Data Science Project: Progress

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

- Time period considered: from 2008 to 2017 (10 years)
- Education spending = Government expenditure on education, total (% of GDP)
- Countries: UN member states

- Data sources:
    - V-DEM dataset
    - 'World Bank' dataset (for education budgets)

V-Dem High-Level Democracy Indices

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 ideal of liberal democracy achieved?
Participatory Democracy Index	To what extent is the ideal of participatory democracy achieved?
Deliberative Democracy Index	To what extent is the ideal of deliberative democracy achieved?
Egalitarian Democracy Index	To what extent is the ideal of egalitarian democracy achieved?

Scale: Interval, from low to high (0-1)

Lower-Level Democracy and Governance Indices

Electoral Democracy Index

- Freedom of expression and alternative sources of information index
- Freedom of association index
- Share of population with suffrage
- Clean elections index
- Elected officials index

. . .

Obtaining Data

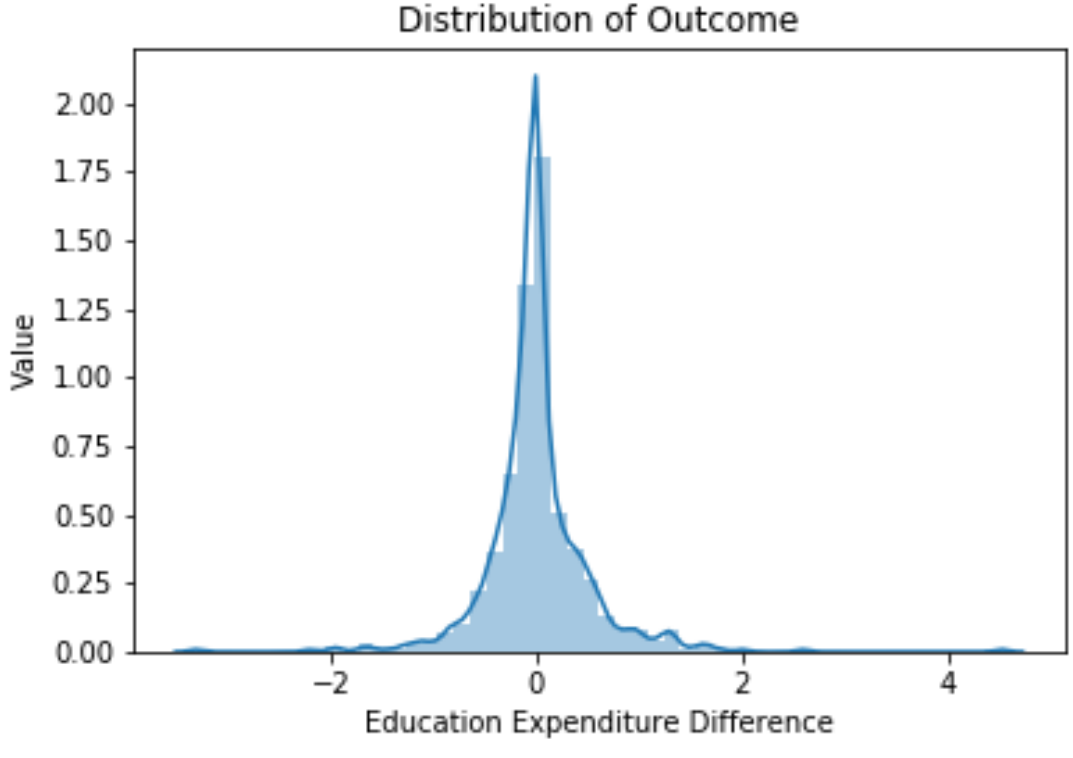
- V-DEM dataset (downloaded)
- WB dataset (scraped from indexmundi.com)
- UN member states (scraped from un.org)

Creating Main Dataframe

- Merged cleaned V-DEM dataset with cleaned WB dataset on `country` and `year`
- 989 rows, 29 columns

First Machine Learning Model

- **Outcome**: Difference in Education Expenditure
- **Predictors**: High-Level Indices

