive some insights from the presented data.



<Figure 16> Power BI Dashboard: Select Year and given country.

In order to improve the base model performance, we plan to use the most important features from the Random Forest Model to predict GDP Per capita. Furthermore, we plan to involve the US Embassy for domain knowledge to better understand the process and identify areas where this solution can fit. Also, we intend to continue with topic modeling which includes the identifying areas how the U.S Embassy should go about using the results thus

creating policies for different countries. For any model changes and suggestions, there is a plan to come up with a survey monkey questionnaire where the different stakeholders at the US Embassy can review any model changes or amendments. Lastly, we intend to customize, and improve the dashboard according to the US Embassy requirements, and also use the best analytical practices to avoid discrepancies within the data and modelling part.

References

[1] Edward W. Wagner (1961), "Failure in Korea", Foreign Affairs.
[2] Morgenthau, H. (1962). A political theory of foreign aid. American political science review, 56(2), 301-309.]
[3] William Easterly (2003). Can Foreign Aid Buy Growth?. Journal of Economic Perspectives—Volume 17, Number 3, 23-48.
[4] World Bank, (2023). World Development Indicators | DataBank (worldbank.org)
[5] ForeignAssitance.gov, (2023). https://www.foreignassistance.gov/
[6] Transparency International, (2021). 2021 Corruption Perceptions Index - Explore the… - Transparency.org
[7] Code repository: https://github.com/kushalnavghare/responsible-foreign-aid