Final Project

STAT 5650

Elsa Jos, Hyrum Hansen, and Zion Steiner

January 10, 2022

Introduction

While the debate over the relative merits of legislative structure is mature, it is fraught with conjecture. Indeed, a model that describes these merits has proven elusive. Theoretical arguments in favor of a bicameral legislature include legislative stability, dual representation, and protection against majority tyranny. Rooted in the 'one person one vote' principle, theoretical arguments in favor of a unicameral legislature include efficiency, transparency, and stability [1].

Linear discriminant analysis has been used to predict the structure of *city* governments in the United States. A 1976 study found that structural and demographic variables may be used to predict government form, focusing specifically on features of the executive branch and voting constituencies [2]. The authors of this paper could find no evidence of a model making similar classifications with international data.

This paper uses international development indicators as a proxy for policy outcome. Though some indicators – such as those which measure population demographics – are probably not related to legislative outcome, indicators that track environmental impact, economic activity, education, and international relationships certainly are. Using these measurements, this paper aims to answer three fundamental questions. First, can international development indicators be used to used to build a model that accurately classifies legislative structure? Second, are there unexpected relationships among variables that provide useful information about corruption? And third, can development

indicators be used to build a model that accurately classifies countries into one of three "freedom" levels: not free, partly free, and and free?

Data

The data were collected from four sources. Predictor variables comprise 540 World Bank development indicators and were gathered from the World Bank website. The World Bank indicators are semi-sparse, with most countries having indicator values for several years and some having no observations at all. To solve this problem, we filtered out any indicators that fewer than 75% of countries had observations for. For the remaining indicators, each country's most recent observation was used. In the worst case, there were decades between the most recent observations for different countries. This makes direct comparisons of indicator values between countries more difficult. However, we assume the number of features we have makes up for this challenge.

There were three response variables: type of legislature, corruption perceptions index (CPI), and democracy index (DI). Scraped from Wikipedia, type of legislature is a binary response variable that reports either bicameral or unicameral describing the legislative structure of a given country. The other two responses were scraped in a similar fashion from their respective sources, Transparency International and Freedom House. When merged, there were 187 observations of 544 variables. The 544^{th} variable is the name of the country described by the data.

Missing values complicated the study. Since there were only 26 complete observations, imputation would be required for some analyses. To demonstrate variable "missingness," a subset of 20 variables was randomly selected. Figure 1 gives a plot of these variables vs. their percentage of missing values. No more than 25% of the values were missing for any variable and the overwhelming majority of variables had data for $> 85\%$ of the observations.

To further illustrate the missing observations, a graphical visualization of missing observations for the data was obtained (fig.2.)

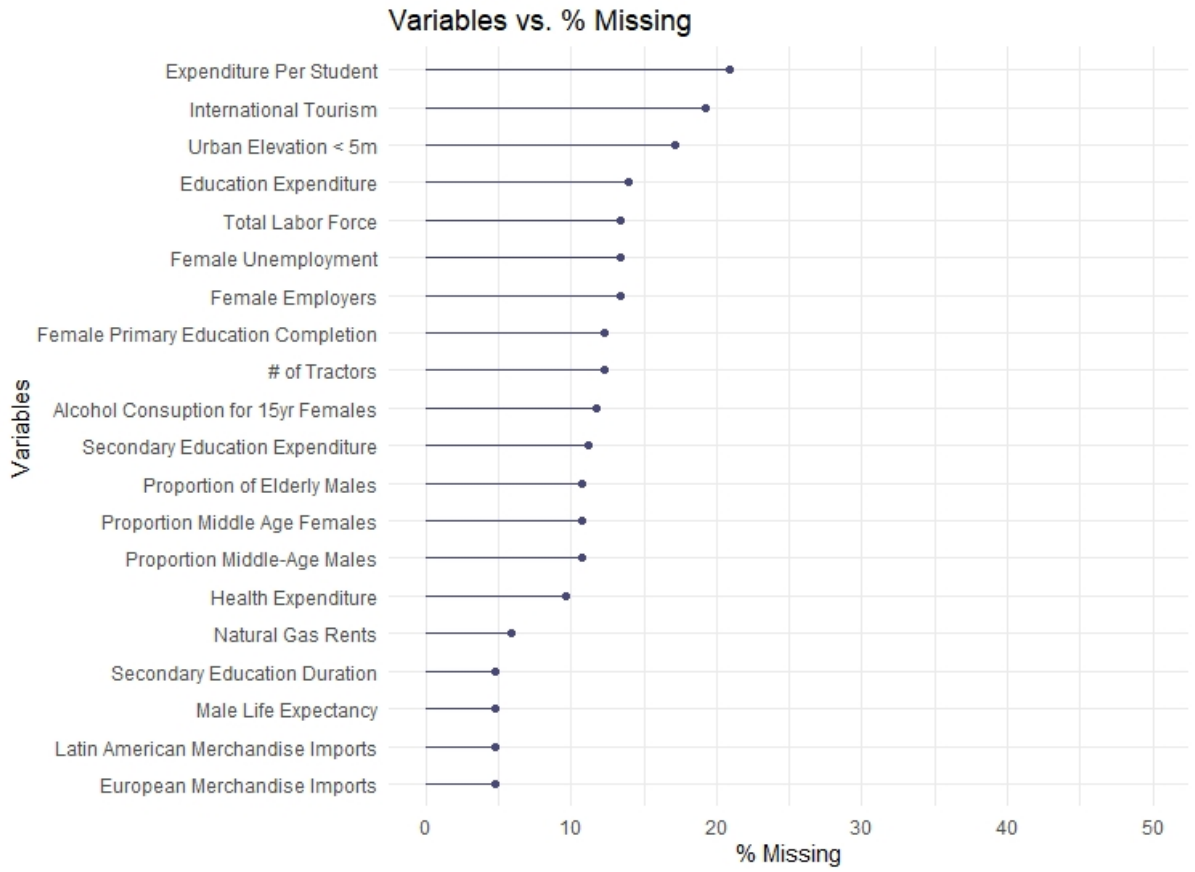

FIGURE 1: *A randomly selected subset of the variables. Most variables are more than 85% complete.*

