Predicting Political Revolutions

Political upheavals have been ever present throughout humanity. Political leaders shape the context for everything we know. For that reason, we as a society need to generate a well-founded understanding of when a revolution or political change is imminent, as well as data-based indicators. This analysis seeks to forecast whether a given protest will lead to a revolution within one year using fundamental characteristics of the protest as well as metrics to understand the government in place at the time of protest.

The stakeholders for this analysis are wide reaching, but it is most directly relevant to political scientists. As a field focused on understanding the science of politics, including to inform decision making and strategy, a data science investigation into the topic of regime change remains highly relevant. Similarly, political organizers or government leadership could also gather insights with the potential to guide strategic decisions and determine where efforts are best focused and with the highest likelihood of creating an impact.

Repository Structure

```
│
├── data
│   ├── processed   <-- SQL files containing processed data
│   └── raw         <-- Original data
│       ├── Mass-Mobilization-Protests          <-- Data source #1
│       ├── Polity-Project                      <-- Data source #2
│       └── Database-of-Political-Institutions  <-- Data source #3
│
├── images                  <-- Figures used in presentation and notebooks
│
├── notebooks               <-- Jupyter Notebooks for exploration and
presentation
│
├── reference_material
│   └── data user manuals   <-- PDFs provided from data sources
│
├── report                  <-- Generated analysis summary
│
├── src                     <-- Custom functions used in notebooks
│
└── README.md               <-- Main README file
```

The main high-level Jupyter notebook for this project can be found here. The in-depth modeling notebook can be found here.

Project Structure

Given the complex nature of combining three separate datasets from distinct sources, the analysis is conducted in five different notebooks, and one additional notebook is used for summary. The first three focus on cleaning the data through data understanding, manipulation, feature selection, and feature engineering: cleaning_protests_dataset.ipynb, cleaning_regime_changes_dataset.ipynb, and cleaning_governments_dataset.ipynb. Each of these files exports the final resulting dataset into a SQL database, structured to cleanly join them all together.

The fourth notebook - MODEL.ipynb - joins these three SQL files and performs feature engineering that could not be done without aggregating the data from multiple sources. This excludes engineering the target, alongside multiple input features. Beyond these steps, the notebook contains all modeling conducted in this analysis. Since it contains data from all sources, **this notebook is recommended as the place to go for the largest exposure to strategic decision making, understanding the ways in which the data are connected, and the testing of model variations.**

The fifth analysis notebook, EDA.ipynb, is used for exploratory analysis of features.

The final notebook, FINAL_SUMMARY_NOTEBOOK.ipynb is a high-level summary of all other notebooks, extracting key points from each and discussing key findings.

Data Sources

The analysis combines three core datasets from different sources to provide a distinctly unique understanding of the subject. Together, they cover the time period from 1990 to 2020 and include 17,000 protests in 167 different countries. There is a total of 131 features available, covering a wide range of topics from government checks and balances to protest location, protester violence, and protester demands. Each dataset is described in more detail below.

The Mass Mobilization Project

The first dataset is described in the source documentation as "an effort to understand citizen movements against governments, what citizens want when they demonstrate against governments, and how governments respond to citizens. The MM data cover 162 countries between 1990 and 2018. These data contain events where 50 or more protesters publicly demonstrate against government, resulting in more than 10,000 protest events. Each event records location, protest size, protester demands, and government responses." (1) The project is sponsored by the Political Instability Task Force (PITF). The PITF is funded by the Central Intelligence Agency (CIA). (1) Throughout the analysis, this dataset is referred to as the "Protests" dataset.

Although the data source does specify that the dataset is not entirely comprehensive of all country across this entire time period, it does contain over 17,000 recorded protests, each composed of 31 features. The data span 167 countries from 1990 to 2020. Seemingly the only large country to be omitted is the United States, which is certainly not a coincidence and can undoubtedly be tied back to the source of the project funding.

