PA

How Cross-Validation Can Go Wrong and What to Do About It

Marcel Neunhoeffer and Sebastian Sternberg

Department of Political Science, Quantitative Methods in the Social Sciences, University of Mannheim, A 5, 6, 68131 Mannheim, Germany. Email: marcel.neunhoeffer@gess.uni-mannheim.de, sebastian.sternberg@gess.uni-mannheim.de

Keywords: data analysis algorithms, binary choice models, forecasting

1 Introduction

With the increasing use of "machine learning" methods in political science new terminology is introduced to our field. While most political methodologists extensively learned how to apply regression models, the application of newly introduced "machine learning" methods and models is often harder. This can lead to serious problems. Even more so, when one term—like cross-validation—can mean very different things. We find four different meanings of cross-validation in applied political science work. We focus on cross-validation in the context of predictive modeling, where cross-validation can be used to obtain an estimate of true error or as a procedure for model tuning. Our goal with this work is to experimentally explore potential problems with the application of cross-validation and to show how to avoid them. With a reanalysis of a recent paper (Muchlinski *et al.* 2016) we highlight that these problems are not only of theoretical nature but can also affect the reported results of applied work.

First, we survey political science articles in the leading journals of the discipline and identify different meanings of cross-validation in applied political science work. Second, we focus on problematic cross-validation in the context of predictive modeling. Using a single cross-validation procedure to obtain an estimate of the true error *and* for model tuning at the same time leads to serious misreporting of performance measures. We demonstrate the severe consequences of this problem with a series of experiments. Third, we use the study by Muchlinski *et al.* (2016) on the prediction of the onset of civil war to illustrate that problematic cross-validation can affect applied work. They claim that their random forest model is more accurate than logistic regression models in the prediction of civil war onsets, even when tested on out-of-sample data. However, our reanalysis shows that they use a single cross-validation procedure for model tuning and estimation of the true error, and therefore report inaccurate performance measures. When we apply cross-validation correctly the authors' conclusions do not hold, including the already prominently cited (see Cederman and Weidmann 2017; Cranmer and Desmarais 2017; Colaresi and Mahmood 2017) main conclusion 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" (Muchlinski *et al.* 2016, 101). We encourage researchers in predictive modeling to be especially mindful when applying cross-validation.

Political Analysis (2019)
vol. 27:101–106
DOI: 10.1017/pan.2018.39

Corresponding author
Marcel Neunhoeffer

Edited by
Jeff Gill

Authors' note: Replication materials are available online as a dataverse repository at https://doi.org/10.7910/DVN/Y9KMJW (Neunhoeffer and Sternberg 2018). We thank Thomas Gschwend, Richard Traunmüller, Sean Carey, Sebastian Juhl, Verena Kunz, Guido Ropers, the participants of the CDSS Political Science colloquium and two anonymous reviewers for their helpful comments. All remaining errors are our own. This work was supported by the University of Mannheims Graduate School of Economic and Social Sciences funded by the German Research Foundation.

2 Cross-Validating Cross-Validation in Political Science

Our survey of the literature suggests that the term cross-validation has four different meanings in applied political science work. We searched JSTOR for the term cross-validation in publications of three leading political science journals since 2010. In total we found 42 articles with the term cross-validation.[1] For a table with all 42 articles, see Online Appendix A available at the Political Analysis dataverse site.

First, the term cross-validation is sometimes—in seven articles—used to describe the process of *validating new measures or instruments*, for instance in the context of survey research (e.g., Cantú 2014). The other three meanings of cross-validation all have to do with resampling as a statistical tool. Generally, cross-validation here means to randomly divide the data set into several about equally sized folds (each fold will contain about $\frac{N}{K}$ observations, where K indicates the number of folds and can be anything between 2 and N the number of rows in the data set). A statistical model is then K times trained on all but the k-th fold, where k runs from 1 to K. Using this logic, the second meaning of cross-validation refers to its use as a *robustness check* of coefficients e.g. in regression analysis. In our survey of the literature this was the case in two articles (e.g. Engstrom 2012).

Third, in the context of predictive models, the term cross-validation is used—in eleven articles—to describe a procedure *to obtain an estimate of true error*. True error is a measure of how well a model can predict outcomes of previously unseen data (see Efron and Hastie 2016; Cranmer and Desmarais 2017). An estimate of true error is important in practice, as it allows one to check whether a model generalizes well to unseen data or just memorizes the patterns in the training data (i.e. overfitting). Cross-validation can be used to approximate true error without setting aside additional validation data. The model is trained according to the resampling scheme described above and then the accuracy (or any other measure) is evaluated on the k-th fold (test fold) that was not part of the training. This process is then repeated for all K folds and the average (across the K folds) accuracy (or the average of any other measure) is reported. An example of using cross-validation to estimate true error from political science is Caughey and Warshaw (2015).

