Is Random Forest Really Better than Logistic Regression for Predicting Civil War Onsets? [*]

Marcel Neunhoeffer and Sebastian Sternberg

Center for Doctoral Studies in Social Sciences, University of Mannheim

July 31, 2017

Abstract

Correctly predicting civil war onset can save many lives. In a recent paper, Muchlinski et al. (2016) compare the predictive performance of different logit model specifications with Random Forest. They find that their Random Forest approach outperforms all logit model specifications in-sample and out-of-sample. However, we show that the impressive superiority of their Random Forest model is an artifact of methodological flaws. We use this opportunity to fix these flaws and provide a checklist for better model comparisons. To move the field forward and to save lives we need to be sure that new models actually perform better.

Word count: 3,075

[*] We thank Thomas Gschwend, Richard Traunmüller and Guido Ropers for very helpful comments. This work was supported by the University of Mannheim's Graduate School of Economic and Social Sciences funded by the German Research Foundation.

1 Introduction

Correctly predicting civil war onset can save many lives. To that end machine learning models, promising better predictions and achieving remarkable results across different domains, become increasingly popular among social scientists.

A recent – and already prominently cited – paper published in Political Analysis claims that "Random Forest offers superior predictive power compared to several forms of logistic regression in an important applied domain – the quantitative analysis of civil war" (Muchlinski et al., 2016, 101). The authors conclude that while "all logistic regression models fail to specify any civil war onset in the out-of-sample data. 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" (Muchlinski et al., 2016, 96).

We show that the impressive superiority of their Random Forest model is an artifact of methodological flaws. Muchlinski et al. (2016) report a biased performance assessment of Random Forest that is overly optimistic due to overfitting and the use of different data. Even more problematic is that the predictive performance of their Random Forest model is never tested on out-of-sample data.

Nevertheless, Muchlinski et al. (2016) make an important contribution. They and we think that machine learning approaches indeed can be useful to political scientists and particularly useful for predicting civil war onsets. Therefore, our goal with this paper is to provide a number of criteria for better model comparisons to political scientists. For prediction, we ideally want to select the models that best predict the outcome.

We organize this letter as follows. First, we line out the flaws in Muchlinski et al. (2016). Second, we fix the flaws in the model comparison by Muchlinski et al. (2016), rerun their analysis and apply a better model comparison to their data and models. Third, based on our application we develop a checklist for better model comparisons. We conclude that in our case Random Forest is not superior to logit models in predicting civil war onsets.

2 Five Problems in Muchlinski et al. (2016)

Muchlinski et al. (2016) analyze a dataset (Muchlinski, 2015) of civil wars with yearly observations for each recognized country in the world from 1945 to 2000 (Hegre and Sambanis, 2006). The dependent variable civil war onset is a binary measure of whether a civil war onset occurred for a given country in a given year[1]. The data set includes 7140 observations and a rich set of variables (while the original data set contains more variables, Muchlinski et al. (2016) use 90 of them).

Muchlinski et al. (2016) compare three different specifications of logit regression models[2] with their Random Forest model:

- the Fearon and Laitin (2003) model specification with **11 independent variables**[3],
- the Collier and Hoeffler (2004) specification with **12 independent variables**
- and a model specification using the **20** most robust **independent variables** identified by Hegre and Sambanis (2006).

From their analysis they conclude: "[T]hat Random Forests offers superior predictive power compared to several forms of logit regression in an important applied domain – the quantitative analysis of civil war. Separation plots, AUC scores, and F_1-scores all demonstrate the superior predictive accuracy of Random Forests in class imbalanced CWD [civil war data]" (Muchlinski et al., 2016, 101). In this section we show that the impressive superiority of their Random Forest model is an artifact of methodological flaws.

First, we think that for comparing models, all models should be build using the same data. Muchlinski et al. (2016) do not follow this rational because they only use a limited set of variables (11/12/20) for each of the logit model specifications but compare it to a Random Forest model that uses all **90 variables**. To us, this is not a very fair comparison: the different models are build on different data and information.