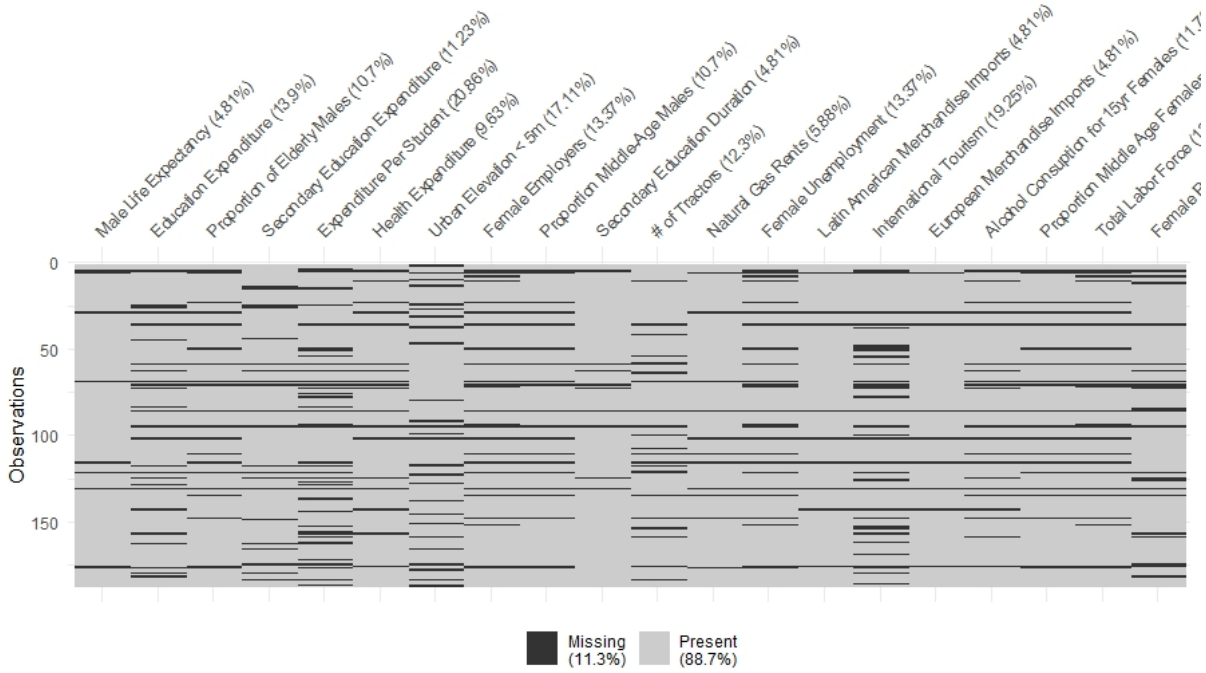

FIGURE 2: *Among the 20 variable subset, the data is nearly 90% complete.*

Principal Components Analysis (PCA)

PCA was applied to the data. Imputation was necessary, so the *imputePCA()* function from the **missMDA** package was used. It took 60 principle components to reach 90% cumulative variability explained and 135 to explain 99%. Table 1 provides information about the first ten principle components. The first ten principle components explained just over 60% of the variability in the data. With well over 500 predictor variables, it is not surprising that many principle components are needed to adequately capture variability.

	Eigenvalue	% of Variance	Cumulative % of Variance
Principle Component 1	154.60	28.79	28.79
Principle Component 2	62.99	11.73	40.52
Principle Component 3	29.07	5.14	45.93
Principle Component 4	20.17	3.76	49.69
Principle Component 5	17.93	3.34	53.03
Principle Component 6	11.79	2.20	55.22
Principle Component 7	11.09	2.07	57.29
Principle Component 8	9.29	1.73	59.02
Principle Component 9	9.29	1.73	60.75
Principle Component 10	8.47	1.58	62.32

TABLE 1: *The first 10 principle components explain just over 60% of the variability in the data. With well over 500 variables this is unsurprising.*

To visualize the cumulative variability explained with each additional principle component, a plot of cumulative variability vs. the number of principle components was obtained (fig. 3). The cumulative variability increases sharply from 1 to 25, then marginally for each additional principle component thereafter. In the second phase of analyses, principle components will be used for model construction. Their predictive power will be compared to the raw variables.

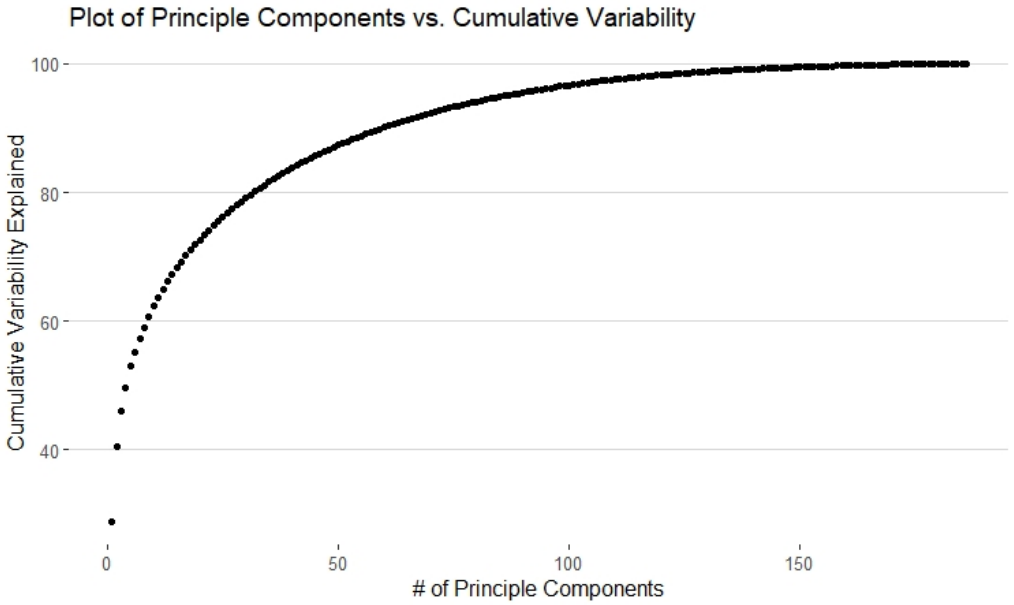

FIGURE 3: *As the number of principal components increases, the variability explained increases logarithmically.*

Legislature Type

Random Forests

The first model fit to the data was a random forest. This model was trained on all predictor variables with legislature type as the response. Table 2 provides summary metrics using cross validation for the model. The model performed adequately when classifying unicameral legislatures, but performed poorly when classifying bicameral legislatures. Compared to the out of bag accuracies, cross validation increased the sensitivity slightly while decreasing specificity. Note that for all analyses in this section, sensitivity describes correct 'unicameral' classifications.

	% Correct	Sensitivity	Specificity	Kappa
Random Forest	67.38	87.04	40.51	0.29

TABLE 2: *Cross validated accuracies for the random forest model using all predictor variables.*

A plot of variable importance for the random forest model was obtained. Education variables were particularly important with the number of secondary education vocational pupils ranking number one and the total number of secondary education pupils ranking seventh. Economic factors also influenced the model, with six of the top twenty variables measuring imports and exports. Environmental factors contributed significantly

as well, two of them measuring dimensions of greenhouse gas emissions.

