Political Analysis (2015) pp. 1–17
doi:10.1093/pan/mpv024

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

David Muchlinski
School of Social and Political Science, University of Glasgow, Glasgow, UK
e-mail: david.muchlinski@glasgow.ac.uk (corresponding author)

David Siroky
Department of Political Science, Arizona State University, Tempe, AZ
e-mail: david.siroky@asu.edu

Jingrui He
Department of Computer Science and Engineering, Arizona State University, Tempe, AZ
e-mail: jingrui.he@asu.edu

Matthew Kocher
Department of Political Science, Yale University, New Haven, CT
e-mail: mathew.kocher@yale.edu

Edited by R. Michael Alvarez

The most commonly used statistical models of civil war onset fail to correctly predict most occurrences of this rare event in out-of-sample data. Statistical methods for the analysis of binary data, such as logistic regression, even in their rare event and regularized forms, perform poorly at prediction. We compare the performance of Random Forests with three versions of logistic regression (classic logistic regression, Firth rare events logistic regression, and L_1-regularized logistic regression), and find that the algorithmic approach provides significantly more accurate predictions of civil war onset in out-of-sample data than any of the logistic regression models. The article discusses these results and the ways in which algorithmic statistical methods like Random Forests can be useful to more accurately predict rare events in conflict data.

1 Prediction, Prediction, Prediction

Prediction is a contentious issue in the discipline of political science. Political scientists are generally taught not to be especially concerned with model fit, but to emphasize the estimation of causal parameters instead. The empirical analysis of civil war onset, a major topic in comparative politics and international relations over the past decade and a half, has mostly followed this prescription. One consequence, which we demonstrate below, is that most statistical models of civil war onset have exceedingly weak predictive power. This fact should be a cause of concern, for two reasons. First, civil wars are incredibly destructive, and accurately predicting their onset is a critical issue for policymakers who must try to anticipate conflicts and find ways to prevent or contain them. Second, the poor predictive power of our models may be a good indication that our existing estimates of causal parameters are not very reliable. Political methodology has long recognized the value of "maximizing leverage" (King et al. 1994, 29–31), or accounting for as much variation in the dependent variable as possible with the smallest number of right-hand-side variables as possible, which is at base a predictive criterion of scientific value.

Author's note: Replication data are available on the Political Analysis Dataverse at http://dx.doi.org/10.7910/DVN/KRKWK8.

While the limited predictive success of civil war onset models may result in part from the need for better theory and data, part of the reason for the limited success of these models stems from the application of relatively restrictive methods to data that are both class imbalanced and exhibit complex nonlinear interactions among covariates. Statistical machine learning methods, such as Random Forests, offer an alternative approach—one currently underutilized in political science—for increasing predictive accuracy. Whereas the utility of Random Forests in a wide range of disciplines has now been firmly established, the method has only recently gained attention in political science (Spirling 2008; Hill and Jones 2014; Blair et al. 2015; Jones and Linder 2015). This article compares the predictive performance of Random Forests with three versions of logistic regression (standard logistic regression, Firth's penalized logistic regression [Firth 1993], and L_1-regularized logistic regression), and finds that the algorithmic approach provides significantly more accurate predictions of civil war onset than any of the logistic regression models.

While identifying causal effects is one essential scientific goal, prediction is another important objective for many substantive issues in political science—civil war onset is certainly one. The analysis below therefore examines how algorithmic methods measure up against models more widely utilized in the study of civil war, particularly logistic regression models designed for rare events and L_1-regularized logistic regression, in terms of out-of-sample predictive accuracy. The vast majority of countries within any given year do not experience a civil war onset, making the occurrence of such an event uncommon. However, because civil wars are often so destructive, accurately predicting their onset is critical to scholars who study civil war and related forms of political instability, as well as to policymakers that must anticipate conflicts and find ways to end or at least contain them.[1]

2 Predicting Civil War Onset

Most statistical research in political science is concerned with identifying causal effects rather than prediction (Beck et al. 2000; Ward et al. 2010). Recently, however, more scholars have advocated using out-of-sample data to evaluate and compare models (Goldstone et al. 2010; Ward et al. 2012; Hegre et al. 2013; Schrodt et al. 2013). There is now a growing literature developing and applying methods to predict occurrences of rare but destructive events such as civil war (Hegre et al. 2013; Shellman et al. 2013; Brandt et al. 2014; Clayton and Gleditsch 2014), interstate disputes (Gleditsch and Ward 2012), and political instability (Goldstone et al. 2010). If we claim that our theories have implications for events with the potential to affect lives, we should examine whether our theories make correct predictions.

Explanation and prediction are distinct, though related, enterprises (Shmueli 2010).[2] Some very well-understood causal processes are largely unpredictable, either because it is too difficult or expensive to measure the relevant independent variables (e.g., earthquakes) or because the causal processes are chaotic (e.g., weather beyond a three- or five-day window).[3] Likewise, quite accurate predictions can sometimes be secured in spite of a relatively poor understanding of the underlying causal processes that generate the outcome or through the use of a causal model that is known, in advance, to be false (Breiman 2001b).

In most understandings of causality, explanations *imply* predictions in at least the minimal sense that the explanation is inconsistent with at least some possible states of the world. Under the now-dominant Neyman–Rubin model of causal inference, causal statements amount to propositions about the difference between the potential outcomes of units under counterfactual treatment and control conditions (Holland 1986). In other words, a causal statement implies a *prediction* about the outcome, conditional on the unit's assignment to treatment or control. If the prediction turns out to

[1]This article is interested in the problem of civil war onset, which is the most common response variable in civil war studies, rather than in its recurrence and termination, which are different targets to be addressed in future research.
[2]We use the terms "explanation" and "causal explanation" interchangeably in this discussion. Although there may be other kinds of explanations, the social sciences are primarily concerned with causal relations. From the point of view of algorithmic modeling, the prediction of new observations and the "retrodiction" of an existing test set are equivalent procedures. We use the term "prediction" to encompass both.
[3]See Silver (2012) for an accessible comparison of predictive modeling across many domains.

be false, it counts against the truth of the causal proposition. Likewise, observational causal models are evaluated on the basis of statistics that summarize the deviation of predicted values from observed values of the outcome. Thus, successful causal explanations, whether experimental or observational, will tend to facilitate empirical prediction (i.e., when a causal statement is true, it tells us what to expect under specific conditions). However, even highly credible causal research designs will often account for small fractions of the overall variation in the outcome; in this sense, even the most rigorously identified causes may not be particularly strong predictors (Ward et al. 2010).

The fact that a statistical model may only explain a small amount of overall variance is often ignored in the context of causal explanation. When making generalizations to the world outside their data, researchers often conflate explanations for predictions (Ward et al. 2010). The proliferation of explanations masquerading as predictions may hamper the effectiveness of policies designed to alleviate human suffering as well as basic knowledge generation regarding the complex causes of civil war onset (Shmueli 2010).

Often, a researcher is more interested in the "causes of effects" than in the "effects of causes" (Gelman and Imbens 2013). Public policy considerations may outweigh the value of basic science in a particular domain. Rigorous causal identification may be infeasible for practical or ethical reasons. Large, multidimensional data sets, coupled with theoretical underdevelopment, may undermine the credibility of causal modeling assumptions. For any of these reasons, or more likely a combination of all three, it may be useful to embrace prediction as the explicit goal of research, rather than solely as a criterion of evaluation for causal models.

The study of civil war onset is particularly ripe for the application of algorithmic modeling, for several reasons. First, our reading of the literature indicates that the field's most influential statistical models of civil war have exceedingly poor predictive power (Beck et al. 2000; Ward et al. 2007, 2010). Second, many theoretically significant variables are not robust across models and specifications (Hegre and Sambanis 2006). Third, and partly as a consequence, theoretical disputes over the most appropriate causal models remain fundamental (Kalyvas 2007). Taken together, these considerations give us good reason to doubt that existing causal models of civil war onset approximate "true" models of the phenomenon, which undermines the usefulness of parameter estimates derived. For these reasons, we make prediction the explicit goal of our enterprise, and utilize statistical learning methods like Random Forests, which have been shown to generate substantially more accurate predictions than traditional parametric methods (Montgomery et al. 2012).

