Midterm – Sheena Sharma

1. Several passengers have NAN values for age. I began by looking to see if there were any other variables given that could predict age. I found that Sex, Fare, and Pclass didn't independently seem to relate to age. I plotted Sex vs Age, Fare vs Age, and Pclass vs Age. I ran a regression on Fare vs Age just to confirm what I was seeing – that there is no relation between fare and age. The r squared was very low, which confirmed that there is no relation. Next, I sought to determine if Sex, Fare, and Pclass collectively could predict age. I did this using a regression. Again the r squared was very low. These three variables do not predict age. In the end, I calculated the mean age for all of the passengers and then replaced the NAN values with this mean age (~29).

2. A logistic model was run, with survival as the target. I initially looked at the features Sex, Age, Fare, and Pclass and found that Pclass and Sex have the largest coefficients, and thus are most important for predicting survival. Being male has a largest negative coefficient. This is perhaps because women and children were evacuated first and so men had a lower survival rate that women. Being part of the first passenger class has a positive coefficient. This is perhaps because the first class cabin was evacuated first or maybe their cabin was closest to life boats.

3. Cross validation was implemented for the logistic regression. I chose 8 folds, because I plotted the accuracy of the prediction model against number of folds in a range of 3-10. Eight folds provided the highest accuracy.

4b. The AUC is .818.
4c. This model shows the relation between how many false positives and true positives. This is achieved by varying the discrimination threshold.
4d. An AUC of 1 is ideal and less than .5 is considered worse than guessing. So, to try and achieve an AUC of more than .818 we could try to vary the discrimination threshold.
4e. Its tricky to find the right threshold value as there are repercussions for the choice. We want to limit the number of false positives (x-axis) we have, because it means we are classifying people as having 'survived' that actually didn't and we might notify their families of this, which could obviously be devastating once they learn the truth. If we choose a lower false positive rate, to avoid this problem, we are also choosing a lower true positive rate which means we are not accurately predicting those who survived. This is also devastating because we want to accurately predict survival. With this in mind, I'd choose a threshold of about .25, because I feel that when the False Positives are at about .25, we have a pretty high true positive rate, of about .8 and still a fairly low false positive rate. better, having more false negatives and therefore I think a threshold of X is a good option.