Biostatistics 140.654
Fourth Term, 2022
Problem Set 2

Instructions: For Section I below, follow the general directions for the class that permits group collaboration on coding and discussion of results. For Section 2, you may work alone or in groups of up to 3 members. If you choose to submit the write-up as a group, each team member is required to disclose their contributions and all team members will receive the same grade.

Due date: Friday, April 29, 2021, 12:00pm EST

Case study in predicting a major smoking caused disease using CART, logistic regression, and random forests

Upon successful completion of this problem, a student should be able to:

- Use CART to identify important predictors and interactions to include in a statistical model for the probability of a major smoking caused disease as a function of smoking history, demographic and socio-economic variables
- Select and estimate a logistic regression models toward this goal
- Build random forest predictions with the same goal
- Check the predicted models for their consistency with the observed data.
- Display the results for a reader interested in smoking caused diseases, not statistics.
- For each model and algorithm, estimate the sensitivity and specificity of a classification model for a major smoking caused disease; calculate the cross-validated ROC curve and its area for each statistical method, i.e. CART, logistic regression model and random forest.

In this problem set you will be using three statistical approaches to predict the probability that a person 40 years of age or older has a major smoking caused disease using the NMES data set. Similar to the prediction models we considered in lectures and lab, candidate predictors are smoking history, age, sex, education, marital status, seatbelt use, and poverty status.

A bit more about the available smoking history variables is provided below. You should use your own judgement to decide which variables to include or not.

- You have available eversmoker (1 – eversmoker, 0 – neversmoker) in addition to current and former smoker indicators.

- In addition, the 1987 NMES survey collected information on ever smokers including: age when started to smoke (AGESMOKE), number of cigarettes smoked per day (CIGSADAY for current smokers and CIGSSMOK for former smokers), age when stopped smoking for former smokers (AGESTOP).

- In our work for the Department of Justice (DOJ) lawsuit against the tobacco industry, we generated additional variables from the data: years since quitting smoking among former smokers (YEARSINCE) and packyears (packs smoked per year of smoking) among ever smokers.

$$packyears = \frac{cigarettes\ per\ day}{20} \times years\ smoked$$

  In the work for the DOJ, we considered separate functions of packyears for current smokers and categories of former smokers (defined by the years since quitting smoking). You may want to explore these variables in your analysis.

Part I: Prediction models for having a major smoking caused disease

1. Explore the key variables of interest in the NMES data set. Note if there are missing values for any of the key variables that you plan to include in your prediction models. If there are missing values, use a random forest approach to impute the missing values.

   HINT: If you created a dataframe, *nmes*, with 6 variables: mscd, age, male, eversmk, packyears and marital status. Then you could impute missing values for any of the 6 variables using *rfImpute* as follows:

   imp.nmes = rfImpute(nmes[, 2 : 6], nmes[,1])

2. Create a training and validation sample for use in Questions 3, 4, 5 and 6 below.

   Our prediction problem is complicated by the fact that only about 10% of the individuals 40 years or older in the NMES have a MSCD. Several phenomenon may occur when constructing a CART or random forest for a low prevalence outcome absent strong predictors: you may obtain a CART with few to no identified predictors OR the individuals in the training and validation sample may be all predicted to be "0" (little to no ability of the model to predict the outcome "1" given the preponderance of 0s and weak predictors).

   Depending on which combination of variables you select to include as predictors, you may run into this problem.

   One way to resolve this issue is to "upweight" the cases so they contribute more information to the prediction model. This can be accomplished in several ways:

   a) Create a weight for each individual and incorporating those weights into the prediction models
   b) Creating a new sample where you over sample the cases

   Both of these approaches result in the same effect.

See the R code below that can be used to create an upweighted sample of the data such that the MSCD cases represent 25% of the data instead of only 10% and subsequently creates a 70-30 percent split for training and validation.

```
# Create indicators for the cases and controls
orig0 = which(dat$mscd == 0)
orig1 = which(dat$mscd == 1)
# Create an upweighted sample of cases
orig1up = sample(1:length(orig1),ceiling(length(orig1)*2.5),replace=TRUE)
# Create a new upweighted dataset
updat = rbind(dat[orig0,],dat[orig1up,])

# From the new upweighted dataset, create a training and validation sample
controls = which(updat$mscd == 0)
cases = which(updat$mscd == 1)
train0 = sample(1:length(controls), floor(length(controls)*0.7))
train1 <- sample(1:length(cases), floor(length(cases)*0.7))
# Name the training and validation samples
dat.train = rbind(updat[train0,],updat[train1,])
dat.test = rbind(updat[-train0,],updat[-train1,])
```

3. Use a CART to predict the probability that a person has a major smoking caused disease in the NMES data set. Be sure to prune your tree and create a graphical display of your final model.

4. Use the CART results and prior health services knowledge to propose and fit a logistic regression model to achieve the same prediction aim. Using the training sample, check the model for consistency with the observations by comparing the observed rates within several bins of predicted rates. Check for extremely influential observations in your final model.

5. Use a random forest to predict the probability that a person has a major smoking caused disease in the NMES data set. Select an appropriate number of trees to include in your forest and an appropriate number of variables to randomly sample as candidates at each split. Use the output to identify the variable importance. Using the training sample, check the model for consistency with the observations by comparing the observed rates within several bins of predicted rates.

6. For each of your prediction methods, use the validation sample to calculate the sensitivity and specificity for classifying a person as having a major smoking caused disease at a threshold of your choosing.

7. For each prediction method, calculate the cross-validated Receiver Operator Curve and its AUC. Compare the AUC across the prediction methods.

Part II: Summarize your findings in a short report

In two pages or less (I think you can do it in a page but giving you some wiggle room), summarize your findings about the prediction of a having a major smoking caused disease and compare the three methods. Be numerate and avoid unnecessary statistical jargon. You may include supplemental figures/tables as you see appropriate; these do NOT count towards the two-page limit for text. In the discussion section of your short report, be sure to include any limitations of your methods or additional analyses beyond what you completed that you think would be important.

Be sure to include the following sections:

- Objective
- Data
- Methods
- Results
- Discussion

If you are submitting as a team, then you should also include the following contribution section on a separate page at the end of your document:

- A list of names defining the team
- A statement about who contributed what to the abstract writing. For example: Initials1 drafted the Objective, Data and Discussion section. Initials2 drafted the figures/tables. All three authors contributed equally to the writing of the methods and results section. Initials3 edited the final abstract.