Journal of Peace Research

Do the robot: Lessons from machine learning to improve conflict forecasting

Journal of Peace Research
2017, Vol. 54(2) 193–214

DOI: 10.1177/0022343316682065
journals.sagepub.com/home/jpr
SAGE

Michael Colaresi & Zuhaib Mahmood

Department of Political Science, Michigan State University

Abstract

Increasingly, scholars interested in understanding conflict processes have turned to evaluating out-of-sample forecasts to judge and compare the usefulness of their models. Research in this vein has made significant progress in identifying and avoiding the problem of overfitting sample data. Yet there has been less research providing strategies and tools to practically improve the out-of-sample performance of existing models and connect forecasting improvement to the goal of theory development in conflict studies. In this article, we fill this void by building on lessons from machine learning research. We highlight a set of iterative tasks, which David Blei terms 'Box's loop', that can be summarized as build, compute, critique, and think. While the initial steps of Box's loop will be familiar to researchers, the underutilized process of model criticism allows researchers to iteratively learn more useful representations of the data generation process from the discrepancies between the trained model and held-out data. To benefit from iterative model criticism, we advise researchers not only to split their available data into separate training and test sets, but also sample from their training data to allow for iterative model development, as is common in machine learning applications. Since practical tools for model criticism in particular are underdeveloped, we also provide software for new visualizations that build upon already existing tools. We use models of civil war onset to provide an illustration of how our machine learning-inspired research design can simultaneously improve out-of-sample forecasting performance and identify useful theoretical contributions. We believe these research strategies can complement existing designs to accelerate innovations across conflict processes.

Keywords

civil war, forecasting, machine learning, methodology, visualization

International relations and the social sciences in general have been increasingly focused on building models that provide useful out-of-sample predictions (Witmer et al., 2017; Ward & Beger, 2017). While forecasting is not new, the recent momentum comes at a time when the weaknesses of traditional applications of in-sample null hypothesis significance testing (NHST) are increasingly apparent across disciplines (Gelman & Loken, 2014; Gill, 1999; Simmons, Nelson & Simonssohn, 2013). The weight of failed replications and exaggerated effect sizes, in addition to well-known worries of overfitting the sample data, have led many researchers to search for new research design strategies that might accelerate breakthroughs in social research. Indeed there are several exciting examples in Geography, Climatology, Natural Language Processing, and other disciplines, where researchers have been able to harness the availability of dense sources of digital information without relying on NHST (Chen & Manning, 2014; Blei, 2014; Raftery et al., 2005; McCormick et al., 2012).

In an important article, Ward, Greenhill & Bakke (2010: 365) argue that the 'search for statistical significance' can be a misleading metric both for how well a model represents the underlying patterns in the data, and how the model will generalize to unseen data. These are two distinct research pitfalls, which we term underperformance and overfitting. A model, by design, is a stylized representation of the underlying process of interest.

Corresponding author:
colaresi@msu.edu

An extremely simple model, such as a linear-additive representation of conflict that only includes a few variables, might capture a small number of patterns in the underlying process. While these patterns might generalize out-of-sample, meaning they are representing observable features of the data generation process, they might also exclude many other signals. We refer to this situation as underperformance on unseen data. Conversely, a model might represent a myriad of patterns in the data used to fit the model, but these patterns may not generalize to new data. This is conventionally known as overfitting the sample data. Using two highly cited models of civil war, Ward, Greenhill & Bakke (2010) provide examples of both underperformance and overfitting. Even when models have many statistically significant coefficients, the resulting predictions on new data will underperform, faring no better than very simple models that take into account only one or two features of the process.[1] These models have learned only a small number of patterns, but no more. Additionally, this research team provides examples where adding statistically significant variables – sometimes with coefficients many times the size of their standard errors – actually 'degrade[s] the predictive accuracy of a model' (Ward, Greenhill & Bakke, 2010: 373). In these cases, the models have overfit by learning patterns in the sample of data used to compute estimates that do not generalize out-of-sample. This research leaves us with the question: if gazing at stars is a poor guide to the future, what research design strategy can take its place and reduce underperformance and overfitting?

Machine learning-inspired research design as an alternative to NHST

In this article, we detail an alternative workflow to NHST that builds on established approaches in machine learning. We also provide an application of this workflow to civil war forecasting, emphasizing the role of model criticism and predictive performance in the model-building process. In contrast to NHST, machine learning-inspired research designs have at their core a set of distinct iterative steps that an applied researcher cycles through to learn generalizable patterns from the available data. Our proposed steps, inspired by David Blei's summary of crucial insights from George Box's loop (Blei, 2014; Box, 1980), include *building* a mathematical representation using domain knowledge; *computing* the unknown parameters and weights from the mathematical representation with training data; *critiquing* the fitted model by identifying theoretically relevant discrepancies between the model and new data; and then using the new knowledge of these discrepancies to *think* of what patterns may have generated them.[2] These research subtasks are then repeated until a researcher is satisfied with the model performance on a specified task.

In this cyclical setup, out-of-sample predictions must be ruthlessly critiqued across multiple scales so that new features and specifications can be innovated to improve the performance in the next build of the model. Model criticism is uniquely beneficial because it identifies discrepancies in the current model. These discrepancies, once identified, can help researchers construct more useful models of conflict processes. Moreover, in thinking about these discrepancies, researchers are by definition updating their domain knowledge, generating new ideas about the relevant data generation mechanisms.

Machine learning for humans

Machine learning is a relatively recent field that has its roots in artificial intelligence research. Successful applications of machine learning over the last several decades in tasks as varied as spam filtering to playing jeopardy have validated the usefulness of this approach (Siegel, 2013). The goal of machine learning is deceptively simple. According to Mitchell (1998), a program is said to learn from experiences E with respect to some tasks T and related performance measure P if its performance on T measured by P increases with E. The canonical example is writing a computer program that can learn to play a game such as checkers. A set of practice games, with known outcomes, is provided as the experiences E. The researcher clearly defines a task T, such as winning future games against humans, and measures the performance on that task with a function P, such as the proportion of games won (Mitchell, 1998).

Since the goal of machine learning is increasing performance as new experiences become available, avoiding both underperformance and overfitting are central concerns (Flach, 2012). Learning a model that overfits the

[1] For example, Ward, Greenhill & Bakke (2010) find that a model only measuring GDP and population performs nearly as well as a model that accounts for 11 features. The models explored are linear and additive on the log-odds scale in this example.

[2] Blei (2014) includes the first three of our subtasks, but since our emphasis is on the contrast between NHST and machine learning-inspired research designs, we highlight the distinction between critiquing a model to identify discrepancies and subsequently thinking about how the discrepancies relate to domain knowledge.

data, such as implementing a rule in the checker-playing model that always moves a piece to the middle of the board, will lead to poor performance on the task in the future since the sides are systematically safer than the center. Similarly, underperformance can occur if an overly simplistic model only learns moves to capture one enemy piece at a time, but not patterns that lead to double jumps and the accumulation of kings.

Machine learning has developed research design strategies that aid in identifying useful features that represent experiences, so that flexible models can learn patterns to accomplish the defined task. This is a crucial difference between machine learning and NHST, as we highlight in more detail below. Instead of concern over bias relative to an unseen parameter or model, as in research based on NHST, machine learning expends its energy improving its performance on a clearly delineated task and related performance metric. NHST conditions what we learn from the data on the assumption that the model is, in some sense, correctly specified a priori. Machine learning approaches, on the other hand, condition what we learn on the ability of a model representation to increase the observable performance metric.

Conflict prediction as supervised learning

While the example of teaching a computer to play checkers may seem far afield for international relations researchers, the definition, goals, and components of machine learning have direct analogues in conflict prediction. In conflict forecasting, the task, T, is to compute a useful forecast of unseen conflicts, within a given territory at a given time, using the features that are available before the period being predicted. This forecast could be a dichotomous prediction of a conflict, a probability of an event, or lie on the real line like a log-odds ratio. E is the set of available data that can be used to develop a mathematical representation of the process that generated the conflicts. These experiences are represented in model form – such as a logit model – by the features or variables that measure the relevant dimensions of instances (for example, economic development, democracy, and previous conflict history). Finally, conflict forecasters use performance measures, P, such as accuracy, precision, recall, or scoring rules, to measure the relative usefulness of their systems at the task.

In fact, conflict forecasting can be seen as a conventional example of a specific type of machine learning problem known as supervised learning (Hastie, Tibshirani & Friedman, 2013; Kotsiantis, 2007). These are a set of machine learning problems where labeled data are available on which to train models. The labels in conflict forecasting correspond to the outcomes, conventionally referred to as the dependent variable in social science research, in the observed data.[3] Therefore, the lessons learned from the application of machine learning to domains across the social and natural sciences over the last decades have significant relevance to conflict research. In fact, machine learning provides a set of research subtasks that can aid conflict researchers not only in building useful forecasting models that generalize outside of a given sample of data, but also in developing and informing theories of conflict processes. The latter benefit of machine learning emerges from its unique iterative cycle of exploration and improvement.

Learning through Box's loop

The definition of machine learning as a trajectory of improved performance on a given task leads practitioners to explicitly and consciously cycle through a set of iterative research subtasks. This workflow contrasts with NHST, which follows an acyclic set of inferential tasks. Moreover, in NHST research design, repeatedly iterating between fitting and building the model undercuts the ability of a researcher to generalize from the available data, leading to cases of both underperformance and overfitting.[4] The two workflows are contrasted in Figure 1.

Build and compute

The first research subtasks of building and computing a model will be familiar to conflict researchers. Once the overall task is defined, a researcher must use their knowledge to build (subtask 1) a mathematical representation of the process under investigation. This includes the familiar task of specifying a model with relevant functions that link the input features to the target, as well as any priors. For example, in the task of forecasting civil war in a country in a given year, the model might map features such as democracy, gross domestic product, and population into a probability of civil war through an inverse logistic function.

[3] Machine learning also deals with unsupervised learning problems, where no labeled data are available, and several other classes of tasks; see Mitchell (1998).

[4] Problems of p-hacking, the garden of forking-paths, and multiple comparisons related to research that uses NHST have been discussed extensively (Levine et al., 2008; Gelman & Loken, 2014; Schrodt, 2014).

