|  |
| --- |
| [Type the company name] |
| 20-PBD-002 Shraddha P Jain |
| End Semester Assignment 3803(B)-MOM |
|  |
|  |
|  |

|  |
| --- |
|  |

The dataset *ESE 2*is ahealthcare dataset which has 14635 observations with several attributes. The objective is to builda model that predicts the presence of complications of surgery of the patient. The data description is given in sheet 2 of excel sheet.

Preprocessing codes:

library(xlsx)

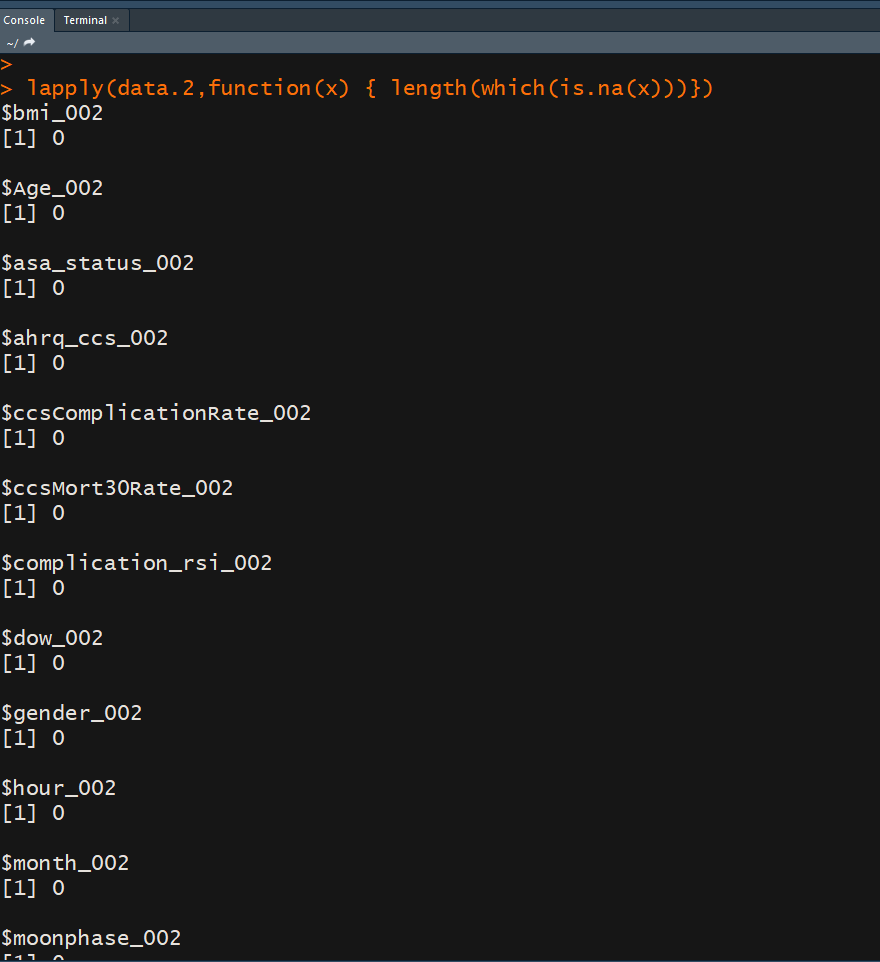
data.2 = read.xlsx("C:\\Users\\Shraddha\\Downloads\\ESE2\_002.xlsx",sheetIndex = 1)

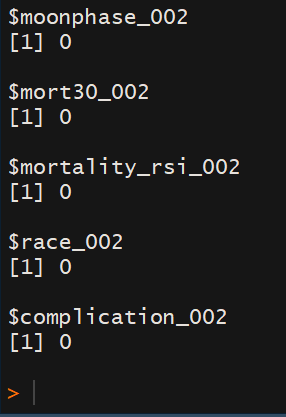
dim(data.2)

str(data.2)

#checking for NA

lapply(data.2,function(x) { length(which(is.na(x)))})





# converting into factor

data.2$asa\_status\_002 = as.factor(data.2$asa\_status\_002)

data.2$gender\_002 = as.factor(data.2$gender\_002)

data.2$dow\_002 = as.factor(data.2$dow\_002)

data.2$month\_002 = as.factor(data.2$month\_002)

data.2$moonphase\_002 = as.factor(data.2$moonphase\_002)

data.2$mort30\_002 = as.factor(data.2$mort30\_002)

data.2$race\_002 = as.factor(data.2$race\_002)

data.2$complication\_002 = as.factor(data.2$complication\_002)

