Final Project Report

Maryam Khalid Shah

12/17/2020

Introduction

The aim of this project is to understand what democratic factors matter the most when it comes to predicting an increase in educational spending in UN member countries. This report explains the sequential steps taken to determine this. As such, it offers a background of the analysis, specifies data sources used as well as data collection and cleaning methods employed, and also describes the methods and tools used to derive insights from the data. The results obtained are then discussed, followed by a concluding discussion about whether the original aim of the project was successfully achieved.

Problem Statement and Background

Problem Statement: The effect of specific democratic institutions and processes on an increase in government expenditure on education.

Education spending is defined as total government expenditure on education as a percentage of Gross Domestic Product (GDP). This analysis will utilize 15 years of data from 2003 to 2017, and would be limited to UN member countries.

Background

As of 2020, more than half of the world's countries are democracies (of which 164 are UN member states). While there has been a lot of research on how education affects democratization, there has been comparatively little research devoted to understanding how democratic institutions affect spending on education. Looking into the effect of specific democratic processes is important because no two democratic countries can be said to have exactly the same democratic institutions, as they may vary on dimensions such as access to power, the distribution of resources, and whether clean elections are held. Moreover, even if a few democratic countries are fairly similar in a few dimensions, the overall mix of democracy defining characteristics may be quite different.

Research supports that in general, greater educational spending improves student outcomes, which in turn can affect employment rates, economic growth and scientific research in a country. Therefore, understanding how democratic institutions affect education spending is important in order to highlight factors that drive greater spending on education. This is of even greater importance in less developed and developing democratic countries where illiteracy and out of school children rates are high - what factor of their democratic processes is limiting their educational expenditure? Questions like these would be interesting to answer because they would provide us insight into the link between democratic processes and educational spending, whether certain predictors can be modified to improve educational spending in a country, and why there are differences in educational spending across democratic countries.

Literature Review

A similar project focuses on the important ways in which the relationship between democracy and education may be influenced by the African context. It puts forth theoretical arguments for why democracy may influence spending on education, focusing specifically on the political incentives generated by multiparty electoral competition (Harding 2019). It also presents related but distinct arguments that focus on how this in turn may affect education outcomes, and on why these dynamics may vary because of factors that are particularly pertinent in many African countries. These factors include variations in the degree of electoral competitiveness and political competition as well as in levels of economic development and ethnic fractionalization (Harding 2019). The paper finds that although evidence suggests that democracy may positively impact access to education, there is fewer evidence for its impact on educational quality and suggests that future work should continue to address these issues while seeking to investigate sources of heterogeneity in the impact of democracy on education in Africa (Harding 2019).

Data

Data Sources

For the purpose of this project, the following 3 data sources are utilized:

1. **Varieties of Democracy (V-Dem) data**: The V-Dem is a multidimensional and disaggregated dataset that is a new approach to conceptualizing democracy. This dataset was used to obtain data on democratic processes and institutions.

2. **World Bank data on government expenditure on education, total (% of GDP)**: The World Bank data, considered to be the most comprehensive and reliable source of developmental and economic metrics, was used to obtain data about educational spending in countries across the 15 years being considered.

3. **United Nations website**: The United Nations website was used to obtain names of member countries.

The unit of observation for this analysis was country-year and the variables of interest were High-Level and Low-Level democracy indices. All code for this analysis was written using the Python programming language (Van Rossum et al. 1995).

Data Wrangling

The names of UN member countries were scraped from the UN website using the packages BeautifulSoup (Richardson 2007) and requests (Chandra & Varanasi 2015). They were then standardized using the country converter package (Stadler 2017). The World Bank Data was obtained by scraping IndexMundi (Index Mundi), a website with World Bank educational spending data. In order to check data accuracy, the World Bank dataset was also downloaded and compared with the scraped data and no discrepancies were found. Therefore, the scraped dataset was used for analysis. The V-Dem dataset on the other hand was directly downloaded from the V-Dem website (V-Dem).