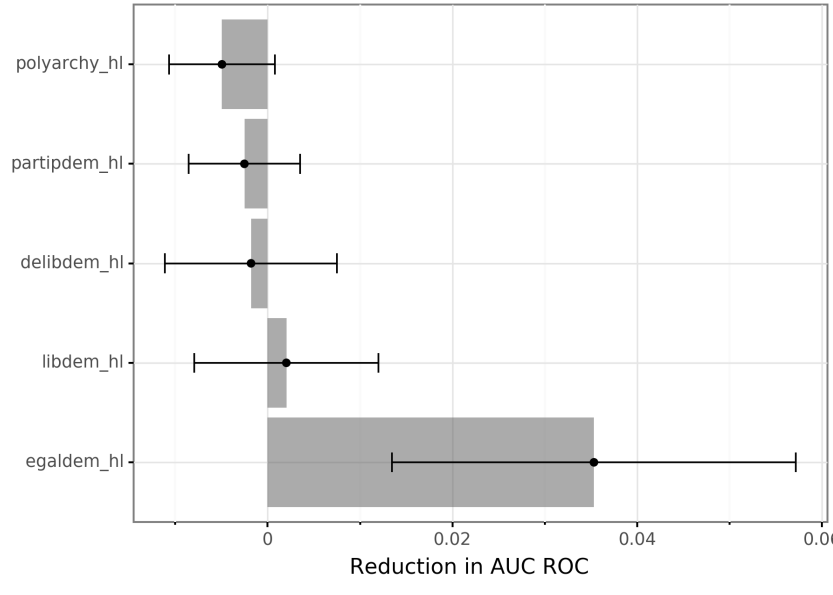


- Best model: Random Forest. (Trees with a depth of 2, spanning across 500 trees)
- Poor test performance: Negative R-squared

Second Machine Learning Model

- Outcome: Increase in Education Expenditure
- Predictors: High-Level Indices

- Best model: **KNN** - Area under ROC (61%)



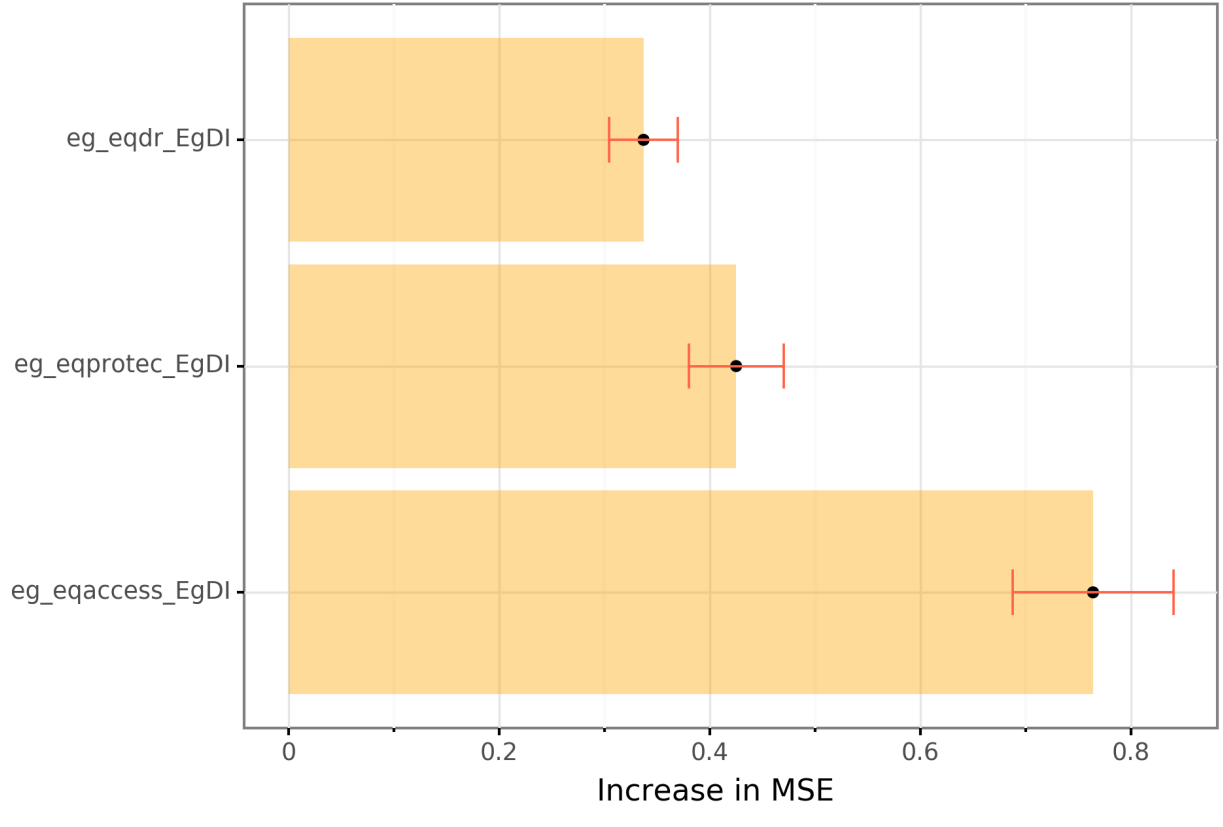
Further Exploration

Equal protection index **Equal access index** **Equal distribution of resources index**

Which one of these Low-Level indices is mostly determining the value of this index?

- Best model: Bagging Regressor
- Performance:
    - MSE: 0.004
    - R-squared: 93%

Variable	Index	Explanation
eg_eqprotec_EgDI	Equal protection index	How equal is access to power?
eg_eqaccess_EgDI	Equal access index	How equal is the distribution of resources?
eg_eqdr_EgDI	Equal distribution of resources index	How equal is the protection of rights and freedoms across social groups by the state?



Lessons Learned

- Interpreting variable importance (continuous data)
- R-squared can be negative

Challenges

- Missing data (even for developed countries)
- Are 10 years enough?
- How to proceed when model performance is extremely poor?