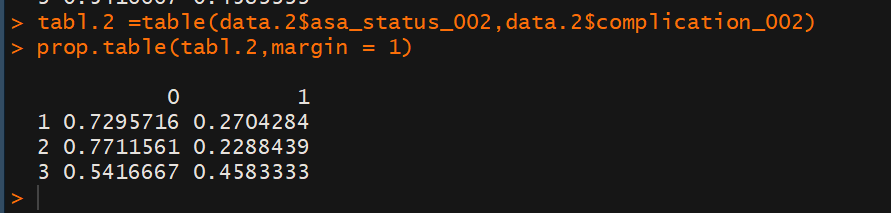
Answer the following questions.

1. Obtain the proportion of patients who had complications based on asa\_status. Which among these seems to have an indication of post-surgery complication according to the given data? Justify your answer.

Ans:

tabl.2 =table(data.2$asa\_status\_002,data.2$complication\_002)

prop.table(tabl.2,margin = 1)



Interpretation: Above is the table of proportions of people who have complications based on their asa status. Around 27% of patients who had asa status as 1 had complications, as compared to 23% of patients with asa status as 2, and 45% of patients with asa status 3.

Ans: Patients with asa status 3 seem to have a higher risk of complication during surgery as compared to others. This can be justified by the above table, as they 45% of patients with asa status 3 had face complications while for patients with asa status 1 or 2, the proportion of complications in surgeries was 27% and 23% respectively

(b)Logistic regression model

Preprocessing

# Creating dummy variables from factors

library(caret)

str(data.2)

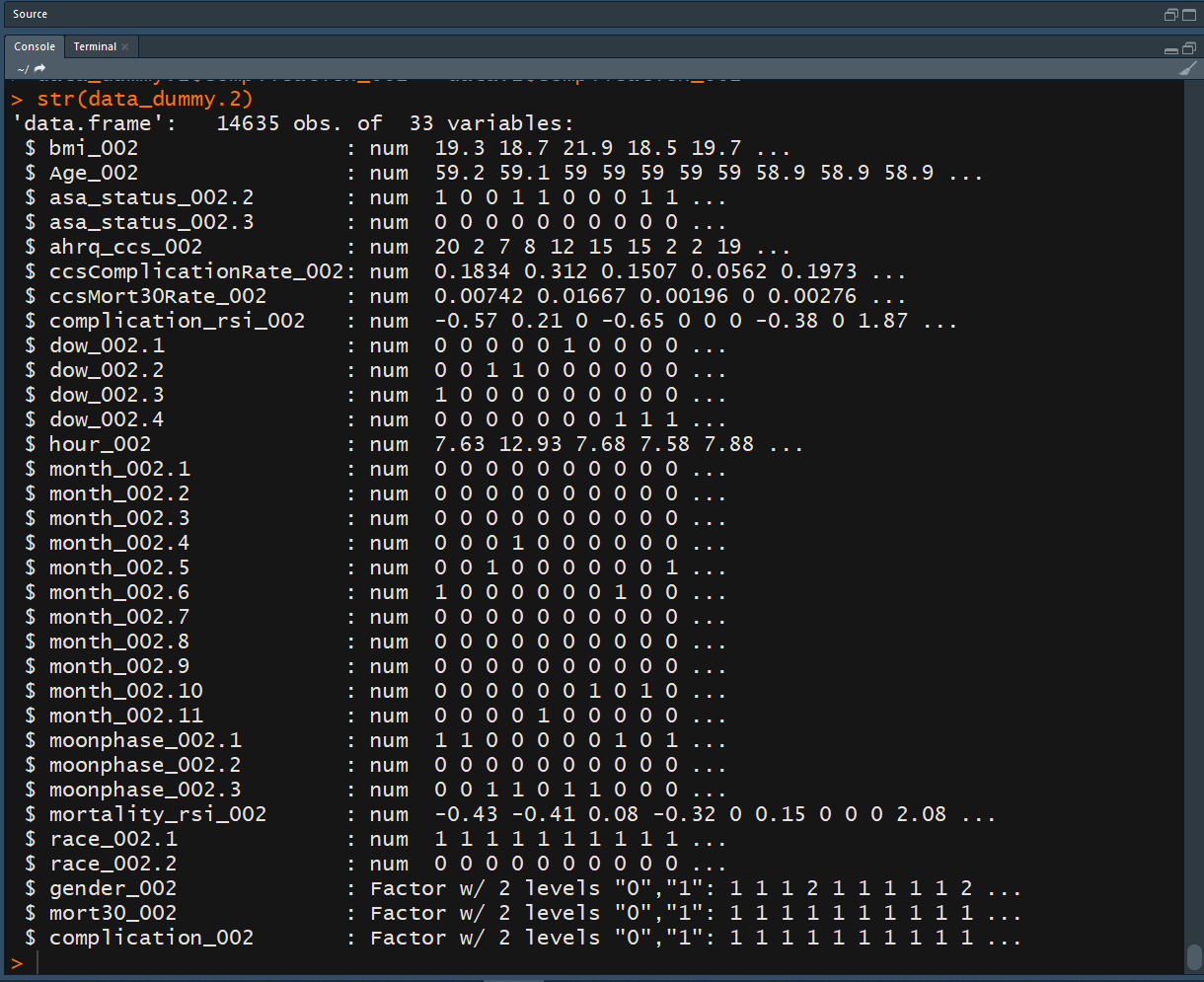
dummy.2 <- dummyVars(" ~ .-complication\_002-gender\_002-mort30\_002", data = data.2, fullRank = T)

