Spike-and-Slab model & Boosted Tree to Predict Conflict

Anh Le
Department of Political Science
Duke University
anh.le@duke.edu

Introduction

Being able to predict political violence is very useful for governments and international organizations to deploy their limited resources effectively. Traditional models of violence in political science have relied on frequentist panel data methods, which have not had good predictive performance because:

- The null hypothesis of $\beta = 0$ being tested is not as important as the substantive effect of the variable
- Many social science variables are highly correlated. For example, *competitive elections*, *civil liberties*, and *property rights* are three conceptually distinct but empirically correlated variables. This results in (logistic) regression with large coefficient variance and meaningless p-value
- The problem of collinearity is compounded by the recent explosion of data available, with hundreds of predictors to be considered

Therefore, this project will use two methods that work well given the large number of predictors:

1. sparse regression with spike-and-slab prior
2. (boosted) classification tree

These models are found to perform better than existing models used in our lab.[1]

1 Description of Data

The dataset contains monthly data of 167 countries from 2001 to present (dimension = $27,000 \times 550$). The training cutoff date is `2013-06-01`.[2]

The labels are binary indicators of whether an event happens. We are interested in four types of events: insurgency, rebellion, dpc (domestic political crisis), and erv (ethnic and religious violence), and mp (massive protest).

The predictors include a country's politics, economics, and financial status. I also include spatial lags (constructed from Gower similarity, 4 nearest neighbors, and centroid distance) and temporal lags (by up to 2 months).

[1] We currently use hierarchical linear mixed-effect models, combined together with Bayesian ensemble. The predictors are hand-picked.

[2] This cutoff date is to comply with existing models used in our lab.

2 Spike-and-Slab

2.1 Model building

I pre-process the data by adding a binary variable that indicates whether an observation belongs to a country. This allows the model to have different intercepts for each country, essentially adding country fixed effect. The spike-and-slab model is fit with the package `BoomSpikeSlab`, running a MCMC chain with iterations = 5000 and burn-in = 500.

2.2 Result

Table 1 and 2a summarizes the predictive performance of my model. The spike-and-slab model performs well on `insurgency, rebellion, and ethnic violence` (97.6% precision and 94.6% recall out of sample), but not well on `domestic crisis` and `massive protest`. Figure 1 and Figure 2 visualizes this performance discrepancy with ROC curves.

Table 2a and 2b shows that the spike-and-slab model perform better across labels in comparison with the ensemble model currently used in my lab.

	insurgency	rebellion	dpc	erv	mp
brier	0.005	0.006	0.042	0.008	0.036
auc.C	0.996	0.999	0.927	0.989	0.764
precision	0.981	0.957	0.789	0.961	0.548
recall	0.767	0.769	0.410	0.681	0.068

Table 1: In-sample predictive performance

	insurgency	rebellion	dpc	erv	mp
brier	0.008	0.020	0.097	0.033	0.024
auc.C	0.998	0.930	0.865	0.975	0.801
precision	0.976	0.907	0.544	0.907	0.647
recall	0.946	0.789	0.548	0.490	0.147

(a) Spike-and-Slab

	insurgency	rebellion	dpc	erv
brier	0.06	0.03	0.12	0.03
auc.C	0.94	0.97	0.78	0.93

(b) Ensemble Linear Mixed Effect

Table 2: Table 2a shows the out-sample performance of the Spike-and-Slab model, which performs better (according to Brier and AUC score) than the currently used Ensemble model in Table 2b