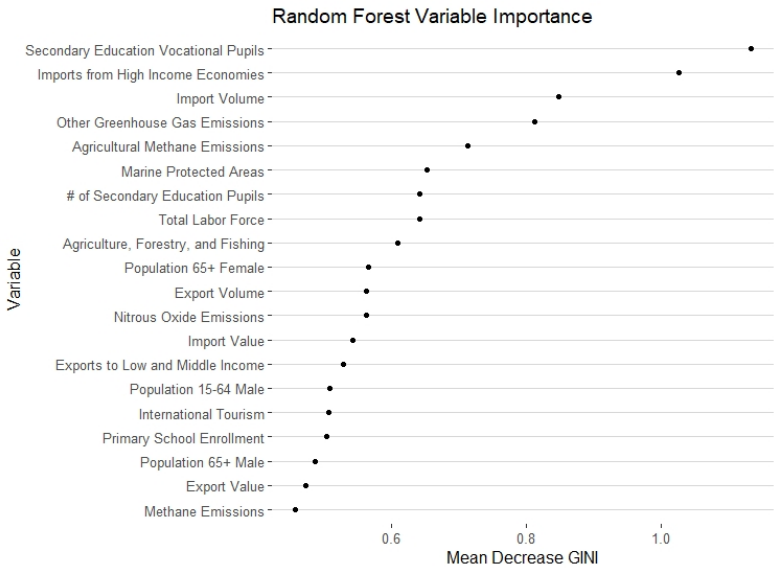

FIGURE 4: *The top 10 variables contribute significantly to the model. The remainder fade to obscurity rather quickly.*

To better understand the predictive behaviour of the random forest model, partial dependence plots were obtained for the two most important variables.

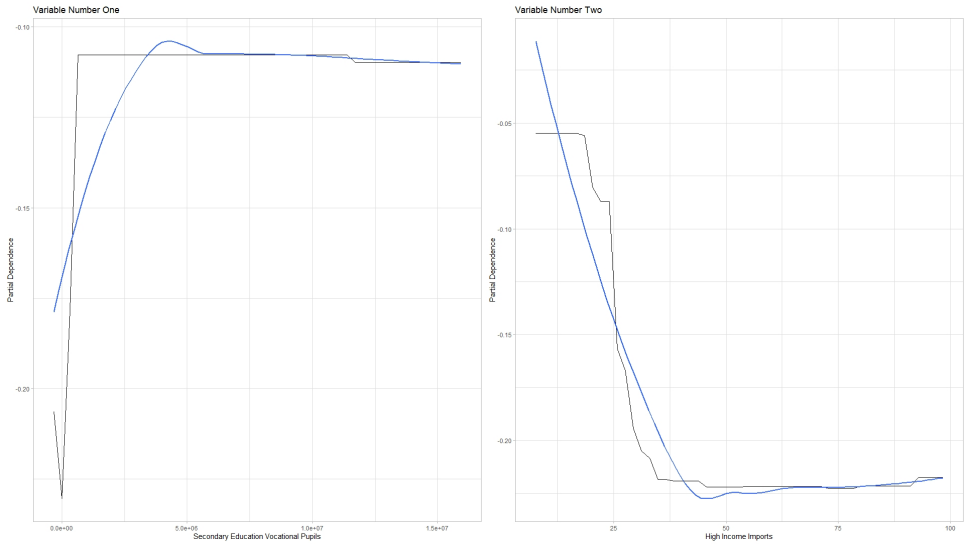

FIGURE 5: *There is an early sharp increase in the model's dependence on secondary education vocational pupils. Dependence on high income imports decreases as high income imports increase.*

Variables number two and three both measured some aspect of merchandise imports. Imports from high income economies were important, as was import volume. To assess the joint effect of these two variables on the random forest predictions, a multi-predictor partial-dependence plot was obtained. Holding one variable at a relatively low

value, the partial dependence is high while the other variable increases.

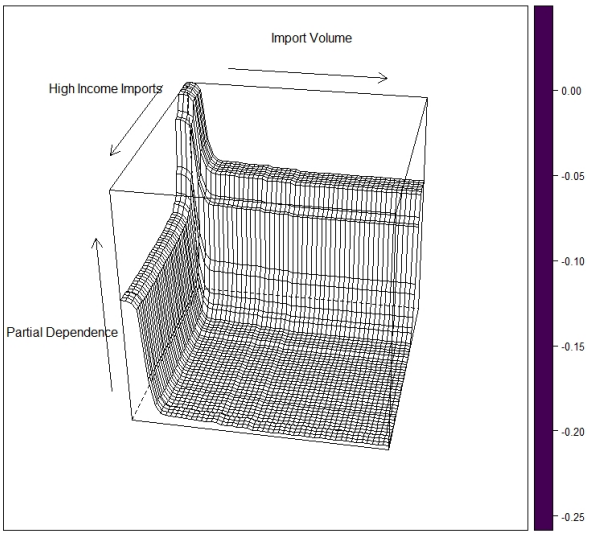

FIGURE 6: *The partial dependence is similar for both variables. The joint partial dependence decreases as high income imports and import volume increases, but the partial dependence is very high when values are low for each variable.*

To visualize the relationships between legislative structure and the most important variables according to the random forests, bee-swarm plots were produced for the top three. These plots showed marginal differences between classes (at best). The lack of obvious separation, suggests that there may not be an algorithm suited to make these predictions. Put differently, no matter how we choose to tune our model we may never significantly improve upon random guessing.

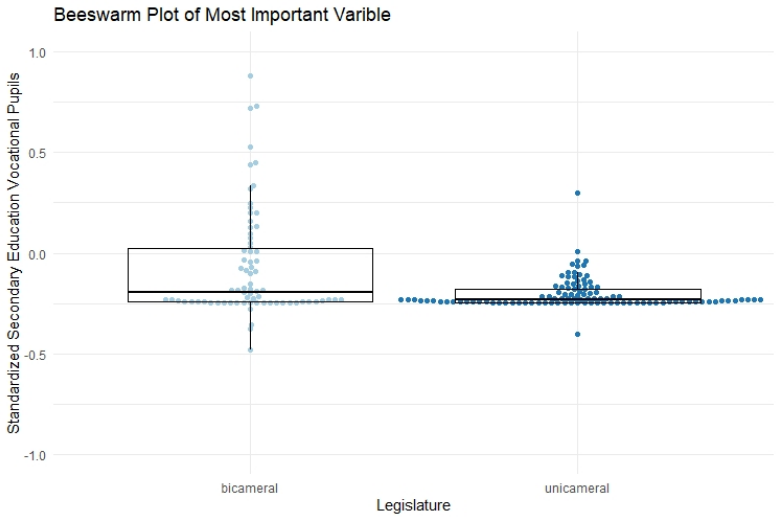

FIGURE 7: *There is no clear separation between classes for the most important predictor in the model.*

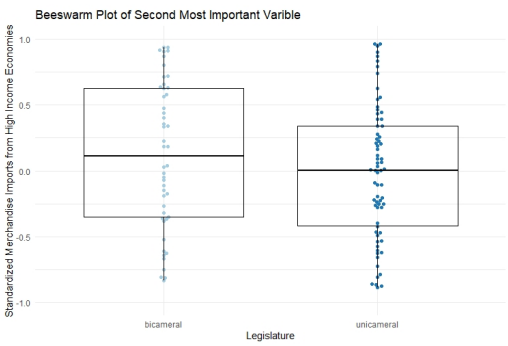

FIGURE 8: 2^{nd} *most important predictor*

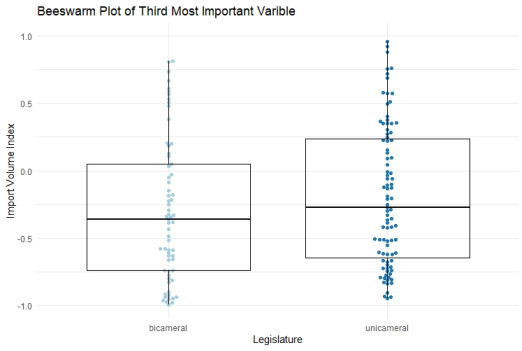

FIGURE 9: 3^{rd} *most important predictor*

Support Vector Machines

For our next test, an SVM model was fit to the data with no tuning. The radial kernel was used. Out of bag prediction accuracy was rather good; however, when cross validation was used model performance suffered. A table of accuracies is given for the model without tuning (table 3). Though this model had a lower sensitivity and percent correctly classified than did the random forest model, the specificity was about 8% higher. Still, 48.1% is no better than random guessing.

