EPPS 6323 Progress Report

Aldex Felix, Wei-chen Huang, Jim Pan, Samuel Adelusi

April 18, 2023

Data Collection:

For our event data, we utilized the ICEWS coded event dataset. For economic and social data, we used the World Bank and IMF as our source. We scraped other variables like freedom scores from reputable sources. We wrangled and merged all our data into a single dataframe which we use to build our machine learning model.

Variables and Model Generation:

As of right now, we have a fairly exhaustive list of socioeconomic variables that we are feeding into a random forest model to predict affinity between the U.S. and other states in the dataset. These include: GDP per capita, GDP per capita growth, health expenditure per capita, military expenditure, proportion of parliament seats held by women, infant mortality rate, freedom scores, deaths due to violent conflict, and literacy rate, among others. Currently there are 26 variables. As we refine the random forest model, we will prune out variables that did not have a strong theoretical reason to be included to begin with and are not contributing very much to the model.

Work in Progress:

We are looking into ways that we can enhance our current dataset, and in turn our model. The primary issue to solve is the amount of countries that had to be removed from the dataset because of a lack of complete data. We are trying to either impute the missing data, or gather the data from other sources. Regaining these countries will increase our model's robustness and make our predictions more relevant. Once we are satisfied with the random forest model we will try a gradient boosted machine model if time permits.

A link to the Github repository with the project files can be found below:

https://github.com/aldenfelix/Predicting-Interstate-Affinity