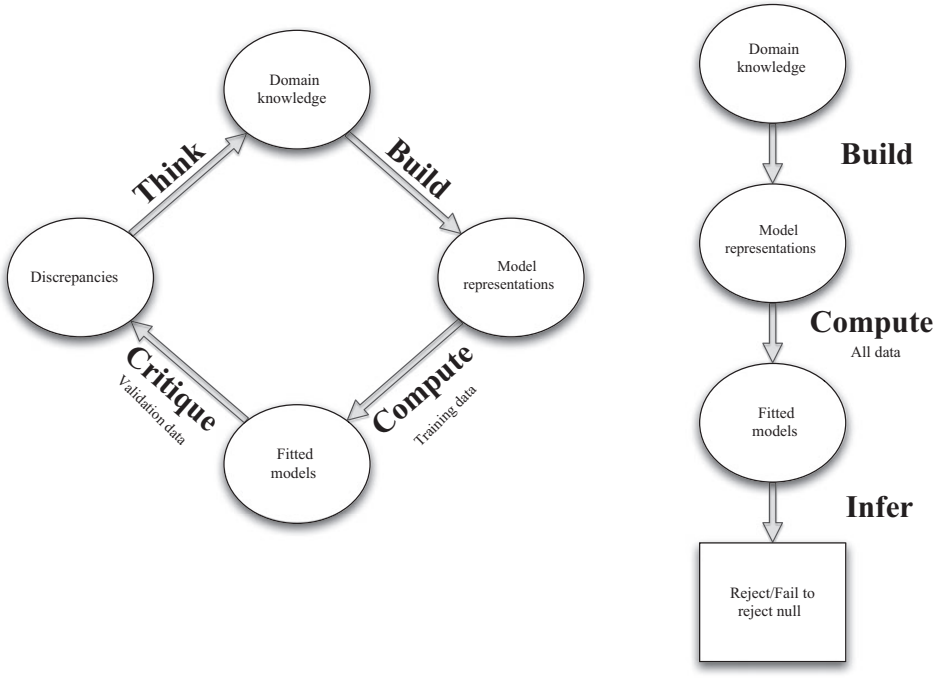

Figure 1. Box's loop (left) vs. null hypothesis significance testing (right) workflows

After a model representation is built, its unknown parameters need to be computed (subtask 2) from the data, just as in conventional conflict processes research. This is known as training the model. However, as we will discuss further below, the computation of the fitted model in machine learning is not typically done on all the available experiences. Some data are held back to evaluate how the fitted model performs on unseen data and allow for future refinements in the representation. Further, the computation of the model is not the final step of the analysis, but merely a means to the end of evaluating the fitted models performance on unseen data.

Critique

With the fitted model computed, a researcher can now analyze the correspondence between the observed data and the model. The goal of this model criticism step (subtask 3) is to find systematic discrepancies between the fitted representation and the data generation process. While it is possible to analyze the fit on the training data, this is usually a suboptimal strategy for improving a model. The trained model has attempted to match the patterns in the data on which it was computed. Some of these patterns are likely to be noise, resulting in the model overfitting the sample data. Thus, in-sample performance measures will often be a poor guide to the ability of the fitted model to generalize out-of-sample.[5] Since overfitting is not easily observed in-sample, separating the data used to fit the model from the data used to evaluate and criticize its performance can be extremely useful. We return to this point below. For example, in applications of civil war forecasting, highly inaccurate individual predictions are simply those cases where the forecast of an event (e.g. unlikely) was distant from the observed value (e.g. war occurred).

The emphasis in model criticism is not on testing and identifying one true model, but rather on observing discrepancies between the current model and the data, which may lead to new features and modeling strategies that can improve performance in the next iteration of the loop. Identifying discrepancies through the explicit step of model criticism is a pivotal subtask in machine learning that differentiates it from NHST research designs. In machine learning, the goal is to build a model that is as useful as possible at accomplishing the task it was assigned. If you do not already know a useful model, criticism is what makes the construction of new and refined representations of the data generation process

[5] From a Bayesian perspective, model criticism on the fitted-data is known as posterior predictive checks.

possible. We provide an empirical application of this process in later sections.

Think

If a researcher is not satisfied with the performance of a model, or set of models, then the next step is to think (subtask 4) about what patterns in the data generation process might have led to these observed discrepancies. Box (1980: 383) reminds us that 'scientific knowledge advances by a practice–theory iteration' that 'employs two inferential devices: Criticism and Estimation'. Thinking about how discrepancies can be explained within existing domain knowledge provides the connection between criticism and re-estimation in Box's formulation.

The researcher can creatively use their knowledge of the process of interest to re-engineer features or the model representation to improve the performance in the next loop. It may be the case that countries in a particular region are forecast particularly poorly by the model, and the researcher knows that countries in this region share a trait such as high rates of HIV. If available, these health related features can be added to the experiences used in the training data. More fundamentally, a researcher might notice that forecasts deteriorate over time, potentially suggesting a dynamic data generation process (Brandt, Schrodt & Freeman, 2014; Greene, Park & Colaresi, 2016).

One of the exciting developments of machine learning-inspired research designs is their explicit invocation for researchers to learn from discrepancies between the data and the current model to build more useful knowledge and theories about the process under investigation. Researchers often consider out-of-sample forecasting and reduced form modeling in general as atheoretical exercises, since they do not claim to estimate unbiased parameters, or to be a complete and full representation of the underlying data generation process.[6] However, creating a model or set of models that is able to learn patterns related to conflict processes necessarily relies upon theoretical knowledge of the data generation process. Since reality is functionally infinite across multiple scales and perspectives, and model representations are finite, theory is essential to filter what is represented. Many decisions made in the creation of the model representation may be ad hoc or justified by convenience, but not only do these decisions have theoretical implications, thinking about discrepancies can help illuminate worthwhile complexities to add to the representation in the next iteration. Incorporating observed discrepancies between the data and the model into knowledge about concepts of interest emphasizes the goal of learning as opposed to confirming what is already known. In contrast to NHST, where the researcher assumes that a hypothetical model – usually nested within the current fitted model – is the data generation process, our proposed research design strategy uses the current model as a tool to add to domain knowledge and learn more about the unknown process that generated the data.

Repeat

Having learned from the discrepancies in the data, a researcher can either move to estimating the performance of the model in the unseen test set (if they are confident in the model) or rebuild the model and iterate through these steps again.[7] This explicit repetition of the research subtasks is perhaps the crucial difference between machine learning and conventional NHST research designs. According to Blei (2014: 224), machine learning views 'model formulation as part of the iterative process of Box's loop'.

The test set is utilized only as a researcher exits Box's loop. This is usually when a researcher is happy with the evaluated performance of a model representation. Even when cross-validation or an evaluation set was used within Box's loop to choose or build a successful model, test data are needed to gauge its generalization performance. It could be the case that through Box's loop, the training instances were overfit substantially, leading to poor test set performance. This has been termed 'second order overfitting' (Ward et al., 2013: 8). While careful sampling and validation in the training set can guard against this, the test set plays an extremely valuable role in identifying optimistic performance in the training instances.

In NHST, model building and computation proceed in one downstream direction. Once a null hypothesis or set of hypotheses are defined along with a confidence level, and the current model built and estimated, a binary decision is made as to whether the given null hypothesis is rejected or not (Gelman & Loken, 2014; Levine et al., 2008). NHST does not provide advice on what to do after these tests have been performed. This fact shows up

[6] See Schrodt (2014) for a discussion of the relationship between reduced-form forecasting and the process of theory development and evaluation.

[7] Below we outline practical advice on how to avoid overfitting the training data during this iterative process.

in the sampling strategy used to fit the data. Most NHST applications use all of the available data to compute the parameters of interest. This makes the implicit assumption that there is not going to be another pass through the data.

Yet there are many signs that researchers do not stop with one estimated model and the set of predefined null hypotheses. Increasingly, applied researchers are providing numerous robustness checks, which aim to convince the reader that the inferences drawn from a set of analyses are not fragile to changes in the model representation. Researchers estimate a model, and then subsequently add features, change the measurement of existing features, or alter the mapping from the input features to the output. Despite a decade of NHST and robustness checks, Ward, Greenhill & Bakke (2010) still found two sets of civil war models to be poor guides to what patterns existed in unseen data. Applying iterative machine learning-inspired research design to these problems has the potential to increase forecasting and generalization performance. However, in order for that potential to be realized, underperformance and overfitting must still be avoided during Box's loop.

How machine learning-inspired research design avoids underperformance and overfitting

Concerns about mistaking noise for systematic patterns in the sample data are extremely important considerations in machine learning research as well as NHST. It may appear awkward at first that a research design strategy aiming for generalization out-of-sample, will risk reusing the data. However, in stepping through Box's loop over several decades across different domains, machine learning researchers provide practical advice on how to avoid the dual threats of overfitting and underperformance. In fact, there are two crucial strategies that inform how researchers use their data to build, criticize, and rebuild models. First, one should judiciously split and sample from the available data in such a way that overfitting the available data will be observed in degraded performance. Second, with increased confidence that overfitting will be detected, a researcher can use more flexible modeling strategies and domain knowledge to attempt to boost performance in each iteration (Flach, 2012).

How splitting the data avoids overfitting

A fundamental idea in machine learning is to separate the data used to compute the fitted model from the information used to critique the performance of that fitted model (Hastie, Tibshirani & Friedman, 2013). In the simplest case, the observations that are available to compute a model are known as the set of all training instances and a discrete test set of experiences is used to judge the performance of the final trained model.[8] In the model building step, a sample from all the training instances is taken to compute the fitted model and a hold-out set is used for criticism and to guide rebuilding. This latter dataset is often referred to as the validation, evaluation, or calibration set. This validation set is usually the set of training instances that were not used to compute the model, so they can be used to identify discrepancies before the test set is assessed. Over several iterations of Box's loop, a researcher can sample from the training instances several times, and examine discrepancies and performance therein.

Figure 2 illustrates how a simple sampling scheme can be implemented through each step of Box's loop to help researchers increase the out-of-sample performance of their models. First, the test set is separated from the data that will be used for training, usually with a simple random partition (e.g. ⅔ for training and ⅓ for testing). The available training instances are then sampled from, either through another simple random partition, k-fold cross-validation or another sampling scheme. This sampling yields a training and an evaluation set. In k-fold cross-validation, there are actually k evaluation sets within a loop, each comprised of $\frac{1}{k}$ of the training instances. Computation of the model is done on a current training set, and criticism on the current evaluation set. Performance can be calculated on both the training and evaluation set for comparison. As discrepancies are identified and new model representations considered, a new set of training and evaluation data is drawn from the available training instances, and the process repeated. The performance of a model across multiple training and evaluation instances can be averaged together to gain a broader perspective on its strengths and weaknesses.[9]

The choice to split the data is practical. If all available data are used to train the model, overfitting the training set is observationally equivalent to faithfully representing generalizable patterns in the data. Figure 3, adapted from Hastie, Tibshirani & Friedman (2013: 220), summarizes the importance of evaluating a model on unseen data outside of the training set. On the x-axis is model

[8] If data are streaming, meaning new instances will arrive in the future, these future experiences can be considered the test set.

[9] We expand on this discussion in the Online appendix. Different approaches to splitting the training instances are discussed in Flach (2012) and Hastie, Tibshirani & Friedman (2013).

