Assignment 2, Replication and extension of Muchlinski et al. (2016) Random forest vs logistic regression

Eve Fleisig, Jordan Klein, Matthew Sun

October 15, 2020

Part 1: Replication

We begin by replicating the separation plots from Muchlinski et al. (2016) using their updated replication materials (Fig. 1). Our separation plots for the logistic regression models by Fearon and Laitin (2003), Collier and Hoeffler (2004), and Hegre and Sambanis (2006) are identical to those from Muchlinski et al.'s original paper, but notably, our separation plot for random forests identifies 2 false positives while that from Muchlinski et al.'s original paper does not identify any. These results align with those produced by Mulchinski et al. in their own replication (Muchlinski, 2019). Our findings refute Muchlinski et al.'s initial assertion that random forests is uniquely impervious to Type II errors compared to logistic regression.

We continue by replicating the ROC curves from Muchlinski et al. (2016) (Fig. 2). Like Muchlinski et al. (2016), we find the random forest model has an AUC of .91, but as explained by Wang (2019), the ROC curve in Muchlinksi et al.'s original paper implies an AUC of .97. Our redrawing of the ROC curve to be consistent with an AUC of .91 aligns with Muchlinski et al.'s replication (Muchlinski, 2019). In contrast, our ROC curves for logistic regression and penalized logistic regression do not change very much compared to Muchlinski et al. (2016). However, we note that Muchlinski et al. do not provide updated calculations of AUC in their replication (Muchlinski, 2019). We therefore recalculate AUC for all models and find those for uncorrected and penalized logistic regression to be slightly higher than those shown in Muchlinski et al.'s original paper. These results demonstrate that the gap in performance between random forests and logistic regression is substantially smaller than they report.

As suggested by Neunhoeffer and Sternberg (2019), the out-of-sample analysis reported by Muchlinski et al. (2016) does not match that from the article's original replication materials (Muchlinski, 2015). In replicating the analysis from Muchlinski et al.'s updated materials (Muchlinski, 2019), using a threshold for positive prediction of 0.5, we find that the logistic regression models all fail to predict any civil war, as they originally reported (Table 1). However,

random forests performs slightly better, correctly predicting 10 rather than 9 out of 20 civil wars. Such a small difference however is not of import, especially because of the randomness inherent in both the imputation and random forests procedures used (Muchlinski et al., 2019).