[1] In the replication dataset (Muchlinski, 2015) provided by Muchlinski et al. (2016) the ratio of 1s (civil war onsets) and 0s (peace) is about 1:100. This class imbalance complicates the prediction.

[2] Also counting the penalized logit regression models makes this six model specifications.

[3] For a list of the variables see the Appendix.

Second, Muchlinski et al. (2016) use down-sampling, which is a popular sampling method to account for class-imbalanced data (Chen et al., 2004). In short, sampling approaches either drop observations (down-sampling) from the dataset or redraw from the minority class (up-sampling)[4] to have a balanced number of events (in the case of civil wars: war onsets) and non-events (peace) in the training dataset. Such a sampling strategy can be can be applied to any classification task, and is not restricted to Random Forest models. Yet, Muchlinski et al. (2016) only down-sample for the Random Forest model but for none of the logit model specifications.

Third, their Random Forest performance assessment is overly optimistic because of a problematic usage of cross-validation[5]. Muchlinski et al. (2016) use 10-fold cross-validation to train their models. Combining cross-validation with down-sampling (or any other sampling approach) is complicated (Hastie et al., 2011). Cranmer and Desmarais (2017, 152) note that: "It is still possible to overfit the data using hold-out methods such as cross-validation." Analogous to Hastie et al. (2011) there is a "Wrong and Right Way to Do Cross-validation"(245) when re-sampling (up-sampling or down-sampling) the training data. The right way is to make sure to sample inside the cross-validation procedure. Muchlinski et al. (2016) down-sample their data before (outside) the cross-validation procedure which in turn leads to serious overfitting of the training data. Therefore, they present overconfident in-sample performance measures for their Random Forest models.

Fourth, and most importantly, the predictive performance of Random Forest is never tested on out-of-sample data. In fact, the authors' model comparison is based on in-sample measures only. Muchlinski et al. (2016) claim that they train three logit model specifications and the Random Forest model on the civil war dataset with observations from 1945 to 2000, and then update this dataset for all countries in Africa and the Middle East from 2001 to 2014. They state that this gives them additional 737 observations with 21 civil war onsets, which

[4] Other methods include hybrid approaches where you do a little of both and possibly add synthetic data for the minority class (i.e. Synthetic Minority Oversampling Technique (SMOTE), (Chawla et al., 2002).

[5] In short, cross-validation generally involves randomly dividing the data into several folds, fitting the model of interest on all data not in a fold partition, and then testing the model with the held-out fold. This process is then repeated for all partitions.

are then used for their out-of-sample predictions. From that they conclude that "All logistic regression models fail to specify any civil war onset in the out-of-sample data. Random Forests correctly predicts nine of twenty civil war onsets in this out-of-sample data [...]" (Muchlinski et al., 2016, 96). This main conclusion of their paper is already prominently cited (see Cederman and Weidmann, 2017; Cranmer and Desmarais, 2017).

However, when we tried to replicate the results we found that the dataset used for the out-of-sample predictions does not contain any variables that were initially used to train and build the models. With this data it is, thus, not possible to get to any out-of-sample predictions[6].

So how did they come to their predicted probabilities of civil war onset? The simple answer is: The authors did not make out-of-sample predictions. They randomly draw 737 probabilities from the in-sample predictions and merge them to the out-of-sample observations of civil war onset. With these random numbers they want to predict civil war onsets in Africa and the Middle East from 2001 to 2014. This is certainly not an out-of-sample prediction of civil war onsets as the table (Table 1, Muchlinski et al., 2016, 98) with the predicted probabilities of civil war onset[7] in their paper suggests. With this procedure it is impossible to conclude that "Random Forests correctly predicts nine of twenty civil war onsets in this out-of-sample data"(Muchlinski et al., 2016, 96).

Fifth, another issue with this table of alleged predicted probabilities is that it only focuses on instances with civil war onsets and leaves aside the peaceful country-year observations. Thus, this table[8] only tells us half of the story. Based on their table one might assume that Random Forests is indeed a very good model to predict the onset of civil wars. The problem with this is that the authors only focus on their true positive (predicting war when there was war) and false negative (predicting peace when there was war) predictions. Assessing the performance of a model by only looking at the true positives and false negatives is misleading.

