PA

Insufficiencies in Data Material: A Replication Analysis of Muchlinski, Siroky, He, and Kocher (2016)

Simon Heuberger

Doctoral Candidate, Department of Government, American University, Washington, DC, 20016, USA. Email: sh6943a@american.edu

Keywords: imputation methods, classification, regression, RandomForest, model selection, replication

1 Introduction

This is the 2018 *Political Analysis* in-house replication of Muchlinski, Siroky, He, and Kocher (2016), henceforth MSHK. This work, "Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data," was published in *Political Analysis* in Volume 24, Issue 1 in 2016.[1] It was accompanied by `Dataverse` replication material as required by the journal.[2] While this material was checked upon submission in 2015, recent replication efforts show that it does not support the claims made by MSHK.

Shown here specifically is that MSHK conducted in-sample predictions instead of out-sample predictions in their use of `RandomForest` as stated in the paper. `RandomForest` is a machine learning algorithm that constructs multiple decision trees to obtain more accurate predictions. The higher the number of trees in the forest, the higher the prediction accuracy. A `RandomForest` model needs to be fitted, or trained, on a data sample. This model can then be used to forecast, or predict, observations. If this prediction is made for an observation that is part of the training sample, it is an in-sample prediction. If this prediction is made for an observation that is external to the training sample, it is an out-sample prediction. By definition, predicting observations from the fitting sample based on a model derived from the same sample, i.e., in-sample predictions, will provide highly accurate results: We are predicting within the same sample that we trained on. To assess the viability of a `RandomForest` model, it is necessary to predict observations that were not used for the model fitting, i.e., to conduct out-sample predictions.

I am the current replicator for *Political Analysis*. I have been in this position since August 2017. I was not involved in the original assessment of MSHK's submitted replication material. I walk through MSHK's 2016 `R` code step by step. I start with the loaded source files, move on to model building, and finally address the out-sample analysis and insufficient output for Table 1. All `R` code, including comments and typos, is copied verbatim from material provided by MSHK. Some code lines in this replication analysis have been omitted for space reasons while others have been rearranged to fit page margins. These alignments do not affect the substantive content of the analysis.

2 Loaded Files

MSHK load three imputed `.csv` files: `SambnisImp.csv`, `Amelia.Imp3.csv`, and `AfricaImp.csv`. The first two are loaded as pre-imputed source files. The latter is imputed by MSHK in a separate `R` script. `SambnisImp.csv` is loaded into the R object `data`, which is further subset into `data.full`.

Political Analysis (2019)
vol. 27:114–118
DOI: 10.1017/pan.2018.43

Corresponding author
Simon Heuberger

Edited by
Jeff Gill

© The Author(s) 2018. Published by Cambridge University Press on behalf of the Society for Political Methodology.

https://doi.org/10.1017/pan.2018.43 Published online by Cambridge University Press

1 Publicly available here: https://doi.org/10.1093/pan/mpv024.
2 Publicly available here: http://dx.doi.org/10.7910/DVN/KRKWK8.

Amelia.Imp3.csv is loaded into data2, myvars, and newdata. AfricaImp.csv is loaded into data3. These R objects are confusingly named, which makes the replication of the material more complex than it needs to be. While renaming them to separate them more clearly from each other would solve this, I have retained MSHK's original names here in the interest of transparency. The R code to load the .csv files into the respective R objects is shown below.

```
data = read.csv(file="SambnisImp.csv") # data for prediction
data.full<-data[,c("warstds", "ager", "agexp", "anoc", "army85", "autch98",
       "auto4", "autonomy", "avgnabo", "centpol3", "coldwar", "decade1",
       "decade2", "decade3", "decade4", "dem", "dem4", "demch98", "dlang",
       "drel", "durable", "ef", "ef2", "ehet", "elfo", "elfo2", "etdo4590",
       "expgdp", "exrec", "fedpol3", "fuelexp", "gdpgrowth", "geo1", "geo2",
       "geo34", "geo57", "geo69", "geo8", "illiteracy", "incumb", "infant",
       "inst", "inst3", "life", "lmtnest", "ln_gdpen", "lpopns", "major",
       "manuexp", "milper", "mirps0", "mirps1", "mirps2", "mirps3", "nat_war",
       "ncontig", "nmgdp", "nmdp4_alt", "numlang", "nwstate", "oil", "p4mchg",
       "parcomp", "parreg", "part", "partfree", "plural", "plurrel", "pol4",
       "pol4m", "pol4sq", "polch98", "polcomp", "popdense", "presi", "pri",
       "proxregc", "ptime", "reg", "regd4_alt", "relfrac", "seceduc",
       "second", "semipol3", "sip2", "sxpnew", "sxpsq", "tnatwar", "trade",
       "warhist", "xconst")]
data2<-read.csv(file="Amelia.Imp3.csv") # data for causal machanisms
myvars<-names(data2) %in% c("X", "country", "year", "atwards")
newdata<-data2[!myvars]
data3<-read.csv(file="AfricaImp.csv") # Reading in the Africa Data from 2001-2014
```