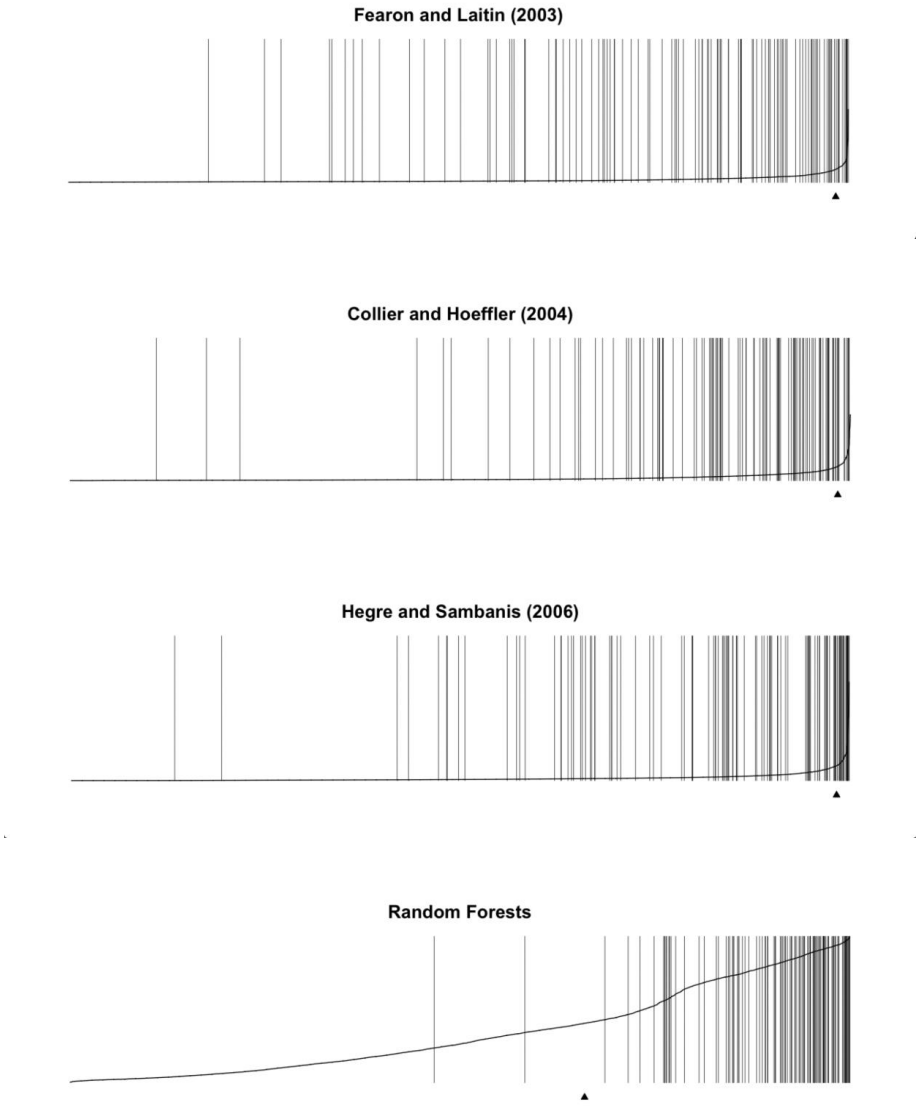

Figure 1: Separation plot for all classifiers.

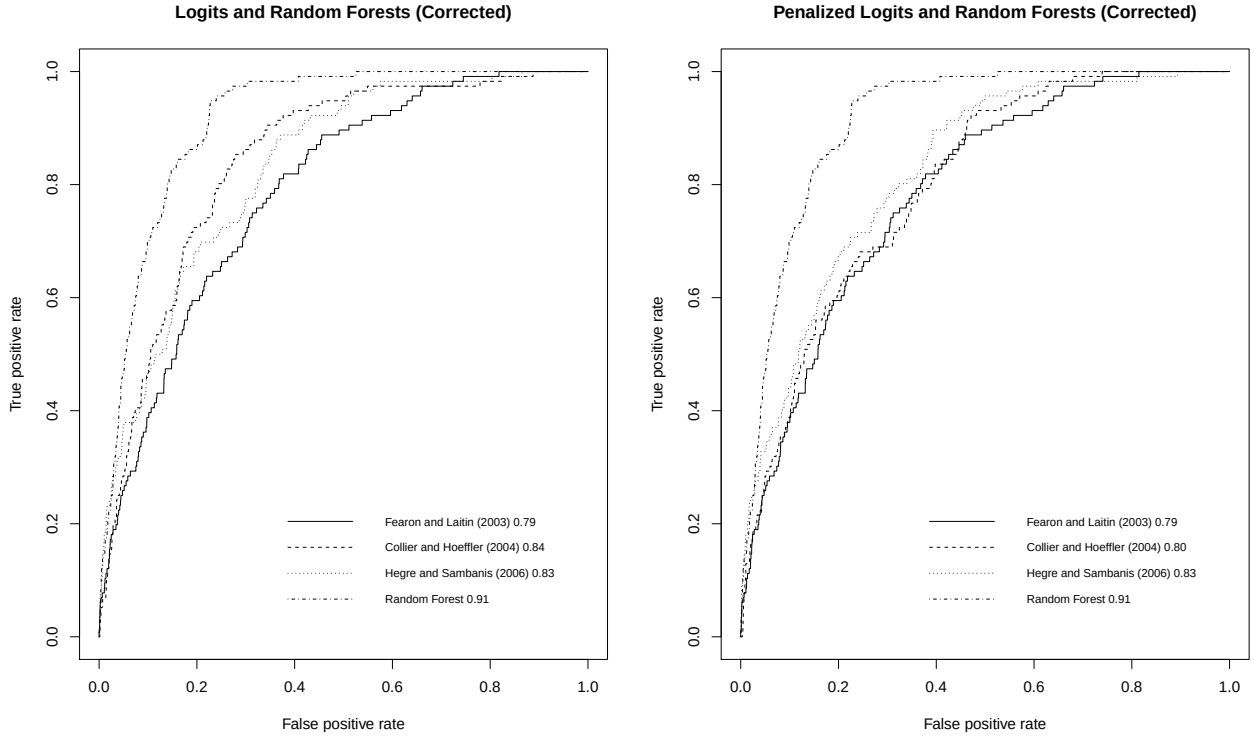

Figure 2: ROC curves for all classifiers.

