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. 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. For all further analysis we already removed the variables that compose the index, so that they don't interfere with our analysis.

For that we first started with a correlation heatmap of the different indices with each other as well as a ranking with corruption as our target. To get an overall impression of what the data looks like in context to each other.

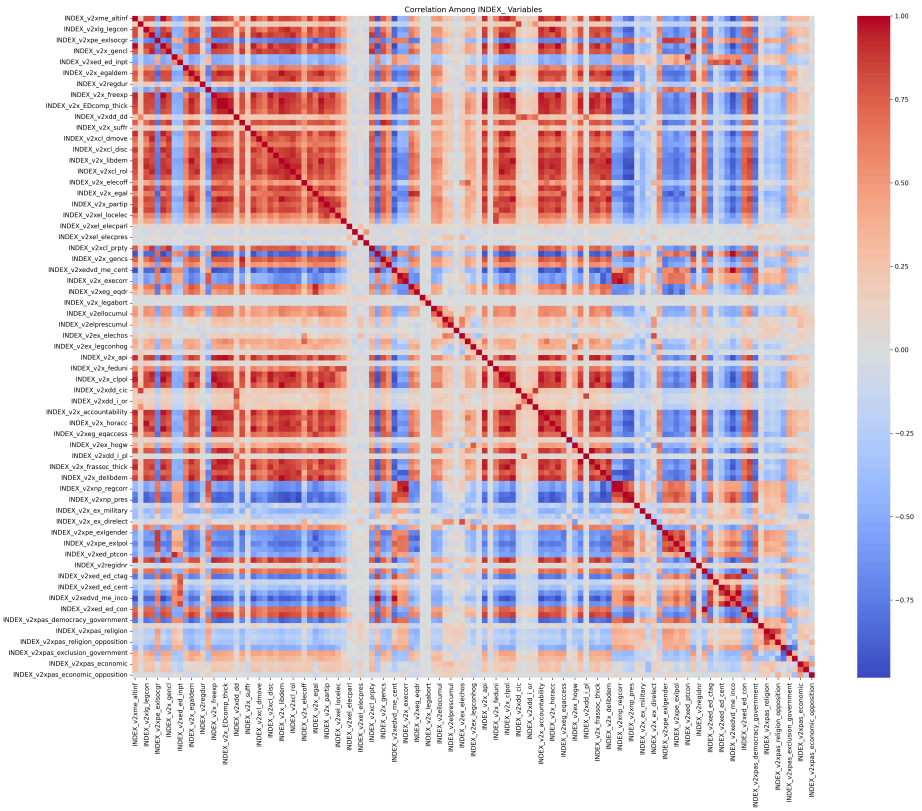


Figure 1: Correlation heatmap of the different indices

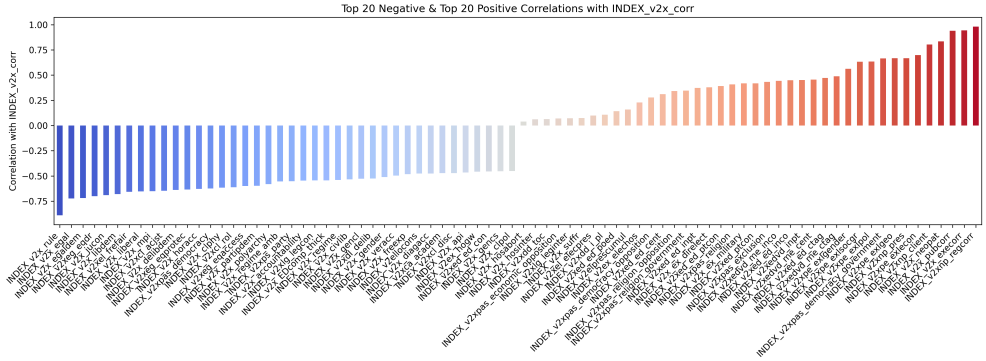


Figure 2: Top 80 Correlation of indices with the Target

Follwoing we also did that for the individual variables to assets that better.