Although there are numerous statistical learning approaches, which are also widely implemented in several statistical software platforms, we focus the comparison in this article on the Random Forest algorithm for three reasons. It is easily interpretable (Hastie et al. 2009), which has contributed to its popularity among analysts. Further, its decision tree growing procedure is robust to outliers and other influential observations. Finally, downsampling procedures allow the analyst to easily correct for class-imbalanced data (Chen et al. 2004). In the remainder of this article, we compare the ability of Random Forest to predict true onsets of civil war to three logistic regression approaches (Firth 1993; Lee et al. 2006).

3 Data and Methods

The Civil War Data (CWD)[4] are measured annually for each recognized country in the world from 1945 to 2000 (Hegre and Sambanis 2006). The dependent variable $Y^{ij}_{cw=[0,1]}$ is a binary measure of whether a civil war onset occurred for a given country, i, in a given year j. N is equal to 7141 county-years. $\mathbf{X}$ is a matrix of eighty-eight predictor variables. Using ten-fold cross-validation, we trained our models on nine of the ten folds of the data, then passed the predictions made in the training sets to the test data. Cross-validation allows researchers to examine the predictive abilities

[4]For replication materials, please see Muchlinski (2015).

of statistical models and algorithms without collecting new data, and has been shown to produce unbiased and accurate error rates (Hastie et al. 2009).[5]

The CWD are extremely unbalanced. The ratio of conflict years to peace years in the data is roughly 1:100. It is likely that the predictions of the logistic regressions will be biased toward the majority class, making logistic regression a poor predictor of civil war onset. In political science, class-imbalanced data are usually corrected for by utilizing a correction for rare events within logistic regression (Firth 1993; King and Zeng 2001). Just as a single or a small number of predictor variables can perfectly predict $Y = 1$ in small data sets, a similar problem can obtain when there is a small ratio of positive cases (1s) to negative cases (0s) in the data (Zorn 2005). Imbalanced classification has been studied extensively in the machine learning and data mining community (Chawla 2005). Existing techniques include sampling methods, like downsampling, which we utilize in this article (Ling and Li 1998; Chawla et al. 2002; Cieslak and Chawla 2008; Köknar-Tezel and Latecki 2011), and ensemble-based methods (Chawla et al. 2003; Sun et al. 2006), among others.[6]

In the following subsections, we discuss our statistical procedures.

3.1 *Logistic Regression*

In logistic regression, a dependent variable is given by $Y_i(i = 1, \ldots, n) \sim Bernouli(Y_i|p_i)$, so that it takes on a value of 1 with probability p_i and 0 with probability $1 - p_i$ over n number of trials. The vector of input variables is given by $\mathbf{x_i}$, and p_i varies over this explanatory space such that

$$p_i = \frac{1}{1 + e^{-x_i\beta}}. \tag{1}$$

If Y_i is conceived of as a latent continuous variable Y_i^* (e.g., the probability of a country experiencing the onset of a civil war) distributed according to a logistic density function with mean μ_i, then

$$\begin{aligned} Y_i^* &\sim \text{Logistic}(Y_i^*|\mu_i) \\ \mu_i &= \mathbf{x_i}\beta, \end{aligned} \tag{2}$$

where $\text{Logistic}(Y_i^*|\mu_i)$ is the one-parameter logistic PDF,

$$Y_i^* = \frac{e^{-Y_i^* - \mu_i}}{(1 + e^{-Y_i^* - \mu_i})^2}, \tag{3}$$

then the probability of observing the dichotomous realization of Y_i^* is

$$\begin{aligned} &\Pr(Y_i = 1|\beta) = p_i = \Pr(Y_i^* > 0|\beta) = \\ &\int_0^\infty \text{Logistic}(Y_i^*|\mu_i) dY_i^* = \frac{1}{1 + e^{-x_i\beta}}, \end{aligned} \tag{4}$$

which is the more general binomial case of n Bernoulli trials of Y_i over the vector $\mathbf{x_i}$. The parameters of the model are estimated by maximum-likelihood, where the likelihood function is

$$-\sum_{i=1}^{n} \ln(1 + e^{(1-2Y_i)x_i\beta}). \tag{5}$$

[5] We also estimated all models and algorithms with five-fold cross-validation to determine if the size of each fold affected the accuracy of predictions. There were no substantive differences between ten- and five-fold cross-validation.

[6] Because prediction of every case in the data required a complete data set, missing data were imputed three separate ways. For space considerations, we present comparisons using only one method of imputation in the text. Methods of imputation included Amelia II (Honaker et al. 2011), Multiple Imputation Using Chained Equations (Buuren and Groothuis-Oudshoorn 2011), and Random Forest (Breiman 2001a). We present the results of imputations using random forests in the text. Predictions were not affected by imputation procedure.

If the data are balanced between the two classes, maximum likelihood estimates are consistent and asymptotically efficient. However, this is not the case when data are extremely unbalanced between the classes, as in data with a small number of rare events. The estimation of class-imbalanced data returns low estimates of $\Pr(Y_i = 1|\mathbf{x_i}) = p_i$ due to the structure of the variance matrix shown below:

$$V\hat{\beta} = \left[\sum_{i=1}^{n} p_i(1-p_i)x_i'x_i\right]^{-1}. \tag{6}$$

The part of this matrix that is affected by class imbalances in the data is $p_i(1-p_i)$. Thus, it can be difficult to predict whether a country might experience a civil war onset because the predicted probabilities of true events returned by the model will be closer to 0 than to 0.5. There are some ways to correct for this bias. We explore two that have been used variously in the literature: rare event logistic regression (Firth 1993; King and Zeng 2001), and L_1-regularized logistic regression (Lee et al. 2006; Park and Hastie 2007; Ravikumar et al. 2010). The researcher can also alter the threshold for positive prediction τ. Given the rarity of events in the data, however, τ might need to be set at an extremely small value in order to generate any true positive predictions—suggesting that the statistical model itself is a poor predictor of civil war onset.

3.2 *Rare Event Logistic Regression*

Previous research by Firth (1993) and by King and Zeng (2001) has demonstrated two equivalent ways to approximate equation (1) in the presence of class-imbalanced data. Let $\mathbf{x}_0$ be a $1 \times k$ vector of values along the predictor variables. The method for computing the probability, given $\mathbf{x}_0$, is given by estimating the true value of the parameter estimate β with $\hat{\beta}$, a biased estimate:

$$\Pr(Y_0 = 1|\hat{\beta}) = \hat{p_0} = \frac{1}{1+e^{-x_0\hat{\beta}}}. \tag{7}$$

King and Zeng (2001) demonstrate that $\hat{\beta}$ is biased downward in class-imbalanced data and that it generates predicted probabilities that underestimate the actual probability of a rare event. To correct for the bias arising from class-imbalanced data, a correction to the likelihood function should be utilized.

Firth (1993) suggests adopting such a correction to the likelihood function. Given $\mathbf{x_i}$, the maximum likelihood is found where

$$\nabla l(\mathbf{x_i}) = 0,$$

where l is the log-likelihood function given in equation (5). Firth corrects the bias in the likelihood function though application of the Jeffrey's invariant prior such that

$$L^*(\mathbf{x_i}) = L(\mathbf{x_i})|i(\mathbf{x_i})|^{\frac{1}{2}},$$

where i is the Fisher information matrix. Firth shows that this method returns estimates that are unbiased in class-imbalanced data.

