BSAN 450 Assignment 18

1) In this problem we will review using Naïve Bayes on the loan acceptance data that was used in the most recent video. This problem will cover the R commands to create a Naïve Bayes classifier. This data is taken from Shmueli et at (2010). The data set contains information of 5000 loan applications. The response is whether or not an offered loan had been accepted on an earlier occasion. The explanatory variables are:

Age = age of the customer

Exp = professional experience in years

Inc = income of the customer

Fam = family size of the customer

CCAve = average monthly credit card spending

Educ = three categories of education level: 1 = undergraduate, 2 = graduate, 3 = professional

Mort = size of mortgage

SecAcc = 1 if the customer has a securities account and otherwise = 0

CD = 1 if the customer has a CD account and otherwise = 0

Online = 1 if the customer has an online account and otherwise = 0

CreditCard = 1 if the customer has a credit card and otherwise = 0

The name of the response variable is Response.

In this assignment we will use Naïve Bayes classification with this data.

a) Read the data into R Studio and create the training and test sets. Note that these are the same training and test sets that were used earlier. The variables Response, Educ, SecAcc, CD and Online need to be changed into factor variables because they are all categorical variables that are codes using numeric values. If they are not recoded, they will be treated incorrectly.

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

loan$Response=factor(loan$Response)

loan$Educ=factor(loan$Educ)

loan$SecAcc=factor(loan$SecAcc)

loan$CD=factor(loan$CD)

loan$Online=factor(loan$Online)

loan$Fam=factor(loan$Fam)

set.seed(1)

train=sample(5000,4000)

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

loantrain=loan[train,]

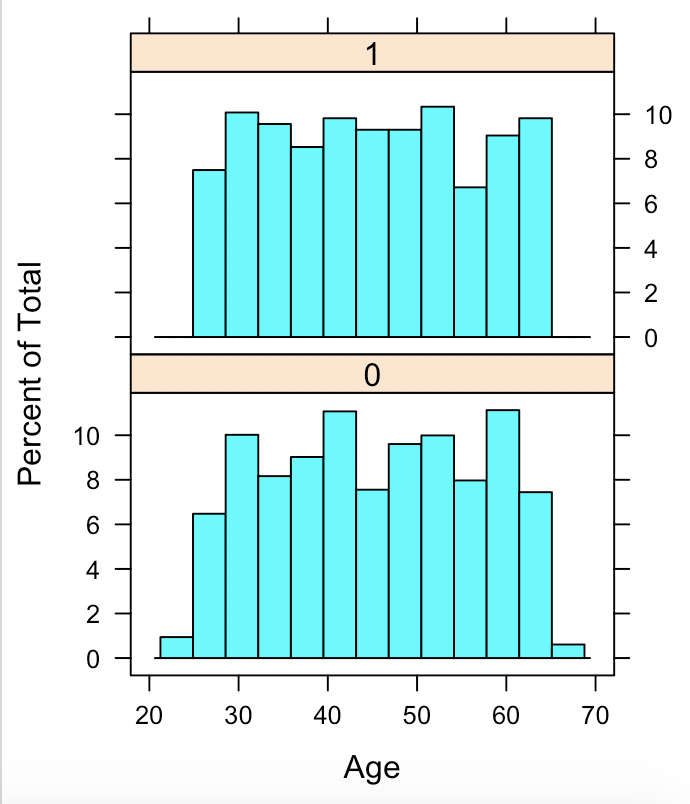
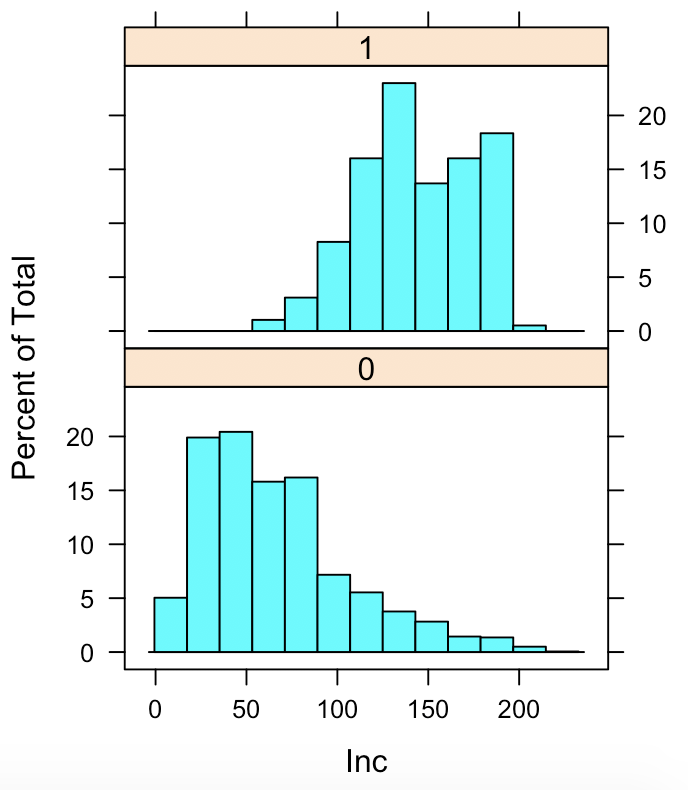
loantest=loan[test,]

a) The following R command will reproduce the histograms for the variable Age that was shown in the Naïve Bayes video. Execute this command.

library(lattice)

histogram(~Inc|Response,data=loantrain,layout=c(1,2))

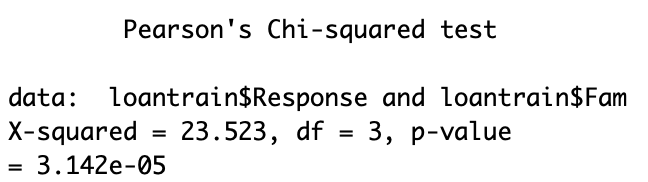
Plot a set of histograms for the different values of Response for the continuous variable Inc.



b) The following R commands will perform a chi-square test with the null hypothesis that the proportions for the categorical variable CD are the same when Response =1 as when Response = 0. Execute this command.

chisq.test(loantrain$Response,loantrain$CD)

Perform the chi-square test for the variable Fam.



c) After examination of the plots and the chi-square analysis in the second video this week, it was decided that the variables Inc, Fam, CCAve, Educ, Mort, and CD should be included in the model. Note that in the dataframe loan, Inc is variable 2, Fam is variable 4, CCAve is variable 5, Educ is variable 6, Mort is variable 7 and CD is variable 10.

