Part 1:  Understanding the ROC curve.

Given below logistic regression probabilities and the ground-truth classifications. Plot the ROC curve via calculating the TruePositive rate vs. FalsePositive rate.

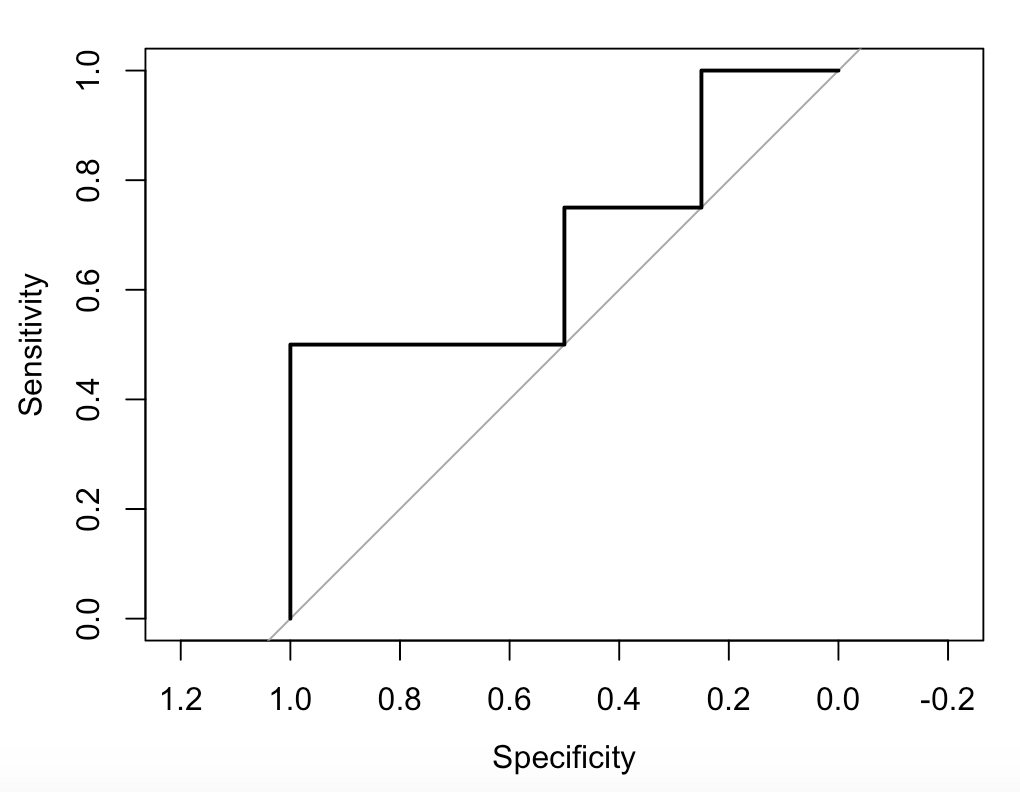
Probabilities = [ 0.967, 0.448, 0.568, 0.879, 0.015, 0.780, 0.978, 0.004]

Classifications = [1, 0, 1, 0, 1, 0, 1, 0]

