STAT 796: Homework 9

Due Friday, April 12 at 11:59pm on Canvas. Please append your code to your responses.

For this assignment we will use data from the Burn Study to predict whether patients hospitalized for burns will survive until discharge (variable `death`). These data are described in HL and are available on Canvas in the file `burn.csv`.

1. Split your data into training, validation, and test subsets with 500, 250, and 250 observations, respectively. (Nothing to report for this).
2. Compute descriptive statistics for the training data (and **only** the training data):
   a. What proportion of males did not survive through discharge? What about females?
   b. What proportion of those of white race did not survive through discharge? What about non-white individuals?
   c. What proportion of those with a flame injury did not survive through discharge? What about those without flame involved in their injury?
   d. What proportion of those with an inhalation injury did not survive through discharge? What about those without an inhalation injury?
   e. Explore the distributions of age and burn surface area (`tbsa`) and how they differ by `death`. Briefly ($\leq$ 3 sentences) describe any trends you identify.
3. Write out at least three candidate logistic regression models for outcome prediction. When identifying candidate models, consider the descriptive statistics from the previous question. (You may describe the model qualitatively by listing the variables and their form; a mathematical equation is not needed).
4. Fit each of the candidate models using the training data. For each model, report the AUC from the training model fit.
5. Qualitatively, do your models seem to be predicting well? If you want to change any of your models, do so now and re-do Question 3 and 4. Briefly describe how you changed your models and how your original choices differ from what you are turning in.

Note: From this point on, you may not change your candidate models.

6. Use the validation data to select among the candidate models. Report the AUC for each of the models evaluated on the validation data and identify which model performs the best.
7. Write out the equation ($\text{logit}(p) = \ldots.$) for the best model.
8. Use the test data to calculate the AUC for the best model. Report the AUC and provide an ROC curve.
9. For a cutoff value of $c = 0.5$ and one other value of your choice, report the following measures of classification accuracy using the test data: (a) sensitivity, (b) specificity, (c) positive predictive value, and (d) negative predictive value.
10. In the context of predicting `death`, which is a more appropriate measure of performance: sensitivity or positive predictive value? Explain your reasoning.