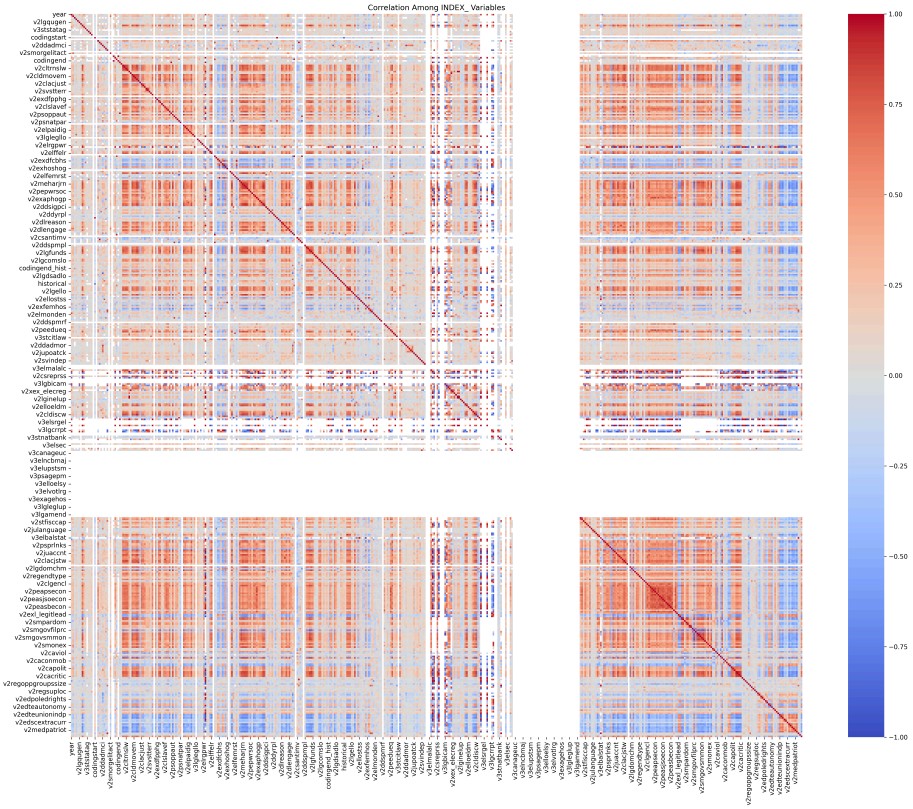


Figure 3: Correlation heatmap of the different indices

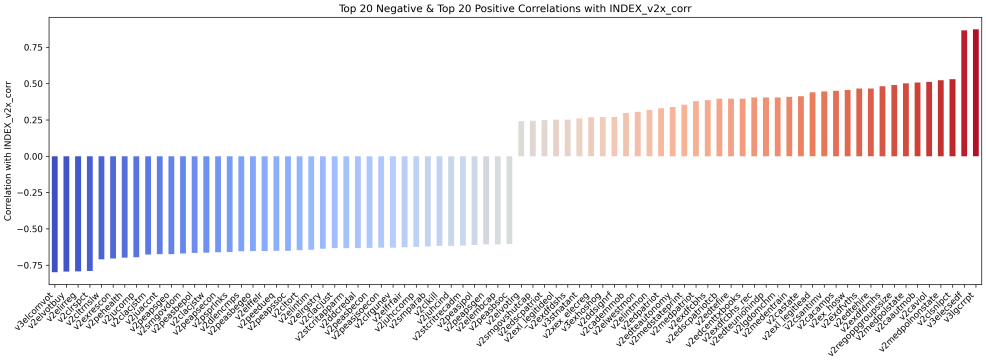


Figure 4: Top 80 Correlation of indices with the Target

These two analysis show us important indicators, first of all, the western liberal democracy index has a very high correlation with the corruption index and the components of it as in our variables as well.

