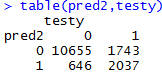
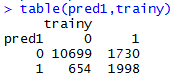
We first used the function glmnet in the package "glmnet", and produced a logistic regression model with overall training misclassification rate of 15.81% and overall test misclassification rate of 15.84%. Since the goal of our model is to predict whether a person makes over 50K a year, we need to examine specifically the proportion that people were misclassified to the group >50K, i.e. type I error. We get the training misclassification rate is 46.41% and test is 46.11%, which is not so good.



Thus we take advantage of function cv.glmnet to find the best  for the best subset. From the fit we get the minimal =0.0001957427. Plug this value in the previous model we can get the overall training misclassification rate is 15.27% and test misclassification rate is 15.18% . Again, we check the type I error and get training misclassification rate of 40.18% and test of 39.68%, which are much better than before. Note that the test error rate is actually slightly lower than the training error rate. It may be that our model actually fits better with the test data.