[6] The corresponding author was not able to provide additional data or code to clear this up.

[7] The Appendix includes the original R code used in the Muchlinski et al. (2016) paper, taken from the Political Analysis dataverse. At least with the available code and data we were not able to replicate the out-of-sample prediction and Table 1 from the paper.

[8] Even if the numbers were proper out-of-sample predictions.

To see this consider a simple example. A crude classifier that always predicts the onset of a civil war in any country in any year would have gotten twenty out of twenty right. But is this a good model of civil war onsets? Certainly not. A complete confusion matrix[9] of out-of-sample predictions could help clear this up.

3 Is Random Forest Really Better at Predicting Civil War Onsets?

We now turn to test whether Random Forest indeed outperforms standard logit models when our methodological concerns have been accounted for. Because the out-of-sample data provided by Muchlinski et al. (2016) does not allow for a strict out-of-sample prediction, we split the civil war data into one training and one test set, leveraging the time dimension. We created a training dataset containing all observations prior to 1990 (5299 observations, 88 civil war onsets (1.6%)) and a test data set with all observations between 1990 and 2000 (1841 observations, 28 civil war onsets (1.3%)).

We train two Random Forests using 10-fold cross-validation on the training data – one with the eleven variables specified in Fearon and Laitin (2003)[10] and one using all 90 variables in the dataset by Muchlinski et al. (2016)[11]. We do the same for the logit models[12]. We made sure that all models were trained on the same folds of the data. We also ensured by using the civil war onset variable as stratification variable that every fold contains roughly the same amount of civil war onsets.

After training the models we make out-of-sample predictions and base all our performance measures on these predictions. We present out-of-sample confusion matrices of the models at a threshold of 0.5, separation plots (see Greenhill et al., 2011), as well as ROC curves (Fawcett, 2006).

[9] A 2×2 matrix tabulating the observed vs. the predicted outcomes. See more in the following sections.

[10] We also implemented the Collier and Hoeffler (2004) and Hegre and Sambanis (2006) model specifications. The conclusions we draw from these comparisons are the same as the conclusions presented here. The detailed results can be found in the Appendix.

[11] We also implemented these models using the proper order of cross-validation and up-sampling to account for the class imbalance. The results are in the Appendix.

[12] Note that for the logit specification with all variables, not all coefficients can be estimated because some variables have zero variance or are perfectly collinear with other variables. However, for prediction those not available coefficients are set to zero.

