Sam Cohen
PPOL 5204
Professor Ziogas
5/5/2024

Final Project: Stage 5

Research Question

What factors predict instances of electoral violence? How can they predict the likelihood of electoral violence occurring?

Background

Violent disputes over elections are not a new phenomenon. Though the January 6th, 2021 insurrection at the US Capitol proved that even relatively stable, wealthy and developed states are not immune from electoral violence, there are many different causes, effects, and variables that determine the chance of violent events occurring.

Flores and Nooruddin, for instance, find that during times of civil unrest, incumbent political authorities are more likely to commit violent accounts during election cycles (Flores and Nooruddin 2022). Fjelde determines that political party strength also plays a role, with stronger political parties ensuring peaceful elections more often due to the lack of force needed to maintain legitimacy (Fjelde 2022). In the modern age where social media plays an ever-important role in daily life, online presence has major impacts on instances of electoral violence. In fact, Burch and Muchlinski et. al. found that neural networks helped accurately predict and categorize instances of electoral violence (Burch and Muchlinski et. al. 2020).

While previous research confirms the significance of electoral conflict and its causes, there are few studies that use machine learning and multivariate analysis to take into account the many variables that may play a role in violent election seasons. Determining the atmosphere and bellwethers around electoral cycle—which, in turn, could predict electoral violence-- may allow institutions and political actors to ensure free, fair, and safe voting practices and a stable democracy.

Data

The data used for this research comes from the Uppsala Conflict Data Program (UCDP) database. UCDP collects conflict data from events and conflicts spanning multiple decades, countries, conflict types, etc. The Deadly Electoral Conflict (DECO) data set includes observations on incidents of electoral violence from 1989 to 2017, and contains metadata on the country, geolocation, actors, violence type, and violence reason in question (UCDP). The set is made up of over 4,000 observations.

Accounting for additional variables that may play a role in election violence (e.g. GDP per capita, inflation, unemployment, etc.), I merged a data set of country-year development indicators from the World Bank.

a. Cleaning and preprocessing

While the data from both the DECO and World Bank data sets was relatively clean, several pre-processing steps were needed. For instance, because of the significant amount of categorical variables, one-hot encoding was used substantially. While this generally worked well for strictly validated categorical variables, there were several features that were categorical integers, some that could take on more than one value (e.g. "item1;item2…"), and others that had both of these characteristics. To generate these dummy variables, lists were created housing unique values for each of these columns separated by delimiter (usually a semi-colon), and a for loop was then employed to check whether or not a given item was located in the original data (1) or not (0). Results were then saved to a new dummy variable.

Some categorical variables were particularly messy. Election violence purpose (electoral_purpose) in particular presented a challenge— there were many different purposes listed, some combinations of purposes, and often, there were spelling differences and errors for observations. After creating a list of cleaned, stripped, and unique purposes, dummy variables were generated if the original column contained a given string in this list.

Because of the cleaning done by UCDP on the back end before preprocessing, there is little to no missing data. Only several columns had any instances of missing data, and none of these were used in the final analysis,

b. Preliminary analysis and descriptive statistics

The final data set contains 4,233 rows of data. Before preprocessing, the DECO data set contained 41 columns. After cleaning, one-hot encoding and merging with World Bank indicators, this increased to 183 columns. Of all 4,233 instances of electoral violence, over 2,233 led to at least one death, according to best estimates. The highest recorded number of deaths for any instance of electoral violence was 1,000 people. India had the highest number of instances of electoral violence (648), as well as the highest number of total estimated deaths (2,404). The Gambia had the lowest number of instances of violent events (1). Of the types of elections where violence took place, a plurality (932) were national parliamentary elections, followed closely by sub-national elections, and then presidential elections.

The most common purpose for electoral violence was to disrupt elections, followed by violence due to diverging affiliations, securing elections, and crackdowns on protests. More than 100 instances had an unclear motive.

The majority of electoral violence occurred before elections took place (2,398), post-election violence (1,032), violence concurrent with elections (522), and pre-and-post violence (237). 32 instances were unclear regarding timing of events.

The most common targets of electoral violence were voters and/or party supporters, the politician or candidate themselves, members of security forces, insurgents/soldiers, non-designated civilians, and election related protesters, in that order respectively.

Some of the most important features included unemployment, GDP per capita, inflation, and sub-national elections (dummy).

