Online Appendix

for

Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data: A Comment

Yu Wang*
ywang176@ur.rochester.edu
University of Rochester

Author's note: The replication materials (Wang, 2018) for all the figures and tables in the paper and in this online appendix are available at the *Political Analysis* dataverse site.

1 Appendix A: Nested Cross Validation

In k-fold cross-validation, hyper-parameters, such as the penalty term in LASSO and Support Vector Machine, the number of trees in random forest and the number of neighbors in k-Nearest-Neighbor, are usually selected by minimizing the prediction error on the testing fold.[1] As the same data is used to select these hyper-parameters and to evaluate model performance, it has been observed that there will be a downward bias in the estimate of prediction error (Cranmer and Desmarais, 2017; Cawley and Talbot, 2010; Tibshirani and Tibshirani, 2009). One way to solve this problem is to use nested cross-validation (Varma and Simon, 2006). In nested cross-validation, the outer loop always keeps a portion of the data for final testing and uses only the remaining observations for the inner cross validation. To compare the performance of two competing algorithms, such as penalized logistic regression and random forest, the average of the respective algorithm's performance for each iteration of the outer loop is calculated.

Nested Cross-Validation for Performance Evaluation

Input: N observations, K^{outer} for outer CV and K^{inner} for inner CV

shuffle the N observations and divide them into K^{outer} folds

for i = 1 **to** K^{outer} **do**

use observations in the ith fold as testing set

use observations in the other (K^{outer}-1) folds for the inner cross validation

tune hyper-parameters using K^{inner}-fold cross-validation

use the model to predict on the testing set

calculate the performance metric based on predicted probabilities and true labels

end for

Output: the mean performance metric over the K^{outer} iterations

---

In Table 1, I compare the random forest model with three penalized logistic models (Fearon and Laitin, 2003; Collier and Hoeffler, 2004; Hegre and Sambanis, 2006) using nested cross validation. I compare the random forest model with penalized logistic models because these penalized models have hyper-parameters. Same as in Muchlinski et al. (2016), I use the area under the ROC curve (AUC-ROC) as the performance metric. K^{outer} for the outer loop is set to 10, and so is K^{inner} for inner loop. It can been seen that these mean AUC metrics are all quite similar to those reported in Muchlinski et al. (2016), which is re-assuring. The random forest model achieves by far the highest mean AUC at 0.915. The second best classifier is Hegre and Sambanis (2006), achieving an AUC of 0.800. It should also be noticed that in addition to achieving the highest mean AUC, the variance of random forest over the 10 iterations is also the smallest.

[1] Note that not all models have hyper-parameters. For example, ordinary least squares (OLS) does not have hyper-parameters, nor does logistic regression.

Table 1: Nested Cross Validation for Model Comparison

Iteration	Fearon, Laitin	Collier, Hoeffler	Hegre, Sambanis	Muchlinski et al.
1	0.9224586	0.8003152	0.9670607	0.936249
2	0.7431633	0.8057143	0.7781633	0.8822449
3	0.6387707	0.8026793	0.7560284	0.9486998
4	0.8463173	0.8401204	0.9135977	0.9436969
5	0.7167944	0.8068327	0.785281	0.9374601
6	0.6499646	0.6501416	0.7907224	0.9411296
7	0.7789796	0.6291837	0.7245918	0.8054592
8	0.8384943	0.8539773	0.8059659	0.9362926
9	0.6975146	0.7922813	0.7407557	0.9242271
10	0.7266944	0.7836426	0.7379023	0.8920485
Mean(AUC)	0.7559152	0.7764888	0.8000069	0.9147508
Var(AUC)	0.0082407	0.0056717	0.0062905	0.0019705

2 Appendix B: Out-of-Sample Prediction

In this appendix, I discuss in more detail the error in performing predictions on out-of-sample data found in Muchlinski et al. (2016).[2] Muchlinski et al. (2016) first train the random forest model with the whole dataset.

```
RF.out<-randomForest(as.factor(warstds)~., sampsize=c(30, 90),
importance=T, proximity=F, ntree=1000, confusion=T, err.rate=T,
data=data.full)
```

Then, instead of feeding the model with out-of-sample data, i.e., the 737 new observations focusing on Africa and the Middle East, Muchlinski et al. (2016) use the predictions obtained during training the random forest model. Put another way, the prediction of the civil war onsets in Africa and the Middle East between 2001 and 2014 should be based on their observed independent variables, but were not implemented so in Muchlinski et al. (2016).

```
yhat.rf<-predict(RF.out, type="prob") #taken from RF on whole data
###We used original CW data for training data here for all
models/algorithms###
Yhat.rf<-as.data.frame(yhat.rf[,2])
```

Muchlinski et al. (2016) proceed to randomly sample without replacement 737 predicted probabilities of civil war onsets out of 7140.

```
###Selecting random samples to make pred and actual lengths equal###
set.seed(100)
predictors.rf<-Yhat.rf[sample(nrow(Yhat.rf), 737),]
```

[2]The discussion is based on the original codes from Muchlinski (2015).

Muchlinski et al. (2016) match the sampled probabilities against the true labels of their extended dataset and perform evaluation.

```
pred.rf.africa<-prediction(predictors.rf, data3$warstds)
perf.rf.africa<-performance(pred.rf.africa, "tpr", "fpr")
plot(perf.rf.africa, lty=4, add=T)
```

To summarize, when making predictions on a new dataset, Muchlinski et al. (2016) do not feed the trained model with new information, but randomly sample predicted probabilities from the model's training examples. This error has affected the random forest model, all the logistic models and the discussion on out-of-sample predictive accuracy.