	% Correct	Sensitivity	Specificity	Kappa
Support Vector Machine	63.1	74.07	48.1	0.23

TABLE 3: *Cross validated accuracies for the SVM model.*

Using the **EZtune** package, optimal parameters for the SVM model were found. The model was refit with cross-validation. Interestingly, after several iterations of the SVM model with parameter adjustments, the tuned model was consistently outperformed by the model that was not tuned. The model's abysmal performance is not unexpected. Support vector machines rely on data with many more observations than we have available. An SVM model suffers when there is insufficient training data and our data has fewer observations than variables. To determine if the kernel had an influence on prediction accuracies, an SVM model was fit with four different kernels. Table 4 summarizes the models. None of the models performed better than guesswork.

Our inability to tune a model that performs better than guesswork forces us to conclude that support vector machines do not adequately describe the data.

Kernel	% Correct	Sensitivity	Specificity	Kappa
Linear	55.61	72.22	32.91	0.05
Polynomial	56.68	83.33	20.25	0.04
Radial	58.29	74.04	36.71	0.11
Sigmoid	52.94	82.41	12.66	-0.05

TABLE 4: *Cross validated accuracies for four SVM models with optimal parameters and variable kernels.*

Classification Tree

One final model was built to test the power of world development indicators to predict legislative structure. To find the optimal model parameters, a tree was grown with 1 as the minimum number of observations required to make a split. Figure 10 provides a visualization of the cost-complexity parameter's behaviour as the tree grows.

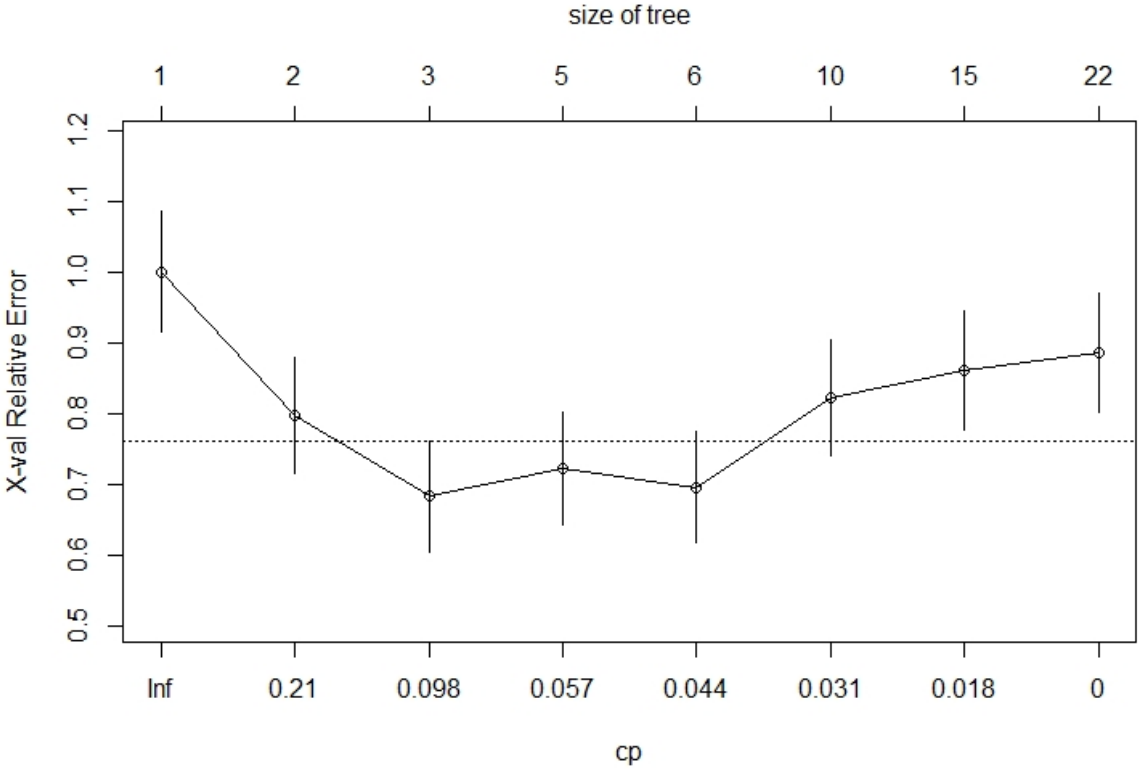

FIGURE 10: *The relative error is minimized when cp = 0.098 and 0.044*

Because there was some ambiguity concerning the optimal *cp* value, two models were fit. The first model used $cp = 0.098$ while the second $cp = 0.044$. The former had a high specificity and low sensitivity while the latter had roughly equal accuracy between the two classes.

	% Correct	Sensitivity	Specificity
$cp = 0.098$	50.8	34.18	62.96
$cp = 0.044$	47.59	49.37	46.3

TABLE 5: *The model which used* $cp = 0.98$ *was superior.*

With just three terminal nodes, the classification tree algorithm was able to make predictions that were about as good as random guessing. Figure 11 provides a visualization of the pruned tree. Predictably, the two most important splits were made on the top two variables according to the random forest algorithm.