Fourth, and most often—in 17 articles—cross-validation is used to describe a procedure for *model tuning*. A model can be tuned, for example, by repeatedly testing different (hyper-) parameter values and selecting the value that had the lowest error on a test fold. Hainmueller and Hazlett (2014) for instance use cross-validation to find the best regularization parameter λ in a Kernel Regularized Least Squares model. Model tuning can take many forms, including (hyper-) parameter tuning, feature selection or up-/down-sampling of imbalanced data prior to model training.

3 Experimental Exploration of Problematic Cross-Validation

A problematic use of cross-validation occurs when a single cross-validation procedure is used for model tuning *and* to estimate true error at the same time (Cawley and Talbot 2010; Hastie, Tibshirani, and Friedman 2011; Efron and Hastie 2016). Ignoring this can lead to serious misreporting of performance measures. If the goal of cross-validation is to obtain an estimate of true error, every step involved in training the model (including (hyper-) parameter tuning, feature selection or up-/down-sampling) has to be performed on each of the training folds of the cross-validation procedure. Hastie, Tibshirani, and Friedman (2011, 245) refer to this problem as the wrong way of doing cross-validation. We take down-sampling of imbalanced data as a simple example of this problem. Down-sampling the data set, e.g. to balance the dependent variable, prior to the cross-validation procedure implies that the fold that is used for testing

1 The search was conducted on April 25 2018. The three journals in our search are APSR, AJPS, and PA. Since the time period covered by JSTOR is different for each journal, we supplemented the JSTOR search with manual searches on the journal websites for the term cross-validation in the period after the last result in JSTOR and before April 25 2018.

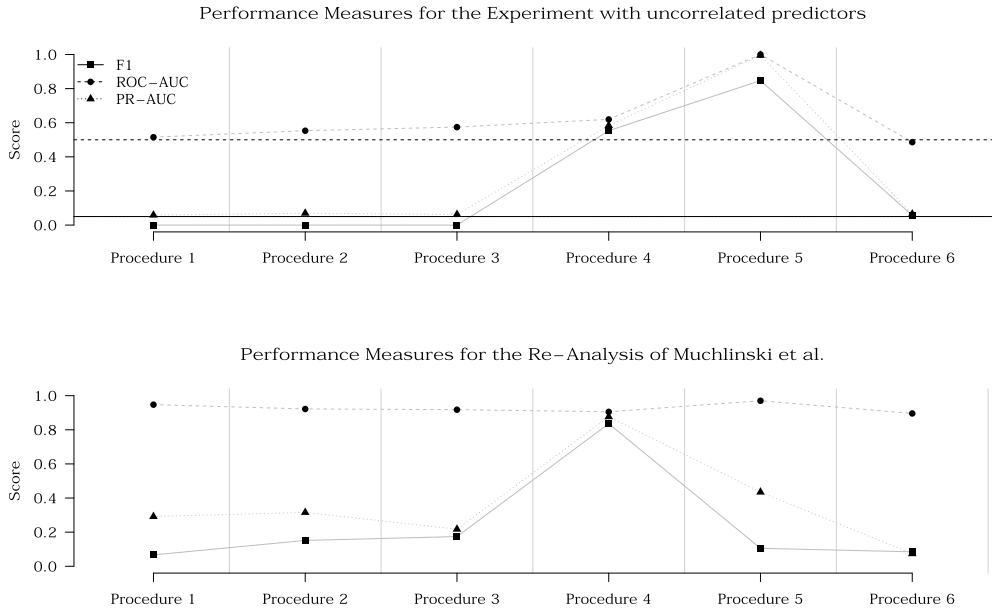

Figure 1. Cross-Validation and Performance Measures. Top Panel: Experiment. Note that the lines for the true F_1 score and the true PR-AUC score overlap. Lower Panel: Reanalysis of Muchlinski *et al.* (2016).

in each iteration of the cross-validation procedure is also balanced (like the training set). It is straightforward that errors calculated on these test folds cannot serve as an estimate of true error, where the data will always be imbalanced. The right way of combining down-sampling of imbalanced data with cross-validation would be to first split the entire data set into the k folds and then only down-sample the folds that are used for training. The test fold should remain imbalanced to reflect the imbalance in unseen data. It is even more problematic if researchers apply cross-validation for model tuning and report the performance of, for example, the best model on the training set—the so called apparent error, while believing they are reporting some cross-validated error. This apparent error should not be used as an estimate of true error as it most likely dramatically overestimates the performance of a model.