Model and Methodology

Uncertainty of electoral violence ("electoral_vio_uncertainty") was chosen as the target feature for my research. This variable assesses the uncertainty that a violent even should or should not be classified as electoral violence. It is measured on a scale of 0 to 2, with 0 being "low uncertainty," 1 being "some uncertainty" and 2 being "high uncertainty." My intent was to use the uncertainty of election violence as a conduit for predicting whether or not it will occur considering the other variables in the final data set.

I employed decision trees to help answer my research questions, though random forest models were used in the final iteration of my analysis. I chose these specific models due to the prevalence of categorical data, as well as the limited observations in this data set. My original intention was for the root node to be a country, with each sub-node sub-setting the data by further indicators, namely: type of election (i.e. federal, local, etc.), timing of event, uncertainty of electoral violence, intended targets, perpetrators of electoral violence, and, if possible, national indicators such as GDP per capita, Gini coefficient, and whether or not the individual or party in power is incumbent. These last several features are not provided by Uppsala, so further work would be required to merge this data onto the existing data frame.

Due to the limited data, variable types, and dimensionality, random forests were chosen as the final model to minimize overfitting provide more robust results. Both tuned and untuned models were tested for both the random forest and decision tree algorithms.

Results and Evaluation

Overall results

Given the multiple limitations within the pre-processed DECO dataset (dimensionality, limited observations, difficulty interpreting target variable, etc.), the decision tree algorithm delivered relatively robust results, with a 0.813 accuracy score on the test data.

Further, random forests (both tuned and untuned) were instituted to develop models less prone to overfitting. Both the unpruned random forest and random forest with hyperparameters performed better than the decision tree, with accuracy scores of 0.876 and 0.861 respectively.

Evaluation

Cross-validation for the decision tree models indicated a mean cross validation score of 0.614 for the unpruned decision tree model. While this is smaller than anticipated, omitting several unnecessary variables from the data set (year, dyad ID, already-one hot encoded variables like country of violence type, etc.) helped increase the score (up from 0.54 originally). Despite the lower-than-expected mean cross validation score, the unpruned model was able to deliver relatively promising results. A weighted average precision score of 0.81 was a welcome

sign: this indicates that the model could correctly classify positive instances around 81 percent of the time out of all true positives and false positives in the test data. In addition, the accuracy score for the unpruned model was 0.813, meaning that over 81 percent of test data was correctly classified by the model. Interestingly, this model had a much high precision rate for categories 0 and 2 (0.85 and 0.9 respectively) than it did for 1 (0.46). This shows that the model is better and predicting incidents of low and high violence uncertainty, but falls short when the instance of violence is not clearly related to elections. The Gini scores (that is, the "purity" score of each leaf) varied for terminal nodes: many had Gini scores of 0, but scores as high as 0.6 were reported. This shows that depending on the branch, final classifications might include many instances of incorrectly classified data.

Though less prone to overfitting, the pruned decision tree model did not perform as well as the unpruned model. With an average weighted precision score of 0.37, this model is only about half as good at predicting true values of the target variable as our unpruned tree. In addition, after omitting a max depth component to this tree, the second model developed much deeper branches, with a depth of over 20. Precision for 0 was 0.61, while precision for both 1 and 2 was 0.0. Even when the condition of a maximum depth of 10 was added, the pruned tree gave similar results. The "best alphas" in these models were 0.0005 for the unpruned tree with no max depth, and 0.0004 for the unpruned tree with a max depth of 10.

In order to minimize overfitting, taking randomness into account, and increase accuracy, I also opted to employ random forest models. After fitting an untuned random forest model to the data, an accuracy score of 0.876 was determined. This already surpassed the pruned and unpruned decision tree models in terms of both precision and accuracy. After hyperparameter tuning, the accuracy dropped slightly to 0.861 (due to the model theoretically being less overfit). A confusion matrix illustrated that much like the previous decision tree models, the tuned random forest was much better at classifying 0s (low uncertainty of electoral violence) than either 1s (medium uncertainty) or 2s (high uncertainty). This model was better at classifying 2s than 1s as well. In terms misclassifications in the random forest model, 1s were classified as 0s at a higher rate than they were classified as 1s.

After fitting the random forest models, feature importance was then analyzed. The top 20 features in terms of importance, as well as their importance scores, were generated. It was found that some of the most important features were unemployment (the most important), followed by the South Africa dummy variable (i.e. if the violence occurred in South Africa or not), GDP per capita, electoral type 1 and 3 (i.e. parliamentary and sub-national electios), and Gini index. Given these features, I created a new pruned random forest using only these features as inputs. In this final model, accuracy was 0.85, and precision was 0.79. This model was also much better at classifying 1s as 1s instead of 0s, a significant issue in previous models.

