Gov 2018: Lab 5 Random Forests

Your name:

February 22, 2022

This exercise is based off of Muchlinski, David, David Siroky, Jingrui He, and Matthew Kocher. 2016. "Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data". *Political Analysis*.

Descriptions of the relevant variables in the data file `data_full.rds` are:

Name	Description
`warstds`	Factor, `peace` and `war`
`year`	Numeric for year of obs

← labels (y_i)

predictors (X_i)

And a list of 90 covariates from the Sambanis dataset: "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".

We will compare ① (classic) logistic regression, ② penalized logistic regression and ③ Random Forests

Advance Access publication October 22, 2015

Political Analysis (2016) 24:87–1
doi:10.1093/pan/mpv0

Original paper :

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

rare events

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

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

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

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

Edited by R. Michael Alvarez

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

Before we start,

We're going to use the cross-validation function from the **caret** package. Set aside the years 1999 and 2000 for testing data.

```r
# caret::trainControl() controls parameters for train
tc<-caret::trainControl(method="cv", # the resampling method
                        number=10, # the number of folds
                        summaryFunction=twoClassSummary, # a function to compute performance metrics ac
                                                         # twoClassSummary computes sensitivity, specif
                        classProb=T, # class probabilities be computed for classification models
                                     # (along with predicted values) in each resample
                        savePredictions = T)

# Set train data
data.train<-subset(data.full,year<1999)
```

need this for tuning parameters
(in "plr" & "rf")

Overview

what we want to predict

year	y_i	x_{i1}	$\cdots$	x_{i90}
1945	peace	$\vdots$	$\vdots$	$\vdots$
1945	war			
1946	peace			
1946	peace			
$\vdots$	$\vdots$			
1999	peace	$\vdots$	$\vdots$	$\vdots$
1999	war			
2000	peace			

train set
- for fitting the model (Q1)
- in-sample evaluation (Q2)

test set
- for out-of-sample evaluation (Q3)

Question 1 (Fit the models 1 - 7)

We're going to compare several model specifications using classic/penalized logistic regressions with a random forest model.

for rare events

(a) FL model specification

as.factor(warstds) ~ warhist + ln_gdpen + lpopns + lmtnest + ncontig + oil + nwstate + inst3 + pol4 + ef + relfrac

1 logistic regression: `method="glm", family="binomial"`

2 penalized logistic regression: `method="plr"`

(b) CH

as.factor(warstds) ~ sxpnew + sxpsq + ln_gdpen + gdpgrowth + warhist + lmtnest + ef + popdense + lpopns + coldwar + seceduc + ptime

3 l.r

4 p.l.r

(c) HS

as.factor(warstds) ~ lpopns + ln_gdpen + inst3 + parreg + geo34 + proxregc + gdpgrowth + anoc + partfree + nat_war + lmtnest + decade1 + pol4sq + nwstate + regd4_alt + etdo4590 + milper + geo1 + tnatwar + presi

5 l.r

6 p.l.r

(d) 7 Random Forest (RF) (※ Star w/ ntree =10 ⇝ After you're done, change it to ntree = 1000)

E.g. model 1

```
### (a)
# Fearon and Laitin (2003) + classic logistic regression
model.fl.1<- caret::train(as.factor(warstds)~warhist+ln_gdpen+lpopns+lmtnest+ncontig+oil+nwstate+inst3+ ...
                          metric="ROC", method="glm", family="binomial", trControl=tc, data=data.train)
summary(model.fl.1)
```

In-sample eval.

Question 2

(2 Figures < ROC of model 1, 3, 5, 7
" " " 2, 4, 6, 7)

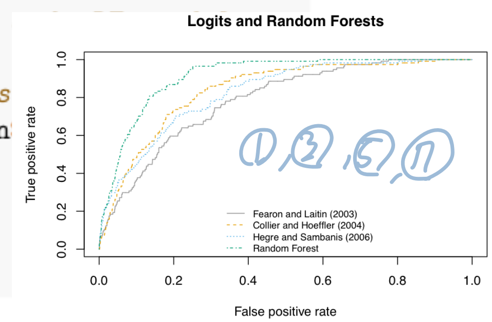
We will now create ROC plots for different models:

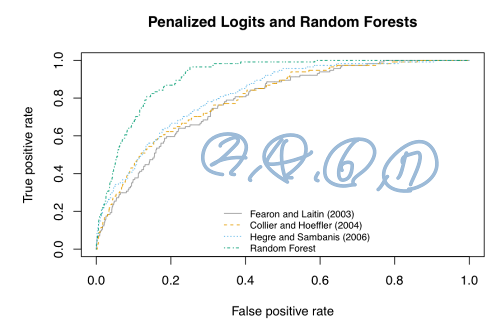
- Collect the predicted probabilities for the outcome from each of the above models (note these should be for the highest AUC score in the caret CV procedure, `your-logit-model$finalModel$fitted.values`). For the random forests model, requires a call from `predict()` with type set to "prob" (i.e., `predict(your-rf-model$finalModel, type="prob")`).
- Follow the sample code below to create a `prediction` object from which to calculate the performance of the classifier in terms of true positive and false positive rates.
- Plot the ROC curves of all the unpenalized models (= classic logistic regression) and the RF model. 1, 3, 5, 7
- Then separate, plot the ROC curves of all penalized models and the RF model. How does the RF model compare? 2, 4, 6, 7