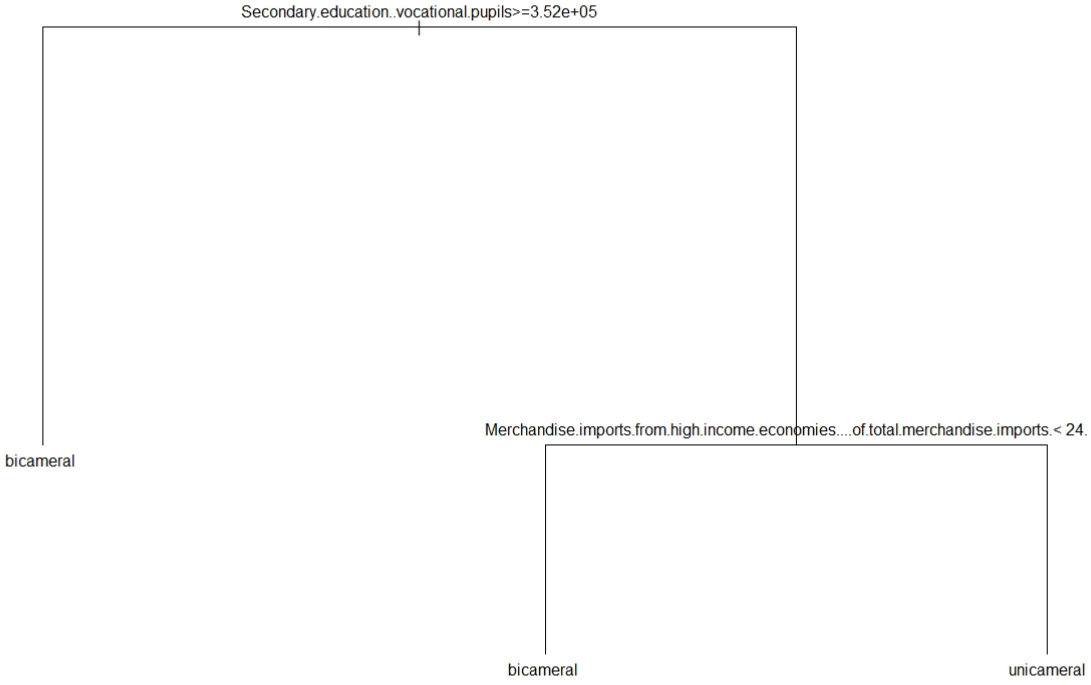

FIGURE 11: *The classification tree with* $cp = 0.98$ *performed best; however, it was still unable to make predictions that were better than random guesses.*

There seems to be a trade-off between sensitivity and specificity. Models that perform well classifying bicameral legislatures suffer when classifying unicameral legislatures. These results indicate that there may be no significant differences between the two legislative structures in terms of policy outcome.

Corruption Perceptions Index

The Corruption Perceptions Index attempts to measure public sector corruption on a scale from 0-100. High scores indicate high governmental transparency, while low scores indicate frequent abuses of power. The index itself is informed by 13 expert surveys compiled by various institutions [3].

The goal of this analysis is to determine if a nation's CPI score can be predicted given the collected World Bank data. If so, a secondary objective is to determine which indicators are most useful for making this prediction.

Linear Regression

We tested linear regression on the dataset as a baseline.

Because of the number of features we have, we will omit the process of ensuring the models have normal residuals. Four linear regression models were fit:

- Linear regression of 537 features
- Linear regression of 30 principal components
- Tuned LASSO regression of 537 features
- Tuned LASSO regression of 30 principal components

Table 6 summarizes the results. Mean R2 values increase and R2 value standard deviations decrease on both datasets when LASSO regularization is used. Both LASSO models achieved the highest R2 score when the regularization coefficient is 1.0.

At this level, 35 and 23 features are kept for the 537 feature LASSO model and the 30 PCA feature models, respectively. This translates to $35/537 = 6.5\%$ and $23/30 = 76.7\%$ of features being kept. Figure 12 shows the 10 features corresponding to the greatest 10 LASSO coefficients.

Interestingly, the coefficient with the highest magnitude is customs clearing efficiency. This feature relates to trade and therefore indirectly to the economy of a country. Efficient market economies have historically had relatively high quality governmental institutions. Still, this is a stretch of reasoning to justify such a high reliance on a seemingly unimportant feature. Most of the remaining top features related to education and healthcare

spending, which are associated with better democratic outcomes. Export documentary burden is closely tied to customs clearance efficiency. These features must contain mostly redundant information, but the model must have seen enough discriminative power unique to each to select for both of them. The most surprising feature is the number fish species threatened.

Model	R2
LinReg	0.513 +- 0.191
LinReg (PCA)	0.686 +- 0.142
LASSO	0.787 +- 0.036
LASSO (PCA)	0.770 +- 0.072

TABLE 6: *R2 of linear regression models on CPI*

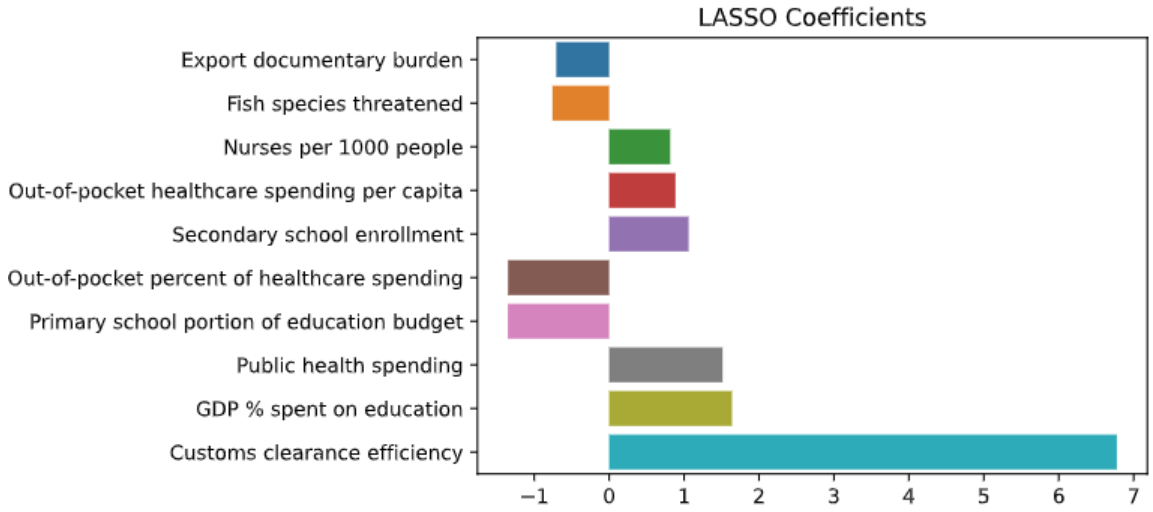

FIGURE 12: *Features corresponding to greatest magnitude LASSO coefficients*

XGBoost

We began by tuning an XGBoost (gradient booster implementation) regressor on the standardized World Bank indicators. We compare this model's cross-validated performance to another tuned XGBoost fit to the first 30 principal components, which explain 80% of the data variance. Table 7 shows the results.

Model	R2
XGBoost	0.771 +- 0.048
XGBoost (PCA)	0.722 +- 0.055

TABLE 7: *XGBoost trained on 537 raw features outperforms XGBoost train on 30 PCA components*

Although tuned, both XGBoost models are just on par with both of the linear regression LASSO models.

Figure 13 shows the 10 features that provided the highest average information gain when fitting XGBoost. Interestingly, 3 out of 10 are measures of health spending per capita. The LASSO model also had healthcare spending and customs efficiency features in its top 10. The biggest difference in important features between LASSO and XGBoost are the population and air pollution features used by XGBoost.