The R commands to fit a Naïve Bayes model to classify the variable Response with the 6 inputs listed above is:

x.vars= c(3,4,5,6,7,10)

x=loantrain[,x.vars]

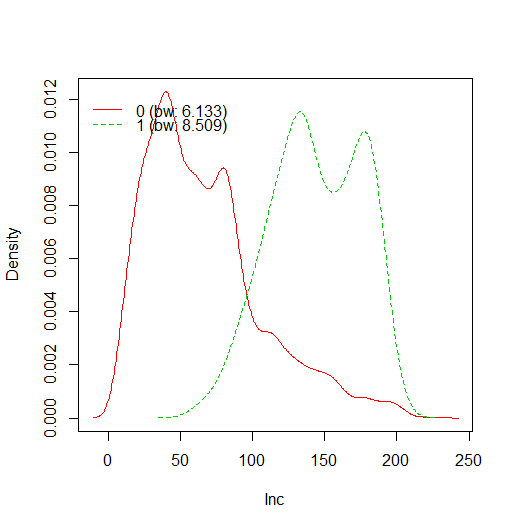
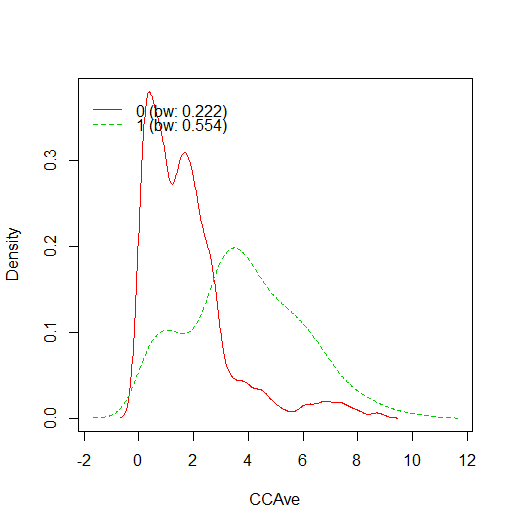
y=loantrain$Response

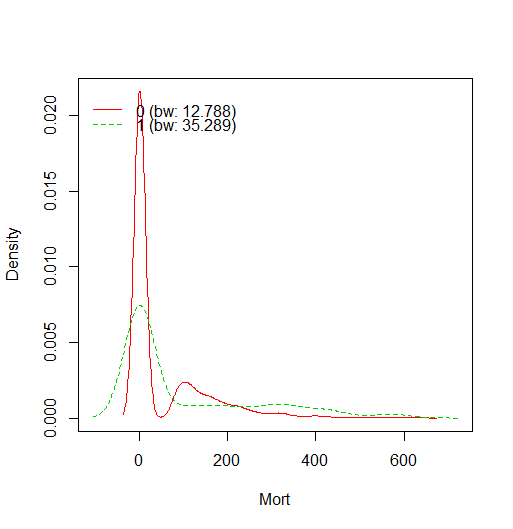
library(naivebayes)

model=naive\_bayes(x,y,usekernel=TRUE)

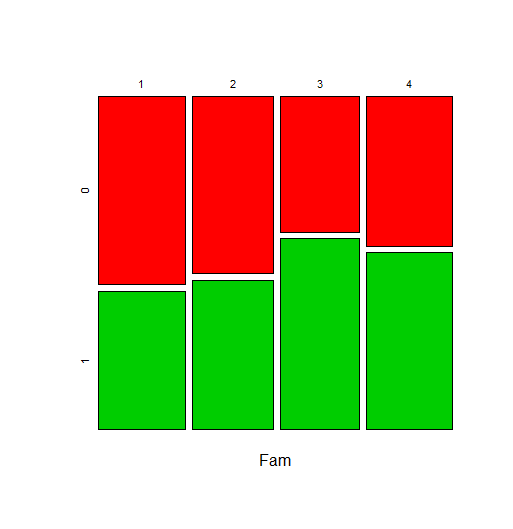
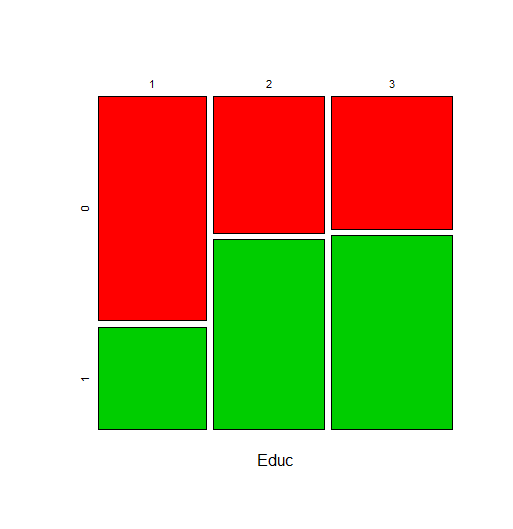
plot(model)

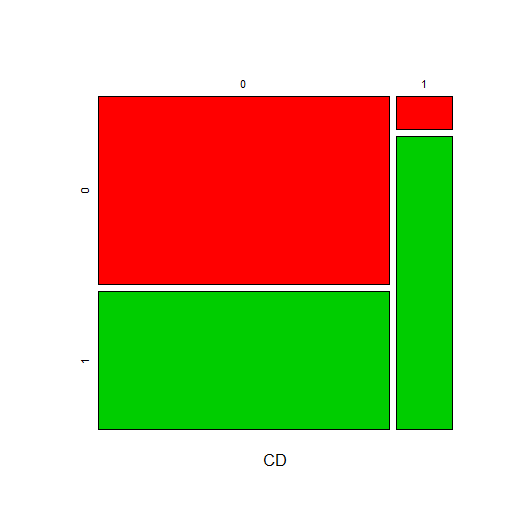
Execute these commands to produce the output given below.