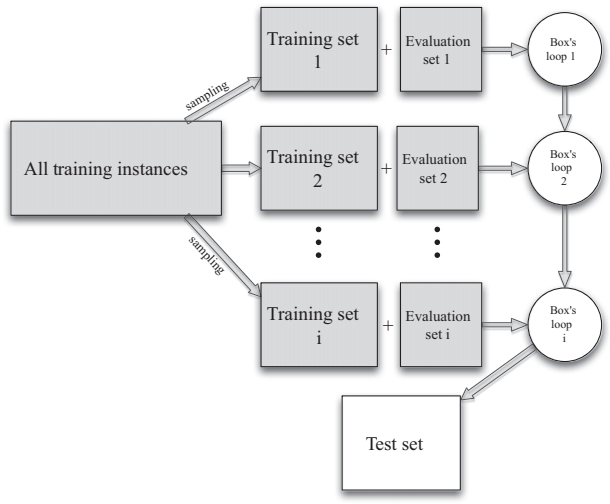

Figure 2. A simple sampling strategy for use with Box's loop

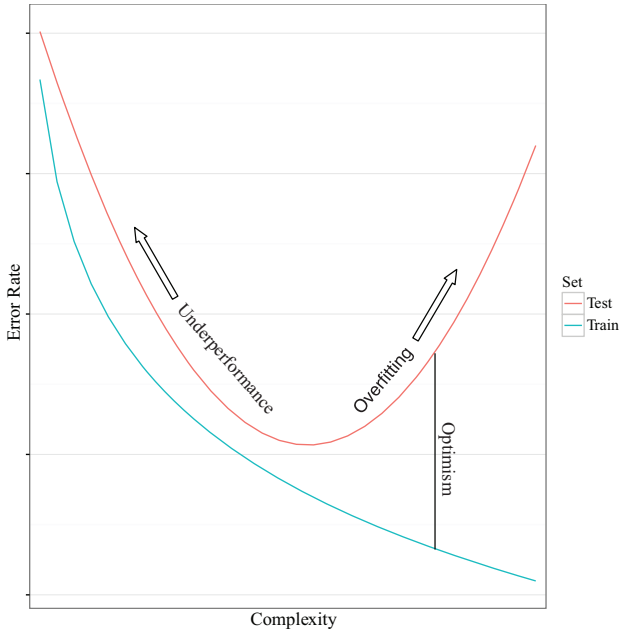

Figure 3. Hypothetical error rates in training and test sets, adapted from Hastie, Tibshirani & Friedman (2013: 220)

complexity, which can be thought of as the number of features included in the model or the non-linearity of the mapping between the features and output. The y-axis represents observed error. This could be operationalized as mean squared error between the forecast and the observed value for regression problems; 1 minus accuracy; or the area under the ROC curve for classification tasks. If we were to fit an ever more complicated model to a fixed dataset, the error as measured in the training set would continue to drop as more and more patterns are fitted. This fact is represented by the monotonically decreasing line. However, in the test set, which was not used to fit the model, increasing complexity will at first decrease the error, as underperformance is avoided, but then increase the error rate, since the trained model mistook noise for patterns, which by definition is highly unlikely to be represented in the test set. Thus, we can observe the generalizability of a computed representation in the relative performance of that model within the test set, information that is not available in training data alone.

How flexible mathematical representations avoid underperformance

Figure 3 also illustrates the tight relationship between overfitting and underperformance. Using separate training and testing data has the important implication that it frees a researcher to explore the performance of models that potentially include more flexible components – such as hierarchical structure or conditional relationships – relative to simpler baselines without being blind to potential overfitting. If the more flexible models are mistaking noise for signal, they will not perform well out-of-sample. On the other hand, if the simpler models are underperforming relative to the new models, the added flexibility of the latter will boost out-of-sample performance. In conflict forecasting, the popularity of simpler models such as logistic specifications hint at worries that more complicated specifications might overfit the sample data. Below, we show that separating training, evaluation, and testing data within Box's loop can free researchers to use flexible models that are still able to generalize out of sample, and still contribute to domain knowledge through model criticism and feature selection.

An application to forecasting civil war

We use the task of forecasting civil war to illustrate Box's loop.[10] Our inspiration for this project was Ward, Greenhill & Bakke (2010), who showed that conventional civil war models built for NHST perform poorly in this task. We begin our analysis with the Fearon & Laitin (2003) model they use as one of their examples[11] as well as a simple two-variable model that includes the

[10] Forecasting civil war is an especially challenging task because the outcomes are imbalanced. For a discussion of class imbalance see Japkowicz & Stephen (2002). Imbalanced data have been considered from a different point of view in King & Zeng (2001).

[11] The features in this model can be found in Ward, Greenhill & Bakke (2010) and Fearon & Laitin (2003).

two strongest features they identify: the log of GDP per capita and the log of population from the full Fearon & Laitin (2003) specification. We then present three iterations of Box's loop to explore whether we can improve on the poor performance that Ward, Greenhill & Bakke (2010) identified.

While Ward, Greenhill & Bakke (2010) is our entry point, we will also utilize a recently published model of civil war from Muchlinski et al. (2016) as a benchmark. This work deploys a random forest ensemble model, computed using 88 features to improve on predictions from simpler logistic models.[12] In this case, the strategy appears to pay off in improved out-of-sample performance. Therefore, we can judge our own performance relative to not only Ward, Greenhill & Bakke (2010) but also Muchlinski et al. (2016).

Defining the task, experiences, and performance measures

Our explicit task is to forecast the binary observation of the presence or absence of a civil war onset within each country around the globe in year $t + 1$, using information available up to year t.[13] The set of experiences we use comes directly from Hegre & Sambanis (2006), as utilized in Muchlinski et al. (2016). We set aside data from 1988 to 2000, as our test set, and utilize all countries in the world from 1945 through 1987, as our set of all training instances. As we describe below, it is from this set of all training instances that we draw our training and evaluation sets within a loop. The models we will build first draw on the features from those included in Hegre & Sambanis (2006), although we will expand these features as we progress. To again connect our work to the previous literature, we use the area under the receiver operator characteristic curve (AUC) as our overall performance metric (Ward, Greenhill & Bakke, 2010; Muchlinski et al., 2016). Since prediction is a multidimensional problem, we also at times plot precision-recall (PR) curves, in addition to the receiver operator characteristic (ROC) curves that produce the AUC measure.[14]

In addition to these overall model performance metrics, we utilize graphical tools for model criticism to help identify discrepancies and guide model improvements. In past work, we have found separation plots (Greenhill, Ward & Sacks, 2011), which are devices to visualize whether a model is usefully sorting the observed data based on the relative forecasts, to be useful for this task. Unlike ROC and PR plots, separation plots allow the individual observations to be plotted. Our two novel plot types, the model criticism plot and the biseparation plot, build upon separation plots with respect to the information they convey, serving as visual guides to identifying discrepant observations and placing them in the full context of the distribution of forecasts.[15]

Table I. A comparison of cross-validated AUC values and AUC values computed on the full training set for all loop 1 models

Model	*Cross-validated AUC*	*Training set AUC*
Two-feature	.67	.68
Full Fearon & Laitin	.60	.68
Muchlinski et al. random forest	.88	.91

Loop 1

We begin with the three pre-existing models that are of different levels of complexity: the simple two-feature (GDP and population) logistic model from Ward, Greenhill & Bakke (2010), the full logistic model from Fearon & Laitin (2003), and the random forest representation from Muchlinski et al. (2016). We use 10-fold cross validation with all the available training instances, that is the data from 1945 to 1987, to compute the parameters of the model and produce a fitted model.[16] For the two logistic regression models, there are no tuning parameters and so the models are fit ten times, each to $\frac{9}{10}$ of the data, and the last $\frac{1}{10}$ is used as the validation set in each case. This process gives us out-of-sample predictions for all the training observations, as each observation is left out of a fold once. For random forests, we need to tune the number of variables that are sampled within each tree. Again, we use 10-fold cross-validation for this.[17]

The performance of these models varied significantly, as we would expect (see Table I). The average AUC

[12] Muchlinski et al. (2016) provide more details on the random forest ensembles and how they can be used for conflict forecasting.

[13] We choose this because it builds directly on the models used in Ward, Greenhill & Bakke (2010) and Muchlinski et al. (2016).

[14] Brier scores and F1 would also be reasonable choices for future exploration (Steyerberg, 2009).

[15] We have created an R package, ModelCriticism, to allow researchers to utilize these tools.

[16] We use the caret package in R for the training of our models and calculation of performance measures (Kuhn & Johnson, 2013).

[17] We utilized downsampling in the random forests so that the class imbalance was reduced to 2:1 (Muchlinski et al., 2016). Changing this ratio between 1:1 and 4:1 had very little influence on the results. We grow 1,000 trees in each computation step.

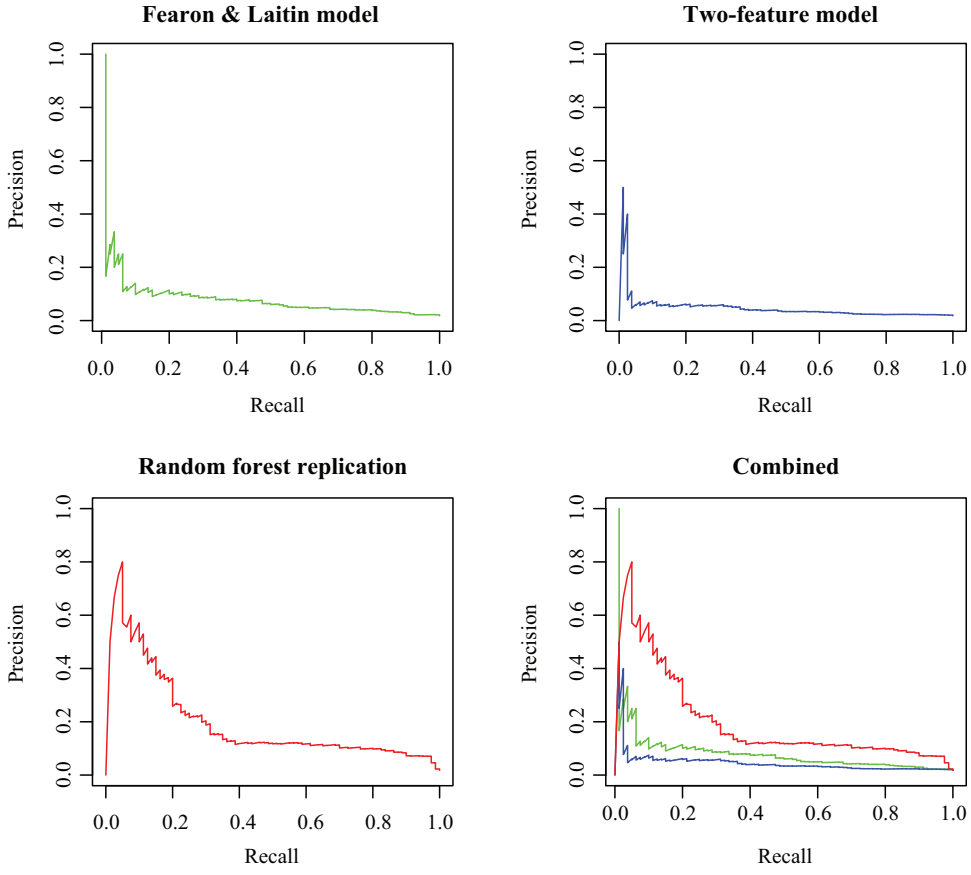

Figure 4. Precision-recall curves for all loop 1 models

across the ten folds for the two feature model was 0.67, and the same value for the full Fearon & Laitin model was 0.60. These differences are consistent with the findings of Ward, Greenhill & Bakke (2010) that the additional features in the full logistic model do not aid in prediction, and can even potentially hurt performance. The flexible random forest model does considerably better with a cross-validated AUC of 0.88.

The precision-recall plots are presented for these models in Figure 4. In PR plots, dominant models are pulled up and to the right. We would like to see a high proportion of cases where, when the model forecast a civil war, one actually occurred (precision). In addition, a higher performing model would predict a civil war in a high proportion of cases where the civil war actually was observed (recall).[18] Here again, the random forest model has the best performance.