3.3 *L_1-Regularized Logistic Regression*

Unregularized logistic regression—or more simply logistic regression—is an unconstrained convex optimization problem with a continuously differentiable objective function. As a result, it can be fairly easily solved with standard convex optimization problems like Newton–Raphson and other common procedures. L_1-regularized logistic regression is widely utilized in the machine learning community as a classifier when data are high dimensional. L_1 logistic regression minimizes $\|\theta\|_1$, where θ is the model's weight vector, and $\|\cdot\|_1$ is the L_1 norm. The L_1 constraint is imposed to prevent overfitting and to aid in feature selection. The L_1 constraint shrinks the estimate of features not contributing to the model's classification accuracy to 0, thus returning only parameters that aid in feature selection (Park and Hastie 2007).

The addition of the L_1 constraint, however, makes the optimization problem computationally more intensive to solve. If the L_1 regularization is enforced by an L_1 norm constraint on the parameters, then the optimization problem becomes an optimization problem with the number of constraints equal to twice the number of parameters to be learned (Lee et al. 2006). A number of ways to solve the L_1 optimization problem have been proposed, but an analysis of this literature is beyond the scope of this article.

For L_1-regularized logistic regression, we consider a supervised learning procedure over N training sets $[(\mathbf{x_i}, y_i), i = 1, \ldots, N]$. Regularized logistic regression begins with modeling the probability distribution of Y given the vector $\mathbf{x_i}$ according to the equation given in (1).

Here, $\mathbf{X} \in \mathbb{R}^N$ are the parameters of the logistic regression model. Under the Laplacian prior $p(\mathbf{X}) = (\beta/2)^N e(-\beta\|\mathbf{X}\|_1)$ with $(\beta > 0)$, the maximum a posteriori (MAP) estimate of the parameters is given by

$$\min \sum_{i=1}^{N} -\log p(y_i|\mathbf{x}_i; \mathbf{X}) + \beta\|\mathbf{X}\|_1. \tag{8}$$

This optimization problem is referred to an L_1-regularized logistic regression. Often, it will be convenient to consider the following alternative parameterization of the L_1-regularized logistic regression:

$$\min \sum_{i=1}^{N} -\log p(y_i|\mathbf{x}_i; \mathbf{X}), \text{ subject to} \|\mathbf{X}\|_1 \leq C. \tag{9}$$

The optimization problems in (8) and (9) are equivalent. For any choice of β, there is a choice of C such that both optimization problems have the same minimizing argument since (8) is the Lagrangian of the constrained optimization problem in (9), where β is the Langrange multiplier (Lee et al. 2006).

3.4 *Random Forests*

Statistical learning methods like Random Forests have rarely been used in political science (for overviews, see Siroky 2009), but have gained some attention recently (Spirling 2008; Schrodt et al. 2013; Shellman et al. 2013; Hill and Jones 2014; Jones and Linder 2015). Unlike logistic regression, where a statistical model that was likely to have generated that data is specified by the researcher prior to estimation, no "model" in the conventional sense is generated by Random Forests. Random Forests grow a forest of classification trees to the data. A classification tree for a binary outcome uses a training sample of n cases. Case i has a vector of covariates x_i that are used to build a tree-structured classification rule. Recursive partitioning splits the training sample into increasingly homogeneous groups by inducing a partition on the explanatory space. Three kinds of splits using the vector of inputs x include

1. Univariate split: Is $x_i \leq t$?
2. Linear combination split: Is $\sum_{i=1}^{p}(w_i x_i) \leq t$?
3. Categorical split: Is $x_i \in S$.

The split searches for the separation that best differentiates the cases in the training sample into two maximally homogeneous groups, which has been defined variously in the literature. Formally, the tree is grown according to the following equation:

$$Y = \sum_{j=1}^{r} \beta_j I(x \in R_j) + \varepsilon, \tag{10}$$

where the regions R_j and the coefficients β_j are estimated from the data. The R_j are usually disjoint, and the β_j is the average of the Y values in the R_j.

Random Forests is a collection of n identically distributed decision trees, where each tree is built using the classification algorithm described above and bootstrap samples from the training set. An "un-pruned" or complete classification tree is formed for each bootstrap sample, meaning that all terminal nodes are "pure" (all zeros or ones). Approximately two-thirds of all observations are used to grow each tree. Once the set of classification trees has been grown on the bootstrapped samples, unsampled cases from the test set are dropped down each tree. These "out-of-bag (OOB)" cases are used to generate predictions and to calculate prediction error rates. The values of the explanatory variables are used to classify each observation into a terminal node, and the tree that most accurately classifies the test set of observations prevails.

The forest uses randomness in the tree building process and the aggregation process, and relies on a random sample of covariates (P) and cases (N). This produces some desirable properties, including highly accurate predictions, robustness to noise and outliers, internally unbiased estimate of the generalization error, efficient computation, and the ability to handle large dimensions and many predictors. It handles missing data well and provides estimates of relative importance of each covariate in the classification rule (Strobl et al. 2008). It is also possible to compute proximities between pairs of cases that are useful in clustering and identifying outliers or influential observations.

The Random Forest algorithm also contains a downsampling procedure that allows the researcher to select a balanced subsample of positive and negative cases. Such a procedure allows the algorithm to weight the relative importance of each class according to such a ratio, and to generate correct predictions of rare events even in extremely imbalanced data (Chen et al. 2004; Weidmann 2008).

4 Evaluating Random Forests versus Logistic Regression: Predicting Civil War Onset

Our evaluation of Random Forests and logistic regression is divided into two parts. The first, presented in this section, evaluates the ability of Random Forests to correctly predict true instances of civil war onset in out-of-sample data. We show the superior predictive power of Random Forests through separation plots, Receiver Operating Characteristic (ROC) Curves, and the F_1 Score, which is the harmonic mean of precision and recall. The second compares the variable importance measures generated by Random Forests to the well-known results of the literature on cross-national civil war onset and discusses what Random Forests can tell us about civil war onset.

We compare Random Forest against three well-known logistic models of civil war onset developed in Collier and Hoeffler (2004), Fearon and Laitin (2003), and Hegre and Sambanis (2006). We have chosen these models of civil war onset because they are well known and were responsible for a paradigm shift in the quantitative analysis of civil war. Although these models generally have poor predictive power (see, for example, Ward et al. 2007, 2010), we believe it is useful to compare the predictive accuracy of Random Forests to these models since these models are well known, and because debate regarding the causes, implications, and predictions of civil war implied by these models is still unsettled (Cederman et al. 2013).

One stringent test of a model's predictive accuracy is how well a theoretically informed statistical model of civil war onset predicts rare events in out-of-sample data. As data on the future cannot exist, we use cross-validation, which has been widely used to assess the relative predictive performance of statistical models in many disciplines (Geisser 1975; Efron 1983; Hastie et al. 2009), and is gaining increasing use in political science (Hoff and Ward 2004; Ward and Hoff 2007; Ward et al. 2007, 2010; Hill and Jones 2014). Cross-validation involves taking the CWD under examination and breaking it up into different "folds." A number of these folds (commonly five or ten) are used to train the model, while a separate fold is held out to test the predictions made by the model in the training data (Hastie et al. 2009). The model or algorithm is first trained on the training cross-validation folds, and then we examine out-of-sample data. We utilize ten-fold cross-validation to examine how well the three logistic models predict civil war onset. We compare the results of this exercise to the predictions made by Random Forest. For Random Forest, we utilize ten-fold cross-validation to maximize comparability as well as Random Forest's own internally generated OOB

estimates of error, which have been shown to closely approximate error rates determined by cross-validation (Breiman 1996).[7]

One way to gauge the predictive accuracy of any classifier is to use a technique known as a separation plot (Greenhill et al. 2011). The separation plot begins by rearranging the data such that the fitted values are presented in ascending order. A model's fit can be visualized by seeing the extent to which the actual instances of civil war onset (in gray) are concentrated in the right-hand side of the plot, while instances of peace (in white) are concentrated in the left-hand side of the plot. We also include a line that represents the predicted probability of civil war onset for each observation using either the logistic regressions or Random Forests. In adding this line, we can observe if the model makes correct predictions throughout the range of the data, or only when actual onsets become clustered together so that the probability of the model identifying an event is quite high. Finally, we add a triangle to the bottom of each plot representing the expected number of events. A model that perfectly separates onsets from peace will show all white to the left of the triangle and all gray to the right.

