STAT 645: Biostatistics - Assignment 6
Due Thursday, October 15, 11:55pm CT

Please obtain the heart disease data (from Course Content > Data). This database from Cleveland clinic (through kaggle) contains 14 attributes. The "target" field refers to the presence of heart disease in the patient. It is an integer, 0 (absent) to 1 (presence). A good description of the attributes can be found here `https://lucdemortier.github.io/projects/3_mcnulty`.

1. There are four categorical variables, cp, restecg, slope and thal. Categorize thal into two groups, 0 (thal = 3) and 1(thal other than 3).
2. Scale all numeric variables. Do not scale the binary and categorical variables.
3. Fit a logistic regression model to target on 13 explanatory variables.
4. Use this fitted model to estimate the probability of the disease (target= 1) for the following set of values of the explanatory variables. For these cases, also obtain the 95% interval for the chance of the disease. Note that before the prediction, don't forget to apply the same transformation on the explanatory variables as you have done before the logistic model fitting to the data in the previous question.

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
68	1	3	145	233	1	0	150	0	2.3	0	0	3
75	0	3	145	150	1	0	150	0	2.3	0	0	1
78	1	0	144	193	1	1	90	0	3.4	1	2	3

5. Check the adequecy of the model using the Hosmer Lemeshow test. Clearly write out the hypothesis, test statistic and p-value, and conclusion.
6. Consider the first 100 and the last 100 subjects of the data and fit the logistic regression based on these data only. You don't need to re-scale the data again. Just take the above subset of the data that you have created previously.
7. Next, apply this fitted model to predict the target variable for the remaining set of observations (test data). Show the confuson matrix for prediction when you use 0.5, 0.6 and 0.7 as the cutoff value and use the cutoff to declare a target equal to one if the estimated probability exceeds the cutoff. Comment on the results.
8. Draw an ROC curve for the test data mentioned in the previous question and then comment on the discriminatry power of the model.
9. Re-do the analysis stated in questions 6 and 8 without ca, cp, and thal. Comment on the discriminatory power of this model?