data\_dummy.2 <- data.frame(predict(dummy.2, newdata = data.2))

data\_dummy.2$gender\_002 = data.2$gender\_002

data\_dummy.2$complication\_002 = data.2$complication\_002

data\_dummy.2$mort30\_002 = data.2$mort30\_002



# train test split

smp\_size.2<-floor(0.7\*nrow(data\_dummy.2))

set.seed(1024)

trainingdata.2 <- sample(seq\_len(nrow(data\_dummy.2)),size=smp\_size.2)

training.2<-data\_dummy.2[trainingdata.2,]

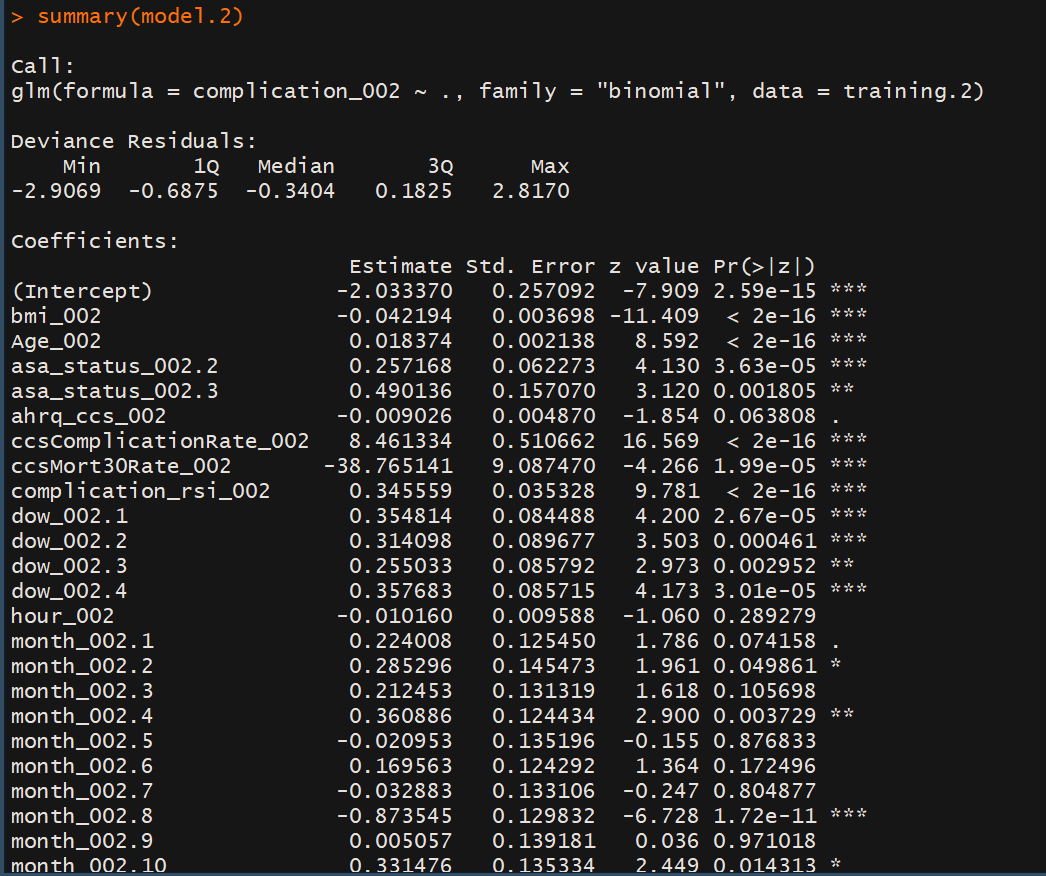
testing.2<-data\_dummy.2[-trainingdata.2,]