Was research question answered?

The generated decision trees and random forest do indeed give researchers a clue into some of the most important variables to consider when predicting election violence. For example, the variable and split for the root node for the decision tree is GDP per capita (<= 18.02 USD). This alone is incredibly telling: it shows that in our sample, elections in countries and years where GDP per capita is low is one of the most important indicators of whether violence in and around election cycles will be classified as electoral violence. In addition, with unemployment having the highest importance scores, this variable alone may be the most telling when it comes to classifying electoral violence.

Though the models generated promising outcomes, their ability to classify election violence in real-world scenarios is still difficult to determine. The model may succeed at classifying violence around elections with low, medium, or high uncertainty, but it cannot predict whether or not election violence will or will not occur as a result of an election. However, they do illustrate many of the factors that can lead to difficulty determining the likelihood of a violent event being related to elections.

Reflections

Limitations

Perhaps the greatest limitation is the number of observations in this data set. This data set only has 4,233 observations, so conclusions drawn from results must be taken with a grain of salt. The use of decision trees may also exacerbate this problem, as they are quite prone to overfitting, (the use of random forests helped ease many of these concerns, however). Future iterations of this model should therefore include updated and additional data (whether from DECO or not).

Another factor to take into account is that each observation in this set is associated with a specific date and year. Though these features were omitted from the final data set before train-test splitting, the existence of autocorrelation and its effect on the variance in our model must not be underestimated.

When it comes to performance, the results from both the pruned and unpruned decision trees are a bit off bewildering. One unexpected outcome was the fact that the pruned decision tree actually performed worse than the unpruned when taking into account precision and complexity. In addition, the splitting rules for some of the branches are much greater or smaller than some of the data in the features they are representing. For example, though most features in this set are dummy variables, splitting cratering may indicating branching if a feature is greater than or equal to 1,000.

Finally, as previously stated, given the interpretation of the target variable, the decision tree models may be better as classifying violence as being related to elections, rather than predicting whether or not election violence will occur. The random forest models may have better accuracy scores and may have less bias associated with them, but having more data to feed this model would give us a better picture.

Learning Curve vs. Objective

The learning curve regarding my model and its performance relates to my prior understanding of the benefits and drawbacks of the use of decision trees and the benefits of random forests. For instance, I knew from the start that decision trees are prone to overfitting. Figuring out how to overcome this obstacle (especially with such a small data set) became a key part of this project. This is why I decided to include random forests in the final stage of my analysis.

In addition, the target variable measure is quite different than what I intended to predict. Specifically, the target variable (election violence uncertainty) reflects the likelihood that an instance of violence is directly related to a specific election. While this is certainly important, it is not a reflection of what I intended to predict: whether or not certain conditions bring about election violence. This model is therefore well-suited to classify violence as election-related, but may not be entirely useful for predicting whether or not election violence will occur given certain conditions.

Substantive outcomes regarding the topic chosen

In terms of substantive contributions and outcomes to the subject matter at hand, the models definitely help illustrate the importance of multiple different features when considering how violence occurs in response to elections. For instance, considering that unemployment is one of the most important features in this model, it is crucial to consider when taking into account the chances of electoral violence erupting. Abasili and Ogu, for instance, researched the link between unemployment and electoral violence, and believe youth unemployment is a primary driver of electoral violence in Nigeria (Abasili and Ogu 2023).

The timing of violence is also an important feature, and much research has been done (and still needs to be done) on this topic. Burton et. al. believe that post-election electoral violence is often more focused on anti-incumbent mass protest, whereas violence before elections often results in incumbent victories (Burton et. al. 2017).

The target of electoral violence, another important feature in the model, is also crucial variable to consider. From their research in Zimbabwe, Daxecker and Rauschenbach find that

violence intended to demobilize opposition voters is a widely used and popular tactic in areas where strong anti-incumbent allegiance lay (Daxecker and Rauschenbach 2023).