The separation plots for each of the uncorrected logistic classifiers and Random Forests are shown in Fig. 1.[8] Each logistic model of civil war onset generally does a poor job of accurately separating the data. There is a large amount of gray on the left-hand side of each plot—indicating that the logistic classifiers fail to accurately predict the vast majority of actual civil war onsets. Instead, the logistic classifiers inaccurately classify these actual instances of conflict onset as negative cases, committing a large number of Type II errors. The separation plot for Random Forests tells a different story. There is only white on the left-hand side of the plot, indicating that Random Forests can learn complex patterns in the class-imbalanced data and more accurately separate instances of peace from instances of conflict onset. There is still some white on the right-hand side of the plot, indicating that Random Forests does make some Type I errors. All gray, however, is on the right-hand side of the plot, indicating that Random Forests accurately predicts nearly every onset of civil war in the data.

Another way to visualize the predictive performance of a binary classifier is with an ROC plot. An ROC graph is a technique for visualizing, organizing, and selecting classifiers based on their performance. An ROC graph illustrates the performance of a binary classifier—like logistic regression or Random Forests—as its discrimination threshold is varied. ROC graphs are especially useful for applications where data are class imbalanced or have unequal classification error costs, since the graph measures the true and false positive rates for a classifier, rather than the total number of true or false positives predicted—which can vary substantially depending on the class distribution in the data (Fawcett 2006). The ROC graph is also easily summarized by a single metric called the area under the ROC curve (AUC). The AUC is a numerical summary of the probability that a given classifier ranks a randomly chosen positive observation higher than a randomly chosen negative one. ROC curves that are pulled to the upper left of the plot, and hence have higher AUC scores, represent superior classifiers. In short, the larger the AUC score, the better the model or algorithm's predictive accuracy.

The ROC curves for all classifiers are shown in Fig. 2. The ROC plot on the left shows the ROC curves and AUC scores for the uncorrected logistic regressions as well as Random Forests. The plot on the right compares Random Forests with the same logistic regressions fitted with a penalized likelihood (Firth 1993). The AUC scores are given for each classifier in the legend at the bottom right corner of each plot. Random Forests outperform all logistic regressions by a sizable margin. AUC scores in the 0.70 s are considered average, AUC scores in the 0.80 s are considered good. Any AUC over 0.90 is considered to be an excellent classifier. The logistic models specified in three of the most influential statistical analyses on civil war onset thus range from average to good, regardless of

[7]We present the results of Random Forests with cross-validation error rates in the text. The out-of-bag error rates were not substantially different from the cross-validation error rates. The cross-validation error rate was 25.16%, while the OOB error rate was 27.25%.

[8]There are no substantive differences for each set of plots, so we focus only on the uncorrected logistic classifiers in the text.

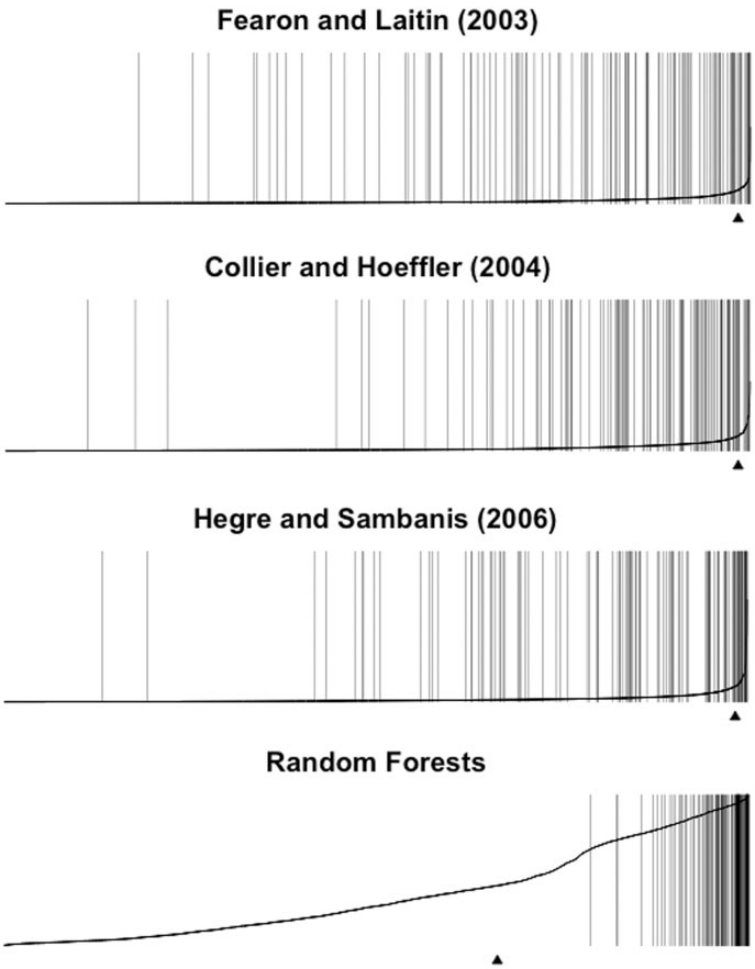

Fig. 1 Separation plots for all classifiers.

whether the uncorrected logistic or the penalized logistic regression is estimated, whereas the performance of Random Forests on the cross-validated data sets is "excellent" (above 0.90).

Also, using the F_1-score as the performance measure, we systematically compare the performance of Random Forests with logistic regression and L_1-regularized logistic regression (Lee et al. 2006). In our experiments, we vary the ratio of the training set such that the percentage of the entire data used for training ranges between 0.2 and 0.8. For each ratio, Fig. 4 compares the performances of all three methods in terms of the average F_1-score (where higher F_1 is better). The error bars (the standard deviation) show the amount of variance with multiple runs of the corresponding method. Random Forests outperforms the other two methods, in terms of both higher average F_1-score and lower standard deviation. This is particularly the case when the ratio of the training set is small. For example, when the ratio is only 0.2, the average F_1-score using Random Forests is close to 0.68, whereas the average F_1-score using the other two methods is less than 0.62; at the same ratio, the error bar for Random Forests is significantly shorter than that for the other two methods, showing that the performance of Random Forests is more stable across multiple runs. In other words, even if trained on limited historical data, Random Forests is able to accurately identify civil war onsets, whereas the performance of logistic regression and L_1-regularized logistic regression is much worse in terms of both the average F_1-score and variance.[9] The F_1-score for Random Forests ranges from

[9]Random Forests has larger variance than logistic regression or L_1-regularized logistic regression when the ratio of the training set to the test set is large (0.8). This is most likely due to the class imbalance in the data and the small size of the test set, which likely exacerbates the class imbalance. Random Forests may not be an accurate classifier if test data

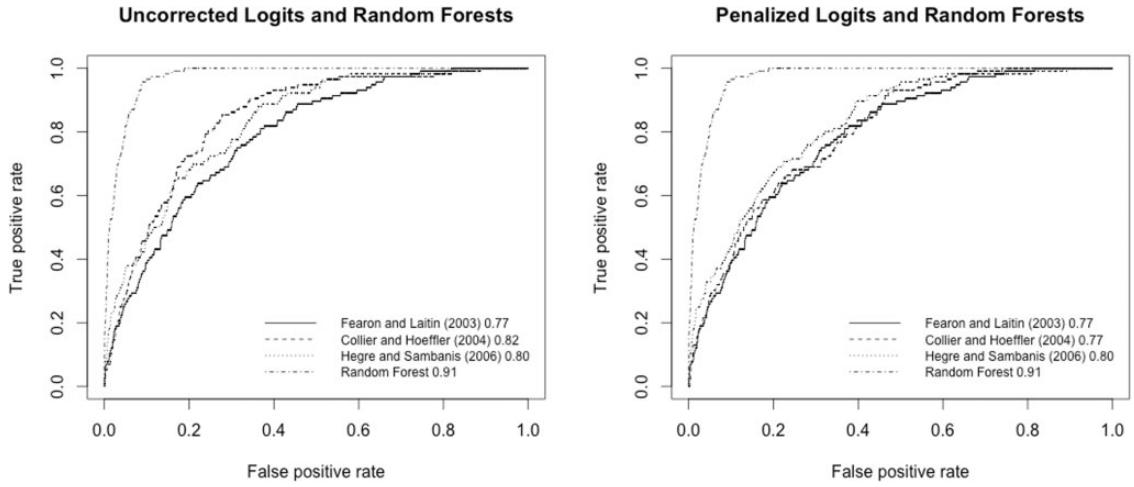

Fig. 2 ROC curves for all classifiers.