Citation:

Clark, David; Regan, Patrick, 2016, "Mass Mobilization Protest Data", https://doi.org/10.7910/DVN/HTTWYL, Harvard Dataverse, V5, UNF:6:F/k8KUqKpCa5UssBbL/gzg== [fileUNF]

The Polity Project

The second dataset codes "authority characteristics of states in the world system for purposes of comparative, quantitative analysis." (2) "The Polity5 dataset covers all major, independent states in the global system over the period [1800-2020] (i.e., states with a total population of 500,000 or more in the most recent year; currently 167 countries. The Polity conceptual scheme is unique in that it examines concomitant qualities of democratic and autocratic authority in governing institutions, rather than discreet and mutually exclusive forms of governance. This perspective envisions a spectrum of governing authority that spans from fully institutionalized autocracies through mixed, or incoherent, authority regimes (termed "anocracies") to fully institutionalized democracies." (2). Most relevant to this analysis, "it also records changes in the institutionalized qualities of governing authority." (2). These changes in governing authority are the target feature of this analysis. The dataset contains 1,693 rows of data, each with 24 features. Throughout the analysis, this dataset is referred to as the "Regime Changes" or "Regimes" dataset.

Citation:

“The Polity Project.” PolityProject, Center for Systemic Peace, www.systemicpeace.org/polityproject.html.

The Database of Political Institutions

The third dataset is provided by the Inter-American Development Bank (IDB). "The Database of Political Institutions presents institutional and electoral results data such as measures of checks and balances, tenure and stability of the government, identification of party affiliation and ideology, and fragmentation of opposition and government parties in the legislature ... [it covers] about 180 countries [from] 1975-2020. It has become one of the most cited databases in comparative political economy and comparative political institutions, with more than 4,500 article citations on Google Scholar as of December 2020." (3) Within the timeframe of this analysis, the data source includes 8,200 rows of data, each with 77 features. Throughout the analysis, this dataset is referred to as the "Governments" dataset.

Citation:

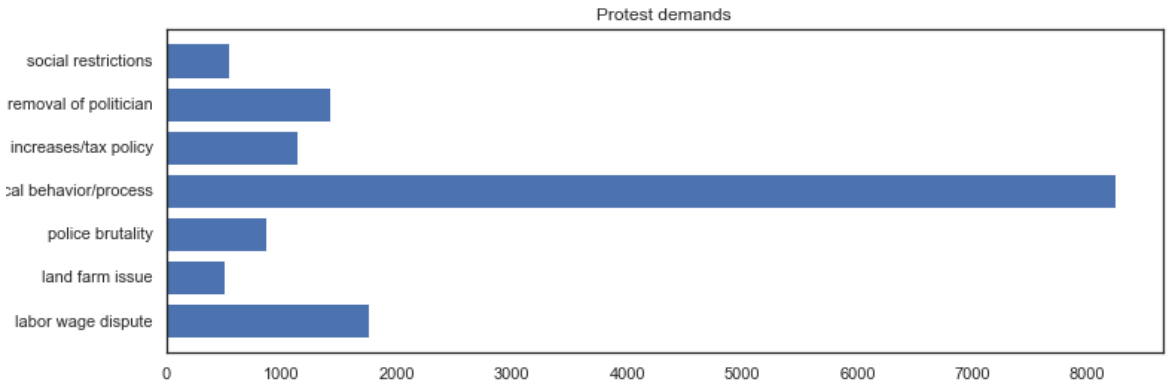
Cruz, Cesi, Philip Keefer, and Carlos Scartascini. 2021. Database of Political Institutions 2020. Washington, DC: Inter-American Development Bank Research Department. https://publications.iadb.org/en/database-political-institutions-2020-dpi2020

Data Analysis

Below is a breakdown of two features that frame the context for the protests: what they aimed to achieve and where they took place.

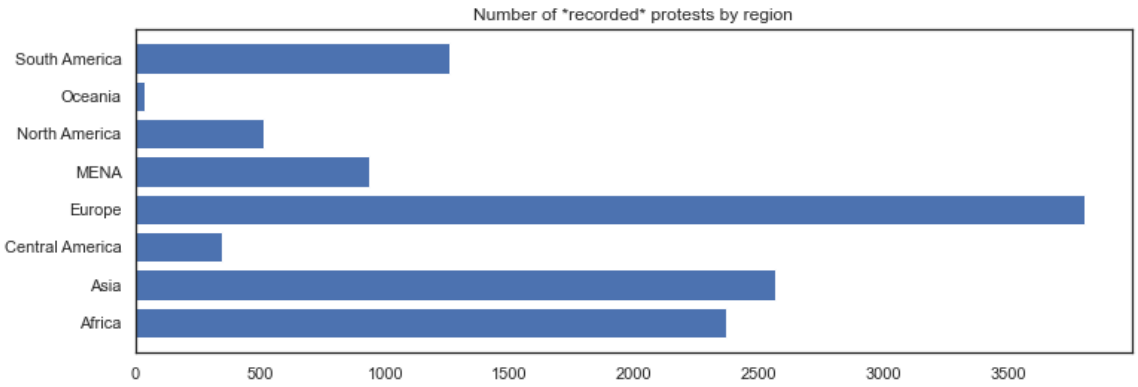
Understanding Protester Demands

The below chart provides insight into the categorization of demands from protesters over the time period of this analysis.