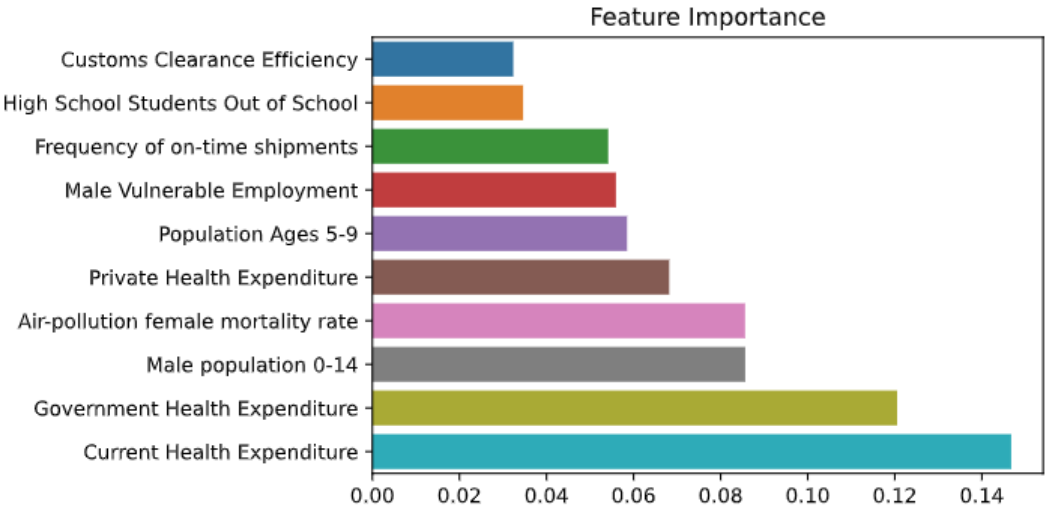

FIGURE 13: *Top 10 features*

Figure 14 shows that CPI is a log of health spending per capita. In fact, countries spending almost $0 per capita on health rank tend to have CPI scores of 0-50. Only when spending reaches about $1200 per capita do countries start having CPI scores of 60 or higher.

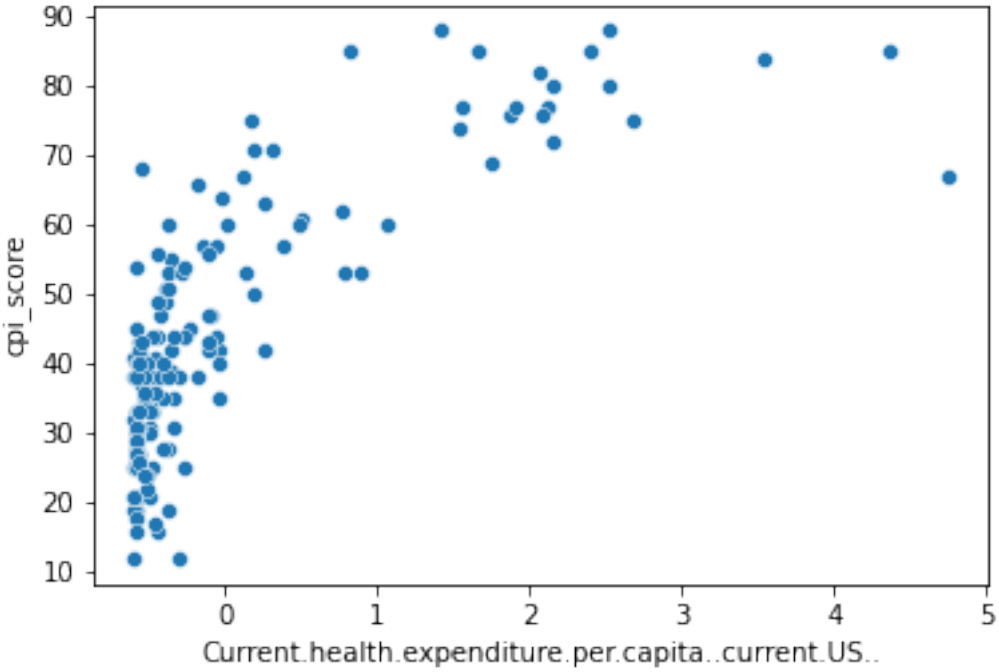

FIGURE 14: *CPI is logarithmically related to health spending per capita*

Income tends to be strongly positively related to quality of government institutions. Because our dataset did not directly include per capita income, health spending might be acting as a proxy for income. Higher income leads to higher spending in general, including on private healthcare. Accounting for tax rates and government budget policy, higher per capita income is also related to higher government tax revenue and public healthcare spending. Considering the sum of private and public healthcare spending equal total spending, it is confusing why all three features are relied on so heavily by XGBoost. One would think the model would use the most informative of the three and discard the others as being redundant. Table 8 shows the correlations between the three health spending features.

	Private	Public	Total
Private	1.000000	0.647185	0.842913
Public	0.647185	1.000000	0.955659
Total	0.842913	0.955659	1.000000

TABLE 8: *Correlations between private, public, and total healthcare expenditure*

In comparison, the XGBoost model trained on the first 30 principal components relied overwhelmingly on the first principal component, as shown in figure 15. Looking at

the first component's coefficients ordered by absolute value, we see many features with coefficients of 0.07. No one component, or even handful of components, dominates the first principal component. Considering we have over 500 features and each nation's situation is very unique, that the principal component is equally comprised of many features isn't too surprising.

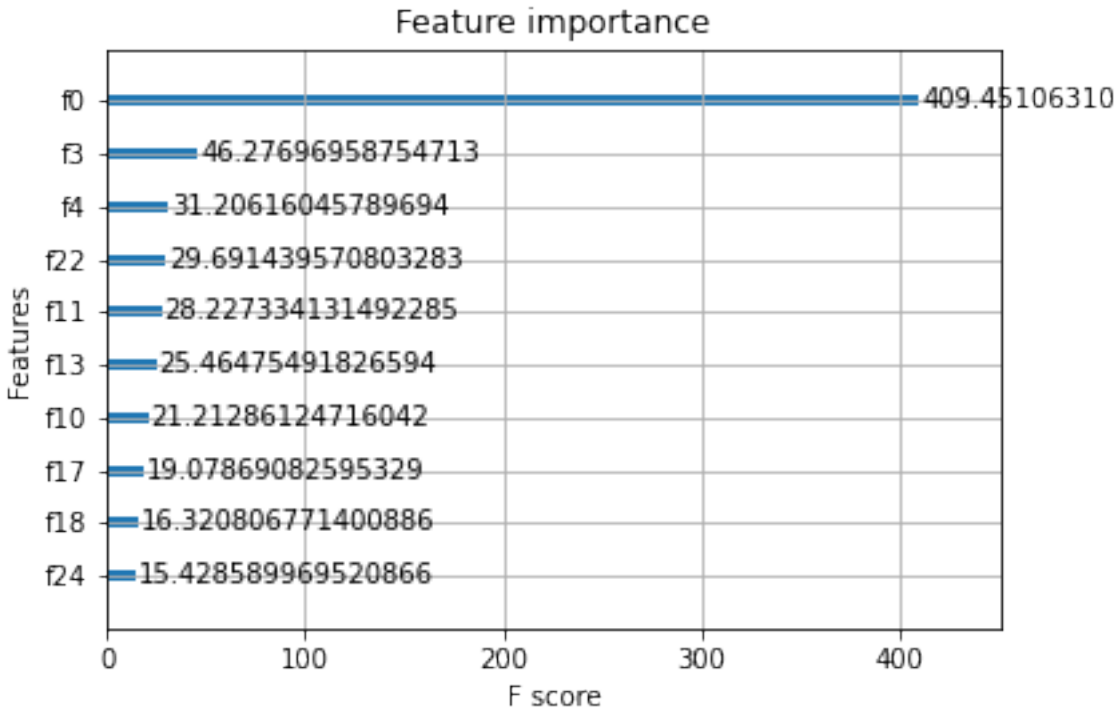

FIGURE 15: *Top 10 PCA features*

Random Forest

The same procedure was followed for random forest regressors as for XGBoost. Performance values are shown in table 9. Again, the random forest regressors are on par with LASSO and XGBoost at predicting CPI, all of which do better when given the non-reduced dataset. Figure 16 shows the top 10 features in terms of average information gain. Comparing the top features from XGBoost and the random forest, we see that 5 features are shared, while the remaining 5 measure similar factors. Again, we hypothesize that health expenditure is being used as a proxy for per capita income.