0.68 to close to 0.75, whereas the F_1-scores for logistic regressions are much lower except when the ratio of the training set is large. This indicates that logistic regression is a poor learner. It requires much more training data to make accurate predictions than does Random Forests.

Finally, we assess the predictive accuracy of the three logistic regression models in Figs. 1 and 2 to Random Forests on out-of-sample data. The CWD described in Section 3 runs from 1945 to 2000. We updated the CWD for all countries in Africa and the Middle East from 2001 to 2014. The updated data give us an additional 737 observations with twenty-one civil war onsets. We trained each model or algorithm on the 1945–2000 CWD and tested on the updated CWD for Africa and the Middle East. We assess each model or algorithm's predictive accuracy using the AUC score defined earlier. Table 1 reports the AUC scores for each model as well as for Random Forests. Random Forests is superior to all logistic models with an AUC of 0.60.[10] The predicted probabilities of civil war for each model are shown in Table 1.

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. Random Forests correctly predicts the onset of civil war in Iraq, Somalia, the Democratic Republic of the Congo, Uganda, Rwanda, and Liberia. It fails to correctly predict the civil wars resulting from the U.S. invasion of Afghanistan, or the civil wars in Syria and Libya that resulted from the Arab Spring. It is possible that civil wars resulting from external intervention or revolutions may have different causes and are thus poorly predicted when civil war is defined only by the number of deaths in battle.[11] We discuss the application of statistical learning methods to the analysis of causality in the following section.

5 Opening the Black Box of Civil War Onset Using Random Forests

In this section, we examine how machine learning algorithms like Random Forests can enhance our understanding of civil war onset. This is primarily an exploratory exercise since algorithms are utilized more for predictive purposes than for identifying causal effects (Hastie et al. 2009), and because the literature surrounding the use of statistical algorithms to make inferences is still being developed (Duncan 2014; Hill and Jones 2014). Statistical analysis can be thought of as divided into

are limited relative to training data. Such a ratio of training set size relative to the size of the test set is larger than common practice would dictate.

[10]The Hegre and Sambanis (2006) model has an AUC score of 0.40, Fearon and Laitin (2003) has an AUC score of 0.43, and the Collier and Hoeffler (2004) model has an AUC of 0.55.

[11]We follow Hegre and Sambanis (2006) in defining a civil war as any intrastate conflict between a government and another violent actor where at least 1000 battle deaths have occurred.

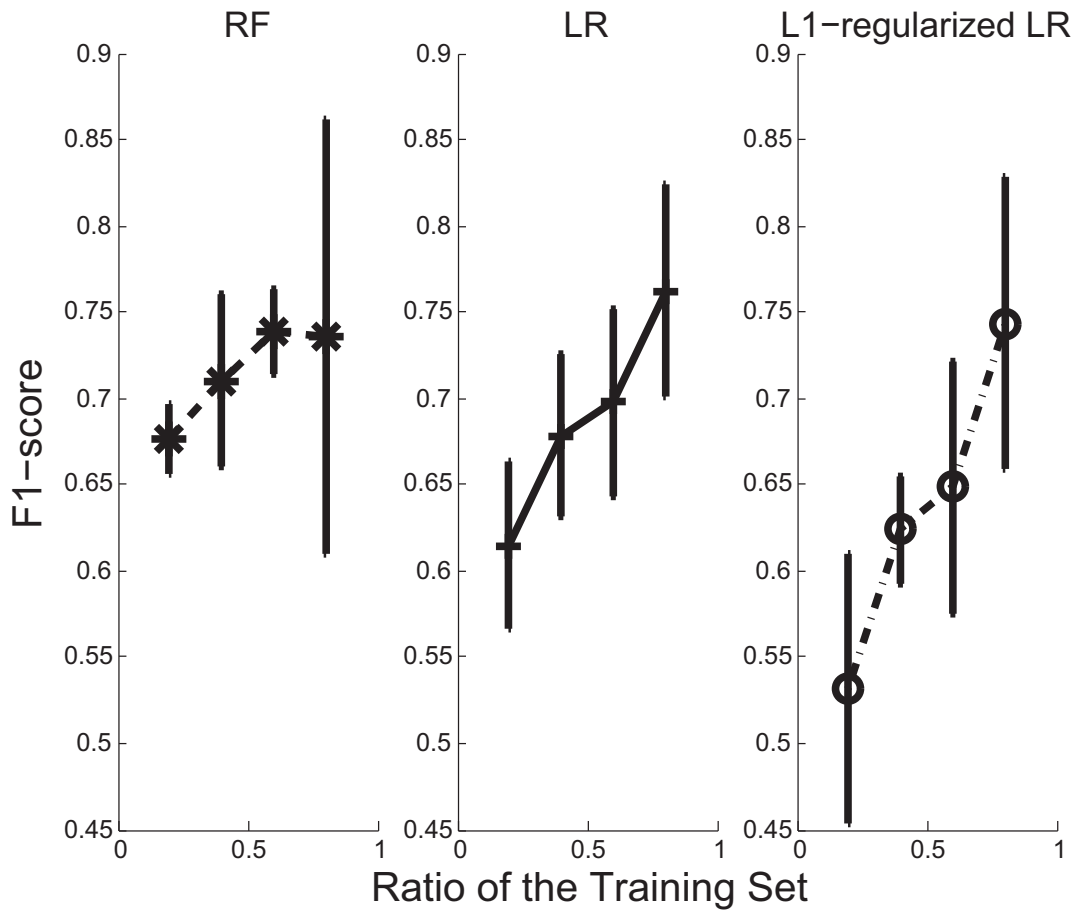

Fig. 3 Comparison of F1 Score with varying training set ratio.

two cultures (Breiman 2001a). The first culture, which is representative of the vast majority of quantitative research in political science, begins by assuming a stochastic data model generated from independent draws from some parametric distribution of data. The values of the parameters theorized to generate the distribution are estimated from the data, and the model is then used to test hypotheses and gather information about the potential effects of the independent variables on the dependent variable. The algorithmic modeling culture, as Leo Beriman describes it, treats the inside of the black box as complex and unknown. Some function $f(x)$—represented by an algorithm–is estimated on the independent variables to predict the response of the dependent variable. Whereas model validation in the first culture is examined by goodness of fit tests, model validation in the second culture is measured by predictive accuracy.

Due to the lack of assumptions about the process that generated the data, it is difficult for researchers using machine learning algorithms to say much about causality. In Rubin's framework, a model has explanatory power to the extent that it makes correct predictions about observations it has not yet seen. If, in the presence of some treatment, a unit of observation undergoes some transformation, but remains the same in the absence of that treatment, the treatment itself can be determined to have a causal effect. Since we cannot observe a unit of observation that is at the same time both treated and untreated, we must look outside our data for untreated units that are similar in all respects to our treated units. Since Random Forests does cross-validation internally through the use of OOB observations, we can use these OOB observations to determine some aspects of causality.