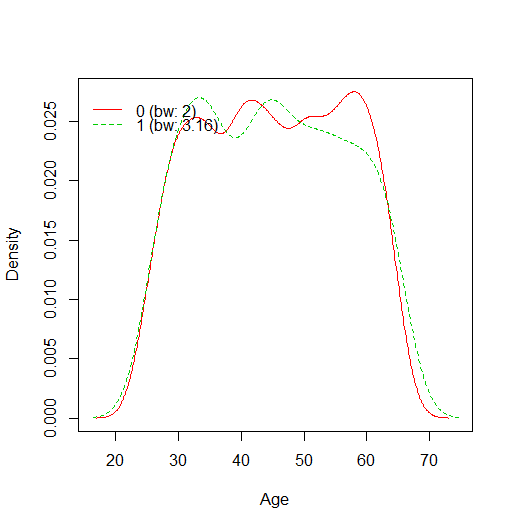
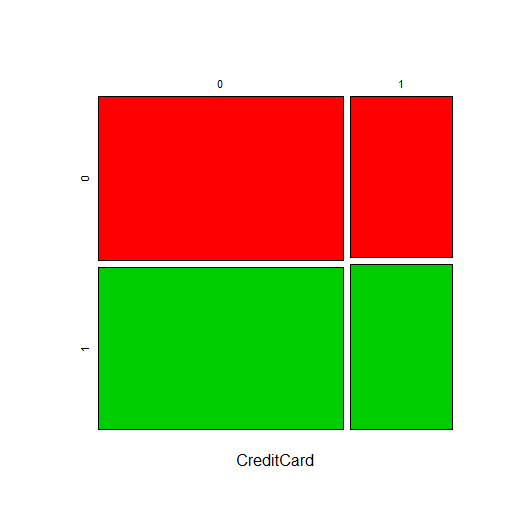
These plot are the separate estimated probability distributions that the Naïve Bayes algorithm estimated for the 3 continuous variables. The red line is the distribution given Response = 0 and the green line is the distribution given Response = 1. It is desired that the two probability distributions are different, if that is true then the variable will be better able to discriminate between the two Response groups. The plots for Inc and CCAVe are quite different, consequently these are good variable to include in the model. There is not as much difference in the 2 probability distributions for the variable Mort, but it still seems like a good variable to include.



The above 3 plots correspond to the 3 categorical variables that are included in the model. The red boxes represent the results when the variable Response = 0 and the green boxes represent the results when the variable Response = 1. If the heights of the green bars change a lot this suggests that the proportion of positive responses changes as the values of the categorical variable change. This is what we want to occur. These plots suggest that all three of these variables are good to keep in the model.

An example of plots of variable that are not helpful in a Naïve Bayes model are:

In the plot on the left, the two probability distributions are so similar that this variable does not do a good job of discriminating between the two responses. In the plot on the right the height of the green boxes in nearly the same, so again this variable does not do a good job of discriminating between the two responses.

b) The following R commands will plot an ROC plot, a plot of the false positive error rate versus the cutoff values, and a plot of the false negative rate versus the cutoff values. Note that these commands are for the most part the same as we have been using for other classification methods in which the probability of the response being positive can be computed. In the Naïve Bayes approach, the posterior probabilities are used to classify the observation with the usual rule being the observation is classified as positive is the posterior probability is greater than the cutoff of .5

The expression prob=predict(model,x, type = "prob") computes the posterior probabilities for the Naïve Bayes output named “model” with input variable x. The statement type = “prob” will produce the posterior probabilities as output.

Since the probabilities in the variable prob is a vector with 2 columns, we need to specify which column contains the correct probabilities in the expression pred=prediction(prob[,2],y). The probabilities are in the second column.

Execute these commands to produce the plots.

library(ROCR)

prob=predict(model,x, type = "prob")

pred=prediction(prob[,2],y)

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

plot(perf,main="ROC plot")