Table 1: Predicted probability of civil war onset: Logistic Regression and Random Forests

Models and predicted probability of civil war onset					
Country	Year	Fearon and Latin (2003)	Collier and Hoeffler (2004)	Hegre and Sambanis (2006)	Random Forest
Afghanistan	2001	0.01	0.00	0.00	0.06
Angola	2001	0.01	0.01	0.02	0.71
Burundi	2001	0.03	0.00	0.02	0.09
Guinea	2001	0.01	0.00	0.01	0.07
Rwanda	2001	0.01	0.00	0.01	0.05
Uganda	2002	0.02	0.02	0.02	0.93
Liberia	2003	0.02	0.04	0.03	0.98
Iraq	2004	0.03	0.01	0.03	0.16
Uganda	2004	0.01	0.00	0.01	0.45
Afghanistan	2005	0.03	0.00	0.02	0.14
Chad	2006	0.02	0.04	0.03	0.98
Somalia	2007	0.06	0.04	0.10	0.96
Rwanda	2009	0.02	0.04	0.03	0.99
Libya	2011	0.02	0.04	0.02	0.95
Syria	2012	0.01	0.00	0.00	0.06
Syria	2012	0.01	0.00	0.00	0.06
Democratic Republic of the Congo	2013	0.01	0.00	0.00	0.04
Iraq	2013	0.02	0.04	0.02	0.96
Nigeria	2013	0.02	0.04	0.03	0.96
Somalia	2014	0.05	0.04	0.11	0.99

Finally, although Mulchinski et al. do not provide the code to do so in their replication materials, we attempt to replicate the F1-score plot from their original paper 2016 (Fig. 3). Our results differ significantly from those they report; we successfully identify few true positives and thus frequently obtain precision and recall of 0, yielding an undefined F1-score, or otherwise low precision and recall, yielding significantly lower mean F1-scores for every method and training set ratio. In contrast to Muchlinski et al. (2016), we find that the random forest model does not perform better than either logistic regression method. The mean F1-scores of random forests are similar to logistic regression when the training set ratio is small, and only slightly higher when the training set ratio is 0.8. However any potential advantage is canceled out by random forests' high standard deviation. Our replication results in full indicate that while random forests may be superior to logistic regression in predicting civil war onset by some measures, its advantage is undoubtedly smaller than was initially suggested by Muchlinski et al. (2016).