Figure 4 shows the mean decrease in a measure of predictive accuracy, called the Gini Score, for the top twenty variables that Random Forests picked. The Gini Score is calculated as follows. Each

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

	Models and predicted probability of civil war onset			
Civil war onset	*Fearon and Laitin (2003)*	*Collier and Hoeffler (2004)*	*Hegre and Sambanis (2006)*	*Random Forests*
Afghanistan 2001	0.01	0.01	0.01	0.09
Angola 2001	0.04	0.01	0.01	0.13
Burundi 2001	0.00	0.00	0.00	0.05
Guinea 2001	0.00	0.00	0.01	0.22
Rwanda 2001	0.02	0.00	0.00	0.56
Uganda 2002	0.03	0.05	0.00	0.81
Liberia 2003	0.01	0.03	0.00	0.94
Iraq 2004	0.04	0.01	0.00	0.68
Uganda 2004	0.02	0.01	0.02	0.52
Afghanistan 2005	0.01	0.02	0.01	0.14
Chad 2006	0.01	0.07	0.02	0.21
Somalia 2007	0.00	0.00	0.00	0.52
Rwanda 2009	0.00	0.01	0.00	0.74
Libya 2011	0.00	0.01	0.00	0.34
Syria 2012	0.00	0.04	0.00	0.25
DR Congo 2013	0.00	0.00	0.00	0.76
Iraq 2013	0.01	0.00	0.00	0.25
Nigeria 2013	0.01	0.00	0.00	0.25
Somalia 2014	0.01	0.04	0.01	0.87

time a given variable is used to split a node into two daughter nodes, the Gini Scores for the daughter nodes are calculated and compared to the original node. The Gini Score ranges from 0 (a completely homogeneous node) to 1 (a completely heterogeneous node). The changes in a node's Gini Score are summed for each variable at each node, and then normalized at the end of the forest growing procedure. Variables that result in nodes with higher purity contribute to a higher decrease in the Gini Score at the end of the calculation. The Gini Score is calculated internally through OOB observations. Each time observation i is OOB, its Gini Score is calculated according to the above procedure. The mean decrease in the Gini Score is the predictive accuracy lost by omitting a given predictor from the tree used to generate predictions about the class of i, where $i \in [0, 1]$. Hence, variables with a greater mean decrease in the Gini Score more accurately predict the true class of observation i when i is OOB.

The results of Fig. 4 reinforce some of the already well-established results common to the last decade of quantitative research into the causes of civil war onset, but cast doubt on some others. The best predictor of civil war onset is national poverty as measured by both the growth rate of national gross domestic product (GDP) and GDP per capita. This reinforces the results of nearly every examination into the causes of civil war onset published in the past decade. Population size and mountainous terrain are also strong predictors of civil war onset, though not nearly as much as economic variables (Fearon and Laitin 2003). Perhaps what is most surprising about Fig. 4 is the variables it returns as having weak predictive power. Variables traditionally thought to strongly influence civil war onset, including anocracy, democracy, political instability, primary commodity exports, and population density, show little predictive power.

To get a sense of the effect each predictor variable has on the response variable, Fig. 5 shows partial dependence plots for nine of the twenty variables contributing most to predictive accuracy. Partial dependence plots give a graphical representation of the marginal effect of a variable on the class probability. The function being plotted is defined mathematically as

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f(x, x_i c), \tag{11}$$

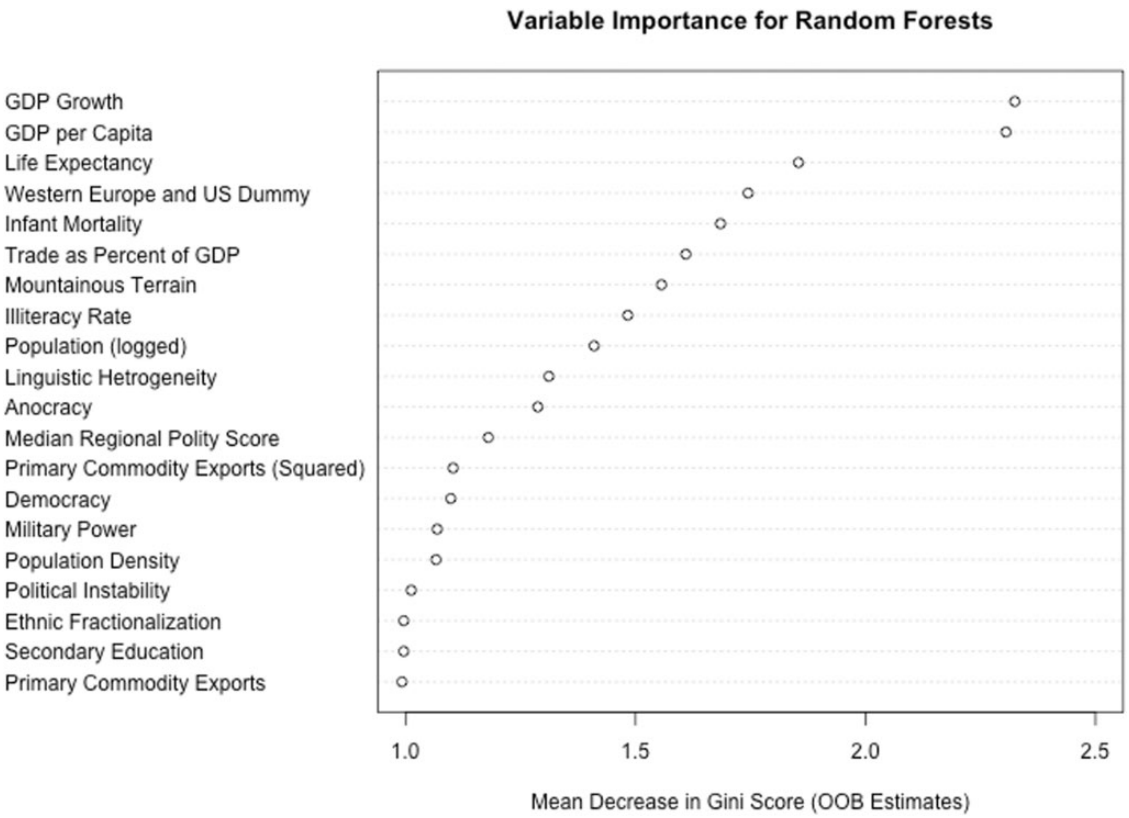

Fig. 4 Plot of variable importance by mean decrease in Gini Score.

where x is the variable for which partial dependence is sought, and $x_i c$ represents the other variables in the data. The summand is the predicted logit (or log of the fraction of total votes) for the classification of y, which is defined according to the following formula:

$$f(x) = \log p_k(x) - \frac{1}{K}\sum_{j=1}^{K} \log p_j(x), \tag{12}$$

where K is the number of classes for y, k is the predicted class, and p_j is the proportion of votes for class j (Liaw 2015). Partial dependence plots can be thought of as the algorithmic equivalent of marginal effects plots, where every other variable is held at its mean and the variable of interest is allowed to vary over its entire range. The values of the y-axis are given in (11) and indicate the change in log-odds for the fraction of votes among all trees for the majority class for a given observation. In other words, a negative slope indicates a greater fraction of votes for peace for that predictor, and a positive slope indicates that the variable predicts a greater fraction of votes for civil war onset.

Figure 5 displays the partial dependence plots. Each variable is labeled on the x-axis. The y-axis shows the change in the fraction of votes for the probability of civil war onset for each variable. The plot shows how the percentage of total votes for civil war onset changes over the range of each predictor. The most outstanding result from a causal perspective is the nonlinearity of nearly all nine predictors. The GDP growth rate is the only predictor that shows the S-curve characteristic of a logistic regression model. The other predictors exhibit significant nonlinear relationships where the probability of civil war onset suddenly increases or declines drastically. These results would not be captured by a linear logistic model.