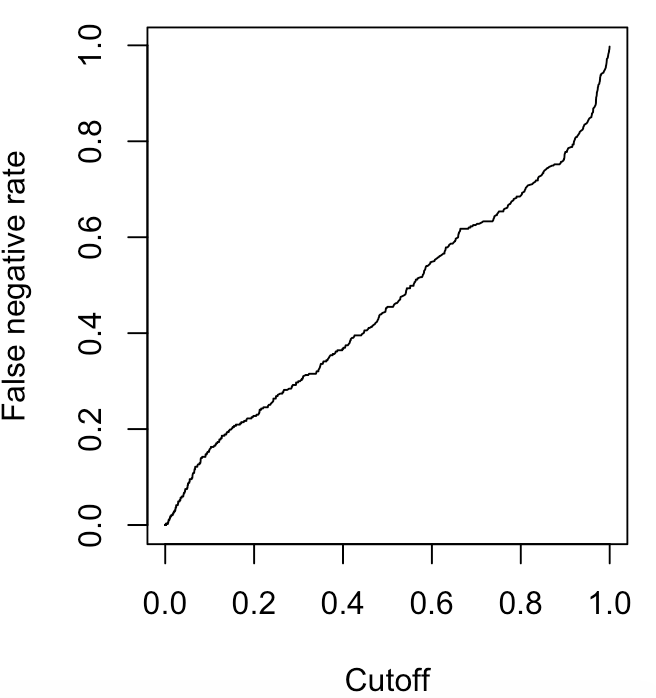
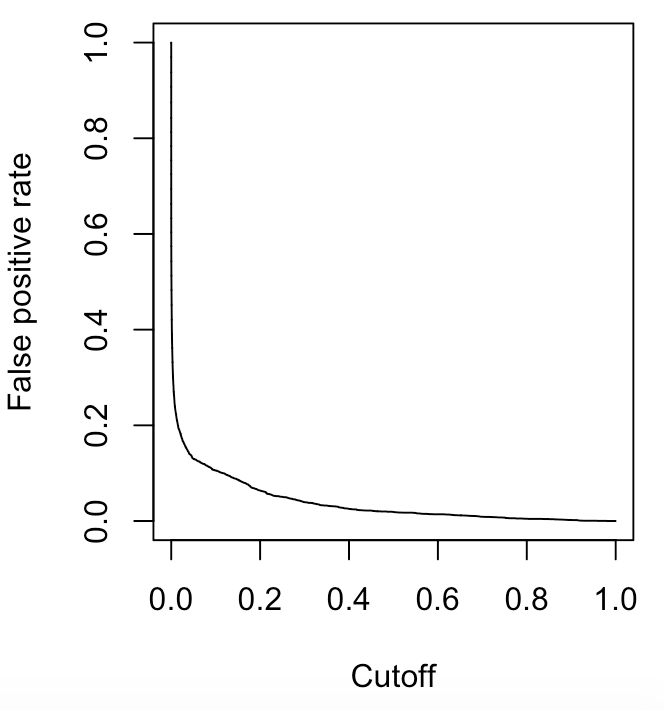
perf=performance(pred,"fpr","cutoff")

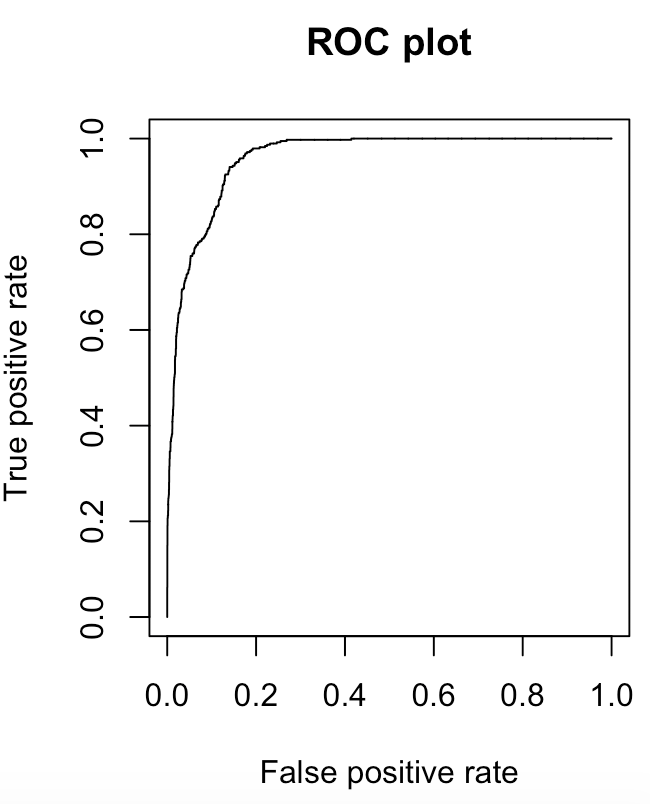
plot(perf)

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

plot(perf)

c) For this example write down an expression of a false positive and a false negative error.



**False positive = number of false positives / total number of positives**

**False negative = number of false negatives/ total number of negatives**

If it is desired to have both the false positive error rate and the false negative error rate small, based on the plots from part b, what is a good cutoff value for classifying a customer as accepting the loan?

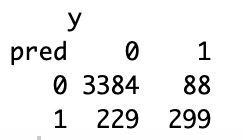
**Cutoff value: 0.2**

d) Suppose we use a cutoff of .2 so we classify the customer as accepting the loan is the probability is greater than .2. The following R commands calculate the confusion matrix for the training data. Calculate the confusion matrix and compute the overall error, the false positive error rate, and the false negative error rate.

prob=predict(model,x, type = "prob")

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

table( pred,y)



**Overall error = (229+88)/(229+88+299+3384) = 0.07925**

**False positive error = (229)/(3384+229) = 0.06338**

**False negative error = (88)/(88+299) = 0.2274**

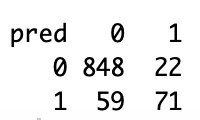
e) Calculate the confusion matrix for the test data and compute the overall error, the false positive error rate, and the false negative error rate.

xtest=loantest[,x.vars]

prob=predict(model,xtest, type = "prob")

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

table( pred,loantest$Response)



**Overall error = (59+22)/(848+59+22+71) = 0.081**

**False positive error = (59)/(848+59) = 0.06505**

**False negative error = (22)/(22+71) = 0.23656**

2) This data consists of a sample of 200 subjects who were part of a much larger study on survival of patients following admission to an adult intensive care unit (ICU). The major goal of this study was to develop a logistic regression model to predict the probability of survival to hospital discharge of these patients and to study risk factors associated with ICU mortality. This data was taken from Hosmer, Lemeshow, and Sturdivant.

The variables in this data set are as follows.

STA: Vital status 1 if lived until discharge and 0 is died prior to discharge

AGE: Patient’s age in years.