Model	R2
RandomForest	0.751 +- 0.059
RandomForest (PCA)	0.728 +- 0.040

TABLE 9: *RandomForest achieves higher CPI score prediction R2 when given full dataset*

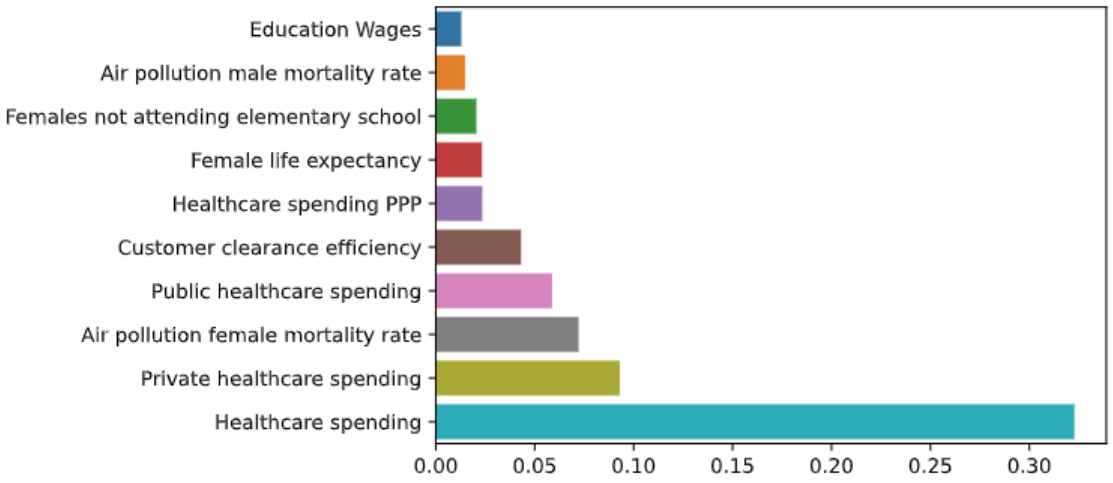

FIGURE 16: *Top 10 features for random forest regressors*

Figure 17 shows a partial dependence plot of healthcare spending on CPI score side-by-side with a healthcare spending vs. CPI scatterplot. The PDP shows a sharp increase in mean CPI prediction around the same level the scatterplot jumps in CPI, $1200.

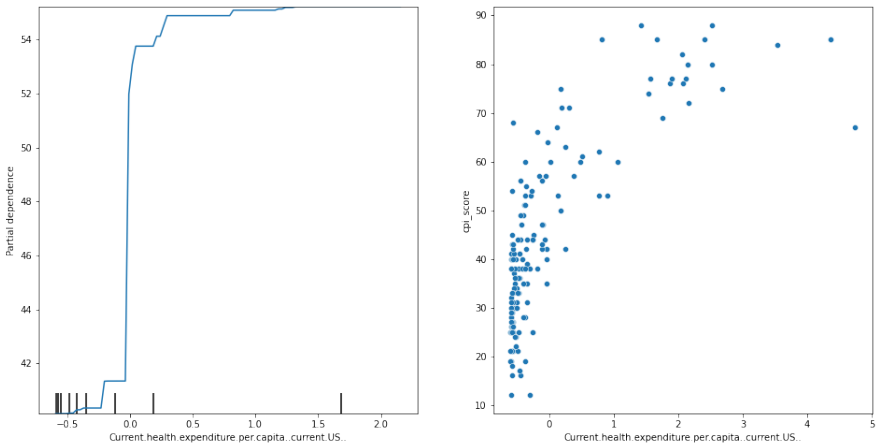

FIGURE 17: *Partial dependence plot of healthcare spending on CPI*

Freedom Score

The freedom scores for countries are from the reports published by the Freedom House. These look at individual freedom for political expression, expression of rights and beliefs, freedom of the press and media, political and civil liberties. Based on the scores the countries have also been classified into 'Not Free', 'Partly Free', and 'Free'. For this response variable, we obtained data for 170 countries, and 537 predictor variables.

Random Forests

We fit random forests to the data, using all the predictor variables with freedom score levels as the response. Table 10 provides summary metrics using cross validation for the model. The model performed poorly at classifying 'Not Free' countries at an abysmal 50 percent error rate.

	% Correct
Random Forests: all predictor variables	73.68
Random Forests: with 5 top variables	71.34

TABLE 10: *Cross validated accuracies for the random forest model.*

The plot of variable importance for the random forest model shows a very low decrease in mean accuracy for even the relatively most important variables. Refugee population, health expenditure, ambient air pollution, seemed important for the classification. Health expenditure, both government and private seemed important, with three of these repeated in the top fifteen variables. Economic factors that influenced the model, include customs clearance, goods and services. Air pollution and mean annual exposure at household and ambient levels showed four times in the fifteen. Some other variables include children out of primary school, and prevalence of anemia maybe again related to overall health sector. It is interesting that many of the variables are related to each other or from similar categories, whereas we would have expected to see different categories showing up. All of these variables seem to be influencing the model only marginally.

To better understand the predictive behaviour of the random forest model, we got partial dependence plots for the two most important variables.