The relationships between the predictors and civil war onset shown in Fig. 5 are generally in line with previous research. Higher levels of wealth dramatically reduce the probability of civil war onset. It appears that severe economic shocks are not necessary to cause a civil war onset, but that a

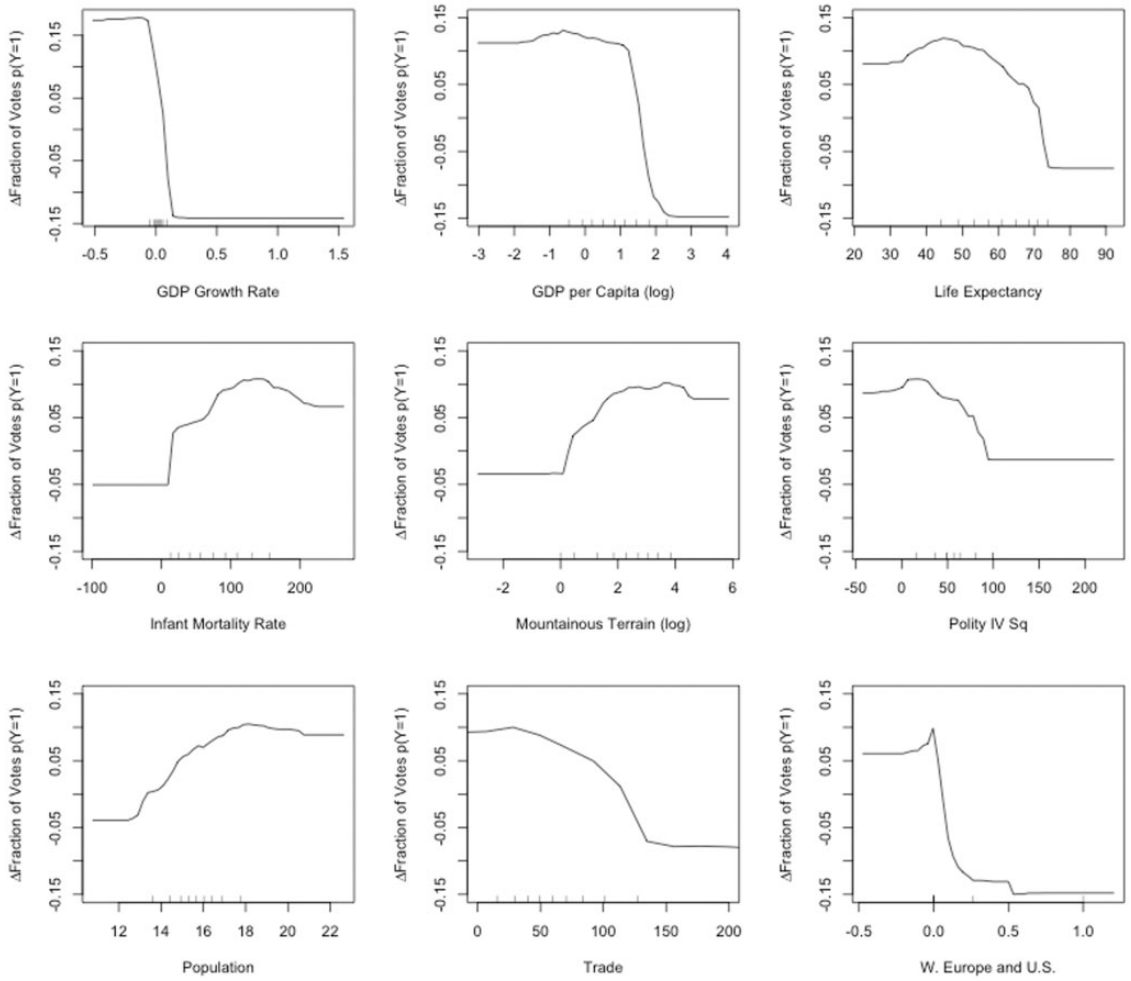

Fig. 5 Partial dependence plots.

decline of only a few tenths of a percent in GDP growth is enough to make such an onset more likely. Mountainous terrain increases the probability of civil war onset—a result in agreement with the literature. Large populations likewise increase the risk of civil war onset. The square of Polity IV—traditionally used to measure anocratic regimes—does not show its characteristic inverted-U shape. Rather autocracies and anocracies appear to be at relatively the same risk for civil war onset, though this risk declines dramatically with democratization.

Although these results should not be treated uncritically, they suggest that statistical learning algorithms such as Random Forests can help us understand the causes of civil war onset. We are therefore cautiously optimistic about the use of statistical learning algorithms to inform observations about the causal processes of rare events like civil war onset.

6 Implications and Conclusions

Prediction is a useful criterion by which to evaluate procedures like Random Forests and logistic regression. Because political scientists often attempt to answer substantively important questions, the field can benefit from the use of statistical methods that allow researchers to make more accurate predictions. The ability of a statistical method to make accurate predictions is just as important as its ability to explain causal processes, and there is enough room for both standards of evaluation within the discipline (Shmueli 2010). When research questions have important policy ramifications, data are known to be imbalanced, or covariates are thought to act in complex and nonlinear ways, statistical learning methods represent a welcome addition to the analyst's toolkit.

By now, most of these techniques are already well established and there is software in the most popular statistical packages.

The analyses presented here show that Random Forests offers superior predictive power compared to several forms of logistic 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. The flexibility offered by nonparametric methods like Random Forests permits more accurate predictions of these important and devastating events. If political scientists are genuinely interested in developing predictive methods, these methods are a useful place to start.

We believe Random Forests can be usefully applied to a range of research problems throughout political science. New developments in Random Forests can be applied to panel data and hierarchical data—two very common data structures studied by political scientists (Freiman 2010; Sela and Simonoff 2012; Hajjem et al. 2014). We leave it to future research to compare the predictive accuracy of Random Forests with and without explicitly modeling the panel data structure common to analyses of civil war onset. Tree-based methods are extremely flexible. The promise of being able to model clustered hierarchical data and time-series cross-sectional data are two areas where Random Forests can be extended in the political science literature. Since these types of data structures are so common to political science, we believe this is a field where political methodologists can provide important and original contributions.

The analyses presented here are specific to the problem of civil war onset and the canonical data sets used to study it. Logistic regression may work quite well as a classifier if the relationship between input and output variables is linear and the data are relatively balanced between classes. If the relationship between the response and predictor variables is truly linear, Random Forests will only approximate linear regression methods like OLS and logistic regression in the limit case of an infinite number of trees. Random Forests exchanges a high degree of variance between each tree for a low bias in predicting the outcome variable. If the assumptions of other methods, including linearity in the parameters, collinearity, and homoskedasticity, are not violated, other methods may give more unbiased estimates. However, when dealing with historical data, as is often the case in comparative politics and international relations, many foundational regression assumptions are routinely violated. If the goal is to develop a predictively accurate forecast of a rare and important event like civil war onset, the flexibility of algorithmic methods like Random Forests can outperform standard methods like OLS and logistic regression. It is difficult to know *a priori* if Random Forests or any other statistical learning method would be a better choice when analyzing other data sets, but they at least deserve more serious consideration than they have received to date in political science. We hope that this article helps move the discipline in that direction and thereby enhances its predictive capacity. Dart-throwing chimps will then have some more serious competition.

Conflict of interest statement. None declared.

References

Beck, N., G. King, and L. Zeng. 2000. Improving quantitative studies of international conflict: A conjecture. *American Political Science Review* 94(1):21–35.

Blair, R., C. Blattman, and A. Hartman. 2015. Predicting local violence. Social Science Research Network. revised url http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2497153 (accessed October 10, 2015).

Brandt, P., J. R. Freeman, and P. Schrodt. 2014. Evaluating forecasts of political conflict dynamics. *International Journal of Forecasting* 30:944–62.

Breiman, L. 1996. Out-of-bag estimation. Technical report, Citeseer.

———. 2001a. Random forests. *Machine Learning* 45(1):5–32.

———. 2001b. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science* 16(3):199–231.

Buuren, S., and K. Groothuis-Oudshoorn. 2011. MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software* 45(3):1–67.

Cederman, L.-E., K. S. Gleditsch, and H. Buhaug. 2013. *Inequality, grievances, and civil war*. Cambridge University Press.

Chawla, N. V. 2005. *Data mining for imbalanced datasets: An overview*, 875–86. Springer.

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 (JAIR)* 16:321–57.

