ORIE 4741 **Midterm Report** **Fall 2021**

Problem Statment

The goal of this project is to analyze factors that may contribute to a countries GDP and find opportunities for growth. We will look at factors such as trading patterns, infrastructure, education, and other economic and social variables to indicate what variables hold the most potential for growth. This project is important to understanding the dynamics of world development and the results can help a country understand how to better utilize its resources to optimize growth.

Dataset

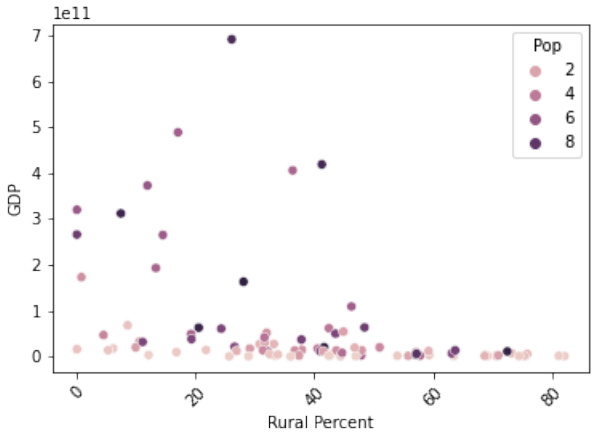
Data is taken from the World Bank World Development Indicators:

https://databank.worldbank.org/source/world-development-indicators

The dataset is a list of countries and statistics on the topics Economic Policy, Education, Environment, Financial Sector, Gender, Health, Infrastructure, Poverty, Private Sector & Trade, Public Sector, Social Protection & Labor, and Health. Examples of these include debt, imports by percent of GDP, health expenditure per capita, and GINI index. There are a total of 1443 of these variables for the years 1960-2020, but there is missing data.

Feature Selection

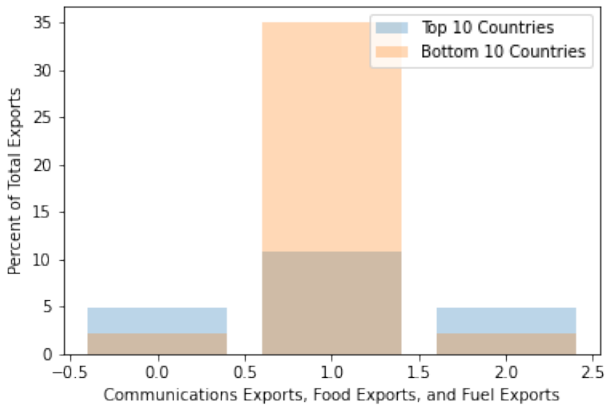
The first step in cleaning the data was to determine how much data was missing for each variable. By going through each variable and computing the percent of countries that had missing data, we were able to determine which variables might be most useful. After analyzing the variables that had the most data available, we looked at variables that are believed to contribute more heavily to a countries growth. Below are some plots that helped us understand the data before running models.



The plot above shows that the relationship between the rural population and GDP changes depending on the population of a country. Countries with high population have a more linear relationship between GDP and the precent of population living in a rural area: GDP is lower when

the rural population percentage is lower. This is not true for countries with lower populations, where the relationship is less linear. When further investigating the relationship between GDP and other variables, we found that variables related to trade seemed to have the strongest relationships. We chose to look at these variables next:

Here, we looked at features related to the type of export a country has. Exports in this dataset



were recorded as a percentage of total exports. The plot above compares a few exports for countries with high and low GDPs. The plot shows that countries with low GDPS tend to have food exports be a large percentage of their merchandise exports, whereas high GDP countries, while still having food exports be a leading percentage, have a smaller difference between percentages.

Preliminary Analyses

We first ran least squares on the two subsets of features. In the first figure, we can see the results of the model on the features that had the most data. The second figure shows the results of the model for the second group of features, which were selected based off of their relevance. In both plot, the predicted GDP was plotted against the actual GDP. The green line is the actual GDP plotted against itself for comparison. The first model does better than the second and there are a few possible explanations. First, the second subset of features was chosen somewhat subjectively, which means they might not actually predict GDP well. Second, the first subset of features were chosen based off of the availability of their data; the second has less data available, and thus, might not make for good predictions.

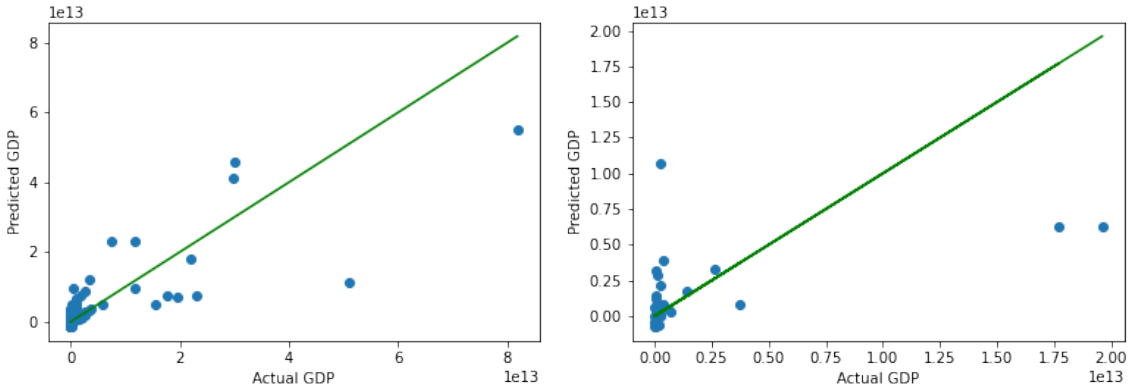


Figure 1: (a) Least squares on feature subset 1 (b) Least squares on feature subset 2

Preventing Over/Underfitting

Because there are a high number of features in our dataset, we chose to look at subsets of variables from different perspectives. First, we looked at a subset of variables where all of the variables had a low percentage of missing data. Second, we looked at only variables that are believed to have a high impact on GDP. Here we focused on variables related to society and the economy, which were grouped as "Financial Sector" or "Economic Policy and Debt" in our dataset. **Choosing a smaller subset of features allowed us to lower the complexity of our model and avoid overfitting**.

Our dataset includes variables from many different categories, which allowed us to choose variables that were unrelated and diverse. **To prevent underfitting, we ensured that the variables chosen represented many different categories. Doing this allowed us to prevent having multiple variables that are highly correlated with each other.**

Future Work

In the future, we plan on adding more features to our models, choosing our features more deliberately, and conducting more complex analyses. Feature selection is one of the larger issues surrounding our project, as we have a large number of features that span many different topics. We plan to make feature engineering and feature selection a larger part of our project as the next step and integrate methods that select for the most useful features.

Another step is conducting more complex analyses. We would like the tailor our methods to our data and problem a little more closely in the upcoming weeks and hope this step will be guided by our feature selection.