To demonstrate the severe consequences of the problematic use of cross-validation, we conduct six experiments. We set up a data set with 2, 000 observations of a binary outcome Y with $p(y_i = 1) = 0.05$ and a set of 90 uncorrelated predictor variables X. We randomly split the data into 1, 500 observations in the training set and 500 observations in the test set. The true error of any classifier on this data set can be expressed by the following performance measures. The true F_1 score is 0.05, the true ROC-AUC score is 0.5 and the true PR-AUC is 0.05. See Online Appendix B.1 for definitions of these three performance measures. The replication code and data are available online as a dataverse repository (Neunhoeffer and Sternberg 2018).

The results of our experiments are reported in the top panel of Figure 1. The true scores are indicated by the horizontal lines. We first train a random forest model[2] without model tuning[3] on the training data and report its performance on the test set (**Procedure 1**). Unsurprisingly, the performance measures for this procedure are close to the true performance measures.

Second, we train a random forest model with 10-*fold* cross-validation—we apply stratified cross-validation such that the distribution of 0 and 1 is similar across all folds—and average the

2 We use a random forest model for consistency with the application to Muchlinski *et al.* (2016). Our results generalize to other statistical models.

3 For random forest models we could tune two parameters, the number of predictors randomly sampled at each split (m) and the number of trees (*ntree*). We set m to the default of $\lfloor\sqrt{(p)}\rfloor = 9$ with $p = 90$ being the number of predictors and the number of trees to 1, 000.

scores across the 10 folds to obtain an estimate of the true error (**Procedure 2**). Again we can observe that the performance scores of Procedure 2 are close to the true scores. This shows that cross-validation correctly applied provides a close approximation of true error. Third, we combine down-sampling and cross-validation correctly as described above (**Procedure 3**). This means we first split the entire data set into 10 folds, and then only down-sample the folds used for training while not touching the test folds. When applied correctly, the error obtained from this procedure is, as expected, close to true error.

Fourth, we combine cross-validation and down-sampling wrongly. This means we first down-sample the data set prior to the cross-validation, resulting in balanced training and test folds (**Procedure 4**). Relying on the results of such a procedure results in a severe misreporting of model performance, as all performance scores are higher than the measures of true error. Fifth, we combine down-sampling and parameter tuning in a single cross-validation and report the apparent error scores of the best model (**Procedure 5**). We set up this procedure for comparison with the application in Section 4. This, of course, is even more problematic. Reporting the results of Procedure 5 leads to substantial misreporting of predictive performance. However, using a procedure similar to Procedure 5 need not be problematic if one uses independent test data to estimate true error. In **Procedure 6** we apply the model from Procedure 5 to out-of-sample data and see that the performance measures are close to the true error.

To summarize, to obtain reliable estimates of the true error researchers can either rely on out-of-sample prediction (Procedure 1 and Procedure 6) or correctly apply cross-validation as in Procedures 2 and 3. Relying on the performance scores from Procedure 4 or Procedure 5 and reporting them as estimates of the true error is wrong and leads to substantial misreporting.

4 Problematic Cross-Validation in Muchlinski *et al.* (2016)

Finally, we show that this problem is not only of theoretical nature, but can also affect the inferences we draw from results in applied work. Muchlinski *et al.* (2016) provide an example of misreported performance. They claim that their random forest model offers an impressive predictive accuracy, even when being tested on independent out-of-sample data.

We find that the performance measures reported in Muchlinski *et al.* (2016) dramatically overestimate the actual performance of their model. Specifically, their analysis suffers from a problematic use of cross-validation. In their article, they report the apparent error scores of their best model (Procedure 5 above). Due to problems with the out-of-sample data from Muchlinski *et al.* (2016) we split the data set into two parts for our reanalysis. One training set with all observations from 1945 to 1989 and a test set with all observations from 1990 to 2000. Descriptive statistics of the training and test set can be found in the Online Appendix C.1.

The results of our reanalysis can be found in the lower panel of Figure 1. We follow the same structure as in the experiments and run the six procedures. For each of the procedures, we calculate the same performance measures as before (F_1[4], ROC-AUC[5], PR-AUC).

In Procedure 5, we run the model described by Muchlinski *et al.* (2016) where they combine cross-validation for model tuning, down-sampling, and then report the apparent error of the model on the training data. From our experiments we expect that reporting performance from such a procedure will lead to serious misreporting of the predictive performance. Indeed, wrongly reporting the values from procedure 5 like Muchlinski *et al.* (2016) would lead to a reported PR-AUC

4 Note that most of the F_1 scores we calculated are substantially smaller than the F_1 scores reported by Muchlinski *et al.* (2016, 97). Unfortunately we could not find code or data to replicate Figure 3 in the original paper. However, since Muchlinski *et al.* (2016, 96) note that "[a]ll logistic regression models fail to specify any civil war onset in the out-of-sample data," the F_1 scores should be close to 0.