[18] PR curves are computed using varying thresholds, like ROC plots (Steyerberg, 2009).

The model criticism plot

One thing to notice about the results so far is that they do not provide obvious markers for how to improve the models, although particularly in the case of the logistic models, they do suggest that improvements are needed. To find the observations that are leading to poorer model performance, we first turn to what we call model criticism plots. We want to easily visualize when a forecast is distant from the observed value. The model criticism plot visualizes discrepancies between a dichotomous target and a forecast by plotting the forecast values on the x-axis for each observation. The observations are then colored by the observed value, in this example, peace (blue) or war (red). In order for all of the observations to be seen, and labeled, the y-axis rank-orders the forecasts from lowest to highest. Highly discrepant positive values appear as red toward the lower left, since the model predicted a low probability of occurrence. Highly discrepant zero values appear as blue towards the top right, since they had high forecasts of war, but remained at peace. As will become apparent, the margins of this simple plot include both the distribution of forecast

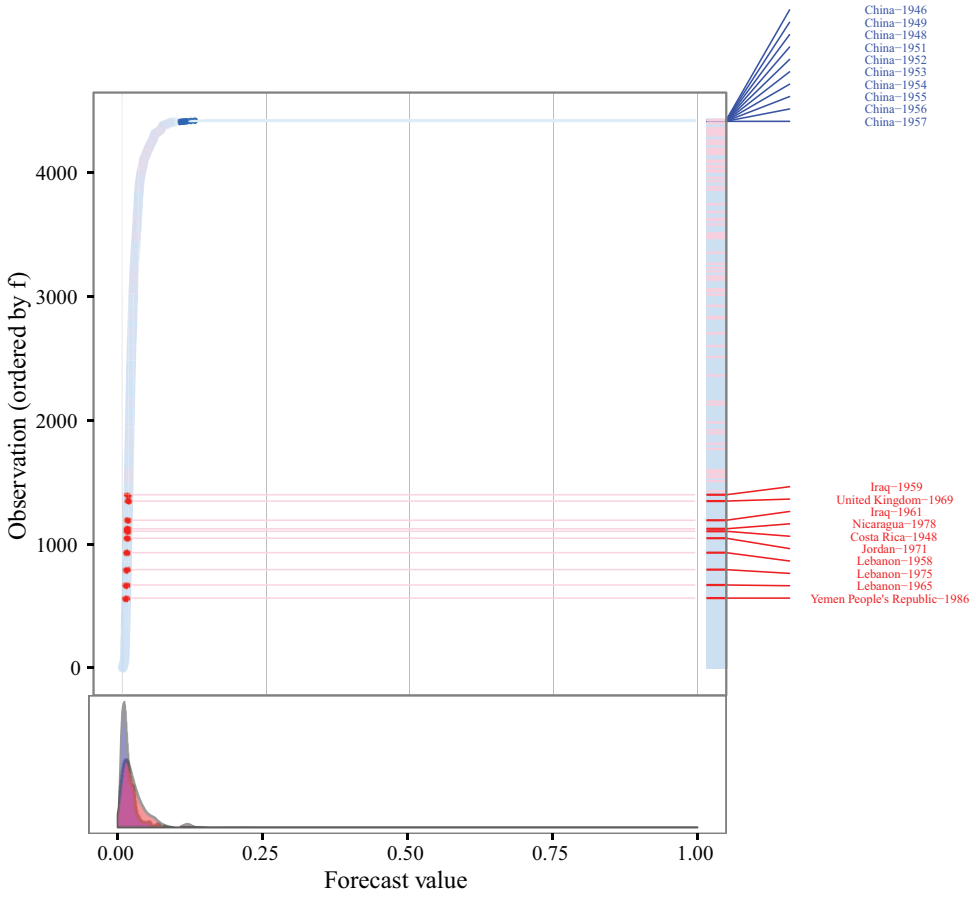

Figure 5. Model criticism plot for the two-feature model (GDP and population)

values on the x-axis – identifying models that made sharply different predictions when war or peace occurs – and a separation plot (Greenhill, Ward & Sacks, 2011) on the y-axis. The most discrepant observations are colored more intensely for emphasis.

Figure 5 presents a model criticism plot for the two-feature model. Like the aggregate measures, it is apparent that overall, this is not a particularly high performing model. There are many civil war (red) observations shuffled towards the bottom left, with low forecasts in absolute and relative terms. Moreover, there are no observations at all in the upper right. Thus, the model makes no sharply distinct forecasts for civil war cases (or any observations). This can also be seen in the overlap of the peace and war distributions below the x-axis. What the model criticism plot uniquely adds is that we can identify the worst discrepancies from the model in the context of the other predictions, using both their ranks and values. Observations are labeled on the vertical separation plot on the right y-axis and they are connected by lines to their associated points. For example, conditioning only on population and GDP makes China seem like a highly probable place for a civil war for the first several decades of its existence. The model also thought, incorrectly, that Lebanon and Iraq were relatively safe places.

The model criticism plots for the full Fearon and Laitin model and the Muchlinski et al. random forest model are presented in Figure 6 and Figure 7, respectively. The random forest model shows quite an improved separation, and greater range of predictions, mirroring the overall performance measure. This model still has a problem with predicting China, with that country showing up in the top ten most discrepant observations in both the civil war and peace cases.

The Fearon and Laitin model criticism plot is illustrative, in that the most discrepant civil war (red) values appear to be places where groups are excluded from

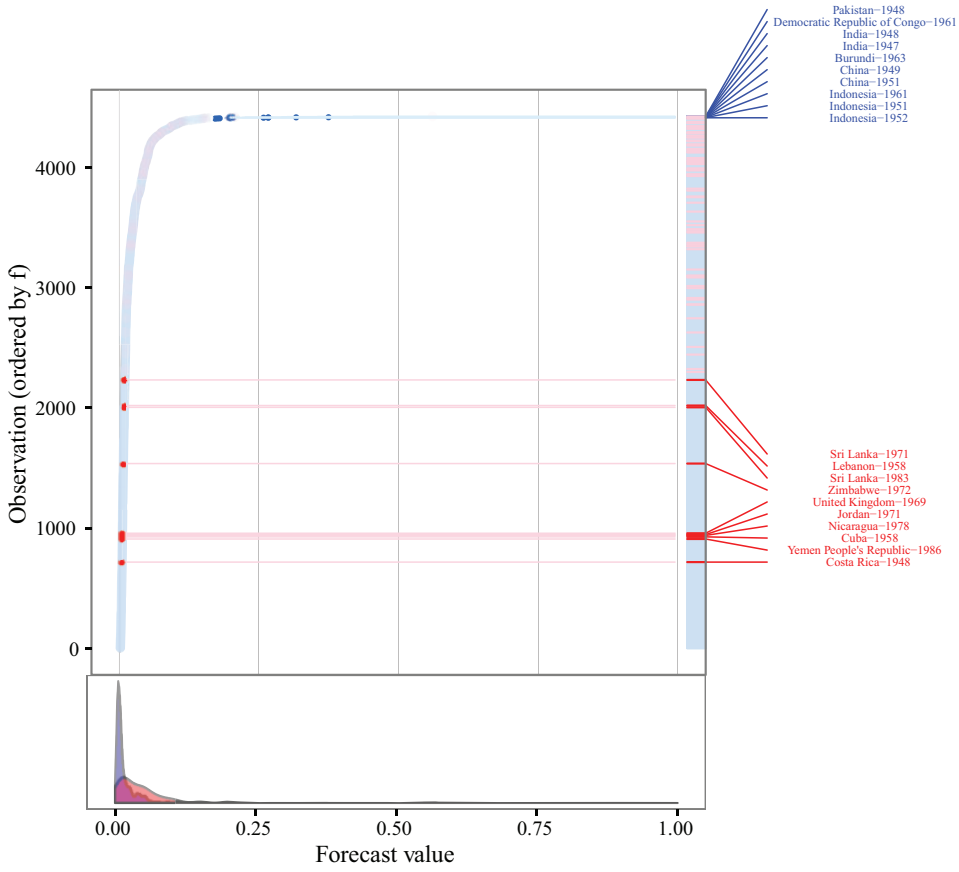

Figure 6. Model criticism plot for full Fearon & Laitin model

power, such as South Africa, Jordan, and Sri Lanka. This suggests a potential set of features that could be added to explore whether they improve the performance of the model. While it is true that the Fearon and Laitin model measures ethnic fractionalization, this does not measure exclusion from political power. Further, research by Wimmer, Cederman & Min (2009) suggests that overt ethnic exclusion and having a history of imperial rule are features that might increase the probability of civil war.

These measures of exclusion mark the first new features that we will explore in our next iteration. In addition, we notice that across the logit models the discrepancies are clustered within a country (China) or region (Africa and the Middle East), particularly for the civil war cases. This suggests that events may not be independent (Buhaug & Gleditsch, 2008). These plots also lead to the conjecture that variables with higher within-country variance could be useful in predicting civil wars. Some of the discrepant observations of peace (blue) appear to reflect structural conditions with high risk – India and Indonesia, for example – without accounting for potential counterbalancing forces, such as economic growth or changes in regime characteristics.

Loop 2

Analyzing the discrepant observations in the previous models led us to create our own model representation. We explicitly build upon the simplest model, the two feature logistic, as it had the lowest optimism in the previous round, suggesting to us that it has captured a small set of generalizable patterns. Therefore, we add two features that represent political exclusion: the log of the ratio of the excluded to the total population, and imperial history. We also add a measure of whether, in the previous year, a neighboring country had a civil war. Finally, to measure changes within countries, we include features that represent economic growth: yearly change in per-capita GDP and political instability, measured as the regime's instability within the previous three years.

