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

1. Introduction . . . 3
2. Vdem in detail . . . 3
3. EDA . . . 4
   3.1. Limitations and subseting . . . 4
   3.2. General Model . . . 5
4. Predicting corruptoin in context . . . 5
5. Prediction modeling . . . 5
   5.1. Linear Models . . . 5
   5.2. Tree based approaches . . . 5
   5.3. Non linear approaches . . . 5
   5.4. Neural Networks . . . 5
6. Motivation and Context . . . 5
7. Our Hypothesis . . . 7
8. Assumpotions and limits . . . 7
9. Reproduceability . . . 7
10. Subset and EDA . . . 8
    10.1. Subset . . . 8
    10.2. EDA . . . 8

Bibliography . . . 10
A Tables and Data . . . 10
B List of Countries . . . 10

Appendix

A Tables and Data 10
B List of Countries 10

List of Figures

Figure 1 average of political corruption over time 6

1. Introduction

Corruption remains a persistent challenge faced by countries across the globe, affecting governance, economic stability, societal trust, and overall quality of life. The severity and scope of corruption's impact, however, vary significantly depending on local contexts, institutional robustness, and sociopolitical factors. With this project, our goal is to systematically identify and analyze consistent predictors of corruption across multiple dimensions and geographic contexts. The inherent complexity of studying corruption emerges notably from difficulties in defining, operationalizing, and measuring corruption itself, as well as ensuring consistent and comparable data across a substantial number of countries and time periods.

To effectively address this complexity, we have selected the Varieties of Democracy (V-Dem) dataset as the cornerstone of our analysis. V-Dem offers comprehensive coverage with a large number of countries, a broad temporal scope, and a rich array of indicators collected using a rigorous, state-of-the-art methodology. The robustness and consistency of this dataset enable us to conduct detailed and nuanced analyses without necessitating reliance on multiple external data sources, thereby maintaining methodological coherence and facilitating clear interpretability of our findings. To further enhance transparency and reproducibility, we developed several helper functions in R, specifically tailored to streamline dataset importation and pre-processing, while mitigating common issues that typically arise from handling large and complex datasets. The complete analytical workflow, including these helper functions, is integrated into a reproducible environment supported by a comprehensive Makefile, ensuring ease of replication and robustness against potential errors or inconsistencies.

Recognizing that mere statistical replication of the V-Dem dataset would yield limited insights, we complemented our empirical model evaluation with a rigorous theoretical framework informed by existing literature. Extensive research has consistently highlighted gender as a significant factor influencing perceptions, prosecution patterns, and predictive models related to corruption. Motivated by these insights, we utilize the gender-related variables provided by V-Dem to develop an additional predictive model specifically designed to examine how gender disparities influence corruption levels across countries. By incorporating gender-focused analysis, we seek to contribute substantively to ongoing debates regarding the intersectionality of corruption and societal inequality.

We further acknowledge that corruption manifests differently depending on national and regional contexts, requiring nuanced differentiation when assessing diverse countries. To address this variability systematically, we adopted a mixed-methods approach complemented by rigorous statistical techniques, detailed comprehensively in Appendix 1. This approach allowed us to group countries strategically based on their socio-political, economic, and institutional characteristics, thereby enhancing the specificity, accuracy, and policy relevance of our predictive models.

An additional methodological consideration was the treatment of missing data, which inevitably arises in large, cross-national datasets like V-Dem. To effectively manage this issue while maintaining both computational efficiency and methodological integrity, we employed a tree-based imputation strategy tailored to country-specific characteristics. This approach minimizes the risk of introducing unintended cross-country correlations, ensuring that our analyses accurately reflect national-level realities rather than artifacts introduced by data handling procedures.

2. Vdem in detail

For the building of the prediction model we used the vdem dataset that we will explain in short summary before we target it in the EDA.

The Varieties of Democracy (V-Dem) dataset implements a large scale anaylis in the context of comaprative political science. Its construction involves a systematic process of aggregating expert evaluations on a wide range of indicators. A global network of country specialists contributes their knowledge by responding to numerous specific questions related to different aspects of political regimes. V-Dem conceptualizes democracy along five distinct yet interrelated dimensions: electoral, liberal, participatory, deliberative, and egalitarian.

Vdem provides different types of indices high and mid level. For the EDA we kept both in and based on the results of the we recived from the EDA decided to remove multiple.

Vdem therefore puts us in a position of asessing the predictors of corruption from various levels. Once it enables us to assets wether we will need to lock the model to a specific region, and or wether that we need to take other percautions, in working with the data.

The dataset also gives some constraints, as this is data that is aggregated by experts in interviews, quilative assesments, and surveys the data collection is impacted by the state of a country, therefore the data can be missing/ not covering enough for partial or complete authoritarian regimes.

As external sources we used economic data on the GDP and the distribution by gender.

3. EDA

We started the EDA by first getting an overview of the correlation of the indices with each other.

We start this EDA by first looking in the development of corruption over time for all countries. as the first step we combine our own knowledge on the impact of political climate and corruption, therefore we ran a simple regresseion between the Liberal democracies index and the corruption index, this regression already shows that there is a high cross correlation between these two. so that the liberal democracy index already explains around 25% of the model.