This leads us to something we already anticipated to not distort our next modeling process we will need handle parts of the data differently. To further manifest that we clustered the different countries.

Based on this knowledge gain we decided to create a subset that focuses on liberal western democracies, as the suggested data would also end in that cluster and as, from a theoretical point of view, corruption is also a sociietal problem which would also factor in with our models. We will perform further analysis for booth datasets.

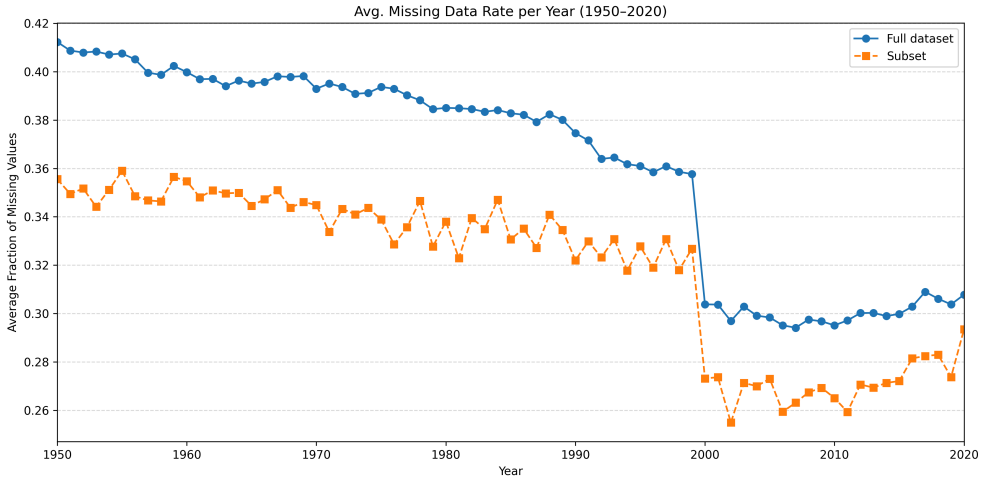


Figure 5: Missing data in Full dataset and subset

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.

4. Data Imputation

To ensure rigorous model construction, we employ a tree-based imputation strategy to address missing entries within our cross-national time-series dataset. This approach offers considerable flexibility in capturing complex, non-linear relationships among governance quality, economic output, and social development indicators, which often interact in threshold-dependent or multiplicative fashions. Unlike global linear estimators or simple mean replacements, a decision-tree imputer partitions the feature space along data-driven boundaries, thereby generating imputations that reflect underlying heterogeneity more faithfully. Moreover, decision trees inherently detect and exploit high-order interactions without requiring the modeller to specify interaction terms a priori; as a result, the imputation algorithm “learns” country-specific patterns whereby, for example, a missing education indicator may be jointly determined by institutional quality and prior GDP growth.

Economic and political variables frequently exhibit heavy-tailed distributions and regime-dependent variance—manifest in crises, transitions, or other structural breaks—yet tree-based methods mitigate the undue influence of extreme observations by constructing splits that isolate homogeneous subgroups. In contrast to

mean- or linear-based imputations, which may be skewed by outliers and produce implausible values during turbulent periods, our imputer preserves the integrity of the data's tail behavior. Crucially, by performing imputations within each country's temporal block—sorting observations by year and grouping on country identifiers—the algorithm respects the panel structure, borrowing strength from related indicators without introducing cross-country contamination and thus maintaining realistic within-country trajectories.

As a non-parametric method, decision-tree imputation obviates the stringent distributional assumptions —such as residual normality or linearity—imposed by parametric techniques like Gaussian expectation–maximization, affording more credible estimates when the true data-generating process deviates from classical families. Finally, from a computational perspective, tree-based imputation scales linearly with sample size and can be parallelized across national units. The resulting tree structures also afford transparency, enabling inspection of the variables driving each imputation and thereby enhancing interpretability relative to "black-box" multivariate models.