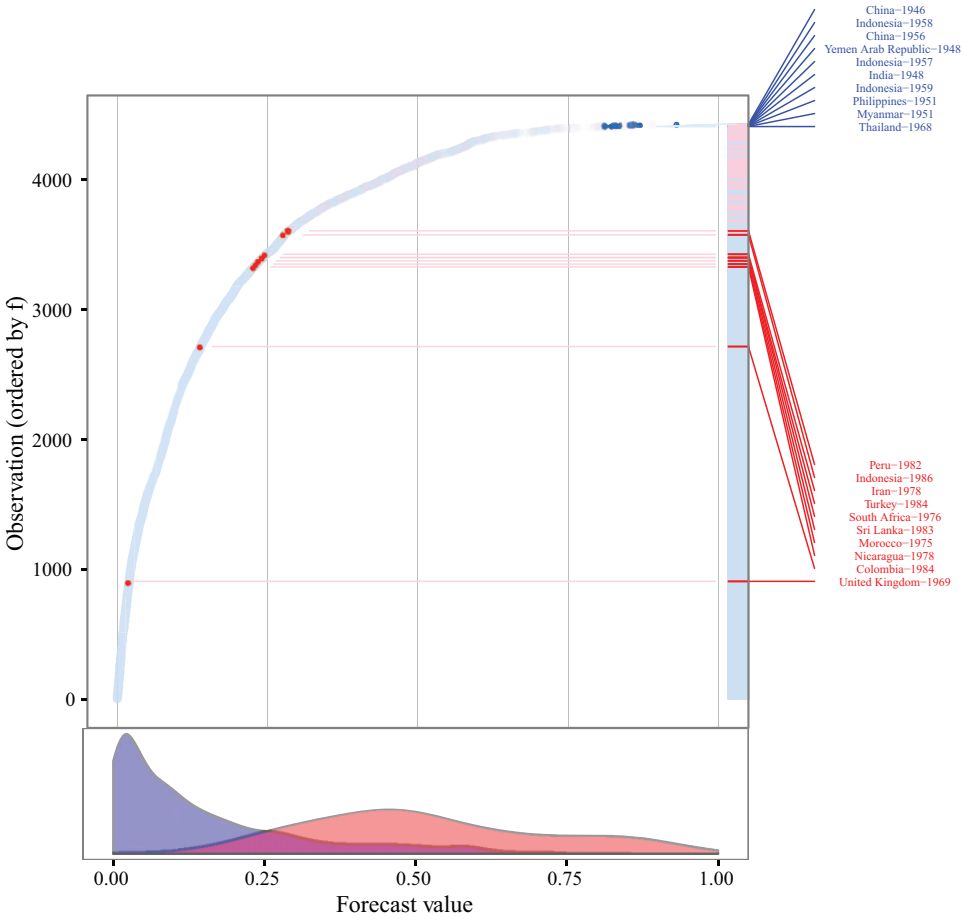

Figure 7. Model criticism plot for full Muchlinski et al. random forest model

Our new model now has a total of seven features. We use two different models to transform the features into predictions, a simple logistic model where the features enter additively on the log-odds scale, and a random forest, to potentially account for conditional relationships. To compute a new fitted model, we resample from the training instances and again use 10-fold cross-validation.

Note that we still have not touched the test set, and that these folds are different samples than the previous loop.[19] The performance for our new logit model is an improvement. We have a cross validated AUC of 0.74. Because the partitions are different, we also have slightly distinct values for the previous models. However, here the stories are similar. The two-feature model, without our additions, has a cross-validated AUC of 0.65, and the full Fearon & Laitin specification, which shares three of our features, has a comparable value of 0.61.[20] Our new random forest model has a cross-validated AUC of 0.86, which is competitive with the Muchlinski et al. random forest model, which scores 0.88.[21]

The model criticism plot for our new fitted logit model, now with seven features, is presented in Figure 8. Costa Rica and Yemen remain discrepant observations,

[19] We ensure this by explicitly setting random seeds. For clarity, while the same data make up the observations across each of the ten folds, there is now a different partition for each fold, and thus different memberships within folds.

[20] Since the set of all training instances are the same, the in-sample AUC, without cross-validation, are the same as those reported in the first column of Table I.

[21] Future research could extend these resampling strategies by using repeated k-fold cross-validation, and then the variables of the performance measures could also be compared. This is beyond the scope of our current project.

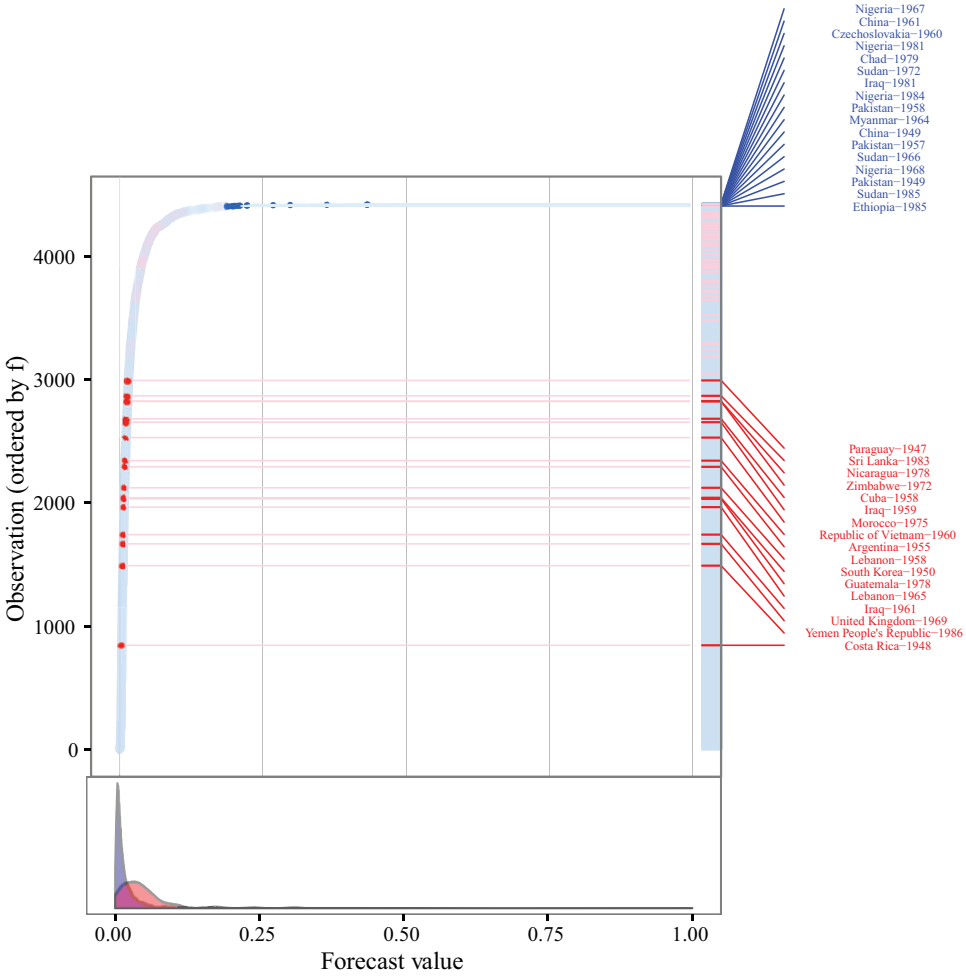

Figure 8. Model criticism plot for loop 2 logistic model

but there is now slightly more separation between the distribution of civil war and peace observations, with the red distribution moved slightly to the right. The new random forest models predictions and discrepancies are illustrated in Figure 9. The flexible representation that random forests provides still boosts performance. We see this in the sharper distance between the peaks of the two distributions, as well as in the shift of civil war observations towards the upper right of the plot.[22]

The biseparation plot

Comparing discrepancies across models is also useful. We would like to see where our rebuilt model improved on the two-feature model, as well as where it fell short. To look at how different models sort the observations, we utilize a new type of plot. This biseparation plot draws a separation plot on each axis, with the resulting scatter plot measuring the relative sort-order of observations. Two identical models would have observations arrayed at a 45-degree angle. If a civil war observation (red) is above the 45-degree angle, this means that the model on the y-axis sorted it more highly than the model on the x-axis. The pull upwards was greater than the pull to the right. Similarly, a zero observation (blue) below the 45-degree angle denotes that the push downwards, a lower ranking for the model on the y-axis, was stronger than the push to the left, for the model on the x-axis. The best improvement in rank for one model is also the worst decline for the other model; the *direction* of improvement depends on the observed value. Improved observations for the x-axis model are downwards and

[22] We do not show the model criticism plots for the other models from loop 1, since they are extremely similar to what was previously presented.

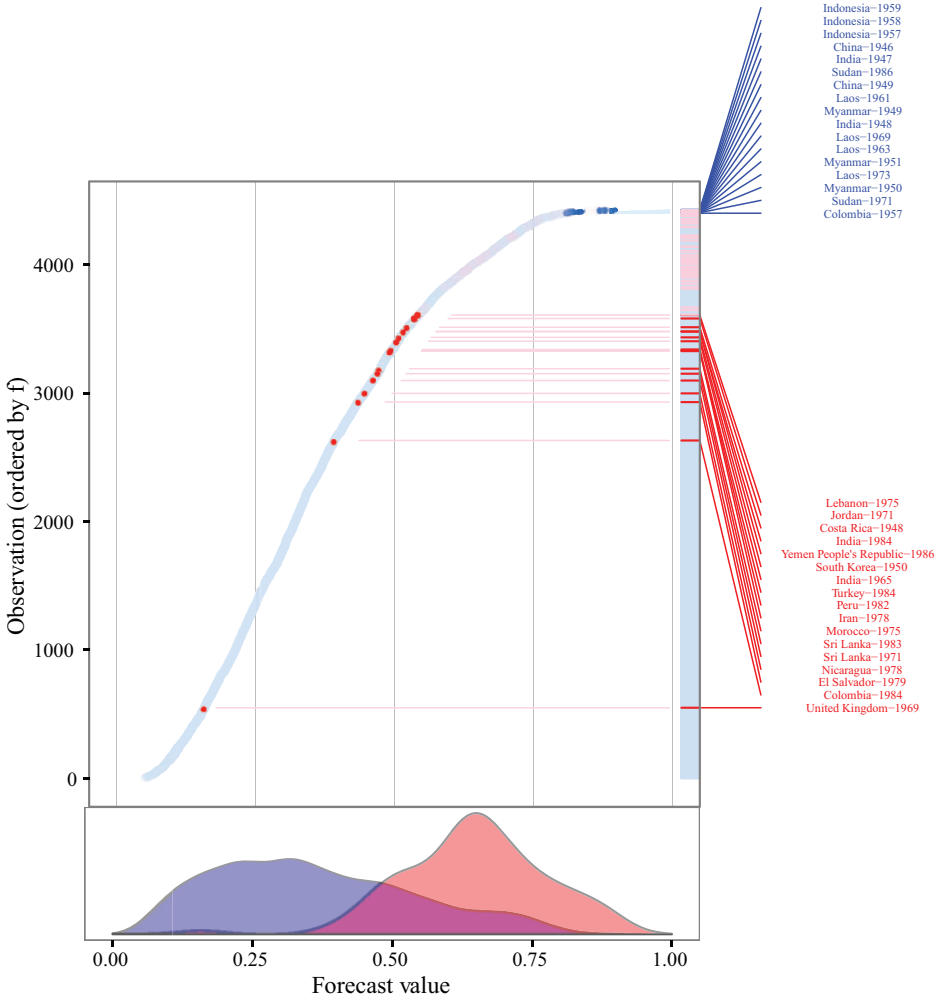

Figure 9. Model criticism plot for loop 2 random forest model

right for civil war cases, and upwards and left for cases of peace.

Figure 10 presents the biseparation plot for our updated logistic model compared to the two-feature minimal model. On the x-axis are the observations, ranked by our updated logistic model. The y-axis is the ranking of the same observations, but now by the two-feature model. We label and highlight the 15 most improved ranks for each model and outcome, relative to the other, and connect those observations to the improving model by a line to aid interpretations. We can see that there are very few light red dots on the left-hand side of the 45-degree divide. In comparison there are many more on the right-hand side of the same division. This signals that the civil war occurrence observations are being sorted to higher ranks in the new model. However, there are several discrepant observations that are apparent. For example, the conflict in South Korea has a lower forecast rank in our new model. Of note is that several of the observations of civil war that are more highly ranked by the simpler two-feature model are primary commodity exporting countries, including Vietnam and Indonesia. In fact, looking back at the model criticism plot for our updated logistic model in Figure 8, we see that many discrepant civil war cases export primary commodity goods such as oil, crude materials, and metals.