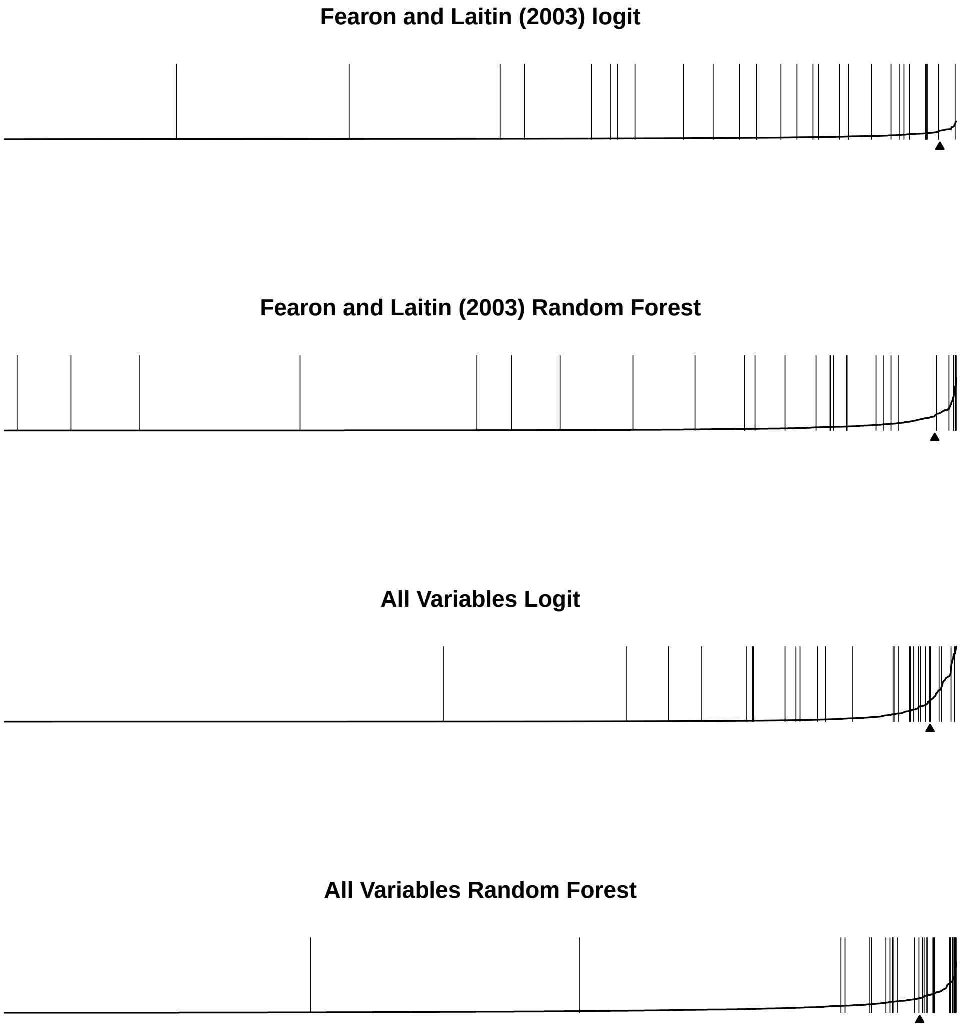

Figure 1: Separation Plots for the comparisons of the logit models and Random Forest with the Fearon and Laitin (2003) Variables and using all 90 Variables.

Figure 1 displays the four separation plots. The predicted probabilities are shown by the line from low to high. Actual civil war onsets are depicted by the dark vertical lines. The little triangle indicates the expected number of civil war onsets in the respective model. Thus, in a good model most of the dark lines should by to the right of the triangle while most white lines should be to the left. These plots also allow us to compare models across all potential thresholds of positive prediction. We can see that the Random Forest model using all 90 variables best separates the white and black lines. But the pattern is not as impressive as