GENDER: 1 if Female and 0 if Male

RACE: 1 if White, 2 if Black, and 3 if Other

SER: 1 if surgical service and 0 if medical service when admitted to ICU

CAN: 1 if cancer is part of the presenting problem otherwise 0

CRN: 1 if a history of chronic renal failure otherwise 0

INF: 1 if infection is probable at ICU admission otherwise 0

CPR: 1 if had CPR prior to ICU admission otherwise 0

SYS: systolic blood pressure at ICU admission

HRA: heart rate at ICU admission beats per minute

PRE: 1 if had been previously admitted to an ICU in the prior 6 months 0 otherwise

TYP: 1 if emergency admission 0 if elective admission

FRA: 1 if a long bone, multiple, neck, single area or hip fracture 0 otherwise

PO2: PO2 from initial blood gases 1 if less than or equal to 60 0 if greater than 60

PH: PH from initial blood gases1 if 1 if less than or equal to 7.25 0 if greater than 7.25

PCO: PCO2 from initial blood gases 1 if greater than or equal to 45 0 if less than 45

BIC: Bicarbonate from initial blood gases 1 if less than or equal to 18 0 if greater than 18

CRE: creatinine from initial blood gases 1 if greater than or equal to 2.0 and 0 is less than 2.0

LOC: level of consciousness at ICU admission “no” = no coma or stupor, “stu” = deep stupor, “com” = coma

a) Read the data into R Studio and recode the variable LOC as we did when we analyzed this data with logistic regression.

icu=read.csv("icu.csv")

library(car)

icu$LOC1=icu$LOC

icu$LOC1=recode(icu$LOC1, "'stu'='com'")

set.seed(1)

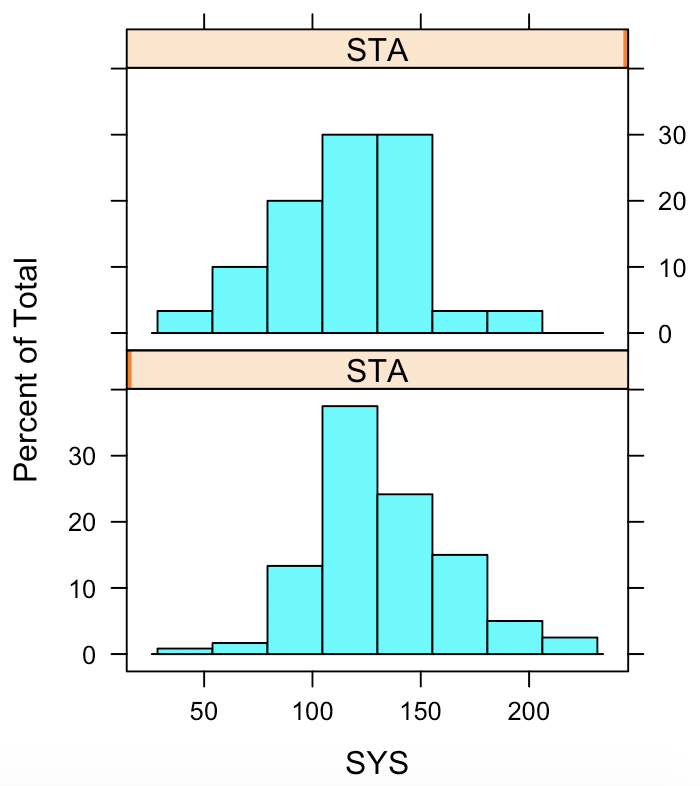
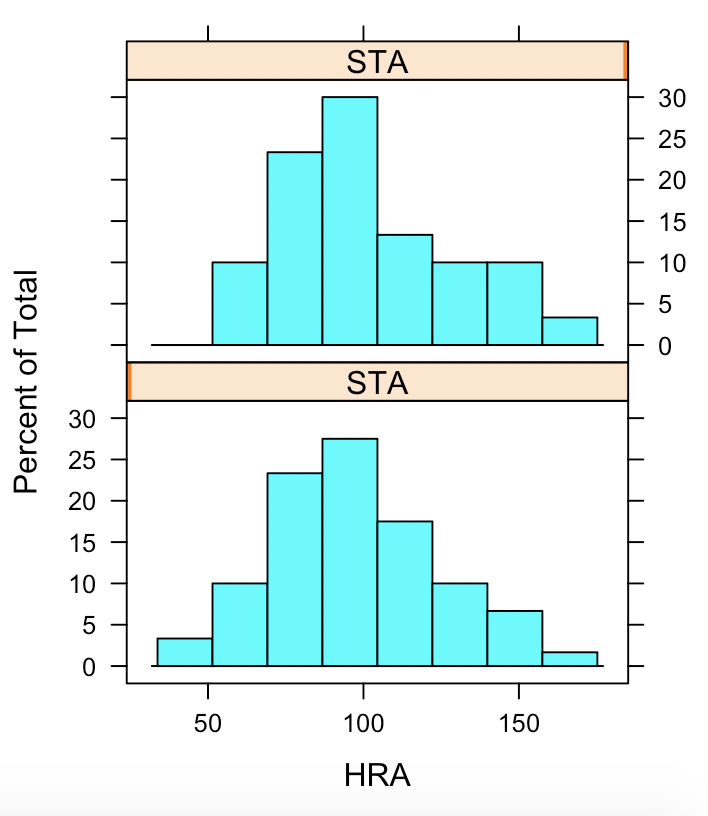
train=sample(200,150)