Loop 3

Using the information from the previous models, we estimated a final model that included a measure of primary commodity exports, along with its square, discussed in Collier & Hoeffler (2004) and collected from the dataset provided by Hegre & Sambanis (2006). While the use of this index has drawn fire (Fearon,

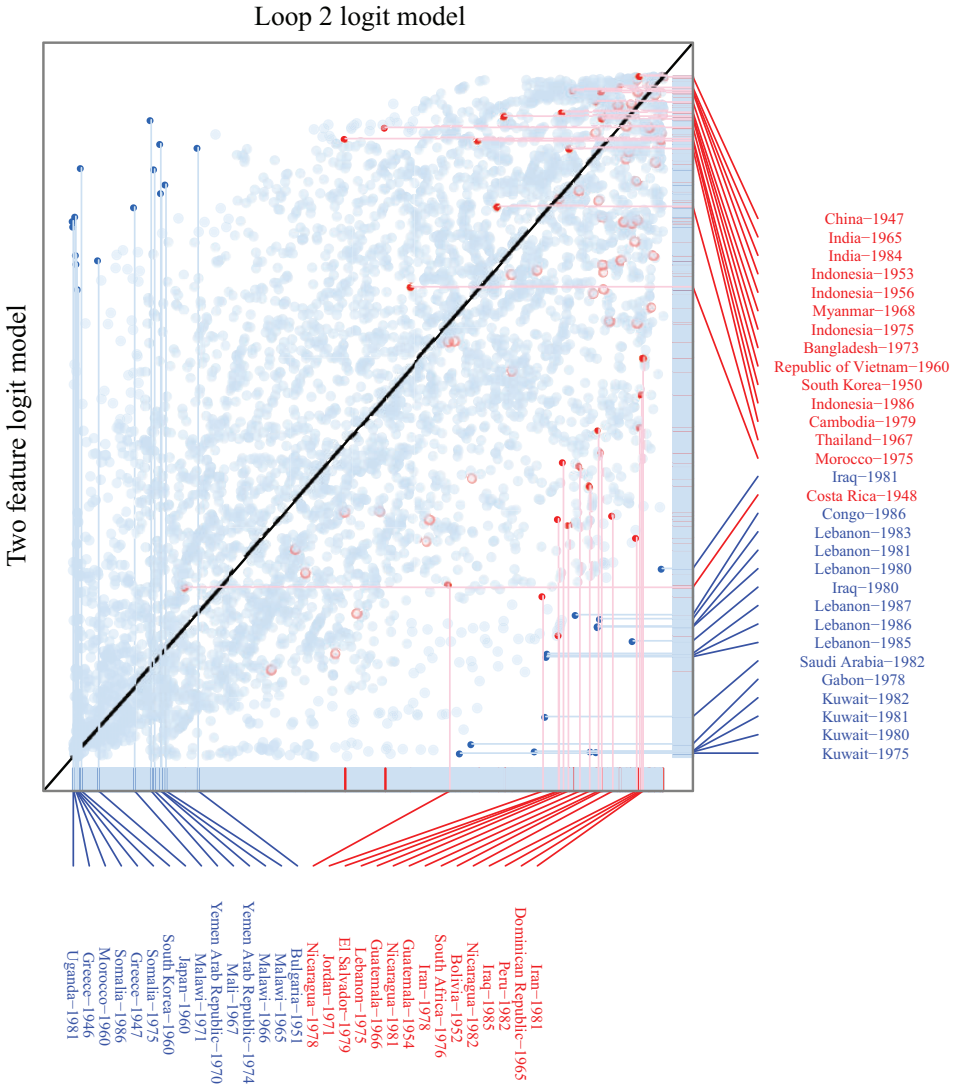

Figure 10. Biseparation plot: loop 2 logistic model vs. two-feature model

2005), the empirical grounds for disagreement have rested on null hypothesis testing and significance. We include these two additional features to see whether performance is boosted in both the logistic and random forest representations. Our random forest representation in this loop only includes the nine features. We again sample ten folds for cross-validation, and fit this new model. The resulting cross-validated AUC for our refined logit model is 0.79, and for the refined random forest is 0.90. This is competitive with the much larger scale random forest that has 88 variables, which has a cross-validated AUC of 0.87. These models are also competitive on precision-recall curves, as presented in Figure 11. The loop 3 logit models are an improvement on the Fearon & Laitin model and the two feature model in cross-validated AUC and on the precision-recall curves. The random forest model using 88 features is slightly better on precision-recall than the smaller scale random forest from our loops. A model criticism plot of the refined random forest model, with the two primary commodities features added, is presented in Figure 12. We only plot the four worst cases for each class here because they are so close to each other, suggesting that there are few discrepancies in the training set. Similarly, the biseparation plot in Figure 13 shows the marked improvement relative to the two-feature model. In particular, the civil war occurrence cases (red) that were given relatively low forecast values in the simpler model (so are on the left) are sorted much higher in the streamlined random forest model (pulled upwards). While more information could be extracted from these plots

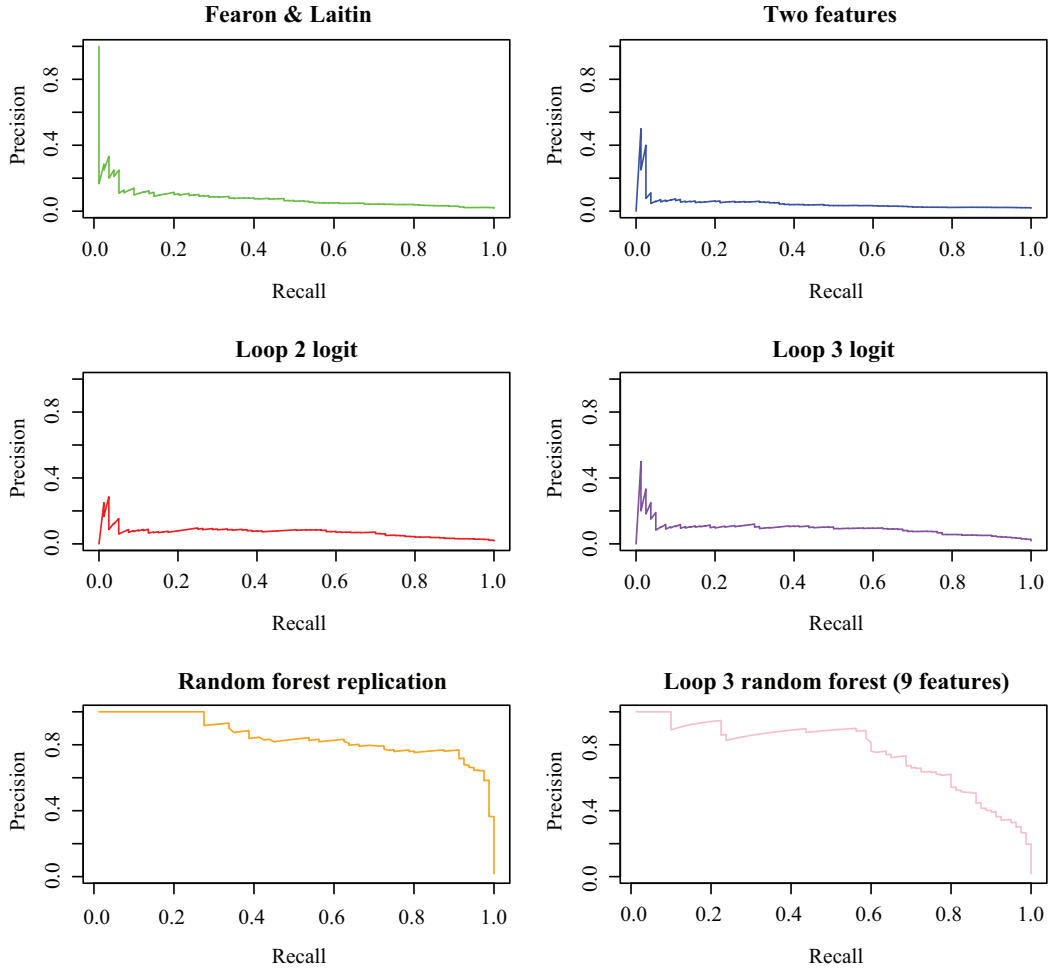

Figure 11. Precision-recall curves for all loop 3 models

(for example, there appear to be country-level factors at work in the civil war absence cases that the simple logit model sorts lower than the random forest model), we move to examining the performance of the models on the as-yet-unused test set.

Performance in the test set

It remains possible that the models that we computed show improved performance in the training set, but not in unseen data. We thus turn to the test set to measure how our models generalize to the post-1988 era. Specifically, we use the models that we developed in loop 3 to make predictions for the test set cases and then compare our predictions to the observed values. We did not use the test set to either develop the models or compute the fitted models. Again, we use AUC, ROC plots, and precision-recall plots to measure our performance. We present the AUC values and ROC plots in Figure 15. The AUC for the loop 3 logit model is over 0.80, which compares favorably to the original models. Further, our random forest model with only nine features has an AUC above 0.90, which is the highest across the representations we tested. A similar story emerges in the precision-recall plots presented in Figure 14. Here we see some improvement in the logistic model from loop 3, including primary commodity exports, political exclusion, economic growth, and whether a neighbor is suffering a civil war, relative to the baseline models. However, the largest improvement comes about in the random forest model that has our smaller set of features (bottom right).

These results suggest that our models are capturing generalizable patterns and not simply noise. Without held-out data, we would not have been able to answer the question of how our models applied to new unseen cases. Further, we were able to build a much more parsimonious representation of civil war with performance

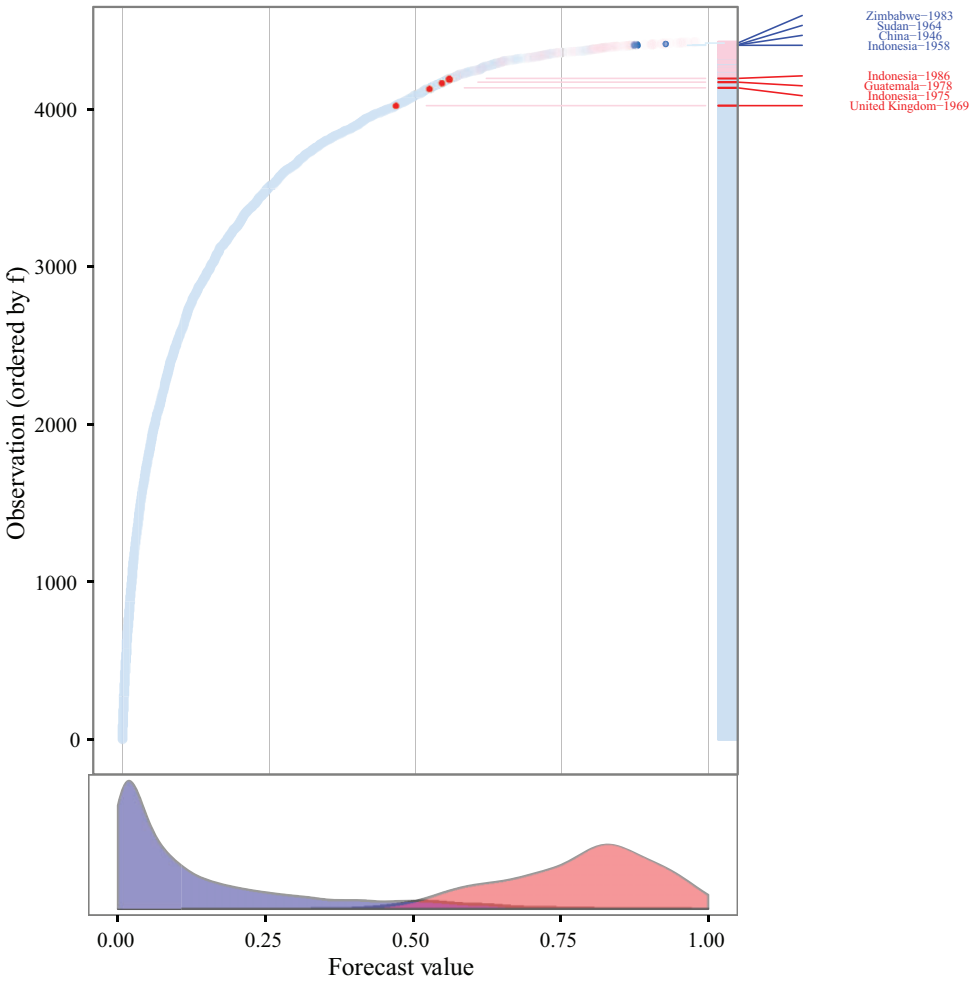

Figure 12. Model criticism plot of loop 3 random forest model

comparable to larger scale, state-of-the-art models such as those of Muchlinski et al. (2016).

