Predicting democratic backsliding with Machine learning

Model comparison

Paul Elvis Otto
249968
p.otto@students.hertie-school.org

Ujwal Neethipudi
248346
u.neethipudi@students.hertie-school.org

Saurav Jha
249354
s.jha@students.hertie-school.org

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et.

Contents

Appendix

List of Figures

1. Motivation and Context

With the rise in populist parties and the posfactual time we live in a lot of problems are accounted to this authoritarianf perspective. one of them is Corruption. Due to the advancements that have been made in Machine learning we have decided to analyse further the interplay between policitacl corrupten and the potential predictors for that. In the sprehre of political science there are multiple datasets that measure corruption and the possible predictors. Allthough most of them dont combine as nicely with the other predicotrs that we want to analyse. Therefore we have decided on using the Vdem dataset as our entry point to the data analysis. In short Vdem is an aggregated dataset that uses multiple sources to combine in one dataset, the dataset is composed of multiple indices, that measuere the state of diffrerent states around the world. In this Final report we want to limit some of the data as we will show in Section 3.

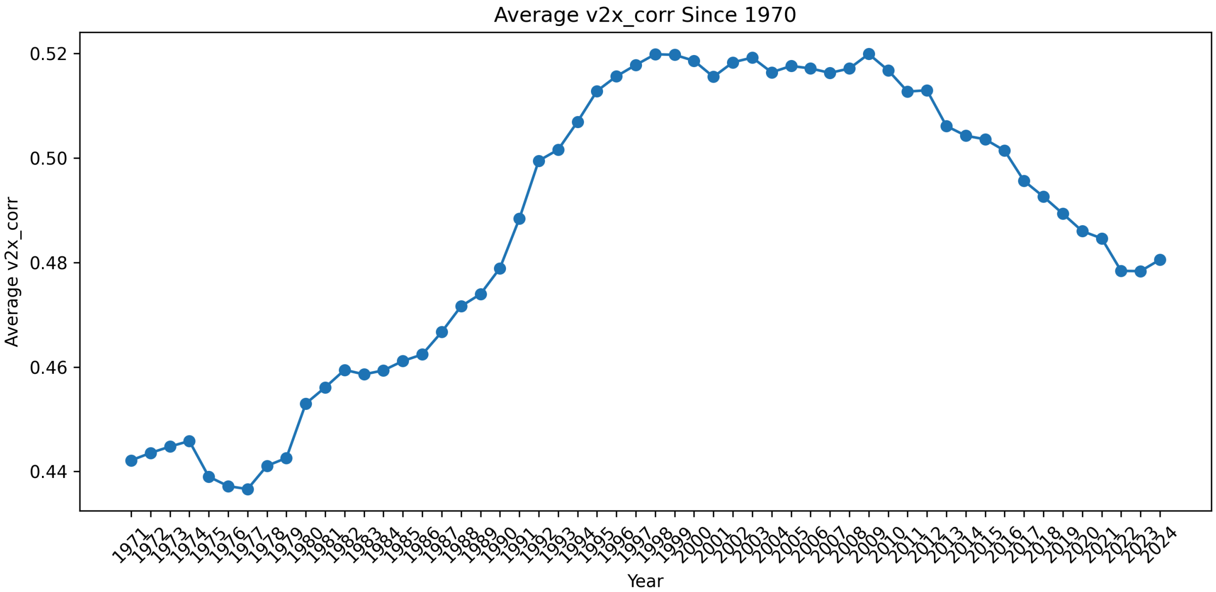

Figure 1: average of political corruption over time

2. Our Hypothesis

Vdem on its own already delivers a properly constructedt corruption index, that takes into account multiple different points of resoureces. The Index is split into multiple sub indices such as Public and politicial corruption. see the figure down below for a dependencie tree of the dataset.

We argue that with the help of machine learning models we can successfully predict the corruption level of a country based on factors that are not included within the current indicators for corruption. The claim lies in the assumption that political parttaking as well as the ability to parttake correlates highly with multiple other datapoints that are already provided and with widely available data such as the GdP of a country.

We argue that especially the social issues can be used as strong predictors for the level of corruption in the Vdem dataset.

R-squared	Adj. R-squared	Residual Std. Error	F-statistic	Model p-value
0.238	0.238	0.247	7940.234	0

3. Assumpotions and limits

We feel the need to explain the stepts that we will take in the further analysis of the dataset as they are not self explainatory in the further cleaning of the data.

1. **We Subset for liberal western Democracies**: We argue that a subsetting of the dataset might not seem to be essential in the context of corruption, but under the inherently different design of States in the sense of authoritatiran perspective, it is essential foR the model success to not only build an environment in which the data, as is, is compareable but also the states behind it. Therefore we have subset the data for Liberal western democracies, as we assume them to be hightrust societies where corruption is not inherent in their state structure as it would be with rent seeking states as shown by Ross (2001) in "Does Oil hinder democracies". The full list of Countires can be found in Appendix B.
2. **Limiting of the Time**: We see the need to limit the data, not only to fill NA values but to keep a compareable time frame of the different countries, such as the deepter integration of the European union as the differnce in justiciable account of corruption has been normalized among the member states, it would make sense to set the limit lower then the 1980s to train a model that can also predict that data but under the lens of non isolated development in political interaction with that topic we choose not do pursue that.

On the matter of *normal* problems in the dataset we need to highlight the non availability of different indeicies, that stems from the fact that the dataset introduced multiple varibales in a progressive manner. Due to that we have decided to limit the numbers variables that we will use for the prediction. Another point that we have taken into account is the dependencies of the data. Vdem gives an overview of the dataset and how the different indices are constucted at Coppedge *et al.* (2025) but that informationis not mapped into data, thus the result of the this project report is also a comprehensive mapping of the Vdem Dataset, made available as a github gist and included in this repository.

4. Reproduceability

To make this analysis reproduceable within the means of it we organized this analysis

5. Subset and EDA

To make t

5.1. Subset

To properly use the model to predict the Target it was necessary to filter the dataset further, for that we implemented a filtering to only keep the bare variables of the dataset without any additional operations as they are included by the authors of vdem.[1] In the next step we filtered for non numeric features and removed as pointed out in the missing values.

5.2. EDA

To get a first overview of the data and potential cross dependencies that we havent captured yet with our dependencie mapping a correlation matrix was created[2] This correlation matrix directly pointed out more highspots that after a manual correction have been removed. The

[1] To make this process reproduceable we organised the dataset subsetting process in a pipeline that can be run from the root of the project dir on github, manual or with a makefile

[2] Script: `./src/corr_matrix_plot.py`

List of Figures

Bibliography

Coppedge, M. *et al.* (2025) "V-Dem Codebook V15."

Ross, M.L. (2001) "Does Oil Hinder Democracy?," *World Politics*, 53(3), pp. 325–361. Available at: http://www.jstor.org/stable/25054153 (Accessed: May 5, 2025).

A Tables and Data

… your appendix content …

B List of Countries

- Austria
- Belgium
- Bulgaria
- Croatia
- Cyprus
- Czech Republic
- Denmark
- Estonia
- Finland
- France
- Germany
- Greece
- Hungary
- Ireland
- Italy
- Latvia
- Lithuania
- Luxembourg
- Malta
- Netherlands
- Poland
- Portugal
- Romania
- Slovakia
- Slovenia
- Spain
- Sweden
- United States of America
- United Kingdom