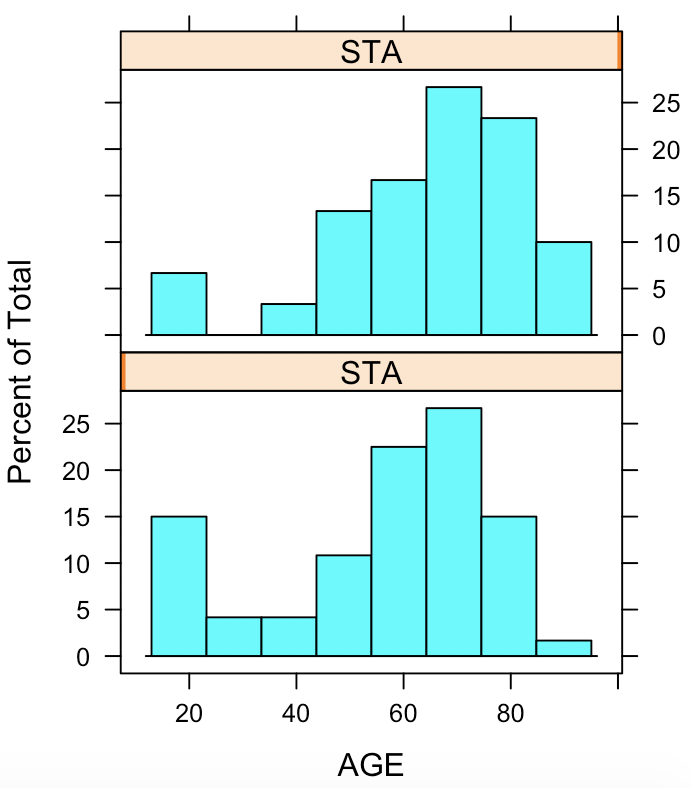
test=(c(1:200)[-train])

icutrain=icu[train,]

icutest=icu[test,]

a) In R plot histogram for the numeric data similar to what was done in problem 1. Which numeric variables should be included in a Naïve Bayes model?

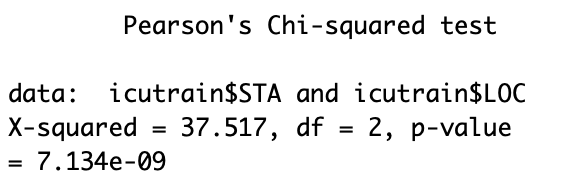
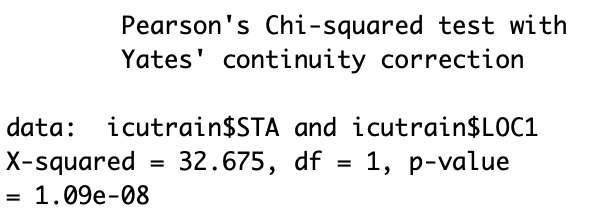
 

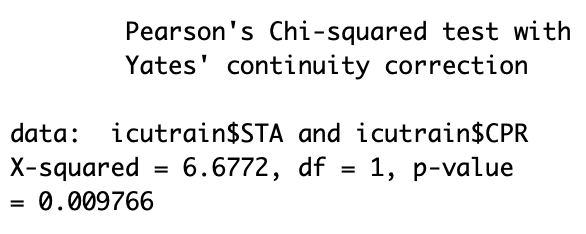
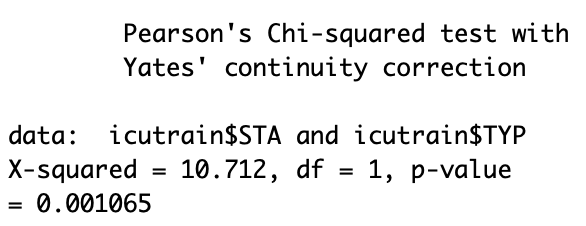


**SYS should be included because the histograms of the two plots look different for SYS but for AGE and HRA the plots look very similar.**

b) In R construct chi-square hypothesis tests for the categorical data similar to what was done in problem 1. Which categorical variables should be included in a Naïve Bayes model?