The relative test set performance of these models has some significant repercussions for theory building and development in civil war studies. While there has been a debate concerning whether primary commodity exports are important drivers of civil conflict (Collier & Hoeffler, 2004; Fearon, 2005), our findings here suggest these are useful leading indicators as they boosted test set performance in the logit specifications and contributed to the top-performing random forest model. In the future, arguments about alternative measures, such as institutional strength, should utilize Box's loop to rebuild a model with new features and test its performance relative to this model, or one like it, on unseen data. In addition, the underperformance of the simpler logistic models, relative to random forests, is of interest. There may be important non-linear and conditional signals in the process of civil war.[23]

Conclusion: Dancing without the stars

In this article, we have attempted to illustrate what is novel and important in machine learning research designs, as summarized by Box's loop (Blei, 2014). Confirmatory analyses such as NHST can be valuable when researchers already have a useful model representation of the problem that they are confident in. However, these

[23] Due to space constraints we do not present partial dependence plots from the random forests. However, these do show numerous non-linearities, mirroring the findings in Muchlinski et al. (2016). We also use an Online appendix to present the relationship between model complexity and error rates within the training and test sets.

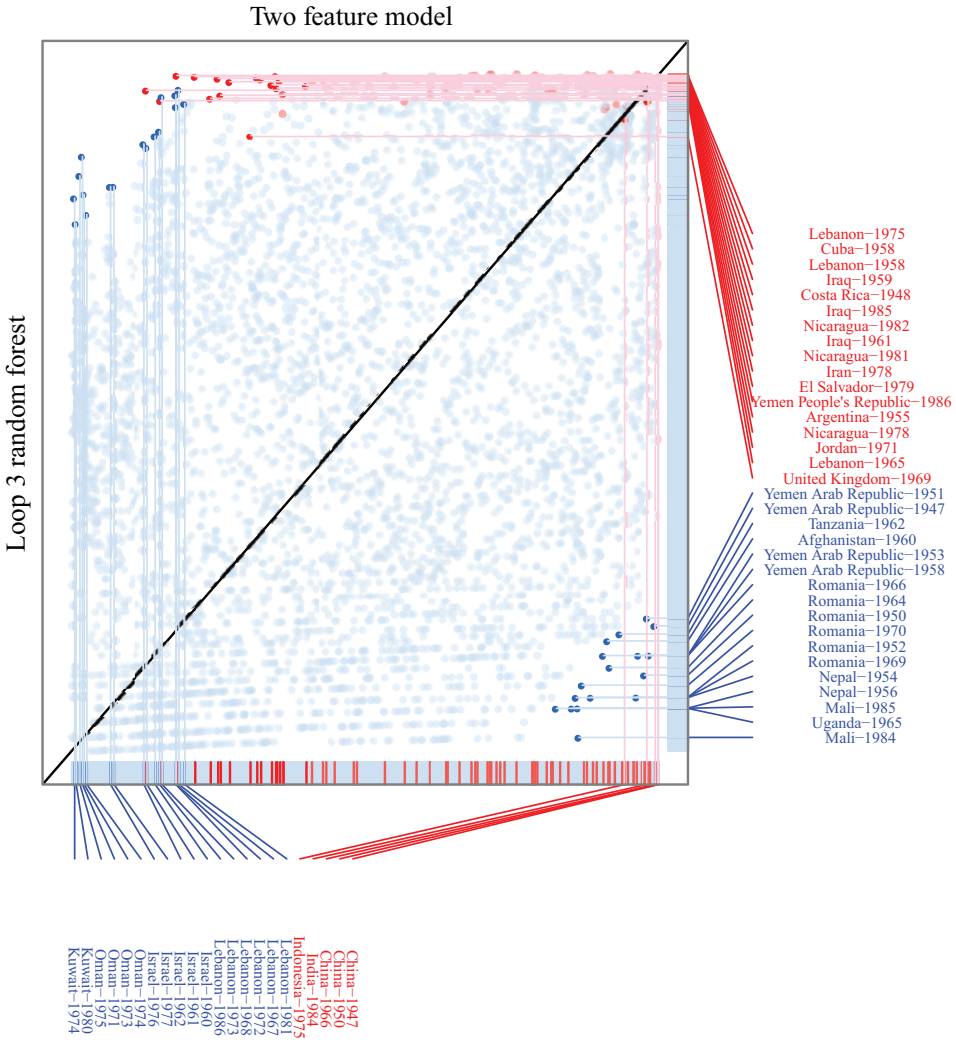

Figure 13. Biseparation plot: loop 3 random forest model vs. original two-feature logit model

types of analysis do not provide advice on how to build that useful representation in the first place. In machine learning, the iterative steps of building, computing, critiquing, and thinking about models provide a strategy for constructing useful models across a variety of tasks, including conflict prediction. Through the careful use of held-out data and flexible model representations, both underperformance and overfitting can be avoided. Instead of assuming that a given model representation is true, we are able to probe and expose its weaknesses to learn how to rebuild a more useful representation in the next iteration.

Machine learning has a number of important principles that can inform conflict processes research. Ward, Greenhill & Bakke (2010) note that highly cited models of civil war built from NHST research designs have failed to predict well out-of-sample. Here, we illustrated how the use of Box's loop enabled us to build modes that not only improved on the performance of these previous models, but even competed with recent top-performing systems (Muchlinski et al., 2016). As conflict processes researchers continue to build forecasting models, the careful sampling and splitting of data and iterative model improvement within Box's loop can accelerate innovations.

It is important to note that Box's loop is not a solution to all research questions. Causal identification strategies and experiments, where practical, offer important windows into conflict data generation processes. The machine learning approaches summarized and deployed here take as their goal building models that are useful for specific tasks, not minimizing the bias of parameters. Prediction and forecasting civil wars fit neatly in definition of a supervised learning problem. Estimating the impact of a peacekeeping mission in Syria, for example,

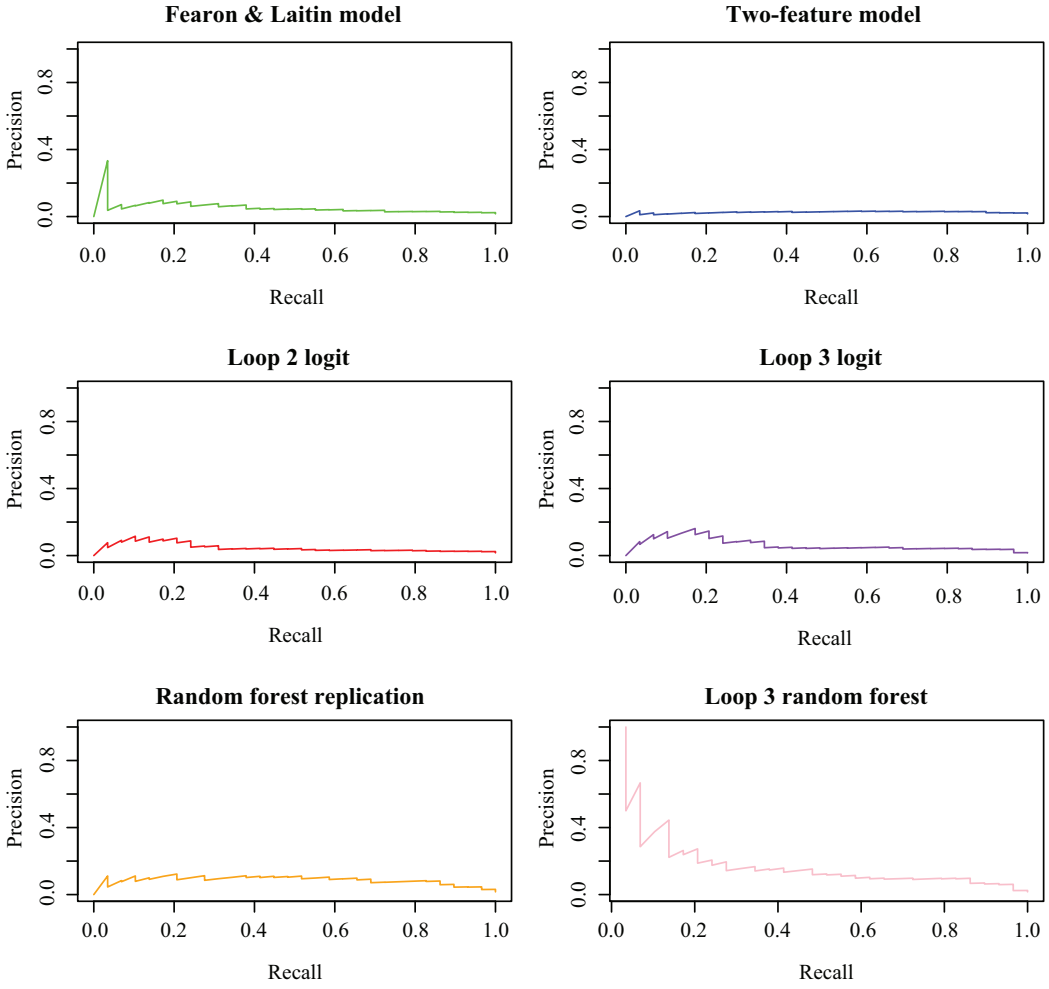

Figure 14. Precision-recall plots for the test set (post-1988)

fits less well. Further, where data are sparse, sampling strategies may not be possible.

However, for many observational studies, Box's loop can help to both improve performance and better understand the data generation process. In NHST, it is possible to simply push the same covariance into different parameters, computing a different constellation of stars on parameters, but doing little to explain new patterns in the data. In machine learning-inspired approaches, new useful theories should increase the performance of a model and that performance gain should generalize out-of-sample. Further, the identification of discrepancies in out-of-sample data can inform theory development and refinement (Daxecker & Prins, 2017; Witmer et al., 2017).