Solution 1：

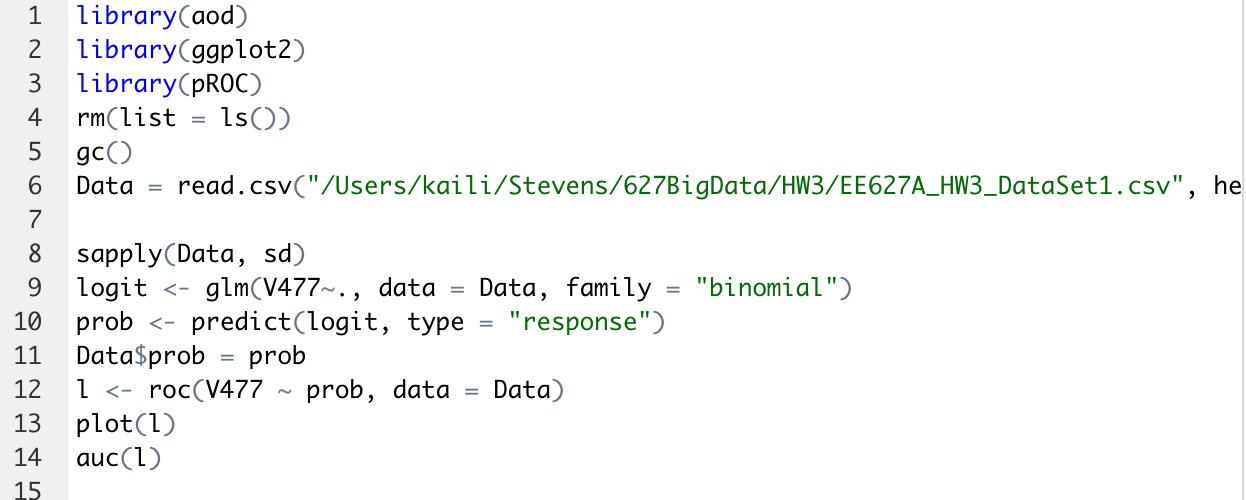
Area under the curve: 0.6875

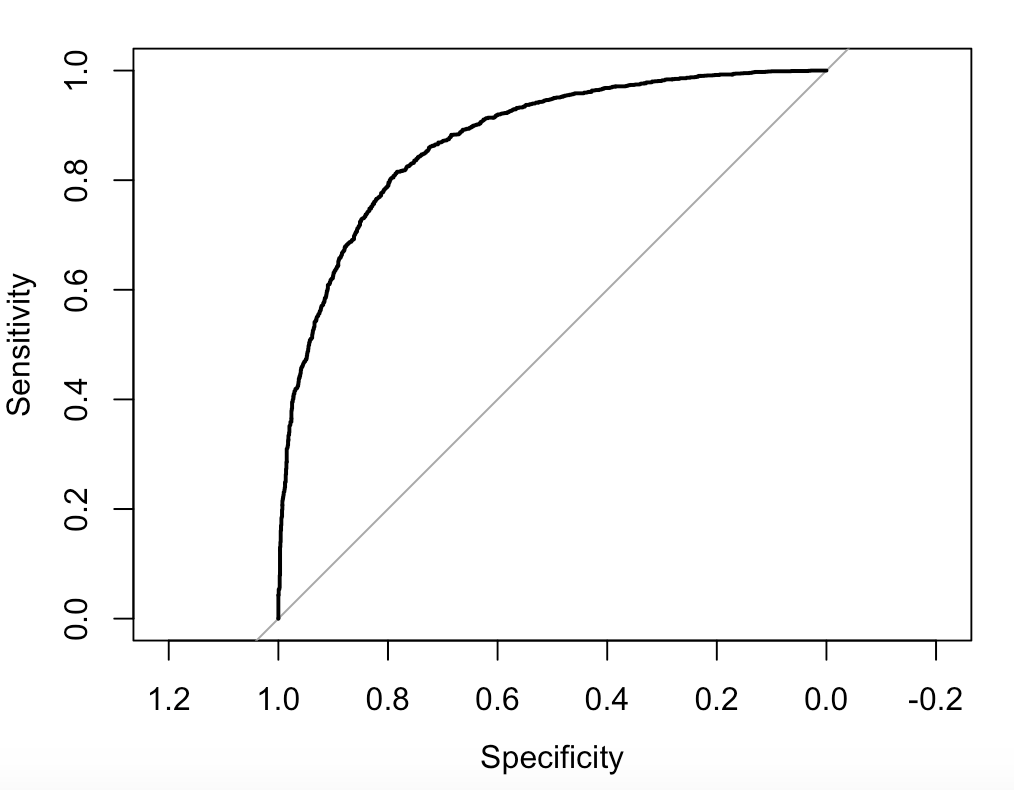




Solution2:

Task1：

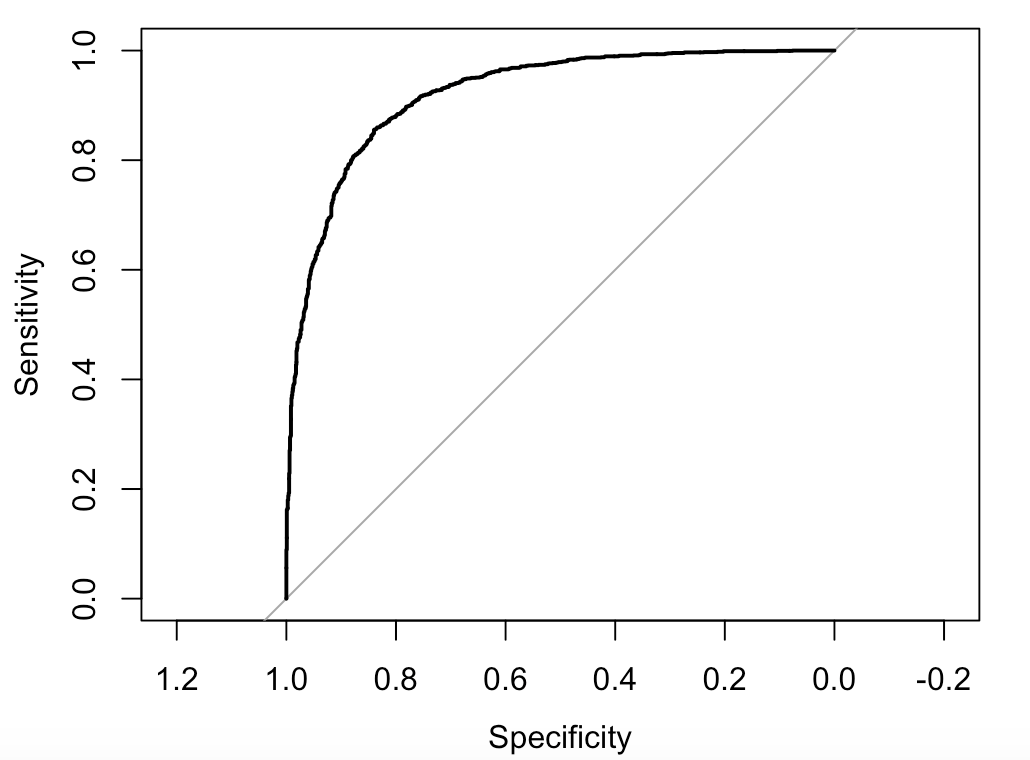
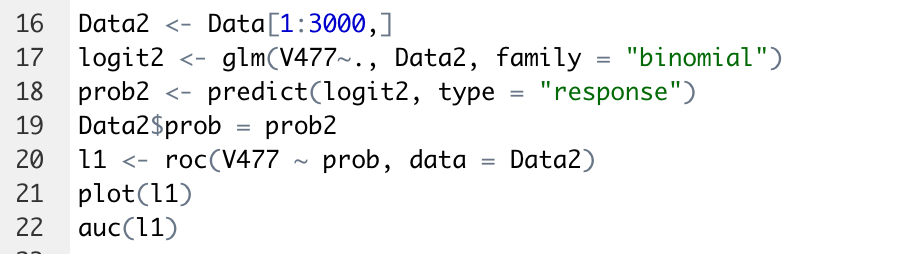