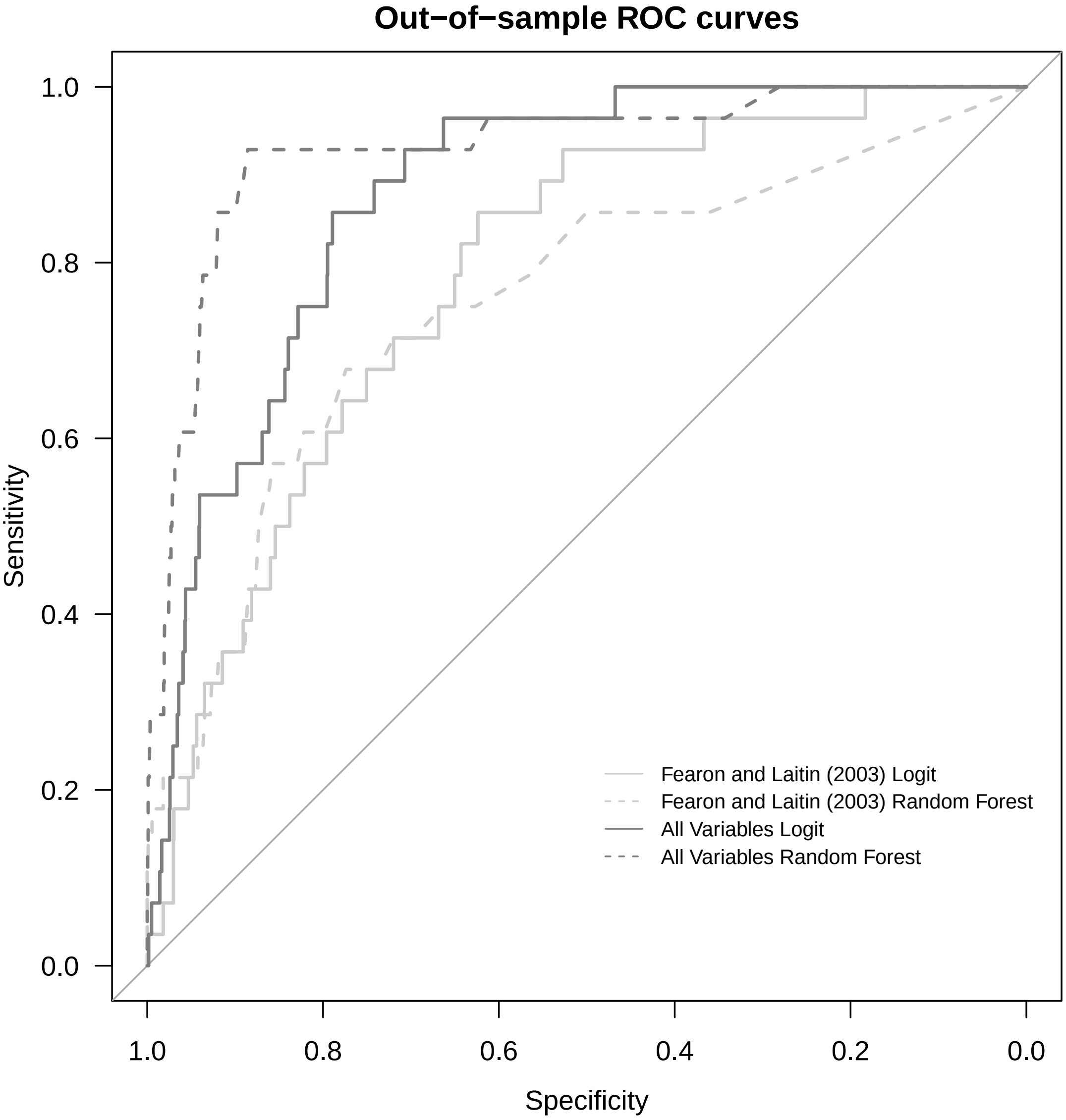

Figure 2: ROC curves for the comparisons of the logit models and Random Forest with the Fearon and Laitin (2003) Variables and using all 90 Variables.

stated by Muchlinski et al. (2016).

Another possibility to visualize the performance of binary classifiers are ROC curves. Figure 2 shows the ROC curves for the fair comparisons. A visual inspection of the plot shows that the Random Forest using the eleven variables of the Fearon and Laitin (2003) specification ($AUC = 0.76$ $[0.65; 0.85]$, 95% confidence intervals in squared brackets[13]) is not better at predicting civil war onset than the Fearon and Laitin (2003) logit specification

[13] These are bootstrapped 95% confidence intervals based on 2,000 stratified bootstrap samples implemented with the pROC (Robin et al., 2011) package in R.

Table 1: Out-of-sample Confusion Matrices.

Fearon/Laitin Logit

		Observed	
		War	*Peace*
Predicted	*War*	0	0
	Peace	28	1813

Fearon/Laitin RF

	Observed	
	War	*Peace*
War	3	2
Peace	25	1811

All Variables Logit

		War	Peace
Predicted	*War*	2	25
	Peace	26	1788

All Variables RF

	War	*Peace*
War	2	1
Peace	26	1812

($AUC = 0.79$ $[0.71; 0.85]$). Thus, we conclude that Random Forest is not superior to logit regression if the algorithmic approach can only rely on a limited set of variables. At first sight a Random Forest using all 90 variables ($AUC = 0.93$ $[0.87; 0.97]$) indeed seems to be better at predicting civil war onsets as compared to a logit model using the same variables ($AUC = 0.88$ $[0.83; 0.92]$). Statistically testing whether the difference in the areas under the ROC curves is significant reveals that the difference in the performance of the two classifiers is not statistically significant ($p = 0.12$)[14]. There might be a small advantage of Random Forest over the logit models, however we note that this advantage is by no means as impressive as stated by Muchlinski et al. (2016).