Potential Issues

It should be noted that the World Bank data (UNESCO 2020) for government expenditure on GDP, total (% of GDP) contains a lot of missing values, even for developed countries such as the United States, which may limit the effectiveness of the data analysis and the usefulness of the insights derived from the results. The V-Dem dataset on the the other hand, contained no missing data, as checked using the package missingno (Bilogur 2018).

Analysis

Once two datasets were formed (one containing educational expenditure figures and the other V-Dem indicators), the datasets were merged such that only those country years for which educational data was available were kept. The dataset was then limited to include only years from 2003 to 2017 i.e. the time period under consideration.

An examination of the educational expenditure data showed a few outliers (observations that are significantly different from the other observations). In order to avoid these data points from biasing results, the dataset was trimmed so as to include only observations with educational expenditure less than or equal to 10%. Since these outliers were few in number and since there is a lot of data, removing them was considered an acceptable option.

Moreover, since there are likely to be cultural and other regional factors that relate to educational expenditure, a continent variable was added to the data in order to control for these factors. For this purpose, continent names for each country in the dataset were created using the py-country package (Stadler 2017). This variable was converted into a categorical type and dummy variables were created for the 6 continents (Asia, Africa, Europe, North America, South America and Oceania), with the last three being kept as the baseline owing to relatively low observations.

The data types of the variables were checked and using hierarchical indexing, the dataset was organized according to country and year. For this high-level data processing, and further modifications of datasets, the packages pandas (McKinney et al. 2010) and numpy (Oliphant 2006) were used.

Further analysis is divided into two parts: High-Level Democracy Indices and Low-Level Democracy Indices.

In both these parts, the merged dataset was used. A new variable was then created in order to calculate the difference between education expenditure in a country by year. The missing values in this 'difference' variable, which indicated the first year available in the dataset for a given country, were replaced with zeroes along with values that were less than 0 (since a negative value denotes a decrease in education expenditure). On the other hand, positive 'difference' values were coded as 1, creating a dichotomous outcome variable. Thereafter, the dataset was split into a training and test dataset in a 3:1 ratio respectively. This is because when machine learning models on the data are run, they need to be tested on data that they have not been trained on (what is referred to as the test dataset). Moreover, a seed of 202011 was set while splitting the data and while setting folds for cross validation in order to generate the same numbers every time the code was executed, and to improve reproducibility of the code. Moreover, the continent variable was kept as a control variable in both the machine learning models.

Both machine learning models were trained on data with a dichotomous outcome. Multiple models, including a Naive Bayes classifier, K Nearest Neighbors (KNN), Decision Tree and Random Forest were used in order to determine the one that performed the best. That model was then tested on test data, and insights were drawn from the results. The same tuning parameters were kept in all machine learning models and a machine learning pipeline was built that pre-processed the data, used K-fold cross validation with 5 folds and incorporated all the models. For machine learning methods and model interpretation, the package sklearn (Pedregosa et al. 2011) was used, whereas for the graphs created, ggplot2 (Wickham 2016), matplotlib (Hunter 2007) and seaborn (Michael et al. 2017) were utlilized.

Part 1: High-Level Democracy Indices

For the first part of the analysis, the aim was to explore what high-level indices best predict an increase in education spending. Before beginning the analysis, the variables were renamed in order to improve ease of readability. A check for missingness revealed a few missing values in two variables which were replaced with the average values of those variables. The predictors for the first machine learning model were only (five) High-Level Democracy indices, as described below:

Title	Explanation
Electoral Democracy Index	To what extent is the ideal of electoral democracy in its fullest sense achieved?
Liberal Democracy Index	To what extent is the idea of liberal democracy achieved?
Participatory Democracy Index	To what extent is the idea of participatory democracy achieved?
Deliberative Democracy Index	To what extent is the idea of deliberative democracy achieved?
Egalitarian Democracy Index	To what extent is the idea of egalitarian democracy achieved?