3 Model Building

MSHK build four models. Three of these stem from previous studies: Fearon and Laitin (2003), Collier and Hoeffler (2004), and Hegre and Sambanis (2006). For each of these three studies' models, they implement uncorrected and penalized logistic regression models. The fourth model is MSHK's implementation of RandomForest. MSHK use these three external studies to showcase the superiority of RandomForest in predicting class-imbalanced civil war onset data. All models are trained on the R object data.full, which is a subset of SambnisImp.csv.

```
tc<-trainControl(method="cv",
                 number=10,#creates CV folds - 10 for this data
                 summaryFunction=twoClassSummary,
                 # provides ROC summary stats in call to model
                 classProb=T)
# Fearon and Laitin Model Specification###
model.fl.1<-train(as.factor(warstds)~warhist+ln_gdpen+lpopns+lmtnest+ncontig
                  +oil+nwstate +inst3+pol4+ef+relfrac, #FL 2003 model spec
          metric="ROC", method="glm", family="binomial", #uncorrected logistic model
             trControl=tc, data=data.full)
###Now doing Fearon and Laitin (2003) penalized logistic regression
model.fl.2<-train(as.factor(warstds)~warhist+ln_gdpen+lpopns+lmtnest+ncontig
                  +oil+nwstate+inst3+pol4+ef+relfrac, #FL 2003 model spec
             metric="ROC", method="plr", # Firth's penalized logistic regression
             trControl=tc, data=data.full)
###Now doing Collier and Hoeffler (2004) uncorrected logistic specification###
model.ch.1<-train(as.factor(warstds)~sxpnew+sxpsq+ln_gdpen+gdpgrowth+warhist
                  +lmtnest+ef+popdense
                  +lpopns+coldwar+seceduc+ptime, #CH 2004 model spec
                  metric="ROC", method="glm", family="binomial",
                  trControl=tc, data=data.full)
###Now Collier and Hoeffler with penalized logistic regression###
model.ch.2<-train(as.factor(warstds)~sxpnew+sxpsq+ln_gdpen+gdpgrowth+warhist
                  +lmtnest+ef+popdense
                  +lpopns+coldwar+seceduc+ptime, #CH 2004 model spec
                  metric="ROC", method="plr", #penalized logistic regression
                  trControl=tc, data=data.full)
```


https://doi.org/10.1017/pan.2018.43 Published online by Cambridge University Press

```
###Now the Hegre and Sambanis Model Specification###
model.hs.1<-train(warstds~lpopns+ln_gdpen+inst3+parreg+geo34+proxregc+gdpgrowth
                  +anoc+partfree+nat_war+lmtnest+decade1+pol4sq+nwstate
                  +regd4_alt+etdo4590+milper+
                    geo1+tnatwar+presi,
                  metric="ROC", method="glm", family="binomial",
                  trControl=tc, data=data.full)
model.hs.2<-train(warstds~lpopns+ln_gdpen+inst3+parreg+geo34+proxregc+gdpgrowth
                  +anoc+partfree+nat_war+lmtnest+decade1+pol4sq+nwstate
                  +regd4_alt+etdo4590+milper+
                    geo1+tnatwar+presi,
                  metric="ROC", method="plr", #penalized logit
                  trControl=tc, data=data.full)
###Implementing RF (with CV) on entirety of data###
model.rf<-train(as.factor(warstds)~.,
                  metric="ROC", method="rf",
                  sampsize=c(30,90), #Downsampling the class-imbalanced DV
                    importance=T, # Variable importance measures retained
                   proximity=F, ntree=1000, # number of trees grown
                   trControl=tc, data=data.full)
```