The confusion matrices at a threshold for predicting civil war onset of 0.5 in Table 1 show the same picture. From 28 civil war onsets that happened between 1990 and 2000, the logit model with the Fearon/Laitin specification predicts no civil war onsets, whereas the Random Forest with same specification predicts three onsets correctly. A Random Forest using all 90 variables correctly predicts 2 out of 28 civil war onsets, the same as a logit regression model using the same 90 variables for prediction. Therefore, with regard to the ultimate goal to correctly predicted civil war onsets, both methods perform equivalently. However, the Random Forest model only produces one false positive prediction (predicting war where we observed peace) and the logit model produces 25 false positive predictions. The confusion matrices in Table 1 communicate these results transparently. To sum up, we find that Random

[14] Testing the difference in AUC-values is also based on bootstrap samples using the `pROC` (Robin et al., 2011) package in `R`.

Forest does not provide significantly more accurate predictions of civil war onset than any of the logit model specifications.

4 Discussion: How to Make Better Model Comparisons?

In this section we offer a checklist for better model comparisons, based on our application of fixing the limitations in Muchlinski et al. (2016).

- **All competing models should be trained using exactly the same data.**

  This implies two points: First, when holdout methods such as cross-validation are used, make sure that the folds of the data are the same across all models so that each model is trained with the same data folds. Second, when comparing two different approaches (such as Random Forests and logit models) use the same data and variables to build the models.

- **Make sure that the order of sampling and cross-validation is correct.**

  If cross-validation and sampling (up-sampling or down-sampling) are done in the wrong order, serious overfitting of the training data can happen. It is crucial to make sure to sample inside the cross-validation procedure. For the right order in an analogous case see Hastie et al. (2011, 245).

- **Always report out-of-sample confusion matrices.**

  Table 2 shows a 2×2 confusion matrix for the example of civil war onset. We argue that whenever different classification models are compared, you should always report the out-of-sample confusion matrices. A confusion matrix gives an impression of how the different types of classification errors are distributed, for instance the amount of *false alarms* (false positives) a model produces when predicting an event which did not occur.

- **Be sure that you make proper out-of-sample predictions.**

  Out-of-sample prediction is the gold standard for the evaluation of predictive models.

Table 2: Confusion Matrix, Sensitivity and Specificity.

Confusion Matrix

		Observed	
		War	*Peace*
Predicted	*War*	True Positive (TP)	False Positive (FP)
	Peace	False Negative (FN)	True Negative (TN)

Performance Measures

Sensitivity (True Positive Rate) = $\frac{TP}{TP+FN}$
Specificity (1- False Positive Rate) = $1 - \frac{FP}{FP+TN}$

The idea is very simple: train the models you want to compare on the same training dataset, and then use the models to predict the outcomes of a test dataset (also holdout sample) containing observations that were not part of the training process. In most classification approaches, each observation of the test dataset is given a certain probability of belonging to a certain class. In our example the classes are a civil war onset and peace. To make a prediction for a certain observation, the test data set must of course contain the same variables that were used to train the model. Otherwise the out-of-sample comparison of the predictive performance becomes infeasible. If it is impossible to collect new data, one can always split the dataset at hand into training and test sample.

- **Calculate the performance measures based on the out-of-sample predictions.** Common indicators of model performance for binary classification include ROC curves[15] or Separation Plots. If the plots and performance measures are calculated based on in-sample predictions one might report overconfident results if the training data was overfitted. Always present these plots and performance measures based on the out-of-sample predictions. The same applies to confusion matrices.

- **Statistically test whether the performance measures are actually different.**

[15] Based on the ROC curves one can calculate the area under the ROC curve (ROC-AUC) as a summary measure.

> Performance measures like ROC curves or the area under the ROC curve are frequently applied to the comparison of binary classifiers. A ROC curve that lies further to the top left is often interpreted as an indicator for a better model. Pure eyeballing and subjective intuition ignores the uncertainty around ROC curves. Without testing the statistical difference, two ROC curves can be incorrectly labeled as similar, and vice versa. Statistical tools to test whether two ROC curves are statistically different already exist and should be applied when comparing binary classifiers (see Fawcett, 2006; Robin et al., 2011).

