BSAN 450 Assignment 15

1) In the last assignment you were asked to build a logistic regression model for the Donner data. To remind you of the background, the Donner and Reed families left Springfield Illinois for California by covered wagon. In July the Donner Party, as it became known, reached Fort Bridger, Wyoming. There its leaders decided to attempt a new and untested route to the Sacramento Valley. Having reached its full size of 87 people and 20 wagons, the party was delayed by a difficult crossing of the Wasatch Range and again in the crossing of the desert west of the Great Salt Lake. The group became stranded in the eastern Sierra Nevada mountains when the region was hit by heavy snows in late October. By the time the last survivor was rescued on April 21, 1847, 40 of the 87 members had died from famine and exposure to extreme cold. This data is reported in Ramsey and Shafer in the Statistical Sleuth. The original data consisted of three variables: Surv = 1 if the party member survived and 0 if the party member died; Gender = 1 if female and 0 if male; and Age = the age of the party member. Abraham and Ledolter constructed a fourth variable MultFam = 1 if other party members with the same family name were present and 0 otherwise. The data used in this analysis consist of the Donner party members that were 15 years or older.

The best logistic regression model with the variable Surv as the predictor had two independent variables: Age and MultFam. The following R commands read the data and estimate this logistic regression model. Execute these commands.

Donner = read.csv("Donner.csv")

fit=glm(Surv~ Age+MultFam,family=binomial, data=Donner)

a) Since Surv = 1 if the party member survived, a success is defined as surviving and a failure as dying. The logistic regression model can be used to compute π = P(Surv = 1 | Age and MultiFam) for all the party members. The R command to compute these probabilities is given below.

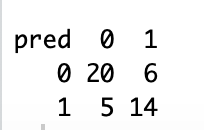
prob=predict(fit,newdata=Donner,type="response")

The probabilities are stored in the variable “prob”. The expression “fit” is the name of the logistic regression model, the expression “Donner” is the input data for which the predictions will be computed, and the expression “response” directs R to compute the probabilities.

The probabilities can be used to predict whether or not each party member will survive or die. The most common rule is: if π > .5, predict that the party member will survive (predict a success) and if π < .5 predict that the party member will die (predict a failure). The following R commands create the predicted classification (Surv = 1 or Surv = 0) and print out the confusion matrix for this data. Execute these commands.

pred=ifelse(prob>.5,1,0)

table( pred,Donner$Surv)



b) Using the confusion matrix for the Donner data answer the following questions.

i) What is the overall error rate?

**11/45 or 0.2444**

ii) What is the false positive error rate?

**1/5 or 0.2**

iii) What is the false negative error rate?

**3/10 or 0.3**

iv) Write a statement describing a false positive error.

**A false positive is predicting a positive result when in actuality it is negative.**

v) Write a statement describing a false negative error.

**A false negative is predicting a negative result when in actuality it is positive.**

c) While the rule predict a success if π > .5 is one way to make the classification, we could consider any rule of the form predict a success if π > k where k is a value between 0 and 1. The value k is called the cutoff value. In other words, changing the cut off value leads to a variety of different rules. It is of interest to investigate what happens by changing the value of k.

The following R commands will produce three plots, an ROC plot, a plot of the false positive error rate versus the cutoff value, and a plot of the false negative error rate versus the cutoff value.

library(ROCR)

pre=prediction(prob,Donner$Surv)

perf = performance(pre,"tpr","fpr")