5 Note that Muchlinski *et al.* (2016, 94) state that "ROC graphs are especially useful for applications where data are class imbalanced", while Cranmer and Desmarais (2017, 152) state and show the opposite.

value of 0.43, which drops to only 0.07 when the same model is used for out-of-sample prediction (Procedure 6).

All of this would not be problematic if out-of-sample testing was performed in the analysis by Muchlinski *et al.* (2016) to estimate the true error of the model (Procedure 6). While the authors claim to report the results of applying their model to an independent out-of-sample test set, we find no evidence that they did so. They present random numbers as predicted probabilities for civil war onset. In our reanalysis, we found that the data set used for the out-of-sample predictions contains fewer variables than initially used to train the model. With this data set it is, thus, not possible to obtain out-of-sample predictions. Our analysis of the replication code shows that they randomly draw 737 probabilities from the in-sample predictions and merge them to out-of-sample observations of civil war onset. The authors then compare those random probabilities with the true values of the out-of-sample data. The corresponding author was not able to provide additional data or code to clear this up. For a detailed explanation of our reanalysis and the original code, see our Online Appendix C.3.

In short, in our reanalysis we find no evidence for the impressive predictive performance of random forest as reported in Muchlinski *et al.* (2016). Given their misunderstanding of cross-validation and based on a wrong out-of-sample prediction it is neither correct to conclude that "Random Forests correctly predicts nine of twenty civil war onsets in this out-of-sample data" (Muchlinski *et al.* 2016, 96) nor 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" (Muchlinski *et al.* 2016, 101).

5 Discussion

We show that the term cross-validation has different meanings in applied political science work. We focus on cross-validation in the context of predictive models and experimentally show that misunderstanding cross-validation can have severe consequences on the results of applied work. Particularly, problematic cross-validation undermines the main conclusions drawn by the authors of a recent article by Muchlinski *et al.* (2016). In our reanalysis we show that this approach offers no substantial improvement in predicting civil wars. We encourage researchers in predictive modeling to be especially mindful when applying cross-validation.

Finally, we want to stress that by just reading the paper by Muchlinski *et al.* (2016) it is really hard to identify the problems. It was only when we read the paper and the replication code side by side that the problems with the analysis and results became apparent. With that in mind we asked ourselves: How can reviewers assess the quality of the results without access to (some form) of the replication code? Answering this question will become more important as new "machine learning" methods are more and more part of research projects in political methodology.

Supplementary material

For supplementary material accompanying this paper, please visit
https://doi.org/10.1017/pan.2018.39.

References

Cantú, F. 2014. Identifying irregularities in Mexican local elections. *American Journal of Political Science* 58(4):936–951.

Caughey, D., and C. Warshaw. 2015. Dynamic estimation of latent opinion using a hierarchical group-level IRT model. *Political Analysis* 23(2):197–211.

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

Cederman, L.-E., and N. B. Weidmann. 2017. Predicting armed conflict: Time to adjust our expectations? *Science* 355(6324):474–476.

Colaresi, M., and Z. Mahmood. 2017. Do the robot: Lessons from machine learning to improve conflict forecasting. *Journal of Peace Research* 54(2):193–214.
Cranmer, S. J., and B. A. Desmarais. 2017. What can we Learn from Predictive Modeling? *Political Analysis* 25(2):145–166.
Efron, B., and T. Hastie. 2016. *Computer Age Statistical Inference,* vol. 5. Cambridge, England: Cambridge University Press.
Engstrom, E. J. 2012. The rise and decline of turnout in congressional elections: electoral institutions, competition, and strategic mobilization. *American Journal of Political Science* 56(2):373–386.
Hainmueller, J., and C. Hazlett. 2014. Kernel regularized least squares: reducing misspecification bias with a flexible and interpretable machine learning approach. *Political Analysis* 22(2):143–168.
Hastie, T., R. Tibshirani, and J. Friedman. 2011. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. 2nd edn. New York: Springer.
Muchlinski, D., D. Siroky, J. He, and M. Kocher. 2016. Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis* 24(1):87–103.
Neunhoeffer, M., and S. Sternberg. 2018 Replication data for: How cross-validation can go wrong and what to do about it. doi:10.7910/DVN/Y9KMJW, Harvard Dataverse, V1.