Future directions and lessons

Future improvements to this model may come from the use of additional features and indicators that previous literature has associated with electoral violence. For instance, according to elections expert Jeff Fischer, countries that are "freer" tend to have less bouts of election violence: countries designated as "partly free" or "not free" constituted the vast majority of places where election violence occurred in the early 2000s (Fischer 2004). Adding in a Freedom Index feature may therefore be beneficial for future models. Other factors, such as polarization, also play a role in elections and associated violence: in his research on elections in Burundi, Sterck finds that areas with high political polarization are more likely to experience pre-election violence (Sterck 2019). Other factors, such as median age, system of governance (i.e. presidential, parliamentary, monarchy, etc.), degree of centralization (federal vs. unitary), and institutional strength all play a role as well, and their addition into the model would be a welcome one.

Additionally, future enhancements to the model, as well as future research on election violence in general, should be much more context dependent. Factors that may cause pre-election violence in Kenya may be completely different from those which increase the likelihood of post-election violence in the United States, for example.

Finally, future machine learning algorithms that can aid in stopping electoral violence should focus more on predicting election violence, rather than classifying it. Doing so would make future models preventative, instead of solely classificatory. However, this is not to say that the original model here will not be useful: algorithms like these can help determine if violence occurring around election is a direct result of said elections.

References

Abasili, Kingsley & Udeoba, & Ogu, Ogechukwu. (2023). Analyzing the Nexus between Youth Unemployment, Poverty, and their Impact on Electoral Violence in Nigeria. Nnamdi Azikiwe University https://www.researchgate.net/publication/374542768_Analyzing_the_Nexus_between_Youth_Unemployment_Poverty_and_their_Impact_on_Electoral_Violence_in_Nigeria

Bakumenko, Stefan. (2022). Kenya's Electoral Violence: Conditions, Challenges, and Opportunities. Wilson Center. https://www.wilsoncenter.org/blog-post/kenyas-electoral-violence

Birch, Sarah & Muchlinski, David. (2017). Electoral violence prevention: what works? Democratization. https://www.tandfonline.com/doi/full/10.1080/13510347.2017.1365841?src=recsys

Birch, Sarah & Muchlinski, David (2020) The Dataset of Countries at Risk of Electoral Violence, Terrorism and Political Violence, 32:2, 217-236, DOI: 10.1080/09546553.2017.1364636. https://www.tandfonline.com/doi/full/10.1080/09546553.2017.1364636

Byman, Daniel & Clarke, Colin. (2022). Why the risk of election violence is high. Brookings Institution. https://www.brookings.edu/articles/why-the-risk-of-election-violence-is-high/

Daxecker, Ursula & Rauschenbach, Mascha. (2023). Election type and the logic of pre-election violence: Evidence from Zimbabwe. Electoral Studies. https://www.sciencedirect.com/science/article/pii/S0261379423000057#sec6

Fisher, Jeffrey. (2004). A Framework for Analysis and Resolution: Electoral Conflict and Violence. Elections During Conflict. https://ciaotest.cc.columbia.edu/olj/et/et_v12n1/et_v12n1b.pdf

Fjelde, H. (2020). Political party strength and electoral violence. Journal of Peace Research, 57(1), 140-155. https://doi.org/10.1177/0022343319885177. https://journals.sagepub.com/doi/full/10.1177/0022343319885177

Flores, T. E., & Nooruddin, I. (2023). Why incumbents perpetrate election violence during civil war. *Conflict Management and Peace Science*, *40*(5), 533-553. https://doi.org/10.1177/07388942221120382. https://journals.sagepub.com/doi/10.1177/07388942221120382

Hafner-Burton, Emilie M. and Hyde, Susan D. and Jablonski, Ryan S. (2016) Surviving elections: election violence, incumbent victory, and post-election repercussions. British Journal of Political Science. https://eprints.lse.ac.uk/64957/13/Jablonski_Surviving%20elections.pdf

Muchlinski, D., Yang, X., Birch, S., Macdonald, C., & Ounis, I. (2021). We need to go deeper: measuring electoral violence using convolutional neural networks and social media. *Political Science Research and Methods*, *9*(1), 122–139. doi:10.1017/psrm.2020.32 https://www.cambridge.org/core/journals/political-science-research-and-methods/article/we-need-to-go-deeper-measuring-electoral-violence-using-convolutional-neural-networks-and-social-media/808197D6EDA72699670325B6320EA5B9

Sterck, Oliver. (2019). Fighting for Votes: Theory and Evidence on the Causes of Electoral Violence. Economica. https://onlinelibrary.wiley.com/doi/abs/10.1111/ecca.12321