Chawla, N. V., A. Lazarevic, L. O. Hall, and K. W. Bowyer. 2003. Smoteboost: Improving prediction of the minority class in boosting. In *Knowledge discovery in databases: PKDD 2003, 7th European conference on principles and practice of knowledge discovery in databases, Cavtat-Dubrovnik, Croatia, September 22–26, 2003, Proceedings, volume 2838 of lecture notes in computer science,* eds. N. Lavrac, D. Gamberger, H. Blockeel, and L. Todorovski, 107–19. Springer.

Chen, C., A. Liaw, and L. Breiman. 2004. *Using random forest to learn imbalanced data.* Berkeley: University of California.

Cieslak, D. A., and N. V. Chawla. 2008. Start globally, optimize locally, predict globally: Improving performance on imbalanced data. In *Proceedings of the 8th IEEE international conference on data mining (ICDM 2008), December 15–19, 2008, Pisa, Italy, 143–52.*

Clayton, G., and K. S. Gleditsch. 2014. Will we see helping hands? Predicting civil war mediation and likely success. *Conflict Management and Peace Science* 31:265–84.

Collier, P., and A. Hoeffler. 2004. Greed and grievance in civil war. *Oxford Economic Papers* 56(4):563–95.

Duncan, G. M. 2014. Causal random forests. http://econ.washington.edu/sites/econ/files/old-site-uploads/2014/08/Causal-Random-Forests_Duncan.pdf (accessed October 10, 2015).

Efron, B. 1983. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association* 78(382):316–31.

Fawcett, T. 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27(8):861–74.

Fearon, J. D., and D. D. Laitin. 2003. Ethnicity, insurgency, and civil war. *American Political Science Review* 97(01):75–90.

Firth, D. 1993. Bias reduction of maximum likelihood estimates. *Biometrika* 80(1):27–38.

Freiman, M. H. 2010. Using random forests and simulated annealing to predict probabilities of election to the Baseball Hall of Fame. *Journal of Quantitative Analysis in Sports* 6(2):1–35.

Geisser, S. 1975. The predictive sample reuse method with applications. *Journal of the American Statistical Association* 70(350):320–8.

Gelman, A., and G. Imbens. 2013. Why ask why? Forward causal inference and reverse causal questions. NBER working paper number 19614.

Gleditsch, K. S., and M. Ward. 2012. Forecasting is difficult, especially about the future: Using contentious issues to forecast interstate disputes. *Journal of Peace Research* 50(1):17–31.

Goldstone, J. A., R. H. Bates, D. L. Epstein, T. R. Gurr, M. B. Lustik, M. G. Marshall, J. Ulfelder, and M. Woodward. 2010. A global model for forecasting political instability. *American Journal of Political Science* 54(1):190–208.

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):991–1002.

Hajjem, A., F. Bellavance, and D. Larocque. 2014. Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation* 84(6):1313–28.

Hastie, T., R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani. 2009. *The elements of statistical learning*. Springer.

Hegre, H., J. Karlsen, H. M. Nygård, H. Strand, and H. Urdal. 2013. Predicting armed conflict, 2010–2050. *International Studies Quarterly* 57(2):250–70.

Hegre, H., and N. Sambanis. 2006. Sensitivity analysis of empirical results on civil war onset. *Journal of Conflict Resolution* 50(4):508–35.

Hill, D. W., and Z. M. Jones. 2014. An empirical evaluation of explanations for state repression. *American Political Science Review* 108:661–87.

Hoff, P. D., and M. D. Ward, 2004. Modeling dependencies in international relations networks. *Political Analysis* 12(2):160–75.

Holland, P. W. 1986. Statistical and causal inference. *Journal of the American Statistical Association* 81(396):945–60.

Honaker, J., G. King, and M. Blackwell. 2011. Amelia ii: A program for missing data. *Journal of Statistical Software* 45(7):1–47.

Jones, Z., and F. Linder. 2015. Exploratory data analysis using random forests. Prepared for the 73rd annual MPSA conference, April 16–19, 2015. http://zmjones.com/static/papers/rfss_manuscript.pdf (accessed October 10, 2015).

Kalyvas, S. N. 2007. Civil wars In *The Oxford handbook of comparative politics*, eds. C. Boix and S. Stokes, 416–34. Oxford University Press.

King, G., R. O. Keohane, and S. Verba. 1994. *Designing social inquiry: Scientific inference in qualitative research.* Princeton University Press.

King, G., and L. Zeng. 2001. Logistic regression in rare events data. *Political Analysis* 9(2):137–63.

Köknar-Tezel, S., and L. J. Latecki. 2011. Improving SVM classification on imbalanced time series data sets with ghost points. *Knowledge and Information System* 28(1):1–23.

Lee, S., H. Lee, P. Abbeel, and A. Y. Ng. 2006. Efficient L1 regularized logistic regression. In *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16–20, 2006, Boston, Massachusetts, USA, 401–8.*

Liaw, A. 2015. Package "randomforest". https://cran.r-project.org/web/packages/randomForest/randomForest.pdf (accessed October 10, 2015).

Ling, C. X., and C. Li. 1998. Data mining for direct marketing: Problems and solutions. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98), New York City, New York, USA, August 27–31, 1998, 73–9.*

Montgomery, J. M., F. M. Hollenbach, and M. D. Ward. 2012. Improving predictions using ensemble Bayesian model averaging. *Political Analysis* 20(3):271–91.

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,V1 [UNF:6:pwv9cSHI53tZqXlrJ9EDaw== (accessed October 10, 2015).

Park, M. Y. and T. Hastie. 2007. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69:659–77.

Ravikumar, P., M. J. Wainwright, and J. D. Lafferty. 2010. High-dimensional Ising model selection using l1 regularized logistic regression. *Annals of Statistics* 38:1287–319.

Schrodt, P., J. Yonamine, and B. E. Bagozzi. 2013. Data-based computational approaches to forecasting political violence. In *Handbook of computational approaches to counterterrorism*, ed. V. Subrahmanian, 129–62.

Sela, R. J., and J. S. Simonoff. 2012. Re-em trees: A data mining approach for longitudinal and clustered data. *Machine Learning* 86:169–207.

Shellman, S. M., B. P. Levy, and J. K. Young. 2013. Shifting sands: Explaining and predicting phase shifts by dissident organizations. *Journal of Peace Research* 50:319–36.

Shmueli, G. 2010. To explain or predict? *Statistical Science* 25(3):289–310.

Siroky, D. 2009. Navigating random forests and related advanced in algorithmic modeling. *Statistics Surveys* 3:147–63.

Spirling, A. 2008. Rebels with a cause? Legislative activity and the personal vote in Britain. Working Paper.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. 2008. Conditional variable importance for random forests. *BMC Bioinformatics* 9(1):307.

Sun, Y., M. S. Kamel, and Y. Wang. 2006. Boosting for learning multiple classes with imbalanced class distribution. In *Proceedings of the 6th IEEE international conference on data mining (ICDM 2006), 18–22 December 2006, Hong Kong, China, 592–602. IEEE Computer Society*.

Ward, M., R. Siverson, and X. Cao. 2007. Disputes, democracies, and dependencies: A reexamination of the Kantian peace. *American Journal of Political Science* 51(3):583–601.

Ward, M. D., B. D. Greenhill, and K. M. Bakke. 2010. The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research* 47(4):363–75.

Ward, M. D., and P. D. Hoff. 2007. Persistent patterns of international commerce. *Journal of Peace Research* 44(2):157–75.

Ward, M. D., N. W. Metternich, C. Dorff, M. Gallop, F. M. Hollenbach, A. Schultz, and S. Weschle. 2012. Learning from the past and stepping into the future: The next generation of crisis predition. *International Studies Review* 15(4): 473–90.

Weidmann, N. B. 2008. Conflict prediction via machine learning: Addressing the rare events problem with bagging. *Poster presented at the 25th annual summer conference of the society for political methodology*.

Zorn, C. 2005. A solution to separation in binary response models. *Political Analysis* 13(2):157–70.