Low-Level indices were not included in this model since including them would have resulted in high multicollinearity. The best model was a Random Forest classifier with four features per tree, a depth of 8, and 1000 trees grown. This model had an accuracy of 81% on the training data, and 60% on the test data.

Part 2: Low-Level Democracy Indices Since High-Level Democracy Indices are not that nuanced, in the second part of the analysis the aim was to explore which Low-Level Democracy Indices best predict an increase in educational spending. Except for deliberative democracy, each of the four High-Level Democracy indices is composed of Low-Level indices (14 in total). These indices were included as predictors in the second machine learning model and only the maximum features of the Random Forest Classifier were set differently than in the first machine learning model.

The best model was a Random Forest classifier with four features per tree, a depth of 8, and 1500 trees grown. This model had an accuracy of 85% on the training data, and 56% on the test data.

Results

In this section, the results obtained by running each of the two machine learning models would be discussed, starting with the first one.

First Model: High-Level Democracy Indices

In order to determine which variables were the most important i.e. which variables the model relies on the most when making predictions, the features were permuted 30 times. The following results were obtained:

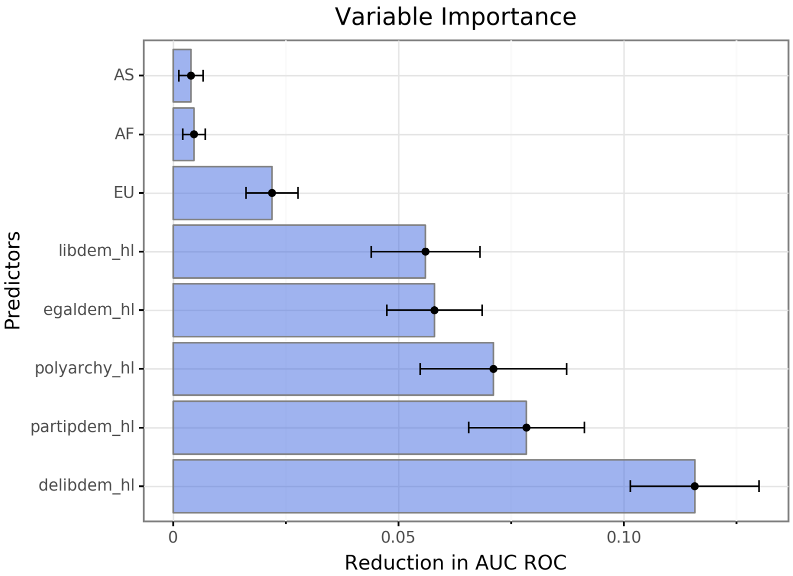

Figure 1: First Model: Variable Importance

As can be seen in Figure 1, among the 5 High-Level Democracy Indices, the most important index for predicting an increase in educational expenditure was the Deliberative Democracy Index, followed by Participatory Democracy and Electoral Democracy Indices.

The index of Deliberative Democracy measures the extent to which the ideal of deliberative democracy is achieved, focusing on the processes by which decisions are reached in a polity (V-Dem Codebook). A centered Individual Conditional Expectation (ICE) plot created using the package pdpbox (Brandon 2017) showed evidence of heterogeneity and interaction. Moreover, partial dependence plots using sklearn (Pedregosa et al. 2011) were created to show the impact of the predictors on the outcome variable. The partial dependency plot for the Deliberative Democracy Index showed that there is no clear direction of the index's impact i.e. at really low as well as really high levels of deliberative democracy, the effect of an increase of the index on the probability of an increase in educational spending was the highest and approximately the same i.e. 45% whereas in between these levels there was a wide variation. Furthermore, an interaction plot between the deliberative democracy index and participatory democracy index created using the pdpbox package (Brandon 2017) showed that at high levels of deliberative democracy, high values of participatory democracy improve the probability of an increase in education expenditure.