5. Machine learning Models

In the ensuing section, we delineate and evaluate the array of machine learning techniques employed in our analysis. To avoid unnecessary repetition, we first describe the sequential steps undertaken throughout the modeling process—encompassing data preparation, feature engineering, subsequently provide a concise overview of the specific models applied to the subset of liberal democracies. This chapter, which also establishes the principal predictors of corruption, serves as a methodological baseline. Its findings will be revisited and contextualized in our theoretical examination of the relationship between gender and corruption, thereby ensuring coherence between empirical benchmarking and subsequent inferential analysis.

5.1. Overall Model

In this chapter, we present a comprehensive synthesis of model performance metrics and benchmarking results derived from the fully imputed variable dataset. We proceed to identify the principal determinants of corruption and conduct a systematic comparison of the various methodological approaches employed. The chapter concludes with a detailed documentation of our benchmarking procedures and an interpretive discussion of the relative strengths and limitations of each modeling strategy.

5.1.1. Linear Models

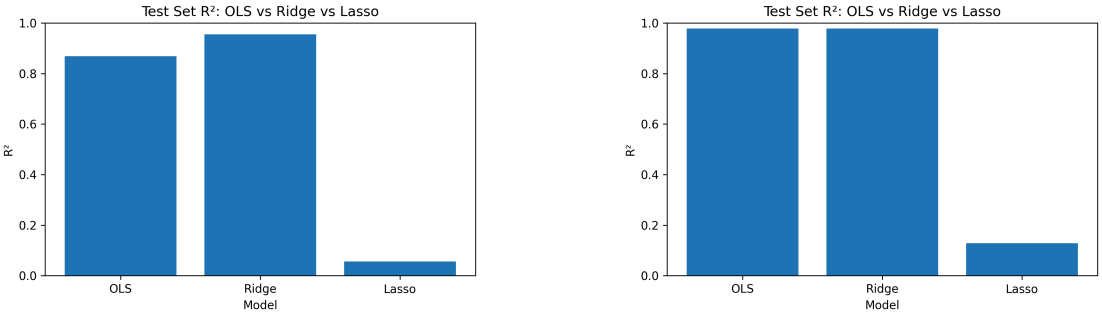
For the unadjusted model we ran we got the following results:

Subset

Model	Test MSE	Test R^2	Best α
OLS	0.000568	0.978477	NaN
Ridge	0.000591	0.977602	46.415888
Lasso	0.023009	0.128624	1.764095

Full Dataset

Model	Test MSE	Test R^2	Best α
OLS	0.010738	0.868308	NaN
Ridge	0.003625	0.955539	0.305386
Lasso	0.076984	0.055894	2.831639