4 Out-Sample Analysis

After training the models, MSHK create three logit models for the external studies by Fearon and Laitin (2003), Collier and Hoeffler (2004), and Hegre and Sambanis (2006) as well as one model with `RandomForest`. All models load the `R` object `data.full` and are thus based on the imputed source file `SambnisImp.csv`.

```
model.fl.africa<-glm(as.factor(warstds)~warhist+ln_gdpen+lpopns+lmtnest
                     +ncontig+oil+nwstate +inst3+pol4+ef+relfrac,
                     family="binomial", data=data.full)
model.ch.africa<-glm(as.factor(warstds)~sxpnew+sxpsq+ln_gdpen+gdpgrowth
                     +warhist+lmtnest+ef+popdense+lpopns+coldwar
                     +seceduc+ptime, family="binomial", data=data.full)
model.hs.africa<-glm(warstds~lpopns+ln_gdpen+inst3+parreg+geo34+proxregc
                     +gdpgrowth+anoc+partfree+nat_war+lmtnest+decade1
                     +pol4sq+nwstate+regd4_alt+etdo4590+milper+
                geo1+tnatwar+presi,, family="binomial", data=data.full)
RF.out<-randomForest(as.factor(warstds)~., sampsize=c(30, 90),importance=T,
       proximity=F, ntree=1000, confusion=T, err.rate=T, data=data.full)
```

Based on these models, MSHK make one prediction per model, turn the predictions into data frames, and subsequently set the seed to draw 737 random units from each predicted data frame. Each separate set of randomly drawn units is saved as a predictor object: `predictors.rf` for `RandomForest`, `predictors.fl` for Fearon and Laitin, `predictors.ch` for Collier and Hoeffler, and `predictors.hs` for Hegre and Sambanis.

```
yhat.rf<-predict(RF.out, type="prob") #taken from RF on whole data
###We used original CW data for training data here for all models/algorithms###
Yhat.rf<-as.data.frame(yhat.rf[,2])
yhat.fl.africa<-predict(model.fl.africa, type="response")
Yhat.fl.africa<-as.data.frame(yhat.fl.africa)
yhat.ch.africa<-predict(model.ch.africa, type="response")
Yhat.ch.africa<-as.data.frame(yhat.ch.africa)
yhat.hs.africa<-predict(model.hs.africa, type="response")
Yhat.hs.africa<-as.data.frame(yhat.hs.africa)
###Selecting random samples to make pred and actual lengths equal###
set.seed(100)
```


https://doi.org/10.1017/pan.2018.43 Published online by Cambridge University Press

```
predictors.rf<-Yhat.rf[sample(nrow(Yhat.rf), 737),]
predictors.fl<-Yhat.fl.africa[sample(nrow(Yhat.fl.africa), 737),]
predictors.ch<-Yhat.ch.africa[sample(nrow(Yhat.ch.africa), 737),]
predictors.hs<-Yhat.hs.africa[sample(nrow(Yhat.hs.africa), 737),]
```

MSHK then create confusion matrices with the predictor objects (based on the imputed source file `SambnisImp.csv`) and the variable `warstds`, which is a column from the data set `data3`, which in turn is based on the imputed source file `AfricaImp.csv`. Subsequently, MSHK load the `R` package `ROCR`. As per the `ROCR` package documentation, the function `prediction()` transforms the input data into a standardized format, while the function `performance()` calculates the area under the ROC curve if set to the parameter "`auc`", as MSHK do in the code shown below.

```
confusion.matrix(data3$warstds, predictors.rf, threshold=.5)
confusion.matrix(data3$warstds, predictors.fl, threshold=.5)
confusion.matrix(data3$warstds, predictors.ch, threshold=.5)
confusion.matrix(data3$warstds, predictors.hs, threshold=.5)
###ROC and AUC scores for out of sample data###
library(ROCR)
pred.fl.africa <- prediction(predictors.fl, data3$warstds)
pred.ch.africa<-prediction(predictors.ch, data3$warstds)
pred.hs.africa<-prediction(predictors.hs, data3$warstds)
pred.rf.africa<-prediction(predictors.rf, data3$warstds)
auc.fl.africa<-performance(pred.fl.africa, "auc")
auc.ch.africa<-performance(pred.ch.africa, "auc")
auc.hs.africa<-performance(pred.hs.africa, "auc")
auc.rf.africa<-performance(pred.rf.africa, "auc")
```

To sum up: For their out-sample analysis, MSHK create models based on `SambnisImp.csv`, make predictions based on `SambnisImp.csv`, draw random samples based on `SambnisImp.csv`, and calculate AUC scores based on `SambnisImp.csv` and one external variable based on `AfricaImp.csv`. In other words: MSHK conduct in-sample predictions, take random samples of these in-sample predicted probabilities, and compare those probabilities with true values from out-sample data. MSHK thus use the same sample to fit the model and conduct the predictions. This is not an out-sample prediction.