Therefore we have decided to limit the scope of the modeling process to only liberal democracies, as we assets that the variation in the liberal democracy index is not a problem for builind a comprehensive model. Another reason for this subset is the quality of the data as that is more consistent and complete for the by us defined liberal democracies for a full list of them see Appendix B.

We then moved on to assesing the data. We removed all the indices that are in the dataset, we only used the individual variables that would compose these indexes.

After that we ran a correlation test between the individual variables and our target, that gave us more insight into the cross dependent factors. Here we included the top 10 predictors. see table below.

3.1. Limitations and subseting

The Vdem dataset brings even though we already subssetted for the somewhat compareable countries additional limitations, most concisting of variableS that only have been introduced in the later releases of the dataset. Therefore we proceded with the following startegies. We imputed the missing values with the help of randomforrest imputing on a country basis to not also introduce cross correlaions between the countries. we also decided not to use mean imputing due to the potential problematic of filling to much data in on the basis on a to small input. We also decided to drop variables that overall only have entries in a fixed amount of years.

3.2. General Model

We then proceded to build a model that used all the available raw variables and assesed wether we can predict corruption with the corruption index components removed and received the following resulst. (To not repear ourself, we will show the detailed methods in the next chapter.)

For this baseline comaprison we chosen linear models as well as tree based approach. See the overall comparison(place figure here)

4. Predicting corruptoin in context

As mentioned in the beginning of this report we dont want to only focus on the prediction of corruption itself but also set in a contextual frame an real life implications. The current and past literarute suggests that the role of women in society has an undenieable effect on corruption itself, overall women percive corruption as more problematic then men, they tend to enforce anti corruption measures harder then men in similar postions, and most interesting and what we decided to reproduce they role in society overall has a direct effect on the occurrence of corruption.

Therefore we decided to subset our dataset again to the following parameters.

1. Liberal democracies
2. GDP
3. Variables that target gender differences
4. Variables that measure explicitly womens access to politics, law and society
5. Time limited to start from 1960

We performed the model building process on the basis on this statistical and theoretical subsetting of the data.

5. Prediction modeling

In the following parts we will explain in more detail the different challenges and outcomes from each modelling steps and end with an overall comparison of the models. The complete list of the variables that we have chosen can be found in the appendix.

5.1. Linear Models

As we also did with the baseline we modeled an OLS regression for the model as

5.2. Tree based approaches

5.3. Non linear approaches

5.4. Neural Networks

6. 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 8.

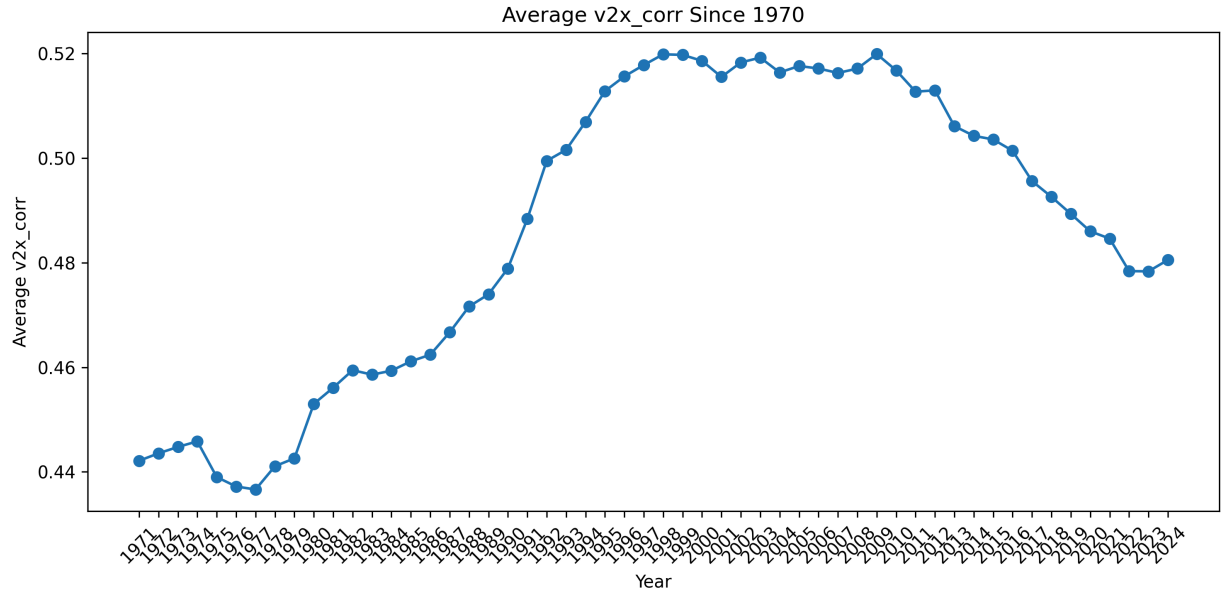

Figure 1: average of political corruption over time

7. 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

8. 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.

9. Reproduceability

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

10. Subset and EDA

To make t

10.1. 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.

10.2. 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

Figure 1 average of political corruption over time 6

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