Lab_4_E2

Lab 4 Exercise 2

```
library(boot)
# grab our bank data
bank = read.csv('/Users/rileylatham/Downloads/SSC442/Lab_Working_Files/data/bank.csv')

# split into training and testing data
# we use 500 samples for testing, about 10% of our data
set.seed(42)
bank_idx = sample(nrow(data), 500)
bank_trn = bank[-bank_idx, ]
bank_tst = bank[bank_idx, ]

# Now we make our logistic regression
logreg = glm(y~., data = bank_trn, family = binomial)

# Perform a 10 fold cross validation
cv_10 = cv.glm(bank_trn, logreg, K = 10)$delta[1]

# Creating our confusion matrix, sensitivity, and specificity functions
make_conf_mat = function(predicted, actual) {
  table(predicted = predicted, actual = actual)
}

sensitivity = function(conf_mat){
  conf_mat[1]/(conf_mat[1]+conf_mat[2])
}

specificity = function(conf_mat){
  conf_mat[4]/(conf_mat[3]+conf_mat[4])
}

# Here we test our logreg model with a confusion matrix
bank_tst_pred = ifelse(predict(logreg, bank_tst) > 0,
                       "spam",
                       "nonspam")

conf_mat = make_conf_mat(predicted = bank_tst_pred, actual = bank_tst$y)
sens = sensitivity(conf_mat)
spec = specificity(conf_mat)
```

In the above code we load in our bank data, split it into training and testing sets, create a logistic regression and perform a 10 fold cross validation. Now we will interpret the results for some of the more intersting and statistically significant coefficients. Interestingly job type, marital status, and education had very little effect on our models predictions. Instead the coefficients which mattered most were campaign, previous, duration, housing status, loan status and on-file method of contact. The campaign and previous parameters are expected to influence the outcome a lot as they are the driving factors for costumers understanding the advertisement with duration playing a large role as well. Interestingly the contact method was also an incredible predictor in part because a person with no contact information is included in the dataset still and had likely not heard of the advertisement at all. This is basically giving the classifier some free-bees. The housing and loan status are also decently strong parameters as they indicate the customers ties to the bank and are likely to be involved in their banking more so than somebody who rents.

I include here the sensitivity and specificity for our model as well as our confusion matrix. The sensitivity was very high at 0.9747706, giving us confidence in the accuracy of the models true positive predictions. The specificity however is quite poor at 0.328125. We have a biased data set with regards to counts of those who say yes and no and this is shown here in our model. It often predicts to the average and gives us our lower specificity. Our confusion matrix is given here. 425, 11, 43, 21