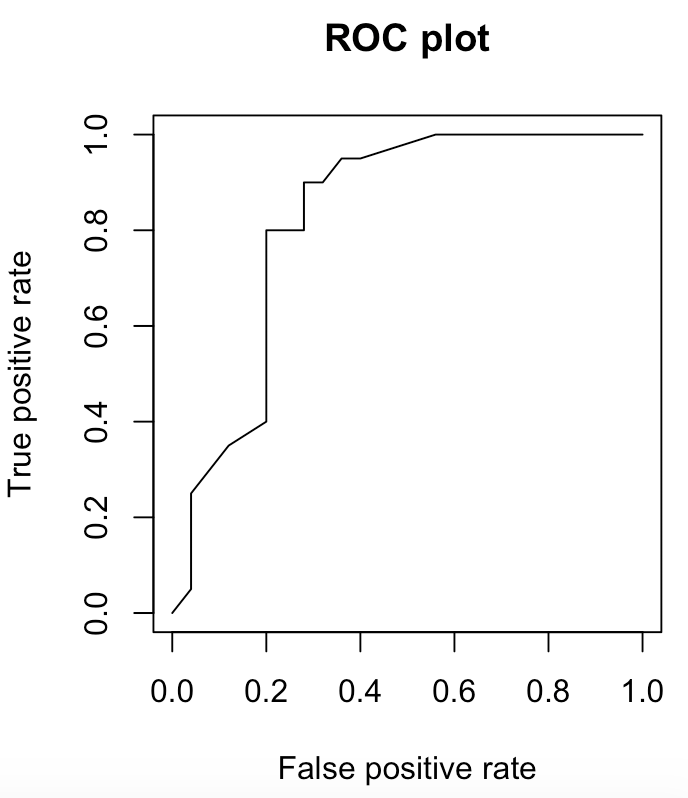
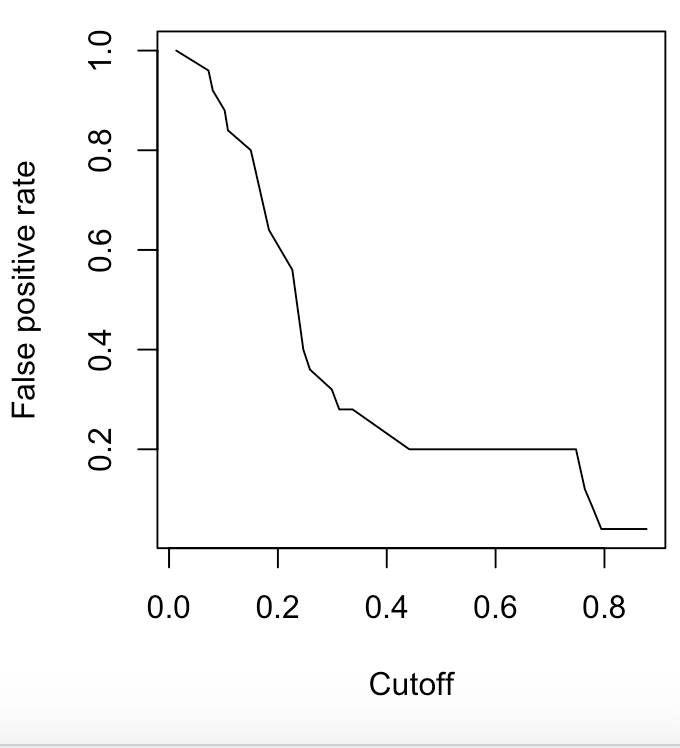
plot(perf,main="ROC plot")

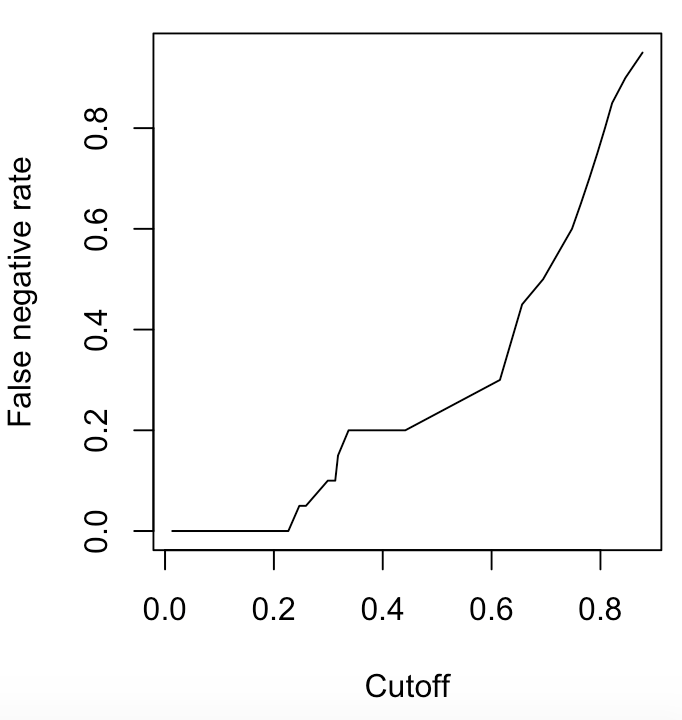
perf=performance(pre,"fpr","cutoff")

plot(perf)

perf=performance(pre,"fnr","cutoff")

plot(perf)



In the expression pre=prediction(prob,Donner$Surv) the expression “prob” is a vector of the probabilities of success for each observation and the expression “Donner$Surv” is a vector of the actual outcomes. The element “pre” is the input to the following commands.