**CPR, TYP, LOC, LOC1**



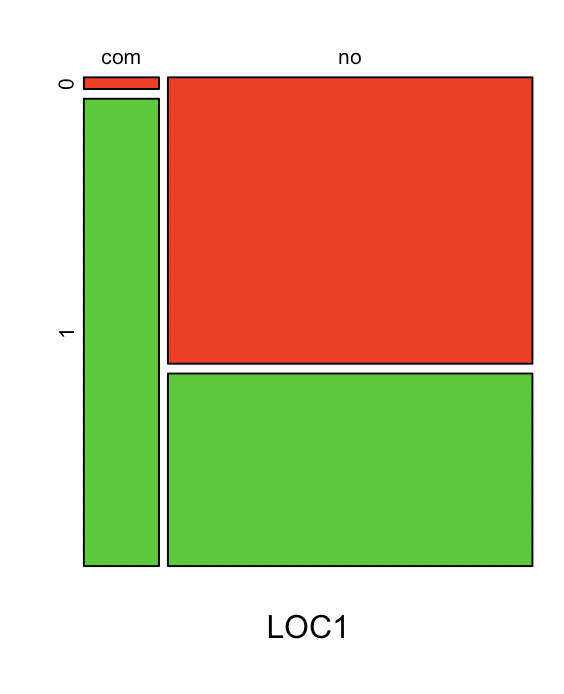
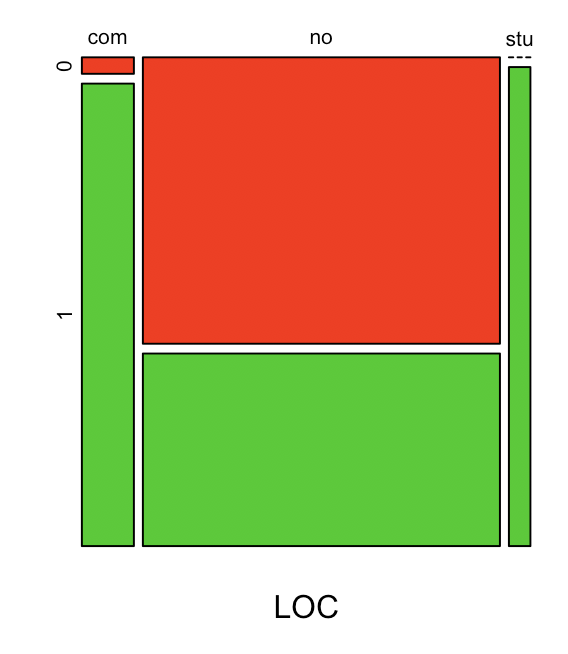
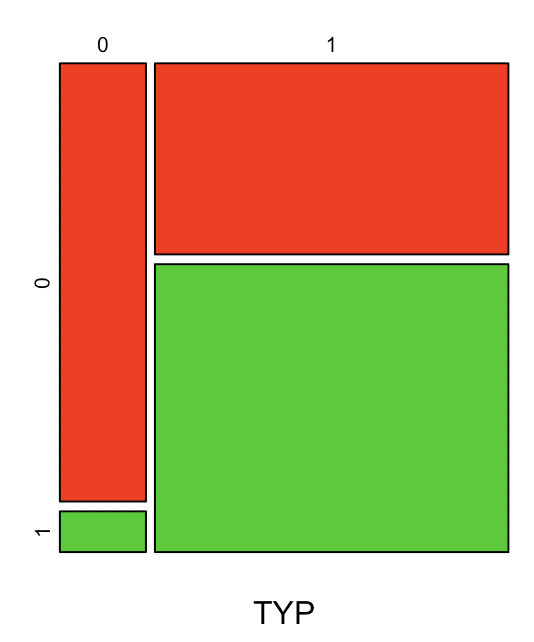
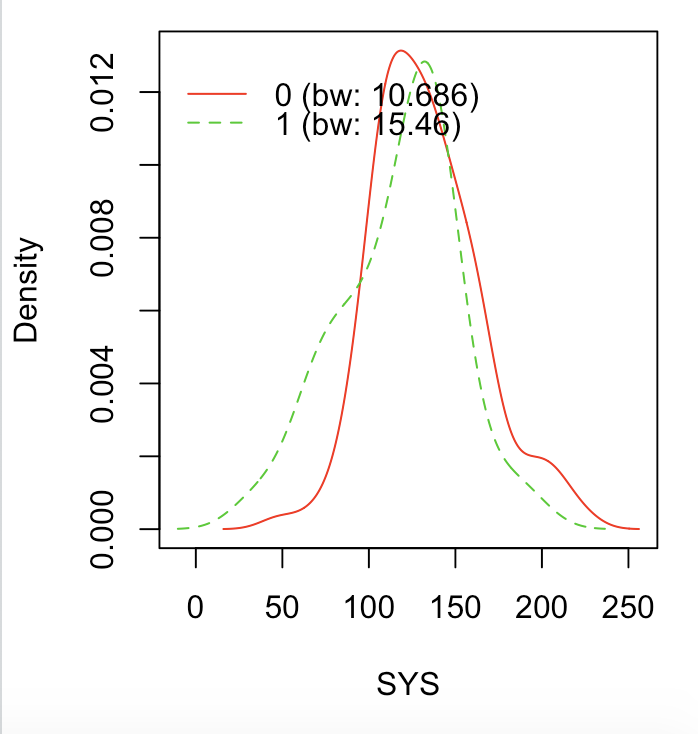
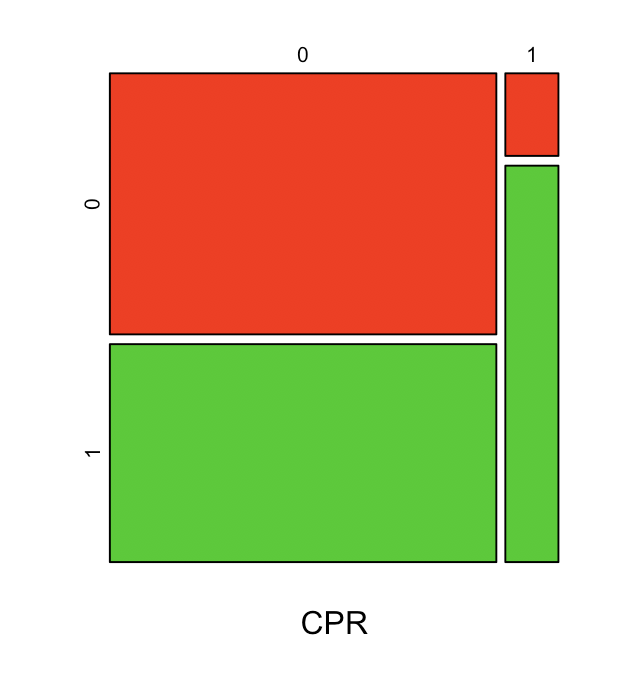
c) All of the categorical variables in this data set except for LOC1 are characterized by values of 1 and 0. In order for these variables to be treated correctly these variables need to be converted to factor variables. Thus, for all the categorical variables that you identified in part b that you wish to include in the Naïve Bayes model change them to factor variables. An example of how to do this is below for the variable SER. Note that you should do this for both the training set and the test set.

icutrain$SER = factor(icutrain$SER)

icutest$SER = factor(icutest$SER)

