# Machine Learning HW 1

Kayla Kahn

March 2021

Hou, Gaibulloev, and Sandler (2020) introduce a new dataset on terrorist groups, based on the Global Terrorism Database (GTD; START 2019). They explain where the data came from and how they cleaned it. Blomberg, Gaibulloev, and Sandler (2011) linked Jones and Libicki's widely used 2008 dataset. The 2020 paper/dataset expands the 2011 data to 760 groups and additionally the dataset is group-year panel data instead of cross-sectional group data. They provide many group characteristics such as the years a group started and ended, group orientation and goals, and different measures of lethality and production such as counts of attacks, fatalities, and injuries. The article gives an overview of trends, a discussion on how the data can be used, and finally, they use logit models to assess factors explaining group failure and negative binomial models assessing fatalities and attacks.

The model that I replicate is the first model in Table 4 of the 2020 paper. This is a logit model for group failure. The explanatory variables that are group characteristics are a variable for duration and duration squared, group orientation - left, right, or national, with religious as the reference category, the proportion of attacks that are transnational out of a group's total attacks, the attack diversity, and the number of bases. Country level characteristics based on the groups base are the log of population, polity and polity squared, ethnicity of a country and ethnicity squared, dummy variables for region with MENA as the reference, and the percent of a country that has elevation and tropics and a dummy for whether the country is landlocked. My replication can be seen in the first column of Table 1.

For part two I split the data into 70% training and 30% testing and reran the model on the training set. Then I tried to find a model specification that predicted better than the original. I had trouble cross validating the models because the outcome is so rare. Even when looking at the precision and recall as metrics, there were still issues with the model trying to predict the same outcome for every observation. I therefore used synthetic sampling in order to be able to cross validate within different models. I also knew that at the end, I would be using the test set to verify the results without any synthetic sampling so I was not too worried about biased introduced from sampling this way. The model I chose performs very slightly better when synthetic sampling is included. For this model, I dropped regional dummies.

The original model has recall .536 and precision .087 and the new model

has recall .524 and precision .091, so there was a trade off and it only slightly improved but the model is still predicting terribly. I do not think that the change in predictive performance is due to it being misspecified. I removed the regional dummies because the model already had variables that I think capture the base country well enough: polity and polity squared, ethnicity and ethnicity squared, elevation, tropics, and landlocked so I don't think it lost very much information and might be getting closer to the true data generating process, although I would like to better understand how to deal with rare events and see if I can improve on this.

I compared the original model and new specification by predicting the test data. Both models perform terribly and I am sure that this is due to the rare events and not dealing with them correctly. The positive rate for the original model on the test data is .086 and on the new specification is .085. I then reestimated the new specification with the full data. Figure 1 shows the precision-recall curves with the test data. The area under the curve is low, showing that the models are not predicting well. Table 1 shows the coefficients of the original specification and new specification. In the new specification, the logged population of base countries gains both magnitude and significance, suggesting that a large population leads groups to be less like to end. The polity and polity squared variables are in the same direction with a similar magnitude but they gain significance, while ethnicity and ethnicity squared lose significance. Group orientation as compared to the reference category religious, the proportion of attacks that are transnational, attack diversity, and duration variables remain the similar in magnitude.

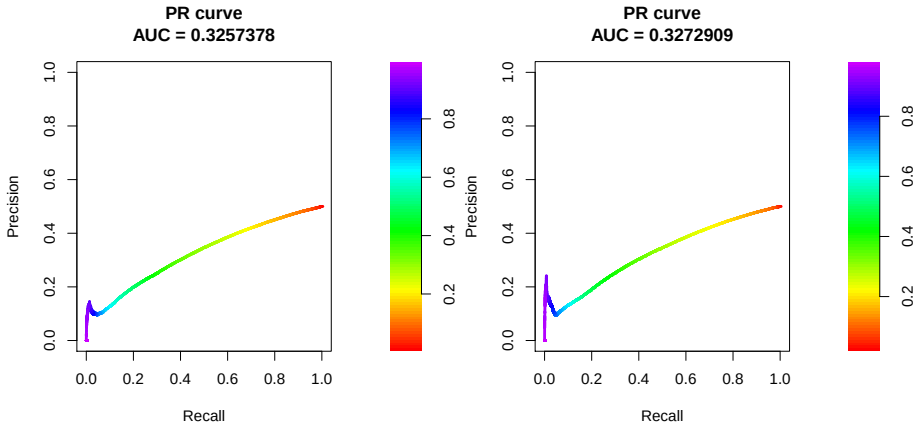
The lack of change for group characteristics and the change in base country characteristics, together with the similar predicting power, leads me to believe that the underlying process being modeled is primarily about group characteristics, and that base country characteristics do matter, but throwing so many country level variables into the model is taking away from the explanatory power as the variables work against each other likely due to collinearity. When regional dummies are removed, other country level variables shift to begin to make up the difference, while group level covariates don't change.

Table 1: Logit Models of Failure

| | *Dependent variable:* | |
|---|---|---|
| | end | |
| | original | new |
| x1 | $-0.082^{***}$ | $-0.088^{***}$ |
| | (0.019) | (0.019) |
| x2 | $0.001^{***}$ | $0.002^{***}$ |
| | (0.001) | (0.001) |
| left | $0.685^{***}$ | $0.983^{***}$ |
| | (0.238) | (0.211) |

| | | |
|---|---|---|
| nat | $0.372^{*}$ | $0.344^{*}$ |
| | (0.206) | (0.198) |
| right | $1.138^{***}$ | $1.375^{***}$ |
| | (0.310) | (0.291) |
| shr_trans | $1.641^{***}$ | $1.725^{***}$ |
| | (0.157) | (0.156) |
| diversity | $-1.676^{***}$ | $-1.555^{***}$ |
| | (0.389) | (0.387) |
| num_base | 0.072 | 0.083 |
| | (0.069) | (0.067) |
| lpop | $-0.030$ | $-0.154^{***}$ |
| | (0.070) | (0.047) |
| polity1 | $-0.071$ | $-0.097^{*}$ |
| | (0.054) | (0.053) |
| polity1sqr | 0.003 | $0.005^{**}$ |
| | (0.002) | (0.002) |
| ethnic | $-2.801^{*}$ | $-1.033$ |
| | (1.484) | (1.225) |
| ethn2 | $3.575^{**}$ | 1.077 |
| | (1.680) | (1.304) |
| EAP | $-0.316$ | |
| | (0.333) | |
| ECA | 0.341 | |
| | (0.235) | |
| LAC | $0.630^{**}$ | |
| | (0.303) | |
| NA. | 0.259 | |
| | (0.321) | |
| SAS | $-0.611^{*}$ | |
| | (0.323) | |
| SSA | $-0.641$ | |
| | (0.421) | |
| lelev | 0.146 | 0.100 |

| | | |
|---|---|---|
| | (0.125) | (0.121) |
| tropics | −0.174 | −0.205 |
| | (0.288) | (0.188) |
| landlock | 0.152 | −0.095 |
| | (0.289) | (0.267) |
| Constant | −3.158** | −1.051 |
| | (1.368) | (1.053) |
| Observations | 8,637 | 8,637 |
| Log Likelihood | −1,109.032 | −1,121.467 |
| Akaike Inf. Crit. | 2,264.064 | 2,276.934 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 | |




(a) Original Model (b) New Model

Figure 1: Precision-Recall Curves