Execute these commands.

1. Comment on the ROC plot.
   1. **The ROC plot looks pretty good. The curve increased quickly and creates the values in the upper left corner.**

ii) The plot of the false positive rate versus cutoff and the plot of the false negative rate versus the cutoff can provide help in determining a good cutoff value. It is desired to choose a cutoff that balances the size of the false positive rate and the false positive rate. Ideally we would like to choose a cutoff that would make both of these rates small. Based on these plots what cutoff would be a good choice for this data?

**Based on the plots it looks like 0.4 would be a good cutoff value.**

2) In the last assignment we also considered the loan acceptance data. A logistic regression model was built to predict the variable Response. Response = 1 if the customer accepted the offered loan and = 0 if the customer did not accept the offered loan. The data was divided into a training set of 4,000 observations and a test set of 1,000 observations. The following R commands read in the data and create the training and test sets.

loan=read.csv("LoanAccept.csv")

set.seed(1)

train=sample(5000,4000)

test=(c(1:5000)[-train])

loan.train=loan[train,]

loan.test=loan[test,]

The best model to predict the variable Response included the independent variables Inc, Educ, CD, Fam, Online, CreditCard, CCAve, and SecAcc.

a) The R commands to estimate this model and to plot the ROC plot, the plot of false positive rate versus cutoff, and the plot of false negative rate versus cutoff are given below. Note this is for the training data.

fit=glm(formula = Response ~ Inc + factor(Educ) + CD + Fam + Online +

CreditCard + CCAve + SecAcc, family = binomial, data = loan.train)

prob=predict(fit,loan.train, type ="response")

library(ROCR)

pred = prediction( prob,loan.train$Response)

perf = performance(pred,"tpr","fpr")

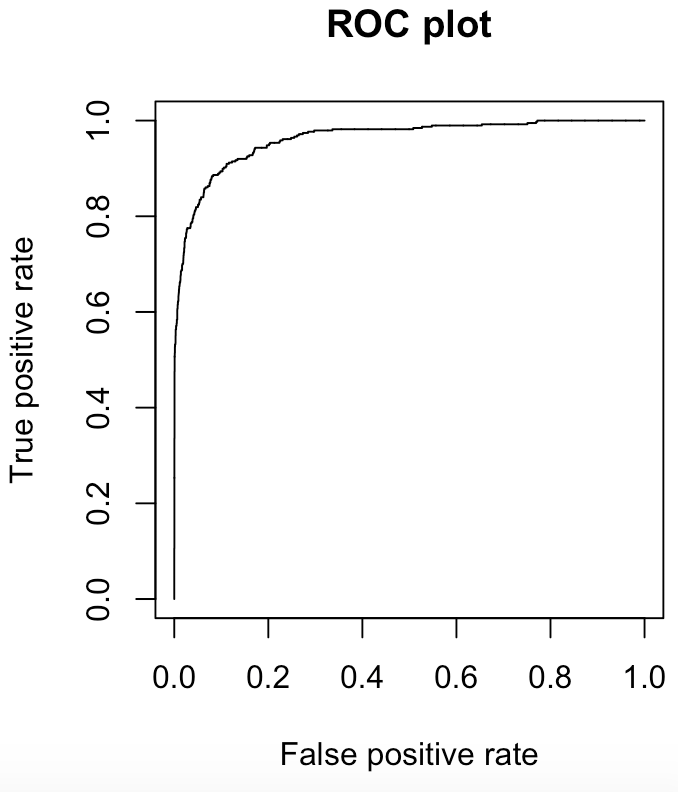
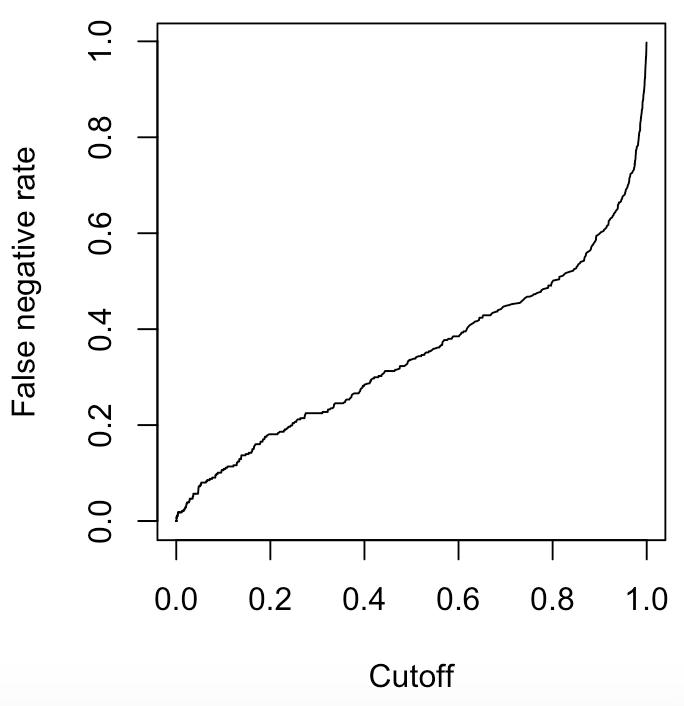
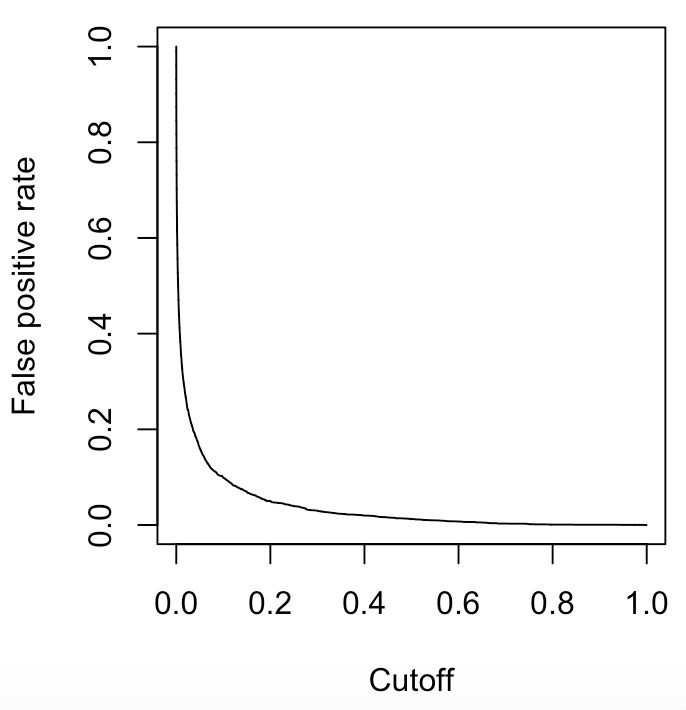
plot(perf,main="ROC plot")

