Homework Assignment #9
SciKit-Learn
DATS 6103 - Spring 2020

How does things change with scikit learn compared to what we did last week?

Titanic data

Question 1
With the Titanic dataset, let us perform similar analysis as last week, with scikit-learn. First, prepare our dataset. Also make a train-test split in 4:1 ratio.

Question 2
Build a logistic regression model for survival with the train set, and score it with the test set. Also do that with CV. How do all these (including last week's result) compared? You can use the same selection of variables as last week to begin with, if you feel that's the best choices of independent variables.

Question 3
Try KNN in scikit learn to model the survival in titanic dataset. Try several k values, and choose the best one that you can find. Compare the scores with logistic regression.

Question 4
Try two other cutoff values at 0.3 and 0.7. You can use Chaelin's solution to change the cutoff (thanks Chaelin!), or there is a function called predictcutoff() in the attached file to help with that. How does the accuracy scores change with the different cutoffs?

Question 5
Also try to score and plot the KNN model and logit model with ROC-AUC.