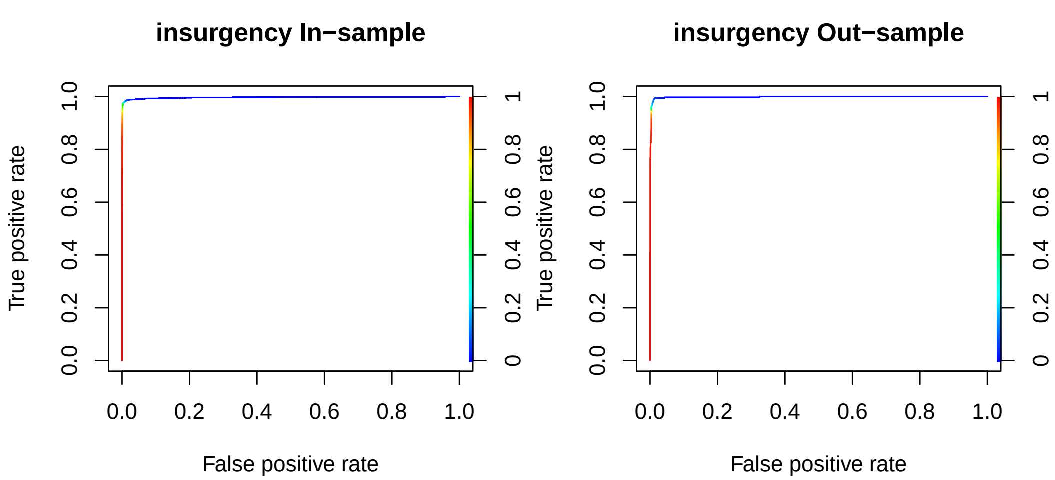

Figure 1: ROC curve of insurgency prediction

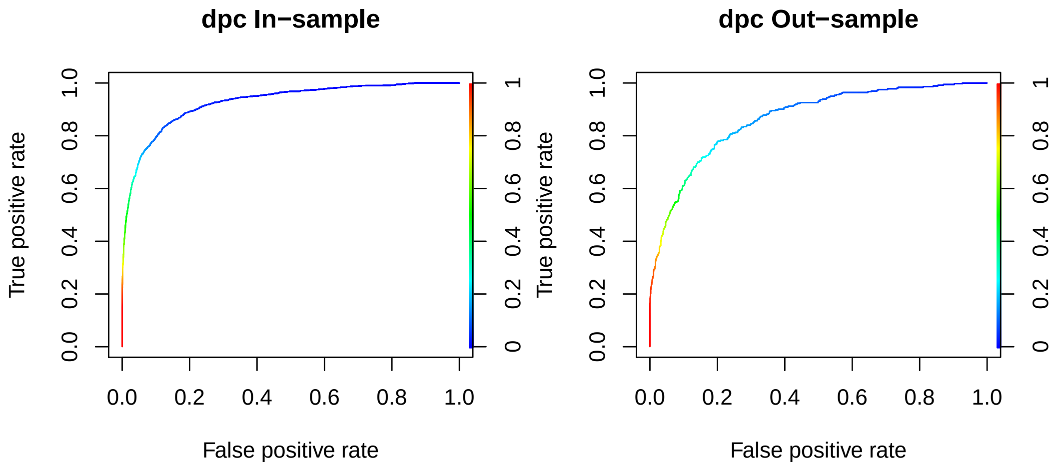

Figure 2: ROC curve of domestic crisis prediction

However, looking at Figure 3, it becomes clear that most of the predictive performance comes from the country dummies. This is because some countries almost always experience insurgency, while some other countries almost never, so it is safe to predict that these countries will follow the same pattern in the future. Therefore, even though the model has excellent performance, it is not necessarily revealing important factors that predict insurgency.

