

Predicting democratic backsliding with Machine learning

Model comparison

Paul Elvis Otto	Ujwal Neethipudi	Saurav Jha
249968	248346	249354
p.otto@students.hertie-school.org	u.neethipudi@students.hertie-school.org	s.jha@students.hertie-school.org

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguere 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	4
2. Vdem in detail	5
3. EDA	6
4. Data Cleaning	6
5. Machine learning Models	6
5.1. Overall Model	7
5.1.1. Linear Models	7
5.1.2. Tree Based approaches	7
5.1.3. Non Linear Models	7
5.2. Liberal Democracies	7
5.2.1. Linear Models	7
5.2.2. Tree Based approaches	7
5.2.3. Non Linear Models	7
6. Predicting corruption in context	7
7. Prediction modeling	7
7.1. Linear Models	7
7.2. Tree based approaches	8
7.3. Non linear approaches	8
7.4. Neural Networks	8
8. Motivation and Context	8
9. Our Hypothesis	9
10. Assumpotions and limits	9
11. Reproduceability	9
12. Subset and EDA	10
12.1. Subset	10
12.2. EDA	10
Bibliography	12

A	Tables and Data	12
B	List of Countries	12

Appendix

A Tables and Data	12
B List of Countries	12

List of Figures

Figure 1 average of political corruption over time	8
--	---

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

To fully understand the methodological considerations and limitations we encountered during the modeling and data analysis phase of this project, it is necessary to introduce and contextualize the Varieties of Democracy (V-Dem) dataset in greater detail.

The V-Dem dataset is a widely recognized and rigorously constructed comparative political science resource, providing extensive cross-national coverage of democratic institutions, governance characteristics, and political outcomes. Developed through a collaborative effort involving hundreds of international scholars and country experts, the dataset is built upon systematic expert assessments, structured questionnaires, qualitative evaluations, and comprehensive country-specific knowledge. This methodical collection of information allows V-Dem to offer robust, nuanced, and multidimensional measures of democracy.

Specifically, the V-Dem framework conceptualizes democracy along five distinct yet interconnected dimensions: electoral democracy, liberal democracy, participatory democracy, deliberative democracy, and egalitarian democracy. Each dimension is operationalized through numerous indicators that capture various institutional features, governance practices, and societal norms, which, taken together, allow researchers to gain detailed insights into the functioning of democratic regimes and related political phenomena such as corruption.

Within the dataset, V-Dem provides both high-level indices—aggregated measures summarizing broad democratic principles—and mid-level indices, which offer more granular evaluations of specific democratic components. Initially, our exploratory data analysis (EDA) leveraged both sets of indices to gain a comprehensive understanding of the available variables. Subsequently, based on insights drawn from our initial analysis, we identified redundancies and correlations among certain indicators, prompting us to strategically eliminate several indices. This process ensured model parsimony while preserving the explanatory power and interpretability of our predictive models.

Utilizing V-Dem data enables us to examine predictors of corruption at varying levels of granularity, facilitating analyses that range from broad global trends to regionally focused models. This flexibility is crucial in understanding whether particular predictive relationships hold consistently across different geographical contexts or whether region-specific modeling considerations and controls are necessary. Consequently, insights from our initial analyses using V-Dem have significantly guided our modeling choices—such as determining whether to segment our models geographically or to adopt specific methodological precautions.

Despite these notable strengths, the use of the V-Dem dataset also imposes certain constraints that must be explicitly acknowledged. Since the data is primarily aggregated from expert assessments, qualitative interviews, and comprehensive surveys, the accuracy and comprehensiveness of the dataset are inherently sensitive to country-specific political environments and conditions. Particularly in authoritarian or semi-authoritarian contexts, access to accurate and objective information may be limited, posing challenges for data collection and leading to incomplete or partially missing data points. Consequently, analyses conducted with the V-Dem dataset must consider potential biases or gaps resulting from these limitations. To mitigate the influence of such biases, we adopted rigorous imputation strategies tailored specifically for country-level characteristics, as previously described, thereby aiming to preserve data integrity and model accuracy.

Furthermore, since expert evaluations involve inherently subjective judgments, the reliability of V-Dem measures may vary slightly based on the expertise and perceptions of individual assessors. While V-Dem employs robust methodologies to aggregate and validate these evaluations, it remains critical to interpret

findings with caution, maintaining awareness of potential subjectivities inherent to expert-based measurements.

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 regression 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 assume that the variation in the liberal democracy index is not a problem for building 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 assessing 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.