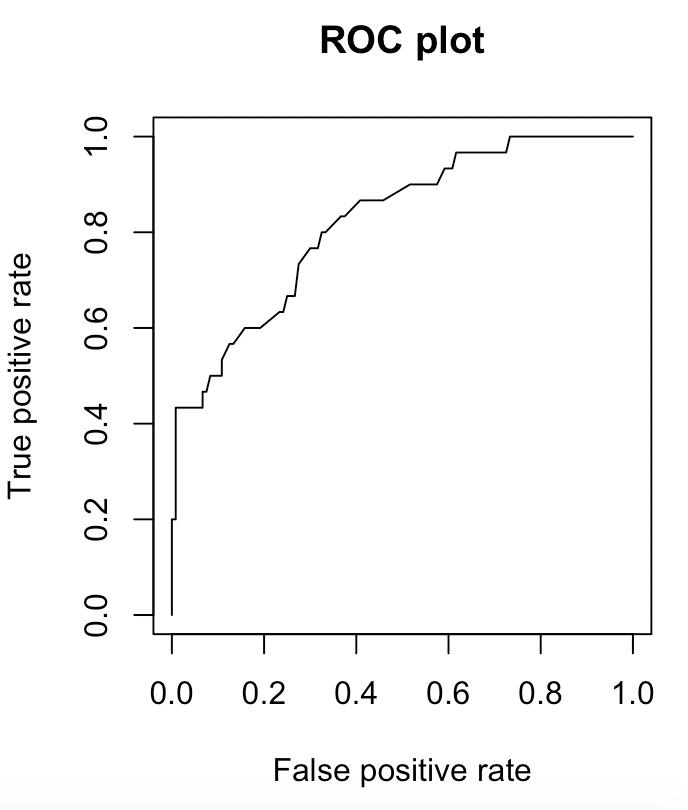
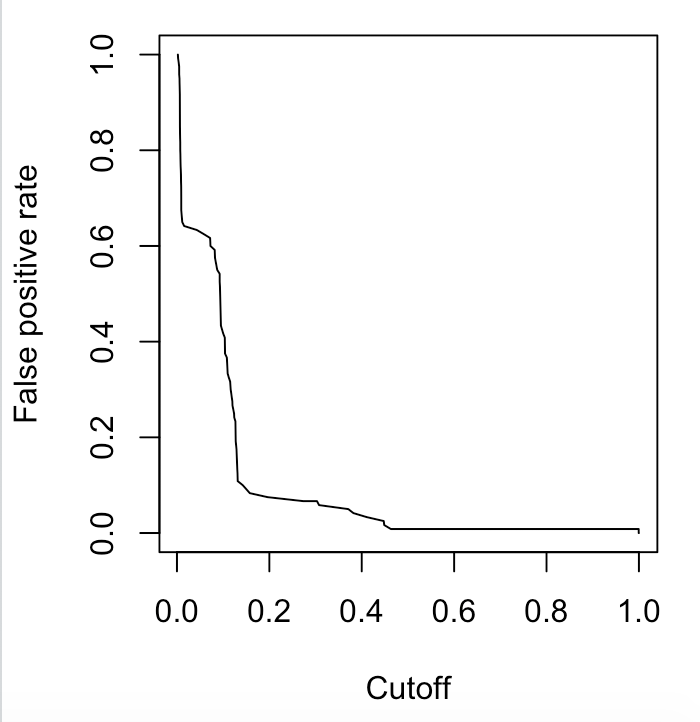
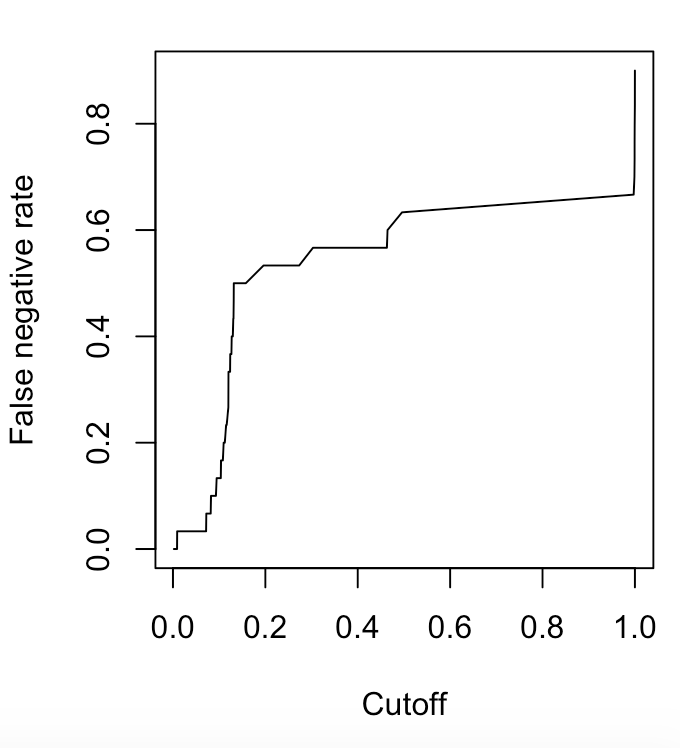
Convert all the necessary variables to factor variables. Estimate a Naïve Bayes model to predict the posterior probability of the variable STA with the variables you have identified in parts a and b as inputs. Use the R commands in problem 1 as an example.

Print out the graphs for the model you fit.

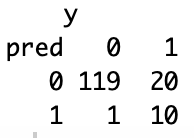
    

**All the categorical graphs looks good. In the numerical graph, the two lines looks very similar and the variable SYS might not be needed.**

d) Plot an ROC plot, a plot of the false positive error versus the cutoff, and a plot of the false negative error versus the cutoff.

  **Cutoff value of 0.5**

e) For a cutoff of .5, print the confusion matrix for the training data. Calculate the overall error, the false positive error rate, and the false negative error rate for this data.

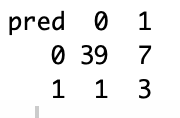


**Overall error = (1+20)/(119+20+1+10) = 0.14**

**False positive error = (1)/(119+1) = 0.00833**

**False negative error = (20)/(20+10) = 0.66667**

f) For a cutoff of .5, print the confusion matrix for the test data. Calculate the overall error, the false positive error rate, and the false negative error rate for this data.



**Overall error = (1+7)/(39+1+3+7) = 0.16**

**False positive error = (1)/(39+1) = 0.025**

**False negative error = (7)/(7+3) = 0.7**