Understanding Protest Locations

Below is a geographical distribution of protests by region. Do note that the Protests data source explicitly excludes some countries, so this figure should not be construed as an understanding of *all* protests globally. Instead, it is the distribution within this dataset.



Modeling

The models explored include K-Nearest Neighbors (KNN), Bayesian classifiers, ADA boost, Decision Trees, Random Forests, and XG Boost. In addition, each type of model is constructed using elements of encoding, scaling, resampling and hyperparameter optimization.

- One hot encoding was essential given the categorical type of some features.
- Standard scaling was essential given the vast array of different numerical feature distributions and ranges. Min-max scaling was tested but proved less effective.
- SMOTE was determined to be essential given the imbalanced nature of the dataset. Only 11% of the target feature values were 1, leaving the other 89% as 0. This is a prime example of the need for resampling, and SMOTE proved highly effective.
- Hyperparameter grid searches are inherently valuable when optimizing a model. Appropriate hyperparameter searches were used for each model type.

The performance of each model is evaluated on four core statistical measures (f1 score, accuracy, precision, and recall), in addition to displaying a confusion matrix for the test data. F1 was selected before the modeling process as the most relevant metric given that it encompasses all possible outcomes, as opposed to the other three metrics which leave out at least one possible outcome from their evaluation.

Evaluation

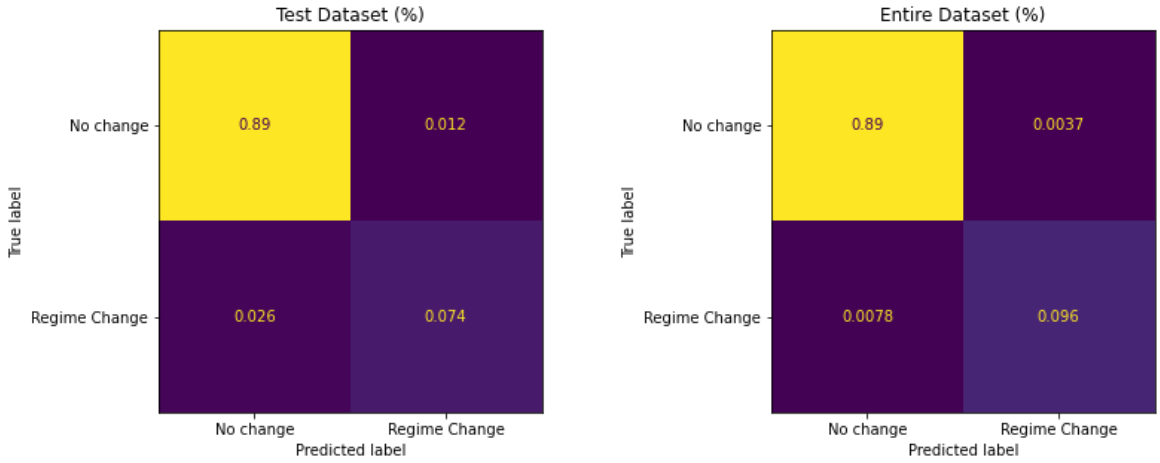
The XG boost model was found to perform the best all around. This decision was based on the F1 score, but the XG boost also outperformed other models on other metrics.

Evaluate performance on test data

The model has a strong performance on the test data:

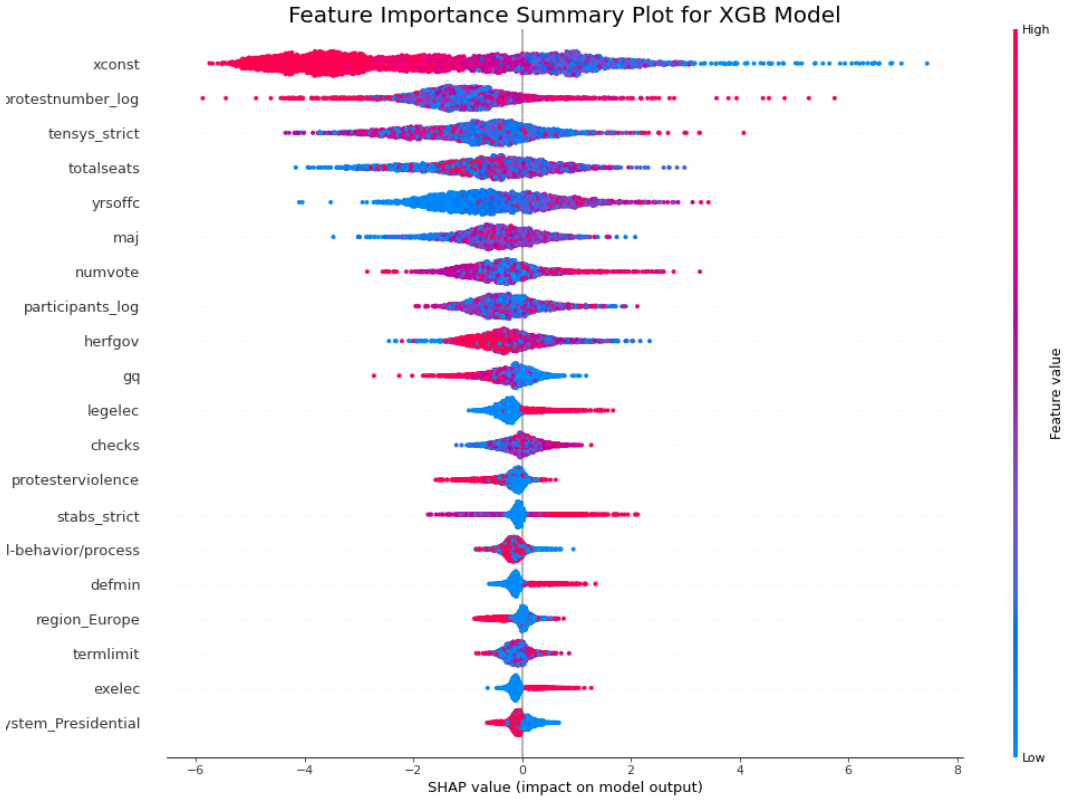
- F1 score: 0.80
- Accuracy: 0.96
- Precision: 0.86
- Recall: 0.74

The confusion matrix below shows the performance on the holdout data (left) and the performance on the full dataset (right), including train and test. In addition to overall performance, the performance discrepancy between the two plots show that overfitting is prevalent, though the performance on the test data remains sufficiently strong for a high performing model.



Feature Importance

Although XG boost models are notoriously difficult to extract meaningful feature importance data from, the plot below does provide an indication per the SHAP summary plot.



Here, we see that the five most significant features are:

1. **xconst**: presence of executive constraints, ranging from "Unlimited Authority" through "Executive Parity"
2. **protestnumber_log**: number of protests that already occurred in the year of the protest, log-transformed
3. **tensys_strict**: length of time the country has been autocratic or democratic
4. **totalseats**: total seats in the legislature
5. **yrsoffc**: number of years the chief executive has been in office

To bring some meaning to the SHAP summary plot, it is worth noting that:

1. Countries with low executive constraint ("Unlimited Authority") are more strongly correlated with a protest overturning the regime.
2. Higher turnouts to protests are associated with regime transitions. However, the data is somewhat split: even very small protests can be associated with regime transition.

3. There is a less distinct divide in the relationship between the length of time a country has been autocratic vs. democratic than in other metrics. That said, the feature remains a strong predictor.
4. Regime transitions happen somewhat consistently across the size of governing bodies.
5. Protests in countries with new executive leadership are less likely to lead to regime change than countries with a long-ruling leader.

Conclusion

Overall, this analysis successfully completes its objective. It creates and tunes a model that helps predict whether a given protest will lead to a regime transition within one year. This incredibly valuable tool can be used by political scientists to better understand regime transitions, including my investigating the model further than was conducted in this analysis. In addition, the model and its findings can be used for proactive or preventative measures by either side of government. With a very strong performing model, it can be trusted to give an accurate estimate of changes to come.

Going forward, this project allows for easy growth as more data is released. Each of the three primary datasets receive regular updates, and this new information can easily be incorporated in order to expand the temporal scope of the project and with more data comes to potential for stronger performance.

Next Steps

A powerful next step could be to investigate this data using time series analyses. Specifically, it would be valuable to understand how protest outcomes are affected by protests before it.

Languages

Jupyter Notebook 99.6% Python 0.4%