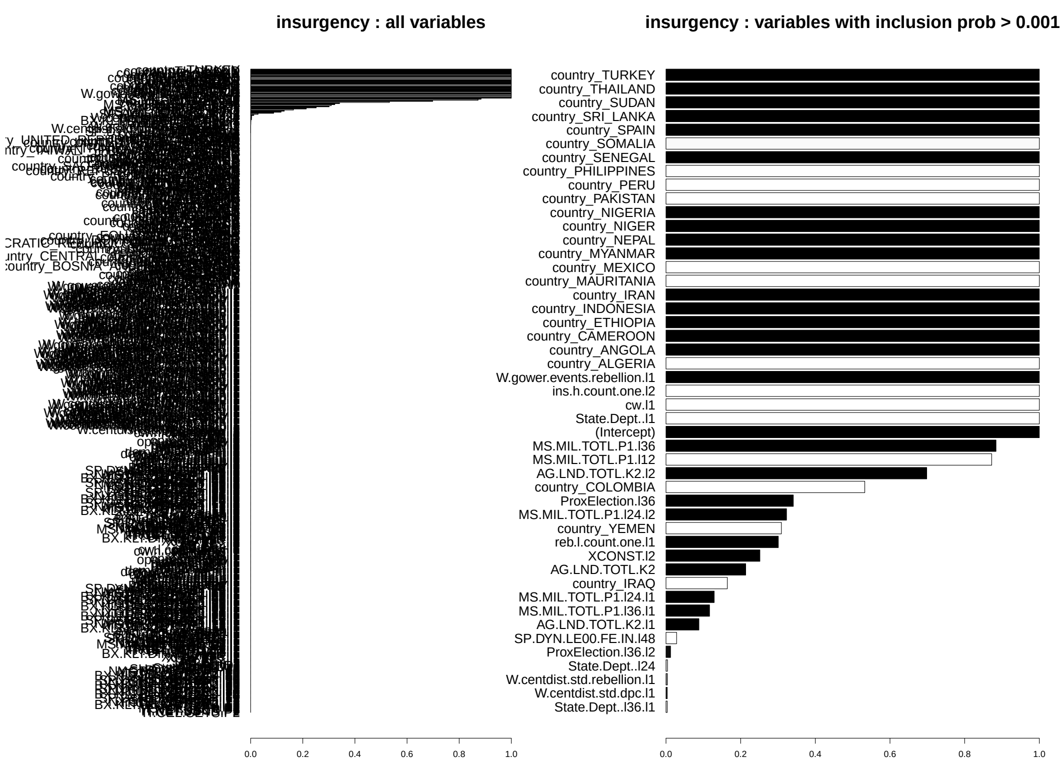

Figure 3: The left figure lists all the variables. The spike-and-slab model selects only a very small subset of the predictors. The right figure lists the variables with inclusion probability larger than 0.001. The majority of these predictors are country dummies.

Indeed, Table 3 shows that the model performance deteriorates sharply without the country dummies. Figure 4 shows the variables selected when there is no country dummy. Many more interesting predictive factors are revealed, such as the number opposition resistance events in neighboring countries, infant mortality rate, international tourism, etc.

	insurgency	rebellion	dpc	erv	mp
brier	0.008	0.020	0.097	0.033	0.024
auc.C	0.998	0.930	0.865	0.975	0.801
precision	0.976	0.907	0.544	0.907	0.647
recall	0.946	0.789	0.548	0.490	0.147

(a) With country dummies

	insurgency	rebellion	dpc	erv	mp
brier	0.080	0.057	0.167	0.037	0.049
auc.C	0.886	0.895	0.688	0.922	0.704
precision	0.519	0.575	0.162	0.508	0.097
recall	0.778	0.586	0.458	0.504	0.033

(b) Without country dummies

Table 3: Table 3b shows the out-sample performance of the Spike-and-Slab model without country dummies, which performs much worse.

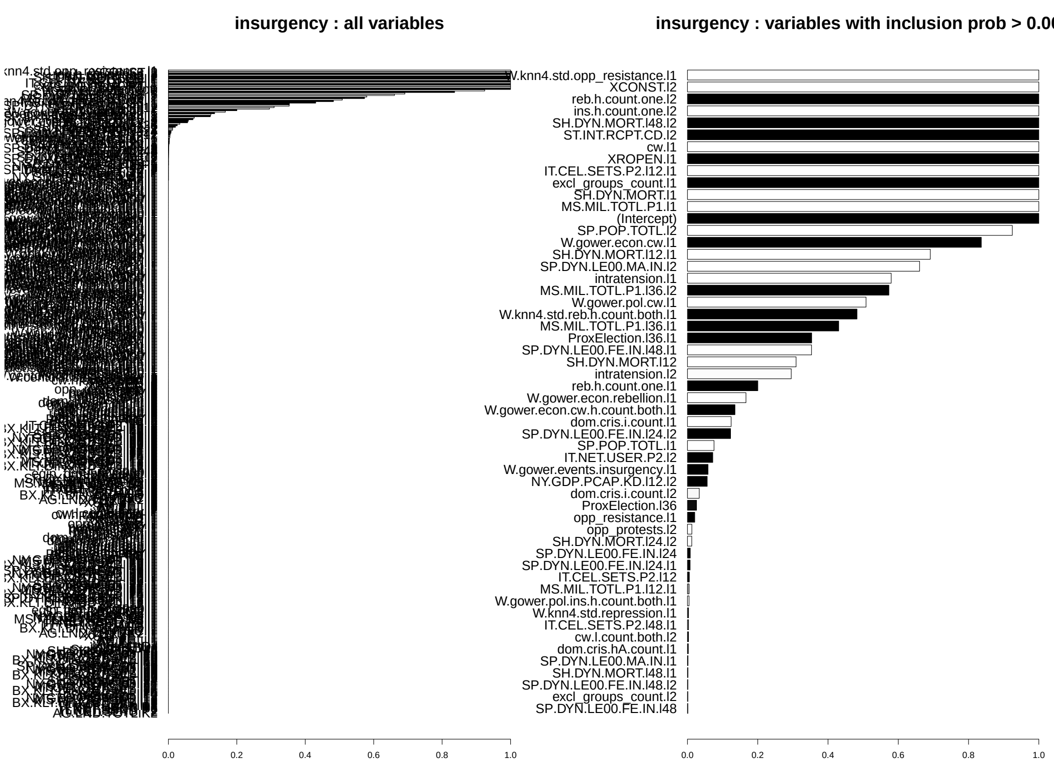

Figure 4: Variable Selection when there is no country dummies

3 Tree-based model

3.1 Model building

Since the spike-and-slab model does not work well without country dummies, in this section I will try tree-based model. From now on, I will also use a month's predictors to predict *next month*'s event instead.[3]