In addition to pointing out this implementation error in making out-of-sample predictions, I also briefly comment on the interpretation of the results. Muchlinski et al. (2016) report that "Random Forests correctly predicts nine of twenty civil war onsets in this out-of-sample data when the threshold for positive prediction is 0.50," whereas logistic regression models fail to specify any civil war onset in the out-of-sample data.[3] As the authors' random forest model tends to give a higher probability of war than the logistic models, it is expected that more predictions by the random forest model will cross the 0.5 threshold than by the competing logistic models. The fact that "Random forests correctly predicts nine of twenty civil war onsets in this out-of-sample data" is impressive, but Muchlinski et al. (2016) fail to mention that out of their 737 observations, they predict 216 observations to be civil war onsets, when there are only 21 of them. This tendency of mis-classifying peace into civil war onsets, which is not obvious in the ROC curve, where the y axis is true positive rate, will become obvious in the Precision-Recall (PR) curve, with precision being the y-axis.

3 Appendix C: Precision-Recall Curve

In this appendix, I report the PR curves of the random forest model, the AdaBoosted model and the gradient boosted trees (Figure 1). It is easy to see that for almost all recall values, the AdaBoosted model and the gradient boosted trees give a higher precision. In terms of area under the PR curve (AUC-PR), the AdaBoosted trees have an AUC-PR of 0.32 and the gradient boosted trees have an AUC-PR of 0.36. By contrast, the random forest model's AUC-PR is 0.14.

[3]There are actually 21 civil war onsets in the out-of-sample dataset not 20. Also, based on the replication data, the random forest model predicts 8 of them not 9.

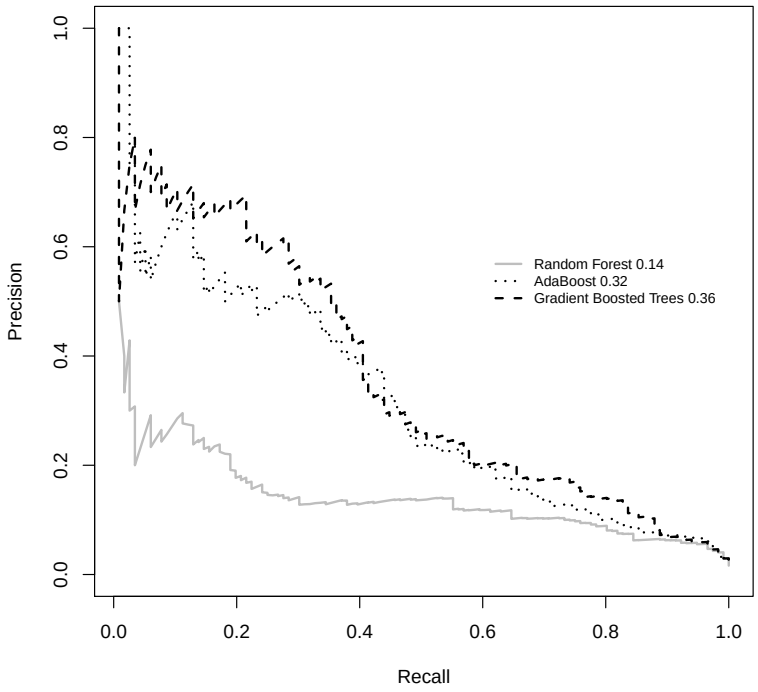

Figure 1: While their AUC-ROC metrics look comparable, the two boosted models substantially outperform the random forest model when evaluated with AUC-PR.

While the difference between the random forest model and the two boosted models appears very small when evaluated with AUC-ROC, the difference becomes very dramatic when evaluated with AUC-PR. To see why, notice that the y axis of the ROC curve is true positive rate, $\frac{TP}{TP+FN}$, where TP stands for True Positives and FN stands for False Negatives, and the y axis of the PR curve is precision, $\frac{TP}{TP+FP}$. In the context of civil war predictions, civil war onset is positive, peace is negative, and there are far more peace observations than civil war onsets. Consequently, a classifier that tends to give higher probabilities of war, as the random forest model does, has a large FP, which then leads to a decrease in the AUC-PR value. For this reason, the PR curve is recommended when the two classes are unbalanced (Davis and Goadrich, 2006; Cranmer and Desmarais, 2017).

References

Cawley, G. C. and N. L. C. Talbot (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research 11*, 2079–2107.

Collier, P. and A. Hoeffler (2004, October). Greed and grievance in civil war. *Oxford Economic Papers 56*(4), 563–595.

Cranmer, S. J. and B. A. Desmarais (2017, April). What can we learn from predictive modeling? *Political Analysis 25*(2), 145–166.

Davis, J. and M. Goadrich (2006). The relationship between precision-recall and roc curves. *ICML '06 Proceedings of the 23rd international conference on Machine learning*, 233–240.

Fearon, J. D. and D. D. Laitin (2003, February). Ethnicity, insurgency, and civil war. *American Political Science Review 97*(1), 75–90.

Hegre, H. and N. Sambanis (2006, August). Sensitivity analysis of empirical results on civil war onset. *Journal of Conflict Resolution 50*(4), 508–535.

Muchlinski, D. (2015). Replication data for: Comparing random forests with logistic regression for predicting class-imbalanced civil war onset data. http://dx.doi.org/10.7910/dvn/krkwk8. *Harvard Dataverse*.

Muchlinski, D., D. Siroky, J. He, and M. Kocher (2016). Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis 24*(1), 87–103.

Tibshirani, R. J. and R. Tibshirani (2009). A bias correction for the minimum error rate in cross-validation. *The Annals of Applied Statistics 3*(2), 822–829.

Varma, S. and R. Simon (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics 7*(91).

Wang, Y. (2018). Replication materials for "comparing random forest with logistic regression for predicting class-imbalanced civil war onset data: A comment". https://doi.org/10.7910/dvn/uiuygy. *Harvard Dataverse*.