Biostatistics 140.654
Fourth Term, 2021
Problem Set 2

Instructions: follow the general directions for the class that permit group collaboration on coding and discussion of results but require each student to write-up his or her own solution.

Due date: Thursday, April 29, 2021, 5:00pm 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, appropriately handling missing data
- 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

In this problem set you will be using three statistical approaches to predict the probability that a person has a major smoking caused disease in the NMES data set. Key predictors are smoking history, age, sex, education, 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.

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

2. Partition the dataset into a training and validation sample for use in Parts 3, 4, 5 and 6 below.

3. Use a CART to predict 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 a logistic regression model to achieve the same prediction aim. Estimate its coefficients and 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.

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

7. Now calculate the cross-validated Receiver Operator Curve and its AUC for each method (i.e. CART, logistic regression model and random forest).

8. In a page or less, 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.