Stat 6021: Guided Question Set 12

For this question set, we will continue using the Western Collaborative Group Study (WCGS) data set, which is from a study regarding heart disease. Data were collected from 3154 males aged 39 to 59 in the San Francisco area in 1960. They all did not have coronary heart disease at the beginning of the study. The data set comes from the `faraway` package and is called `wcgs`. The variables of interest are:

- `chd`: whether the person developed coronary heart disease during annual follow ups in the study, with a '1' indicating the person developed coronary heart disease, and a '0' indicating the person did not develop coronary heart disease.
- `age`: age in years,
- `sdp`: systolic blood pressure in mm Hg,
- `dbp`: diastolic blood pressure in mm Hg,
- `cigs`: number of cigarettes smoked per day,
- `dibep`: behavior type, labeled A and B for aggressive and passive respectively.

From the previous guided question set, we went with a logistic regression model with `age`, `sdp`, `cigs`, and `dibep` as the predictors, dropping `dbp` from the model. We will now evaluate how our model performs in classifying the test data.

Recall that we split the data into a training set and a test set (50-50 split) using `set.seed(6021)`. Be sure to do this split, fit the logistic regression with the training data.

1. Based on the estimated coefficients of your logistic regression, briefly comment on the relationship between the predictors and the (log) odds of developing heart disease.
2. Validate your logistic regression model using an ROC curve. What does your ROC curve tell you?
3. Find the AUC associated with your ROC curve. What does your AUC tell you?
4. Create a confusion matrix using a cutoff of 0.5. Report the accuracy, true positive rate (TPR), and false positive rate (FPR) at this cutoff.

5. Based on the confusion matrix in part 4, a classmate says the logistic regression at this cutoff is as good as random guessing. Do you agree with your classmate's statement? Briefly explain.

6. Discuss if the threshold should be adjusted. Will it be better to raise or lower the threshold? Briefly explain.

7. Based on your answer in part 6, adjust the threshold accordingly, and create the corresponding confusion matrix. Report the accuracy, TPR, and FPR for this threshold.

8. Comment on the results from the confusions matrices in parts 4 and 7. What do you think is happening?