There are two clear directions in which this research could be extended. First, in order to build on the work of Ward, Greenhill & Bakke (2010) and Muchlinski et al. (2016) we utilized a country-year design. However, moving to monthly or quarterly forecasts at smaller geographic levels of conflict might be more relevant and useful. Second, our current approach is limited to binary prediction problems, but could and should be extended to multinomial, ordered, and continuous targets for prediction. Recent research by Brandt, Schrodt & Freeman (2014) convincingly makes the case that forecast accuracy for the continuous case must extend beyond mean squared error. This raises the bar for forecast systems to match not only the central tendency of a process of interest, but also its shape (e.g. variance, skewness, etc.). Building tools to facilitate model criticism and identifying discrepancies in these cases may help to meet that challenge. Additionally, the release of new events data which may be transmuted into categorical, ordered or continuous measures, suggests that tools and approaches to facilitate the use of Box's loop for these types of data will be increasingly in demand.

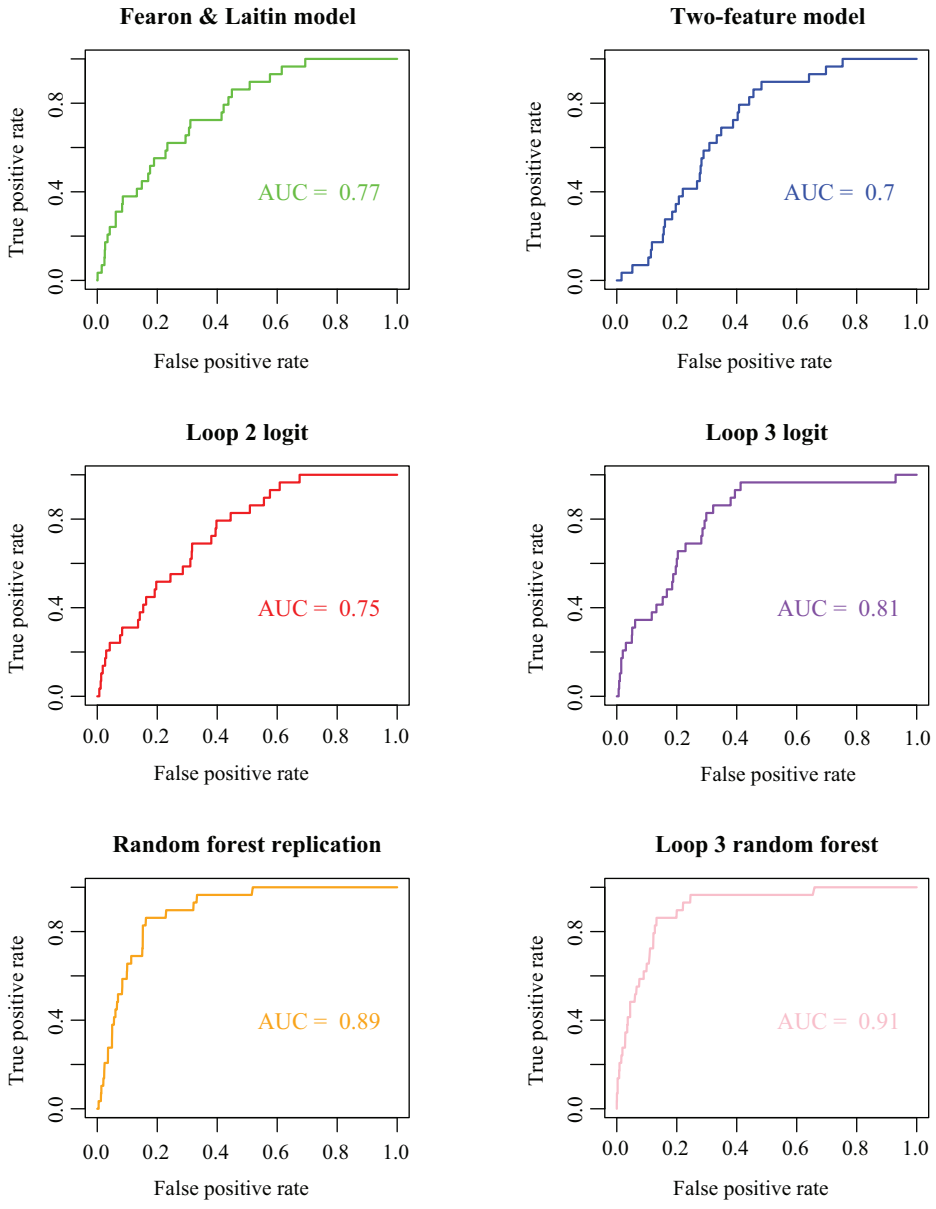

Figure 15. ROC plots within the test set (post-1988)

Replication data

The Online appendix and replication data for this article are available at http://www.prio.org/jpr/datasets.

Acknowledgements

We would like to thank Gerald Schneider, Amanda Murdie, Phil Schrodt, Patrick Brandt, Kristian Skrede Gleditsch, and all of the participants at the March 2016 conference on Conflict Forecasting for their helpful feedback on this project. We are indebted to the special issue and *JPR* editors for their detailed and constructive suggestions as the research evolved. This project also benefited from the suggestions by participants in TextLab at Michigan State University, including details and helpful advice from Baekkwan Park, Lora DiBlasi, and Kevin Greene. Erika Rosebrook, Ezra Brooks, Elizabeth Lane, and Jamil Scott provided crucial reactions to early drafts of the visualizations. The remaining errors and omissions are solely the responsibility of the authors.

References

Blei, David M (2014) Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Applications* 1(1): 203–232.

Box, George EP (1980) Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A (General)* 143(4): 383–430.

Brandt, Patrick; Philip Schrodt & John Freeman (2014) Evaluating forecasts of political conflict dynamics. *International Journal of Forecasting* 30(4): 944–962.

Buhaug, Halvard & Kristian Skrede Gleditsch (2008) Contagion or confusion? Why conflicts cluster in space. *International Studies Quarterly* 52(2): 215–233.

Chen, Danqi & Christopher Manning (2014) A fast and accurate dependency parser using neural networks. *Proceedings of EMNLP* (http://cs.stanford.edu/people/danqi/papers/emnlp2014.pdf).

Collier, Paul & Anke Hoeffler (2004) Greed and grievance in civil war. *Oxford Economic Papers* 56(4): 563–595.

Daxecker, Ursula & Brandon C Prins (2017) Financing rebellion: Using piracy to explain and predict conflict intensity in Africa and Southeast Asia. *Journal of Peace Research* 54(2): 215–230.

Fearon, James D (2005) Primary commodity exports and civil war. *Journal of Conflict Resolution* 49(4): 483–507.

Fearon, James D & David D Laitin (2003) Ethnicity, insurgency, and civil war. *American Political Science Review* 97(1): 75–90.

Flach, Peter (2012) *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. New York: Cambridge University Press.

Gelman, Andrew & Eric Loken (2014) The statistical crisis in science. *American Scientist* 102(6): 460.

Gill, Jeff (1999) The insignificance of null hypothesis significance testing. *Political Research Quarterly* 52(3): 647–674.

Greene, Kevin; Baekkwan Park & Michael Colaresi (2016) Machine learning human rights and wrongs: How the successes and failures of supervised learning algorithms can inform the debate about information effects. Working paper.

Greenhill, Brian D; Michael D Ward & Audrey 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.

Hastie, Trevor; Robert Tibshirani & Jerome Friedman (2013) *Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer.

Hegre, Håvard & Nicholas Sambanis (2006) Sensitivity analysis of empirical results on civil war onset. *Journal of Conflict Resolution* 50(4): 508–535.

Japkowicz, Nathalie & Shaju Stephen (2002) The class imbalance problem: A systematic study. *Intelligent Data Analysis* 6(5): 1–11.

King, Gary & Langche Zeng (2001) Logistic regression in rare events data. *Political Analysis* 9(1): 137–163.

Kotsiantis, SB (2007) Supervised machine learning: A review of classification techniques. *Informatica* 31(1): 249–268.

Kuhn, Max & Kjell Johnson (2013) *Applied Predictive Modeling*. New York: Springer.

Levine, Timothy R; Rene Weber, Craig Hullett, Hee S Park & Lisa LM Lindsey (2008) A critical assessment of null hypothesis significance testing in quantitative communication research. *Human Communication Research* 34(1): 171–187.

McCormick, Tyler H; Adrian E Raftery, David Madigan & Randall S Burd (2012) Dynamic logistic regression and dynamic model averaging. *Biometrics* 68(1): 23–30.

Mitchell, Tom (1998) *Machine Learning*. New York: McGraw-Hill.

Muchlinski, David; David Siroky, Jingrui He & Matthew Kocher (2016) Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis* 24(1): 87–103.

Raftery, Adrian E; Tilmann Gneiting, Fadoua Balabdaoui & Michael Polakowski (2005) Using Bayesian model averaging to calibrate forecast ensembles. *American Meteorological Society* 133(1): 1155–1173.

Schrodt, Philip (2014) Seven deadly sins of contemporary quantitative political analysis. *Journal of Peace Research* 51(2): 287–300.

Siegel, Eric (2013) *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*. New York: Wiley.

Simmons, Joseph P; Leif D Nelson & Uri Simonsohn (2013) False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22(11): 1359–1366.

Steyerberg, Ewout W (2009) *Clinical Prediction Models: A Practical Approach to Development, Validation and Updating*. New York: Springer.

Ward, Michael D & Andreas Beger (2017) Lessons from near real-time forecasting of irregular leadership changes. *Journal of Peace Research* 54(2): 141–156.

Ward, Michael D; Brian D Greenhill & Kristin Bakke (2010) The perils of policy by p-value. *Journal of Peace Research* 47(4): 363–375.

Ward, Michael D; Nils W Metternich, Cassy L Dorff, Max Gallop, Florian M Hollenbach, Anna Schultz & Simon Weschle (2013) Learning from the past and stepping into the future: Toward a new generation of conflict prediction. *International Studies Review* 15(4): 473–490.

Wimmer, Andreas; Lars-Erik Cederman & Brian Min (2009) Ethnic politics and armed conflict: A configurational analysis of a new global dataset. *American Sociological Review* 74(2): 316–337.

Witmer, Frank DW; Andrew M Linke, John O'Loughlin, Andrew Gettelman & Arlene Laing (2017) Subnational violent conflict forecasts for sub-Saharan Africa, 2015–65, using climate-sensitive models. *Journal of Peace Research* 54(2): 175–192.

MICHAEL COLARESI, b. 1976, PhD in Political Science (Indiana University, 2002); Professor, Michigan State University (2013–); Director, Social Science Data Analytics Initiative, Michigan State University (2015–); current interests: domestic politics of international conflict, secrecy in democracies, and computational social science; most recent book: *Democracy Declassified: The Secrecy Dilemma in National Security* (Oxford University Press, 2014).

ZUHAIB MAHMOOD, b. 1989, PhD student in Political Science (Michigan State University); current interests: international conflict and cooperation, and computational social science.