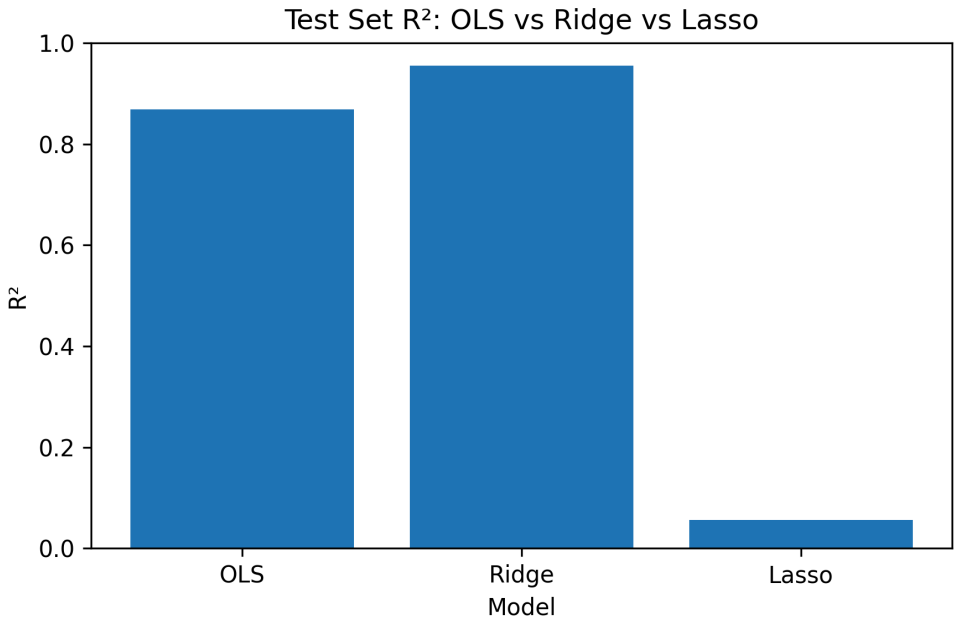


Figure 6: Benchmark Full linear dataset

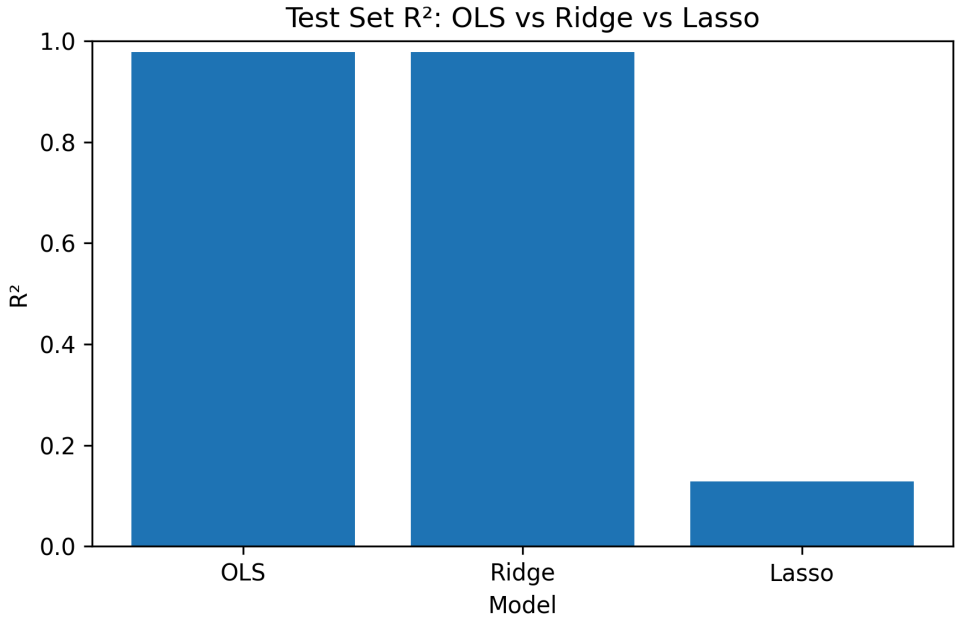


Figure 7: Benchmark Subset linear dataset

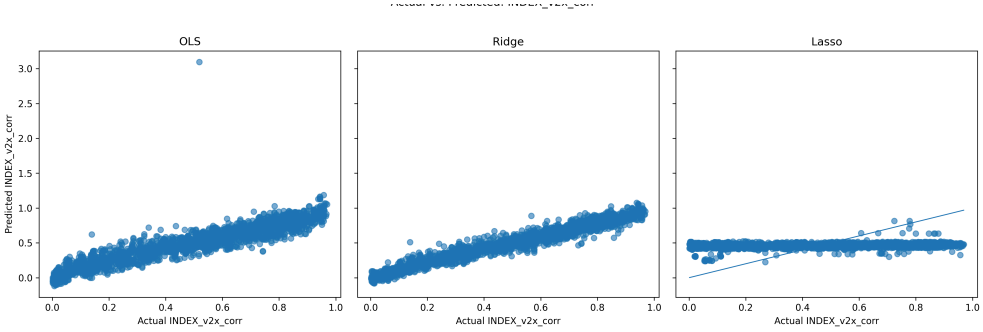


Figure 8: Benchmark Full linear dataset

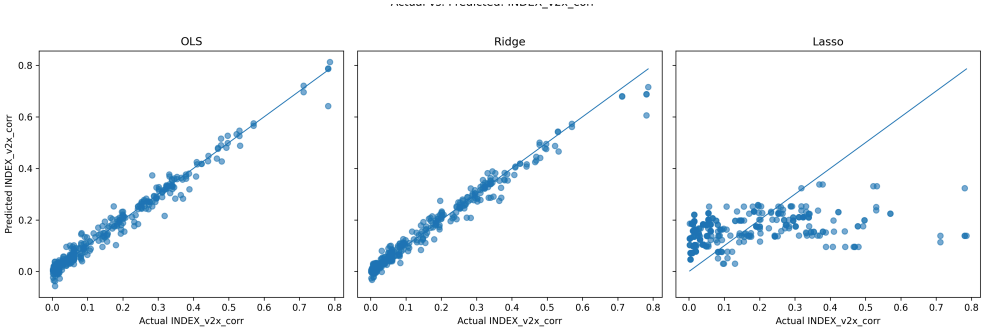


Figure 9: Benchmark Subset linear dataset

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 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.

7. 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.