5 Output for Main Evidence

In their paper, MSHK provide Table 1 as the main evidence for their claim of the superiority of `RandomForest`. This table lists the predicted probabilities for civil war onset for 19 African countries and showcases the superiority of `RandomForest` over logit models in terms of prediction accuracy. MSHK provide `CompareCW_dat.csv` and identify it as the output that forms Table 1.

```
###csv file for Table 1###
d<-data.frame(data3$warstds, predictors.fl, predictors.ch, predictors.hs,
     predictors.rf)
write.csv(d, file="CompareCW_dat.csv")
```

As the `R` code shows, `CompareCW_dat.csv` consists of the random predictor objects (`predictors.rf`, `predictors.fl`, `predictors.ch`, and `predictors.hs`) and the variable `warstds`. The predictor objects are based on the imputed source file `SambnisImp.csv`, while `warstds` is based on the imputed source file `AfricaImp.csv`. If we now juxtapose `CompareCW_dat.csv` and Table 1, we can see that it is not possible to match the information provided in the output `.csv` with the information listed in Table 1, as Figure 1 shows.

https://doi.org/10.1017/pan.2018.43 Published online by Cambridge University Press

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

CompareCW_dat.csv

Filter

	data3.warstds	predictors.fl	predictors.ch	predictors.hs	predictors.rf
1	0	0.036354228	9.380534e-03	0.0209075119	0.29216867
2	0	0.009281192	3.714911e-03	0.0003546772	0.08926780
3	0	0.017708231	8.841568e-03	0.0204368389	0.53253253
4	0	0.034197142	3.136079e-03	0.0107442810	0.19759278
5	1	0.078753771	1.350169e-01	0.0597720247	0.06048387
6	0	0.025539793	1.478751e-02	0.0006688022	0.08625878
7	1	0.013429648	1.852500e-01	0.0073623881	0.71887550
8	0	0.027755230	5.513324e-03	0.0080306531	0.56493506
9	1	0.012395045	5.857035e-03	0.0311944315	0.63453815
10	0	0.006440910	1.213052e-03	0.0475583848	0.67439516
11	1	0.007812825	1.367884e-02	0.0077101553	0.51351351
12	0	0.002278077	5.956352e-03	0.0085521977	0.19719720
13	0	0.011487555	6.777899e-02	0.0045929770	0.84100000
14	0	0.021308705	2.346514e-03	0.0328846433	0.74274274
15	0	0.003904838	9.912065e-03	0.0326538764	0.27437186
16	0	0.008834318	3.621130e-02	0.0023257426	0.10531595
17	0	0.009155091	4.078427e-07	0.0013544994	0.31795386
18	0	0.005292905	6.455024e-02	0.0083213129	0.46339017
19	0	0.011793761	1.542060e-03	0.0121889713	0.27227227
20	0	0.004634562	9.332847e-03	0.0020158494	0.44567404
21	0	0.043097064	6.491715e-03	0.1029864243	0.03006012
22	1	0.005371223	2.583362e-03	0.0351988318	0.71385542
23	0	0.005423841	8.191679e-03	0.0511781647	0.18355065
24	0	0.009828007	1.048320e-02	0.0019674098	0.51055276
25	0	0.009263872	3.037736e-02	0.0149661627	0.06325301
26	0	0.008514706	3.958729e-02	0.0047936840	0.65361446

Figure 1. Actual Table 1 in Paper (left); View of Provided `.csv` in `R` (right).

`CompareCW_dat.csv` and Table 1 should show identical content. This is not the case. Table 1 consists of 19 rows, while `CompareCW_dat.csv` has 737. `CompareCW_dat.csv` does not have any identifiers that make the transition from this source file to the eventual Table 1 apparent and transparent. We do not know which predictor numbers correspond to which countries, as there is no information about the countries in the `.csv` file. Even if we assume that all instances where `warstds == 1` sum up to the number of countries shown in Table 1, the numbers do not add up: There are 21 such instances in the `.csv`, but 19 in Table 1.

6 Conclusion

MSHK create several models (logit, `RandomForest`), make predictions based on these models, and draw random samples from these predictions. The data used for all of this comes from `SambnisImp.csv`. MSHK then create confusion matrices and calculate AUC scores based on data from `SambnisImp.csv` and one external variable from `AfricaImp.csv`. Rephrased in more generic terms, MSHK conduct in-sample predictions and take an in-sample sample, and then compare this in-sample sample with true values from out-sample data. Out-sample data only enters the equation in the final comparison, after the predictions have already been made with in-sample data.

In addition, the provided `CompareCW_dat.csv` cannot be compared to Table 1 because of its lack of identifiers. It is not possible to examine and verify the origin of the numbers in Table 1, which functions as the main piece of evidence in the paper.

https://doi.org/10.1017/pan.2018.43 Published online by Cambridge University Press