Second Model: Low-Level Democracy Indices

In order to determine which variables were the most important i.e. which variables the model relies on the most when making predictions, the features were permuted 30 times. Among the 14 Low-level Indices, the most important was the civil society participation index, followed by the equal distribution of resources index. The civil society participation index measures democratic processes such as whether major civil society organizations (CSOs) are routinely consulted by policymakers and how large the involvement of people in CSOs is (V-Dem Codebook).

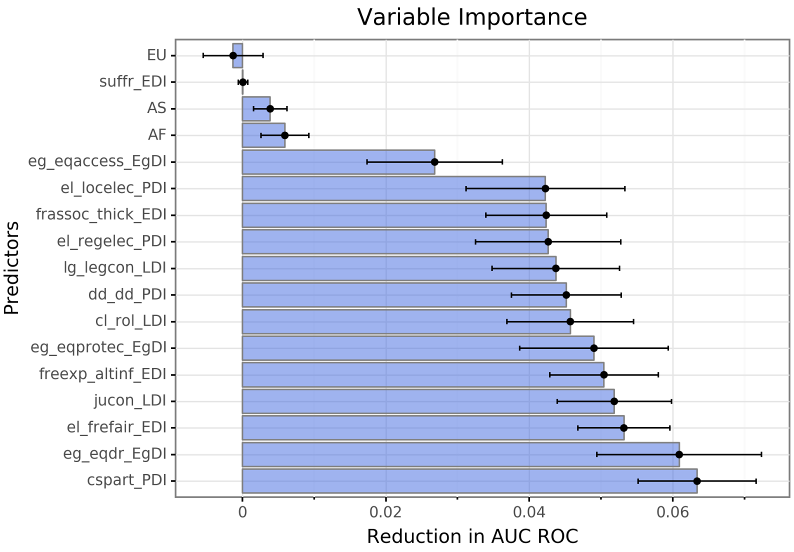

Figure 2: Second Model: Variable Importance

A look at the PDP plots for the two most important indices showed that as the level of civil society participation index increases till about 0.6, the probability of an increase in educational spending decreases. After 0.6 however, there is an increase in the probability of a rise in educational spending as the level of civil society participation increases. The effect of the equal distribution of resources index on the probability of an increase in educational spending is more or less negative till around 0.85, after which it becomes positive i.e. as the level of the equal distribution index increases, the probability of an increase in educational spending increases. Moreover, a centered Individual Conditional Expectation (ICE) plot created using the package pdpbox (Brandon 2017) revealed little heterogeneity, whereas an interaction plot between the two most important indices showed that at high levels of civil society participation, low levels of the equal distribution of resources index improve the probability of an increase in educational expenditure.

Discussion

For this project, I defined success in two stages. The first stage was to obtain data and create a comprehensive, clean dataset that was ready for exploration and analysis. A major component of this stage was web scraping which was used to obtain data on the outcome and a list of UN member countries. Success in the second stage was defined as deriving interesting insights from the data – specifically finding a statistically significant effect of change in the state of democracy on a country's education budget.

In the first stage, success was achieved as the required data was successfully scraped from the web and cleaned. This was enabled by the use of specific packages as mentioned above.

Success in this stage resulted in a comprehensive merged dataset containing both indicators and outcome data that was used for analysis. In the second stage, success was also achieved as machine learning models were run and various insights were drawn from the data. However, the accuracy of the best models on test data was low. Therefore, success in terms of finding models with high accuracy on test data was not achieved, possibly because of a lack of control variables such as the level of economic development.

If given more time, I would expand my analysis by considering change in education expenditure as the outcome, while using the same predictors. This would allow me to understand what democratic factors and processes matter the most when it comes to predicting a change in educational spending in UN member countries. Moreover, I would like to look more closely at the High-Level Democracy Indices, running machine learning models with each High-Level index as the outcome and its Low-Level indices as predictors in order to understand what factors are driving the High-Level indices. Finally, I would be interested to look at more of the indicators present in the V-Dem dataset as predictors of an increase or change in education expenditure.