Machine learning certainly offers a lot of new models and algorithms that could and should be added to the toolboxes of social scientists. However, we need to make good model comparisons in order to asses whether the introduction of a new method is worthwhile. In this paper we make available some criteria to researchers to assess their models' performance. When applying our criteria to a recent and already prominently cited application of Random Forest to the prediction of civil war onsets it, turns out that the impressive performance does not really hold.

This implies one important message to researchers of civil wars. So far, they do not miss much when applying traditional logit models. Others will try to find better models of predicting civil war onset. If you found one be sure the results hold when rigorously challenged. To move the field forward and to save lives we need to be certain that new models actually perform better.

References

Cederman, L.-E. and N. B. Weidmann (2017). Predicting armed conflict: Time to adjust our expectations? *Science* (355), 474–476.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research 16*, 321–357.

Chen, C., A. Liaw, and L. Breiman (2004). Using random forest to learn imbalanced data. *University of California, Berkeley* (1999), 1–12.

Collier, P. and A. Hoeffler (2004). Greed and Grievance in Civil War. *Oxford Economic Papers 56*(4), 563–595.

Cranmer, S. J. and B. A. Desmarais (2017). What can we Learn from Predictive Modeling? *Political Analysis 25*(2), 145–166.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters 27*(8), 861–874.

Fearon, J. D. and D. D. Laitin (2003). Ethnicity, Insurgency, and Civil War. *American Political Science Review 97*(1), 75–90.

Greenhill, B., M. D. Ward, and A. Sacks (2011). The Separation Plot: A New Visual Method for Evaluating the Fit of Binary Models. *American Journal of Political Science 55*(4), 991–1003.

Hastie, T., R. Tibshirani, and J. Friedman (2011). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* (2 ed.). New York: Springer.

Hegre, H. and N. Sambanis (2006). 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.

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.

Robin, X., N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics 12*, 77.

Appendix

List of Variables (variable names taken from the Hegre and Sambanis (2006) codebook):

- Fearon and Laitin (2003):

  warhist (Prior war); *ln_gdpen* (Per capita income); *lpopns* (Population (log)); *lmtnest* (Rough terrain % (log)); *ncontig* (Noncontiguous state); *oil* (Oil exporter); *nwstate* (New state); *inst3* (Political instability); *pol4* (Polity IV Democracy Index); *ef* (Ethnic fractionalization); *relfrac* (Religious fractionalization)

- Collier and Hoeffler (2004):

  sxpnew (Primary commodity exports/GDP); *sxpsq* (Primary commodity exports/GDP (squared)); *ln_gdpen* (GDP per capita (logtransformed)); *gdpgrowth* (Annual change in GDP (percent)); *warhist* (Prior war); *ef* (Ethnic fractionalization); *popdense* (Population density (people per square kilometer)); *lpopns* Population density (logtransformed); *coldwar* Cold War year (before 1990); *seceduc* (School enrollment (secondary)), *ptime* (Peace time)

- Hegre and Sambanis (2006):

  lpopns (Population density (logtransformed)); *ln_gdpen* (GDP per capita (logtransformed)); *inst3* (Political instability); *parreg* (Regulation of participation); *geo34* (Region: Middle East and North Africa); *proxregc* (Year since last regime transition (transformed)); *gdpgrowth* (Annual change in GDP (percent)); *anoc* (Dummy anocracy); *partfree* (Partially free polity); *nat_war* (Neighbor in war); *lmtnest* (Rough terrain % (log)); *decade1* (Decade dummy 1960s); *pol4sq* (Polity IV (squared)); *nwstate* (New state); *regd4_alt* (Median regional polity (using polity 2)); *etdo4590* (Ethnic dominance measure); *milper* (Share of population in military forces); *geo1* (Region: Western Europe and the United States); *tnatwar* (Total number of neighbors at war in a given year); *presi* (Presidential system)