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 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 10.

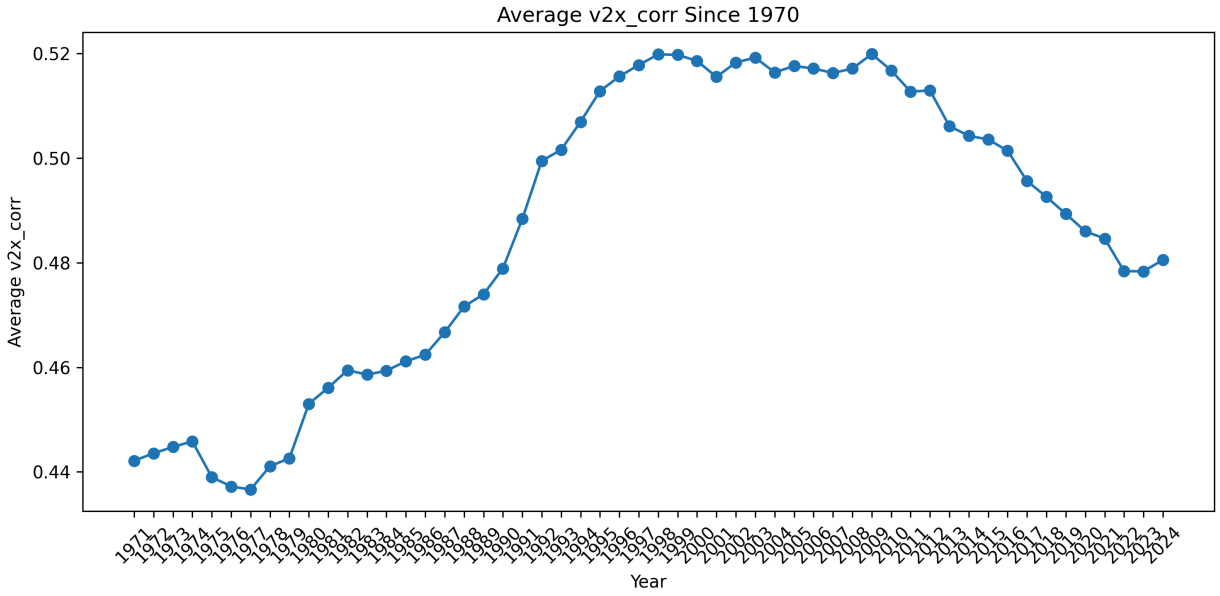


Figure 10: average of political corruption over time

9. 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

10. 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.

11. Reproduceability

To make this analysis reproduceable 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 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