Word Count (including in-text citations & headings): [3143]

Works Cited

Harding, R. (2019, March 26). The Democratic Dividend: Public Spending and Education Under Multipartyism. Oxford Research Encyclopedia of Politics. Retrieved 17 Dec. 2020, from https://oxfordre.com/politics/view/10.1093/acrefore/9780190228637.001.0001/acrefore-9780190228637-e-759.

Richardson, L. (2007). Beautiful soup documentation. April.

Van Rossum, G., & Drake Jr, F. L. (1995). Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam.

Stadler, K. (2017). The country converter coco - a Python package for converting country names between different classification schemes. The Journal of Open Source Software. doi: 10.21105/joss.00332

Chandra, R. V., & Varanasi, B. S. (2015). Python requests essentials. Packt Publishing Ltd.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90–95.

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.

Michael Waskom, Olga Botvinnik, Drew O’Kane, Paul Hobson, Saulius Lukauskas, David C

Gemperline, … Adel Qalieh. (2017, September 3). mwaskom/seaborn: v0.8.1 (September 2017) (Version v0.8.1). Zenodo. http://doi.org/10.5281/zenodo.883859

McKinney, W., & others. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56).

Oliphant, T. E. (2006). A guide to NumPy (Vol. 1). Trelgol Publishing USA.

Bilogur, (2018). Missingno: a missing data visualization suite. Journal of Open Source Software, 3(22), 547, https://doi.org/10.21105/joss.00547

UNESCO Institute for Statistics. Data as of September 2020. Available at: https://data.worldbank.org/indicator/SE.XPD.TOTL.GD.ZS

Brandon M. Greenwell (2017). pdp: An R Package for Constructing Partial Dependence Plots. The R Journal, 9(1), 421–436. URL https://journal.r-project.org/archive/2017/RJ-2017-016/index.html

V-Dem Codebook: Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, David Altman, Michael Bernhard, M. Steven Fish, Adam Glynn, Allen Hicken, Anna Lührmann, Kyle L. Marquardt, Kelly McMann, Pamela Paxton, Daniel Pemstein, Brigitte Seim, Rachel Sigman, Svend-Erik Skaaning, Jeffrey Staton, Agnes Cornell, Lisa Gastaldi, Haakon Gjerløw, Valeriya Mechkova, Johannes von Römer, Aksel Sundtröm, Eitan Tzelgov, Luca Uberti, Yi-ting Wang, Tore Wig, and Daniel Ziblatt. 2020. "V-Dem Codebook v10" Varieties of Democracy (V-Dem) Project.

V-Dem Dataset: Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, David Altman, Michael Bernhard, M. Steven Fish, Adam Glynn, Allen

Hicken, Anna L¨uhrmann, Kyle L. Marquardt, Kelly McMann, Pamela Paxton, Daniel Pemstein, Brigitte Seim, Rachel Sigman, Svend-Erik Skaaning, Jeffrey Staton, Steven Wilson, Agnes Cornell, Nazifa Alizada, Lisa Gastaldi, Haakon Gjerløw, Garry Hindle, Nina Ilchenko, Laura Maxwell, Valeriya Mechkova Juraj Medzihorsky, Johannes von R¨omer, Aksel Sundstr¨om, Eitan Tzelgov, Yi-ting Wang, Tore Wig, and Daniel Ziblatt. 2020. ”V-Dem [Country–Year/Country–Date] Dataset v10” Varieties of Democracy (V-Dem) Project. https://doi.org/10.23696/vdemds20.

and:

Pemstein, Daniel, Kyle L. Marquardt, Eitan Tzelgov, Yi-ting Wang, Juraj Medzihorsky, Joshua Krusell, Farhad Miri, and Johannes von R¨omer. 2020. “The V-Dem Measurement Model: Latent Variable Analysis for Cross-National and Cross-Temporal Expert-Coded Data”. V-Dem Working Paper No. 21. 5th edition. University of Gothenburg: Varieties of Democracy Institute.

V-Dem: https://www.v-dem.net/en/

IndexMundi: https://www.indexmundi.com/