- Classification tree with pruning via cross-validation
- Boosted tree, with
  - tree.complexity = 1
  - learning.rate = 0.01
  - bag.fraction = 0.75
  - 10-fold cross-validation

3.2 Classification tree result

Table 4a and 4b show the predictive performance of pruned tree. There is not much improvement over spike-and-slab model. Figure 5 shows the decision tree for predicting insurgency.

	insurgency	rebellion	dpc	erv
precision	0.680	0.693	0.605	0.823
recall	0.665	0.784	0.064	0.713

(a) In-Sample

	insurgency	rebellion	dpc	erv
precision	0.614	0.600	0.148	0.820
recall	0.713	0.606	0.042	0.575

(b) Out-Sample

Table 4: In- and Out-Sample performance of decision tree

3.3 Boosted tree result

Table 5a and 5b shows the predictive performance of boosted classification tree, which is substantially better than individual tree. The precision and recall improves by 15 - 30 % for `insurgency, rebellion, ethnic violence`. Most dramatically is the improvement in prediction for `domestic crisis`, which goes from 14% to 65% in precision, and from 4.2% to 22% in recall.

4 Conclusion

In this project I predict conflict event with two methods that work well in high-dimensional data: 1) Spike-and-Slab variable selection, and 2) (Boosted) Classification Tree. In both cases, out-sample predictive performance is only slightly worse than in-sample performance, suggesting that the two method successfully avoid overfitting.

The Spike-and-Slab model also reveals the important factor of country dummies in predicting conflict event. It shows that the existing models have excellent performance (up to 97% precision and 94% recall on `insurgency`) simply because some countries always (never) have conflict events. Once the country dummies are taken out, the Spike-and-Slab model performs much worse.

Finally, when there is no country dummy, (boosted) classification tree performs substantially better than single tree and spike-and-slab model.

[3]I am not sure why our lab is "predicting" a month's event with the predictors of the current month. This seems to defeat the purpose of predicting conflict in advance in order to better prepare for it.

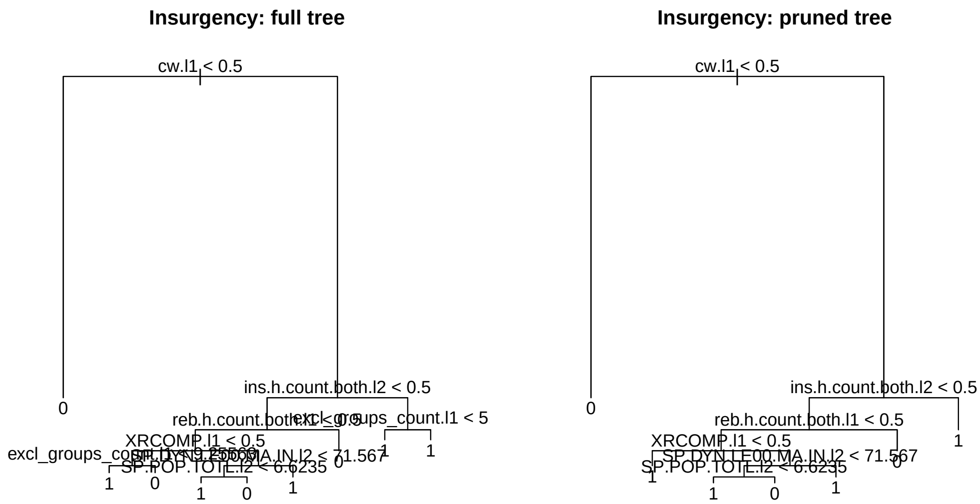

Figure 5: The decision tree to predict insurgency. Left: full tree. Right: pruned tree

	insurgency	rebellion	dpc	erv
brier	0.043	0.023	0.057	0.016
auc.C	0.956	0.984	0.875	0.979
precision	0.781	0.818	0.611	0.833
recall	0.633	0.749	0.261	0.769

(a) In-Sample

	insurgency	rebellion	dpc	erv
brier	0.046	0.030	0.065	0.020
auc.C	0.956	0.978	0.813	0.978
precision	0.694	0.815	0.649	0.837
recall	0.659	0.650	0.224	0.606

(b) Out-Sample

Table 5: In- and Out-Sample predictive performance of Boosted Classification Tree