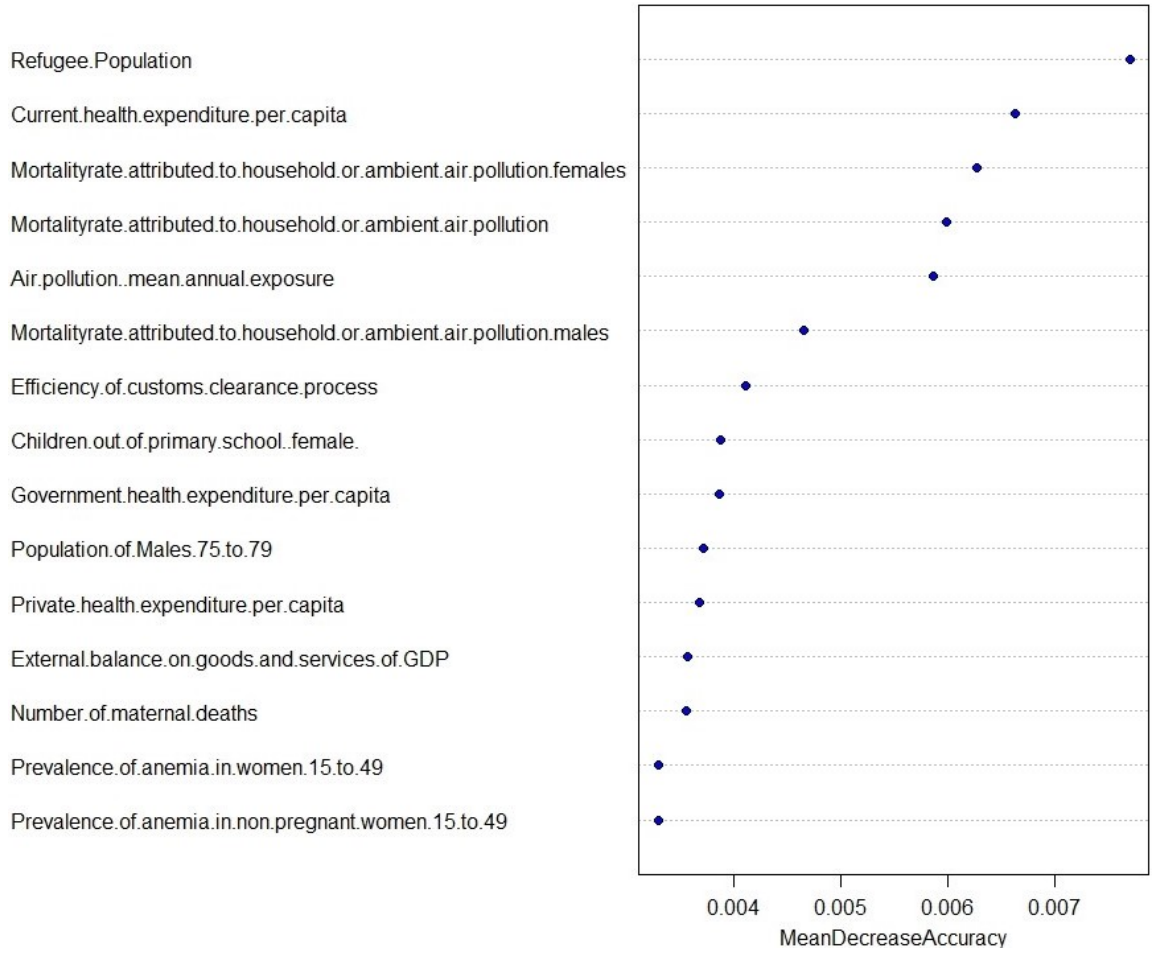

FIGURE 18: *The top 15 variables that contribute to the model, but not significantly much*

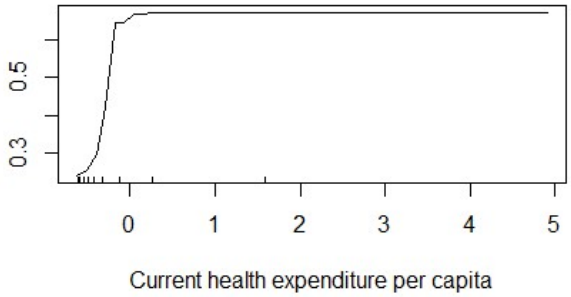

FIGURE 19: *Partial dependence plot of per capita health expenditure on freedom levels*

Support Vector Machines

For our next test, we fit SVM models using different kernels. The linear kernel had a much higher accuracy than radial, with no tuning. an SVM model was fit to the data with no tuning. The radial kernel was used. Whereas out of bag prediction seemed great, we got really low cross validation accuracies for both kernels. A table of out of bag and cross validated accuracies is given for the model with and without tuning for the two kernels (table 11). The SVM has a lower cross validated accuracy than the random forest

model. The linear kernel did perform much better than the radial kernel here.

Kernel	% out of bag	Cross-validated	Tuned
radial	93.56	52.63	38.01
linear	99.82	66.19	40.93

TABLE 11: *Cross validated accuracies for the SVM model.*

We tuned the SVM using **EZtune**, with freedom scores as levels using the radial kernel. The tuned model accuracies were much lower than the not tuned one, at 38% lower than guessing. The linear kernel performed a bit better but nothing to note. Support vector machines do not seem like a model choice for this data, with such lower cross validation accuracies.

Gradient Boosting

We tuned gradient booster to the predictor variables, for freedom scores. The ten most important variables are similar to the random forest model with refugee population and general government health expenditure seeming to be the top two variables. The top variable that shows up is air pollution, with a relatively much higher influence than the rest.

Indicator variables	**Relative influence**
Air pollution	10.876
Refugee population	7.655
Government health expenditure	3.354
Male population 70 to 74	2.66
Mortality rate from household pollution	2.255
Prevalence of anemia 15 to 49	2.01

TABLE 12: *Relative influence of the predictor variables*

The partial dependence plot with GBM for air pollution indicates a decrease in freedom scores with mean exposure to air pollutants. The air pollution index seems to be correlated with freedom scores at around 0.67. The relationship between these two variables do not make a lot of sense, unless air pollution is indicative of some other living conditions.

Conclusion

We were unable to build a viable model predicting legislative structure from world development indicators. The best model was built with the random forest algorithm and had a classification accuracy of 67.38%, but sensitivity was increased at the expense of sensitivity. Our inability to classify legislative structure suggests that there may be no relationship between policy outcome and legislative structure. There are other attributes of government that may be predictable. Future models could test theoretical claims of legislative efficiency. A model that classifies the executive structure may also be useful.

We were able to get better performance on the Corruptions Index regression task. LASSO linear regression, XGBoost, and random forest models were tested. LASSO achieved the highest R2 value among the models of 0.787, although the other models did nearly as well. The LASSO model selected for only 6.5% of features from our 537 feature dataset, indicating large amounts of redundancy throughout the set. One reason why many features were found to be insubstantial may be that the dataset had over twice as many features as examples. The fact we were not able to reach a higher R2 value illustrates how diverse the situations of individual nations are, despite having hundreds of indicators to describe them. Among these indicators, health spending, air pollution levels, and customs efficiency were among the most informative features for these models.

For the freedom levels, random forest models performed best with around 73% accuracy. The support vector machines were relatively not good at predicting the freedom scores. The random forests worked almost just as well with just the relatively more important 15 variables, but surprisingly with many of them being from similar categories such as pollution, health expenditure, health indicators, and population. The gradient boosting model with similar classification as random forest, also had air pollution and health expenditure as the most influential variables. It is interesting that similar variables seem to be informative for freedom and corruption indices.

One limitation of this work is the set of indicators we decided to include. In particular, we omitted data directly relating to the economy and income. Based on our model feature importance results, we believe including economic indicators would increase model performance on the CPI and freedom score regression tasks.

References

[1] Tom Todd. Unicameral or bicameral state legislatures: The policy debate. 1999.

[2] Thomas R. Dye and Susan A. Macmanus. Predicting city government structure. *American Journal of Political Science*, 20(2):257–271, 1976.

[3] Transparency International. Research. https://www.transparency.org/en/research.