Area under the curve: 0.8748

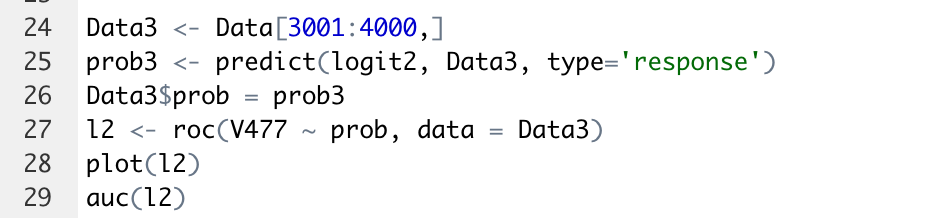
Task2:

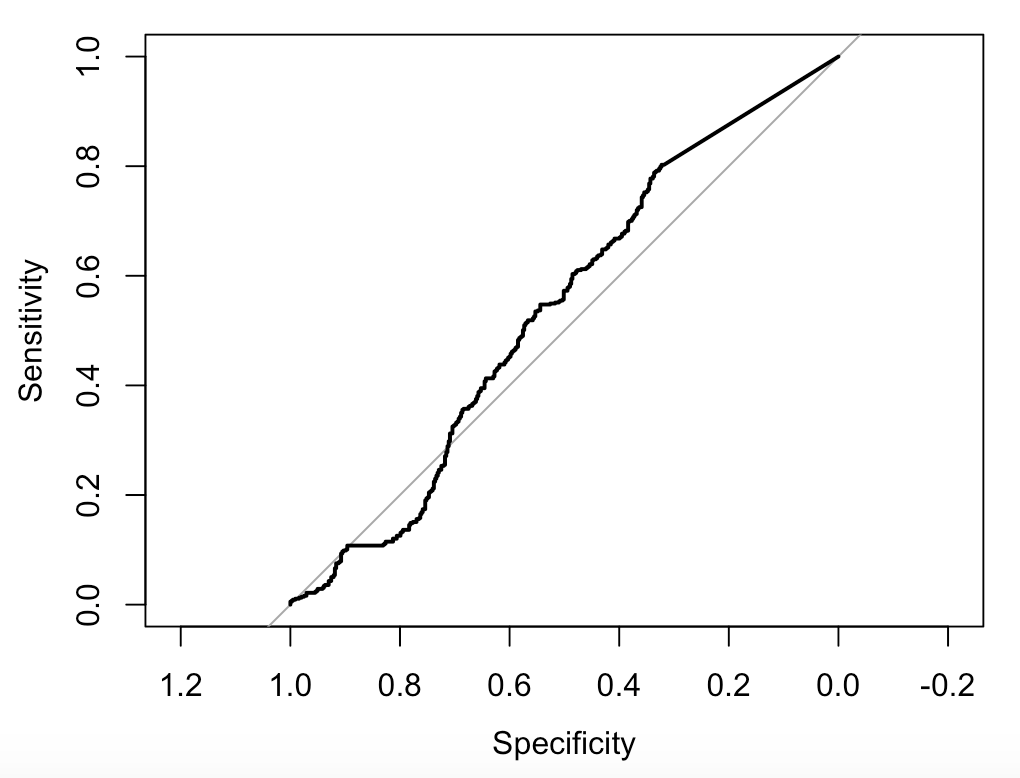
First use training set data to get the logistic regression.



Area under the curve: 0.9214

Then apply the logistic regression coefficients in training set data to validation set.





Area under the curve: 0.5361

As is show above, AUC of training set is closed to 1 but AUC of validation set is closet to 0.5. Obviously, the logistic regression coefficients in training set data is not suitable for validation set. So, this model maybe not good for this data or the first 3000 is not related to the rest 1000 data.