perf=performance(pred,"fpr","cutoff")

plot(perf)

perf=performance(pred,"fnr","cutoff")

plot(perf)

Execute these commands and based on the plots answer the following questions.

1. Describe what a false positive error is for this example.
   1. **A false positive error is the percent of cases in the data set that the model predicted that the person took the loan but in fact the person did not take the loan.**
2. Describe what a false negative error is for this example.
   1. **A false negative error is the percent of cases in the data set that the model predicted that the person did not take the loan but in fact the person did take the loan.**

iii) Based on the plots, what is a good cutoff value for this example?

**0.2 looks like it would be a good cutoff.**

b) Suppose that the cutoff value is changed to .2, then compute and print out the confusion matrix for the training set when you predict that the customer will accept the loan if π > .2. The R commands to fit the logistic regression, compute the probabilities, make the prediction based on these probabilities, and print out the confusion matrix are given below.

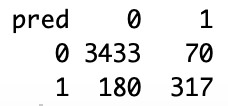
fit=glm(formula = Response ~ Inc + factor(Educ) + CD + Fam + Online +

CreditCard + CCAve + SecAcc, family = binomial, data = loan.train)

prob=predict(fit,newdata=loan.train, type ="response")

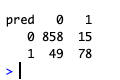
pred=ifelse(prob>.2,1,0)

table( pred,loan.train$Response)



1. What is the overall error rate?
   1. **250/4000 or 0.0625**
2. What is the false positive error rate?
   1. **180/3623 or 0.0497**
3. What is the false negative error rate?
   1. **70/387 or 0.1809**

c) Compute the confusion matrix for the test data. To do this replace the expression newdata=loan.train with newdata=loan.test; this will compute the probabilities for the data in loan.test. Also replace the expression loan.train$Response in the table function with loan.test$Response; these are the actual response values in the test data set.



For the test data:

1. What is the overall error rate?
   1. **64/1000 or .064**
2. What is the false positive error rate?
   1. **49/907 or .054**
3. What is the false negative error rate?
   1. **15/93 or .1613**