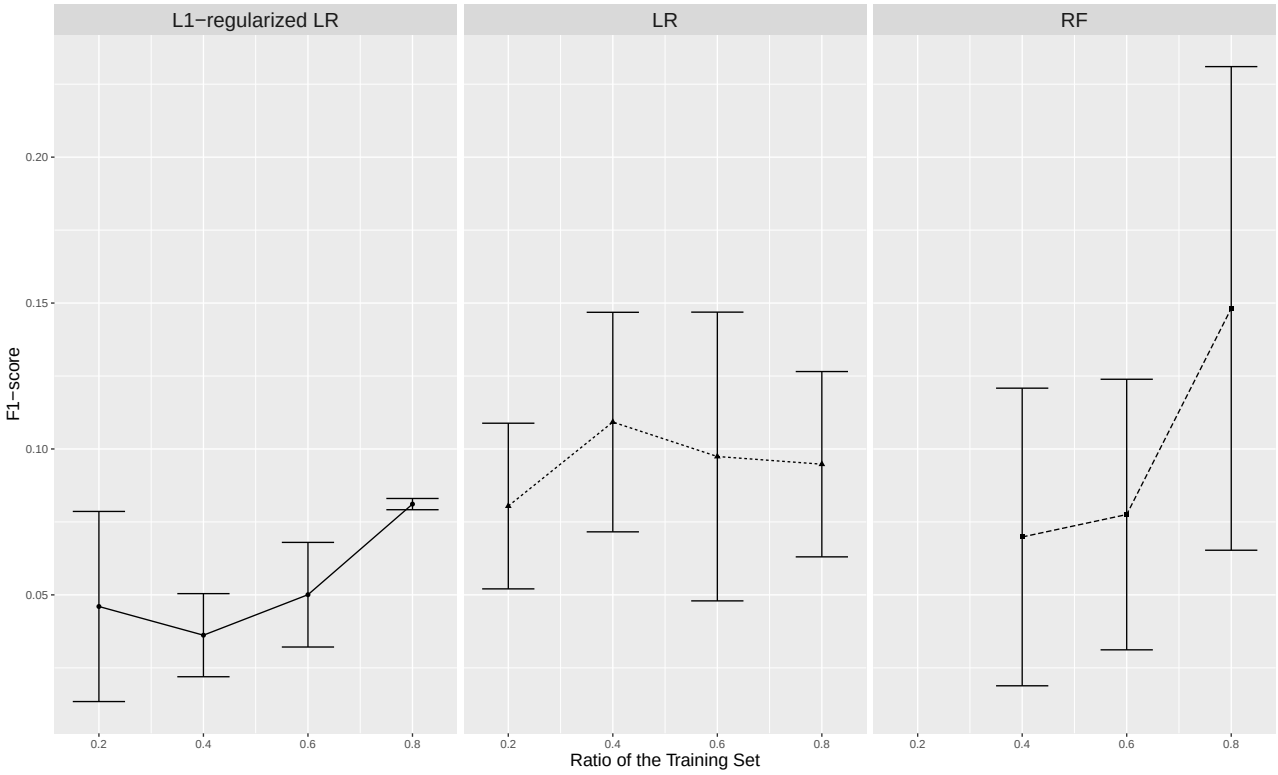

Figure 3: Comparison of F1 Score with varying training set ratio.

https://www.overleaf.com/project/5f8607556236730001cad7bd

Part 2: Extension

Introduction

A central issue in the methodology proposed by Muchlinski et al. to evaluate civil war onset prediction is the assumption that the onset of a civil war is a binary variable, such that civil war either broke out or did not break out in a given year. However, real-life civil wars are rarely so discrete: skirmishes and widespread civil violence often precede the onset of war, and intrastate conflicts may not qualify as civil war depending on the subjective criteria used. We questioned whether the results that Muchlinski et al. obtained might be dependent on their arbitrary cutoff for civil war.

A related question is whether Muchlinski et al.'s results hold when fully taking autocorrelation and lag into account, using violence or other predictors in previous years to predict violence in the current year. This corresponds more closely to real-world applications in which data about the history of a given country is already available.

Finally, we observe that much of the presumed utility of predicting the outbreak of civil wars comes from the possibility of saving lives. If our goal is not to predict war itself but rather to identify areas that will experience high casualties from civil war, then predicting the number of deaths related to intrastate violence seems a more direct measure of a quantity of interest.

To investigate these possibilities, we examined whether continuous prediction of the degree of intrastate violence given data from the previous year(s) would affect the relative predictive power of random forests and regression models. Furthermore, we convert the binary classification task into a time series forecast task, in which we predict a continuous variable: the number of deaths per capita from intrastate violence in a given country in each year. We trained a random forest and a linear regression model[1] to examine their ability to predict the degree of violence.

We also examined whether methods improving on traditional random forests improve classifier results. Discrepancies between the performance of different tree-based models could suggest that arguments about the relative merits of families of classifiers are highly dependent on the specific models used. We compared the performance of the random forest and linear regression models to LightGBM, a faster variant of gradient boosting decision trees (Ke et al., 2017).

Methods

The Uppsala Conflict Data Program (UCDP) measures the number of deaths per country due to several types of conflict. We used the number of deaths per country due to intrastate violence, which the UCDP defines as violence between a government and rebel troops without the involvement of foreign governments with troops, and internationalized intrastate violence, which the UCDP defines

[1]Because this task predicts a continuous outcome (deaths per capita), we use a linear regression model rather than the logistic regression model used by Muchlinski et al.

as violence between a government and rebel troops with some involvement of foreign governments with troops. Note that a necessary limitation of the UCDP data is that the year coverage (1989 to 2000) is significantly smaller than that of the original dataset. In order to convert this into a per-capita measure, we then used United Nations population data (DESA, 2019). We excluded deaths due to extrasystemic violence (between a state and a non-state outside its territory) and interstate violence as inapplicable to civil war.

For each year from 1992 to 2000, the validation set was the prediction for that year and the training set consisted of the data from all previous years. All models used the same training features that Muchlinski et al. used and the number of deaths per capita in the previous year. As a baseline, we used a model that simply predicted the same number of deaths as the previous year in all cases. We evaluate each model by averaging its mean performance across all years used for the validation set, where performance is measured by root mean squared logarithmic error (RMSLE). RMSLE is an approximation for relative error from the true value of per capita deaths in each year.

We train each model (random forest, linear regression, and Light Gradient-Boosted Machine) using several different variants of the feature/target variable. We vary the predictor set by the number of years of lagged outcomes (1 vs. 2) and whether we include the full set of predictor variables used by Muchlinski et al.

Our code is available at `github.com/jordan-klein/muchlinksi_replication`.

Results

Surprisingly, we found that predictions using the previous two or three years performed worse or on par with simply using the previous year alone. We also find that for random forest and linear regression models, the inclusion of the full set of Muchlinski's predictor variables offers no improvement in RMSLE. We report the RMSLE of the best model of each type (linear regression, random forest, and LGM) on the validation set in Table 2. The LightGBM model had the lowest error rate (0.00072), followed by the linear regression model (0.00074) and then the random forest (0.000087). Interestingly, all models experienced a spike in RMSLE in 1997.