4. Data Cleaning

Here Focus on how we filled NAs and what other cleaning measures we did such as the removal of the variables that are inherent to the corruption indices

5. Machine learning Models

In the following section we will explain the different models that we have used and benchmark them. To not repeat ourselfs we will proceed in the following way, we will explain the different steps that we took throughout the modeling process and then in a shorter summary present the models for our split in to different subsets for liberal democracies. Also this chapter with the overall identification of the best predictors we want to view it as a baseline and recall these results for when we apply our theoretical analysis of the relation of gender and corruption

5.1. Overall Model

5.1.1. Linear Models

5.1.2. Tree Based approaches

5.1.3. Non Linear Models

5.2. Liberal Democracies

5.2.1. Linear Models

5.2.2. Tree Based approaches

5.2.3. Non Linear Models

6. Predicting corruption in context

As mentioned in the beginning of this report we don't want to only focus on the prediction of corruption itself but also set in a contextual frame and real life implications. The current and past literature suggests that the role of women in society has an undeniable effect on corruption itself, overall women perceive corruption as more problematic than men, they tend to enforce anti corruption measures harder than men in similar positions, and most interesting and what we decided to reproduce their 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 women's access to politics, law and society
5. Time limited to start from 1960

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

7. Prediction modeling

In the following parts we will explain in more detail the different challenges and outcomes from each modelling step 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.

7.1. Linear Models

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

7.2. Tree based approaches

7.3. Non linear approaches

7.4. Neural Networks

8. Motivation and Context

With the rise in populist parties and the postfactual time we live in a lot of problems are accounted to this authoritarian 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 political corruption and the potential predictors for that.

In the sphere of political science there are multiple datasets that measure corruption and the possible predictors. Although most of them don't combine as nicely with the other predictors 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 measure the state of different states around the world. In this Final report we want to limit some of the data as we will show in Section 10.

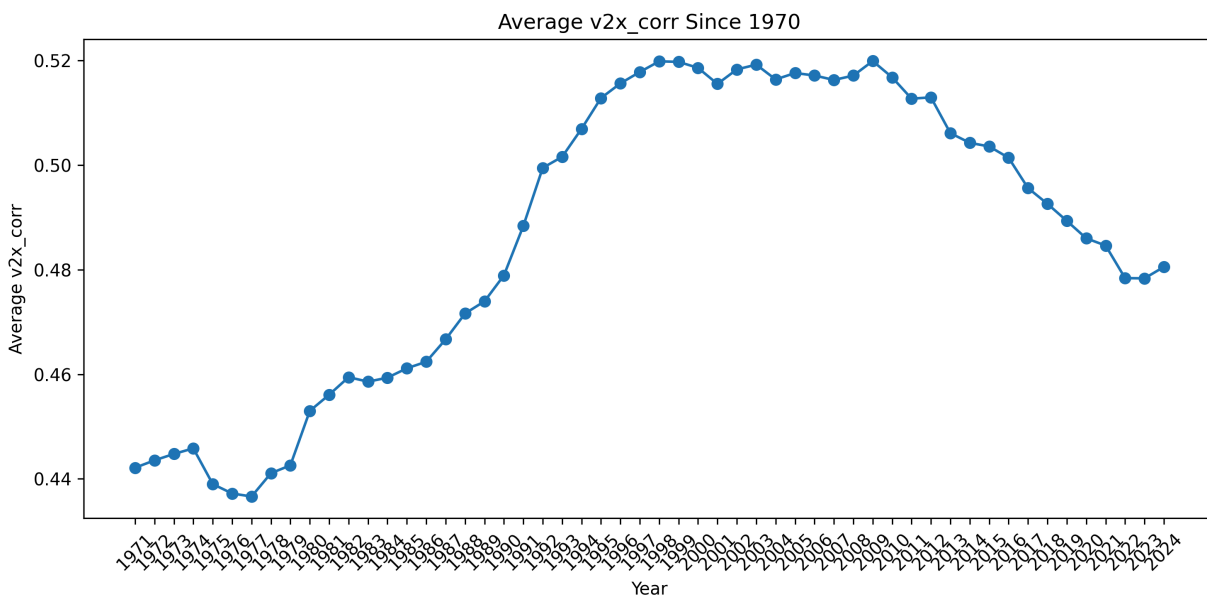


Figure 1: average of political corruption over time

9. Our Hypothesis

Vdem on its own already delivers a properly constructed corruption index, that takes into account multiple different points of resources. The Index is split into multiple sub indices such as Public and political corruption. see the figure down below for a dependence 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 partaking as well as the ability to partake 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

10. Assumptions and limits

We feel the need to explain the steps that we will take in the further analysis of the dataset as they are not self explanatory 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 authoritarian perspective, it is essential for the model success to not only build an environment in which the data, as is, is comparable 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 Countries 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 comparable time frame of the different countries, such as the deeper integration of the European union as the difference in justiciable account of corruption has been normalized among the member states, it would make sense to set the limit lower than 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 to pursue that.

On the matter of *normal* problems in the dataset we need to highlight the non availability of different indices, that stems from the fact that the dataset introduced multiple variables in a progressive manner. Due to that we have decided to limit the number of 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 constructed at Coppedge *et al.* (2025) but that information is not mapped into data, thus the result of this project report is also a comprehensive mapping of the Vdem Dataset, made available as a github gist and included in this repository.

11. Reproducibility

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

12. Subset and EDA

To make t

12.1. 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.

12.2. 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 created² This correlation matrix directly pointed out more highspots that after a manual correction have been removed. The

¹To 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

²Script: `./src/corr_matrix_plot.py`

List of Figures

Figure 1	average of political corruption over time	8
----------	---	---

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