Random Forest vs Extramarital Affairs

01 DATA PREP

Isolate X and Y is a standard process for functions from the sklearn library.

02 VARIABLE TRANSFORMATION

The Y variable is currently not in binary format. Transform it. If someone has cheated, it is 1, otherwise 0.

03 DUMMY VARIABLE

The occupation variable, though numeric, should be a dummy variable since it is a categorical - nominal variable, without a clear order in the values.

04 STRING VARIABLES

All non-numeric variables must be transformed into numeric. The simplest way is usually to transform into dummy variables.

05 RANDOM FOREST

We should have everything set. Then, create a random forest model. The number of trees should be low since there are not a lot of predictors or observations.

06 IMPORTANCE DRIVERS

Like most advanced Machine Learning models, Random Forest focuses on highly accurate predictions. Our main goal is to study the drivers. Create an importance driver graph and visualize what matters most regarding cheating.

Diogo Resende