Table 2: Root mean squared log error of casualty prediction on the validation set.

RMSLE of casualty prediction				
Year	Baseline	Random Forest	LightGBM	Linear Regression
1992	0.00009	0.000044	0.000035	0.000048
1993	0.00007	0.000083	0.000062	0.000072
1994	0.00008	0.000076	0.000058	0.000045
1995	0.00005	0.000051	0.000049	0.000036
1996	0.00005	0.000045	0.000041	0.000039
1997	0.00027	0.000273	0.000272	0.000273
1998	0.00019	0.000114	0.000070	0.000077
1999	0.00008	0.000063	0.000041	0.000049
2000	0.00003	0.000032	0.000022	0.000025
Mean	0.00010	0.000087	0.000072	0.000074

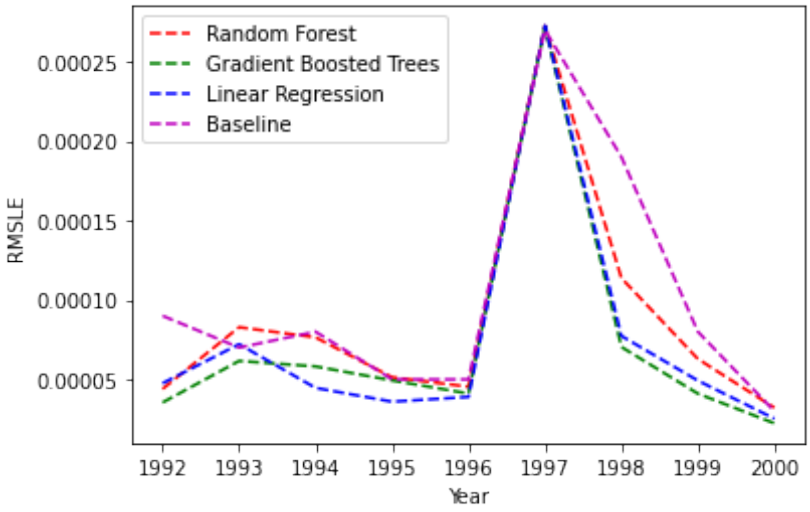

Figure 4: RMSLE of casualty predictions.

These results contrast with Muchlinski et al. (2016), who found that random forests outperformed several types of logistic regression. In fact, the random forest model had the highest or tied for the highest error in eight of the nine years examined (Fig. 4). This discrepancy suggests that Muchlinski et al.'s results hinge on the structure of the problem: the binary variable they predict, the metrics used to evaluate binary classification, and the disregard of autocorrelation. When predicting a continuous variable and taking autocorrelation into

account, random forests no longer outperform regression models. In addition, the fact that gradient boosted trees outperformed the other models suggests that broad conclusions about tree-based methods are dependent on the exact model used in a given scenario.

Ultimately, the differences between Muchlinski et al.'s results and ours underscore the notion that sweeping conclusions about the relative merits of different classifiers rely heavily on the presentation of the problem, particularly the features used by the models and how predicted classes are defined. For phenomena like civil wars, which unfold over time, fit poorly into binary categories, and can be measured with rich, continuous data sources, random forests may indeed not be the best method of prediction.

Bibliography

Uppsala Conflict Data Program. URL `https://ucdp.uu.se/exploratory`.

UN DESA. World population prospects 2019. online edition, rev 1, united nations, department of economic and social affairs. *Population Division*, 2019.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, pages 3146–3154, 2017.

David Muchlinski. Replication Data for: Comparing Random Forests with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data, August 2015. URL `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/KRKWK8`. type: dataset.

David Muchlinski. Political analysis replication files, 2019.

David Muchlinski, David Siroky, Jingrui He, and Matthew Kocher. Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis*, pages 87–103, 2016.

David Alan Muchlinski, David Siroky, Jingrui He, and Matthew Adam Kocher. Seeing the Forest through the Trees. *Political Analysis*, 27(1):111–113, January 2019. ISSN 1047-1987, 1476-4989. doi: 10.1017/pan.2018.45.

Marcel Neunhoeffer and Sebastian Sternberg. How Cross-Validation Can Go Wrong and What to Do About It. *Political Analysis*, 27(1):101–106, January 2019. ISSN 1047-1987, 1476-4989. doi: 10.1017/pan.2018.39.

Yu Wang. Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data: A Comment. *Political*

Analysis, 27(1):107–110, January 2019. ISSN 1047-1987, 1476-4989. doi: 10.1017/pan.2018.40.