Sample code:

```
## ROC plot: Example with a *classic logistic regression* model trained with caret CV procedure
library(ROCR) # We will use prediction() & performance() functions from this package

## 1. Collect the predicted probabilities
pred.(your-mod-name).war <- (your-logit-model)$finalModel$fitted.values
# The above line should be changed for *penalized logistic regression*

## 2. Using in-sample prediction, calculate true positive and false pos
pred.(your-mod-name) <- prediction(pred.(your-mod-name).war, data.train
perf.(your-mod-name) <- performance(pred.(your-mod-name),"tpr","fpr")

## 3. Plot the ROC curves
plot(perf.(your-mod-name), ...)
```

For each 1-7



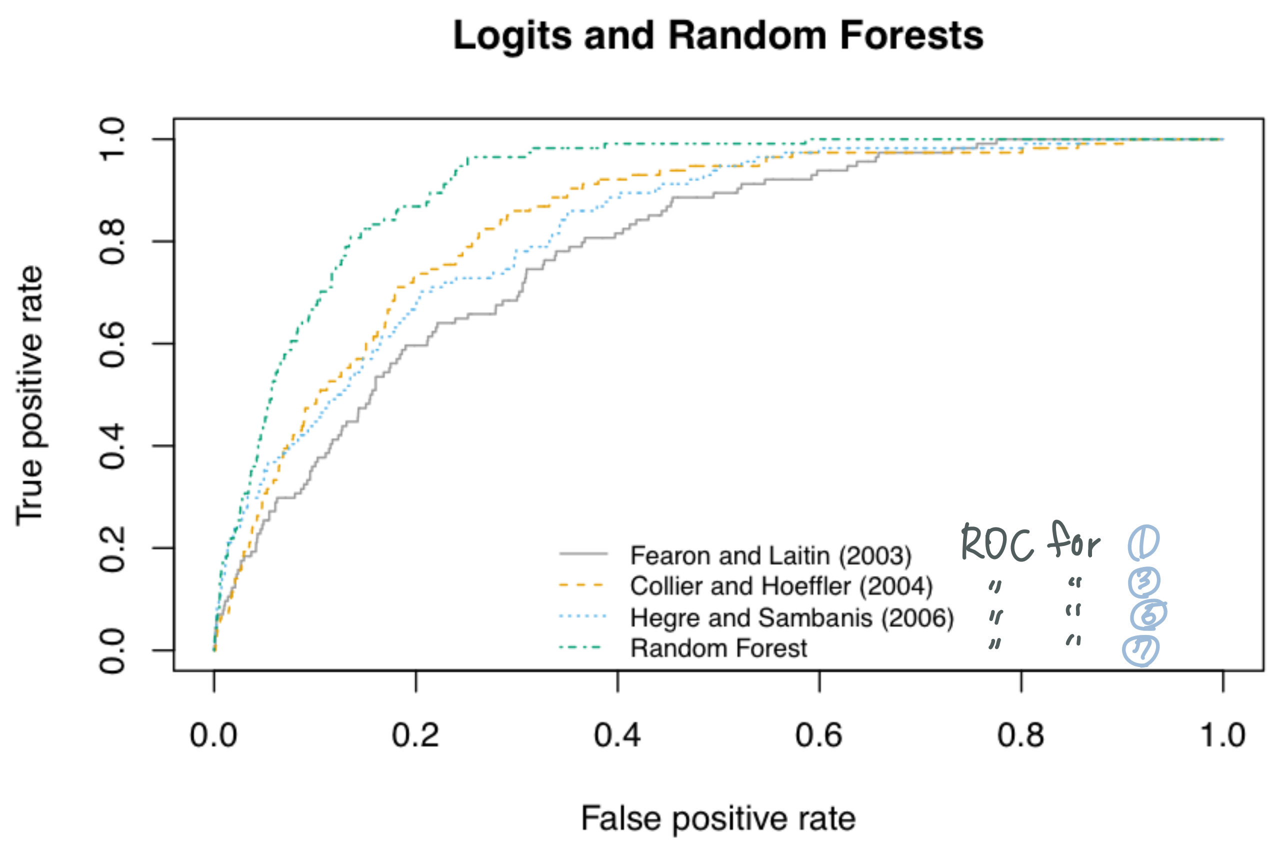


TODO 1 : Do #1 & #2 for model 1 - 7

2 : Do #3 with 1, 3, 5, 7

3 : " " " 2, 4, 6, 7

Example:


Logits and Random Forests
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
Fearon and Laitin (2003)
Collier and Hoeffler (2004)
Hegre and Sambanis (2006)
Random Forest
ROC for 1
" " 3
" " 6
" " 7

Review of ROC (visualization of how a single curve is generated)

