UPPSALA UNIVERSITET

Assignment 4: Explaining and forecasting political violence I

Hegre et al. (2019) introduces the ViEWS system designed to forecast armed conflict. In this assignment we ask you to look into the country-level models in ViEWS. Updated country-month data are available through:

```
data <- read.csv('https://bit.ly/2XHmrMH')
```

Documentation of the variables in the dataset is found in `http://files.webb.uu.se/uploader/1576/Hegre-etal2019-OnlineAppendix.pdf`. The dataset linked to above includes one of the dependent variables in the current version of ViEWS – `ged_dummy_sb`, which notes whether at least one person was killed in state-based armed conflict in the country in the month. The dataset also includes a multi-category version called `severity_sb` which has the value 1 if at least 1 person died, 2 if at least 6, 3 if at least 25, 4 if at least 100, and 5 if at least 500 people were killed.
The existing country-level models in ViEWS are summarized in Table II in Hegre et al. (2019). Some of the studies underlying these models are referred to in the Literature Review and Current Forecasts sections in the article.
We would like you to do the following :

UPPSALA UNIVERSITET

Assignment 4: Explaining and forecasting political violence II

1. Split the data into a training and testing partition. Since the data you have access to is updated through March 2019, using data up to and including 2015 for training might be a good choice, in line with what is currently done in ViEWS. But other partitions are also possible, including cross-validation. Argue for your choice of partitioning, advantages and disadvantages.
2. Choose between the dichotomous `ged_dummy_sb` or the multi-category `severity_sb` variables as your dependent variable.
3. Take one of the thematic models 'demography', 'economics', 'institutions', or 'protest' as your point of departure. Estimate logistic regression models (binomial, ordinal, and/or multinomial) on the training partition, and present, interpret, and discuss the results, comparing what you find with what was found in earlier studies cited in Hegre et al. (2019).
4. PR curves, ROC curves, PR AUCs and ROC AUCs should be included. If you chose the multinomial dependent variable above, the `yardstick` package helps you calculate this for multiple classes automatically. In this case, use the "macro" aggregation procedure, which is explained in detail here (essentially it is a weighted sum of the computed ROC curves): `https://cran.r-project.org/web/packages/yardstick/vignettes/multiclass.html`

Assignment 4: Explaining and forecasting political violence III

UPPSALA UNIVERSITET

5. Discuss whether the observations can be seen as (conditionally) independent in your model, suggest ways to address possible issues related to that, and show results from models implementing these suggestions. The variables in the conflict history theme are likely to be useful for this.
6. Reanalyze your model using random forests. importance scores and look at partial dependency plots to interpret the findings. Compare with the results from the logistic regression model and discuss implications for understanding the thematic model.
7. Generate predictions for the test partition from both sets of models. Compare the results (using confusion matrices, ROC and PR curves) and discuss.
8. Optional: Do a grid search for best hyper-parameters. In a loop, change the number of trees to 50, 100, 250, 500, 1000 and the minimum tree node size of 1, 5, 10 and 50 (testing all possible combinations). Use the same partitioning you describe above for training and testing. Compute AUPR and AUROC for each. Do they change? Discuss!

Assignment 4: Explaining and forecasting political violence IV

UPPSALA UNIVERSITET

9 Optional: Suggest ways to improve the predictive performance of the model, reanalyze the models, and demonstrate whether your ideas improve predictive performance.

We recommend using the `nnet` package for multinomial regression models, `ranger` for random forests (do not use the predict probabilities from `randomForest`!), `pdp` (or manual simulation of QIs/bootstrapping if you are feeling brave!) for PDPs (which, as you remember, are the same as simulated QIs or marginal effect plots) and `yardstick` for model evaluation.
Example code for random forests is posted here: `https://github.com/mihaicroicu/AQM2019/blob/master/R_scripts/random_forest.R`

UPPSALA UNIVERSITET

Bibliography I

Hegre, Håvard, Marie Allansson, Matthias Basedau, Mike Colaresi, Mihai Croicu, Hanne Fjelde, Frederick Hoyles, Lisa Hultman, Stina Högbladh, Remco Jansen, Naima Mouhleb, Sayeed Auwn Muhammad, Desirée Nilsson, Håvard Mokleiv Nygård, Gudlaug Olafsdottir, Kristina Petrova, David Randahl, Espen Geelmuyden Rød, Gerald Schneider, Nina von Uexkull and Jonas Vestby. 2019. "ViEWS: A political Violence Early Warning System." *Journal of Peace Research* 56(2):155–174.
URL: *https://doi.org/10.1177/0022343319823860*