CODERS' CORNER

csae
CENTRE FOR THE STUDY OF
AFRICAN ECONOMIES

RANDOM FORESTS IN STATA

For anybody keen to try some out-of-sample prediction models in Stata, there is a great package for running a random forest algorithm in Stata called rforest.

Random forest models are decision tree based, so are great for contexts where you don't think your data can be modelled as having a linear relationship.

While decision trees are unlikely to perform well in terms of out of sample prediction because they will over-fit the data, random forest models average predictions over many individual trees and use bootstrap aggregating to reduce overfitting, yielding better prediction accuracy.

The Stata syntax for rforest is:

```
rforest depvar indepvars [if] [in] , [ options ]

predict newvar | varlist | stub* [if] [in] , [ pr ]
```

It can be used for either classification (categorical outcome variables) or regression (continuous outcome variables). You first need to split your sample into a training and a testing sample to use this comment. If you want to use a training dataset with a binary dependent variable to classify the observations in a new dataset into the same categories, you can use the option `type(class)` at the end.

Say for example you were trying to predict whether a village would experience flooding in a given year and had a set of appropriate explanatory variables. You could use the code:

```
rforest flood indepvars if dataset==training, type(class)

predict flood if dataset==testing
```

If instead you are trying to predict a continuous variable, use the option `type(regression)`.

You can set the number of trees it will iterate over using the option iterations(x). If you don't specify, the default will be 100. Sometimes it will be a trade-off between prediction accuracy and model running time – if you have a very large sample, rforest can take a while to run. You can also adjust the minimum leaf size, the depth of the tree and the number of variables to randomly investigate.

Katherine Stapleton, DPhil Candidate in Economics, Lincoln College, Oxford
26 November 2019