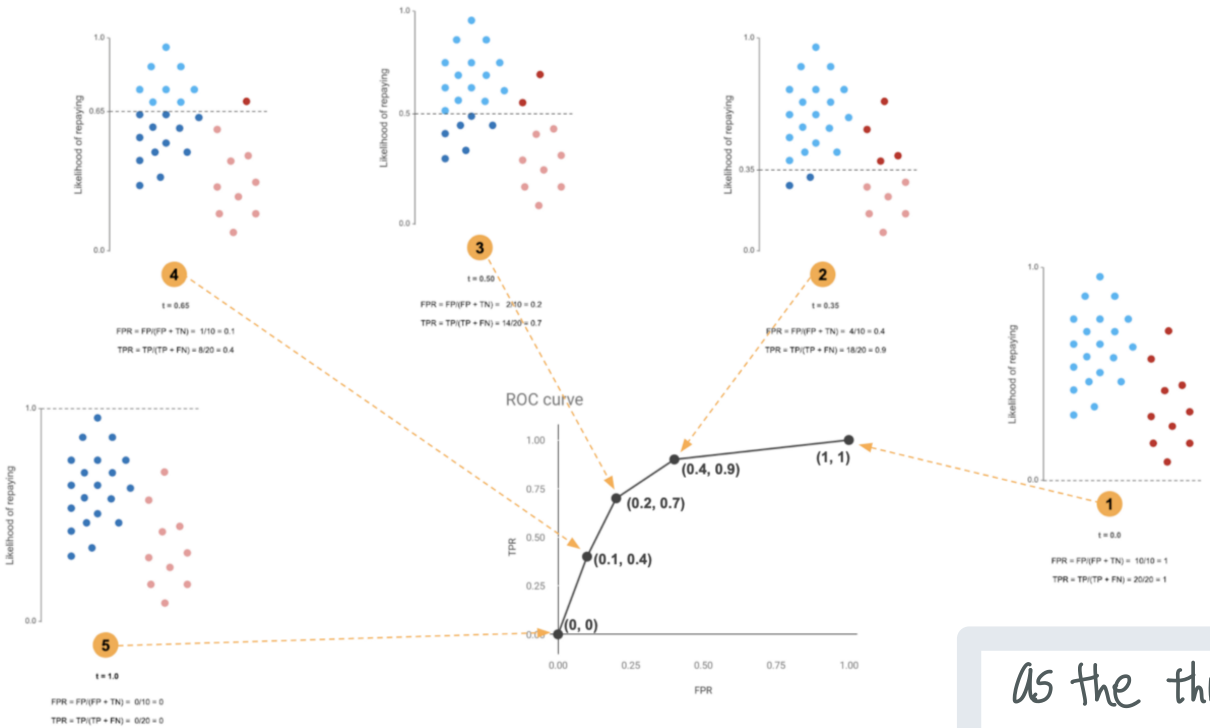


Image source:
https://towardsdatascience.com/understanding-the-roc-curve-in-three-visual-steps-795b1399481c

as the threshold Δ

$\rightsquigarrow$ contingency table Δ

	$y=1$	$y=0$
$\hat{y}=1$		
$\hat{y}=0$		

$\rightsquigarrow (FPR, TPR)$ Δ

connect the dots $\rightsquigarrow$ ROC!

Out-of-sample eval. (use test set!)

Question 3

Finally, we will evaluate out of sample prediction of unpenalized models and the RF model.

Pull out the testing data (1999, 2000) and evaluate each of the model predictions on the testing data. Focus on true positives, or when warstds is (correctly) classified with high probability as a "war".

```
mena<-subset(data.full, data.full$year >= 1999)
table(mena$warstds) # two war cases
```

```
##
## peace   war
##   334     2
```

TODO 1:

```
### Generate out of sample predictions for Table 1
fl.pred<-predict(model.fl.1, newdata=mena, type="prob")
fl.pred<-as.data.frame(fl.pred)
ch.pred<-predict(model.ch.1, newdata=mena, type="prob")
ch.pred<-as.data.frame(ch.pred)
hs.pred<-predict(model.hs.1, newdata=mena, type="prob")
hs.pred<-as.data.frame(hs.pred)
rf.pred<-predict(model.rf, newdata=mena, type="prob")
rf.pred<-as.data.frame(rf.pred)
```

2: create table

```
### Rows 1-5 of the above ###
head(Onset_table, n=5)
```

```
##      year CW_Onset Fearon and Latin (2003) Collier and Hoeffler (2004)
## 138 1999      war              0.06759787                0.0164709598
## 308 1999      war              0.01634902                0.0078920572
## 1   1999    peace              0.01984194                0.0377360013
## 3   1999    peace              0.01989782                0.0009862537
## 5   1999    peace              0.01138666                0.0102527626
##      Hegre and Sambanis (2006) Random Forest
## 138                 0.11004861         0.917
## 308                 0.01099924         0.817
## 1                   0.01599261         0.892
## 3                   0.02659714         0.062
## 5                   0.02027545         0.872
```

y=1

$\hat{y}_{logit}=0$

$\hat{y}_{rf}=1$