Assignment 3: Due by 5:00pm Wednesday, July 17

Machine Learning: Applications in Social Science Research

July 11, 2019

1. Load the `CCES2016_abbreviated.dta` dataset in `R` (this is available in the /assignments/ directory of the class Dropbox folder). The dataset includes the following variables:

   (a) `voted2016` a binary (0,1) validated measure indicating whether the respondent voted in the 2016 election.
   (b) `age` age in years.
   (c) `female` a binary indicator variable.
   (d) `education` a six-category (1-6) variable (No HS, HS graduate, some college, 2-year, 4-year, post-grad)
   (e) `black` a binary indicator variable.
   (f) `hispaniclatino` a binary indicator variable.
   (g) `voted2016primary` a binary indicator variable (did respondent vote/participate in a 2016 presidential primary or caucus?).
   (h) `ideology` an eight-category (1-8) variable indicating ideological self-placement (very liberal, liberal, somewhat liberal, middle of the road, somewhat conservative, conservative, very conservative, not sure).
   (i) `partyid` an eight-category (1-8) variable indicating party identification (strong Democrat, weak Democrat, lean Democrat, independent, lean Republican, weak Republican, strong Republican, not sure).
   (j) `income` a twelve-category (1-12) variable indicating household income (in approximately $10,000 increments).

2. Split the data into training and testing sets using the following code:

   ```
   trainIndex <- createDataPartition(dat$voted2016, p=0.2, list = FALSE, times = 1)
   ```

   and then predict `voted2016` as a function of all (or some) of the predictor variables using the following methods:

   (a) A single decision tree
   (b) Random forest
   (c) GBM

   Report the training/testing accuracy **and** the three most important features from each method.

3. Use a method of your choice to describe the estimated effects of age on voting turnout from the random forest.