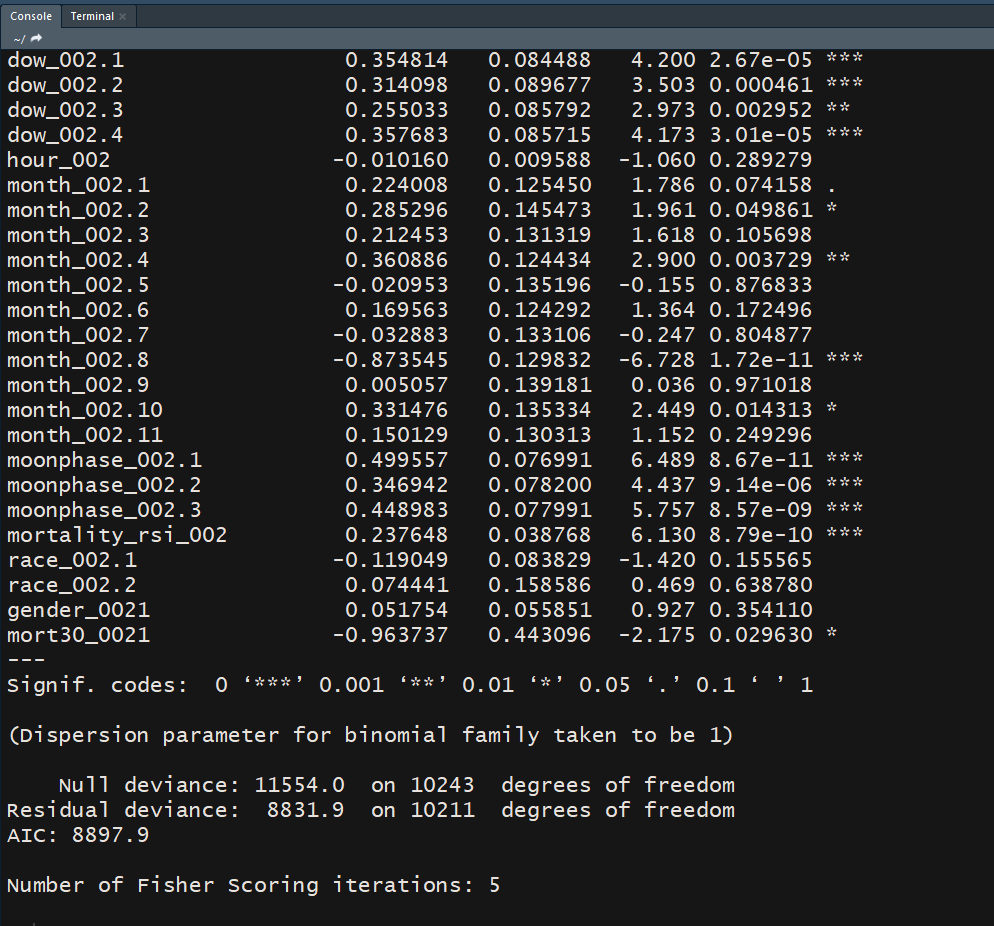
1. Build a binary logistic regression model to predict the probability of having complications of surgery of the patient based on the predictors. Comment on the overall model significance.

Ans:

model.2 = glm(complication\_002~.,data = training.2,family = 'binomial')

summary(model)





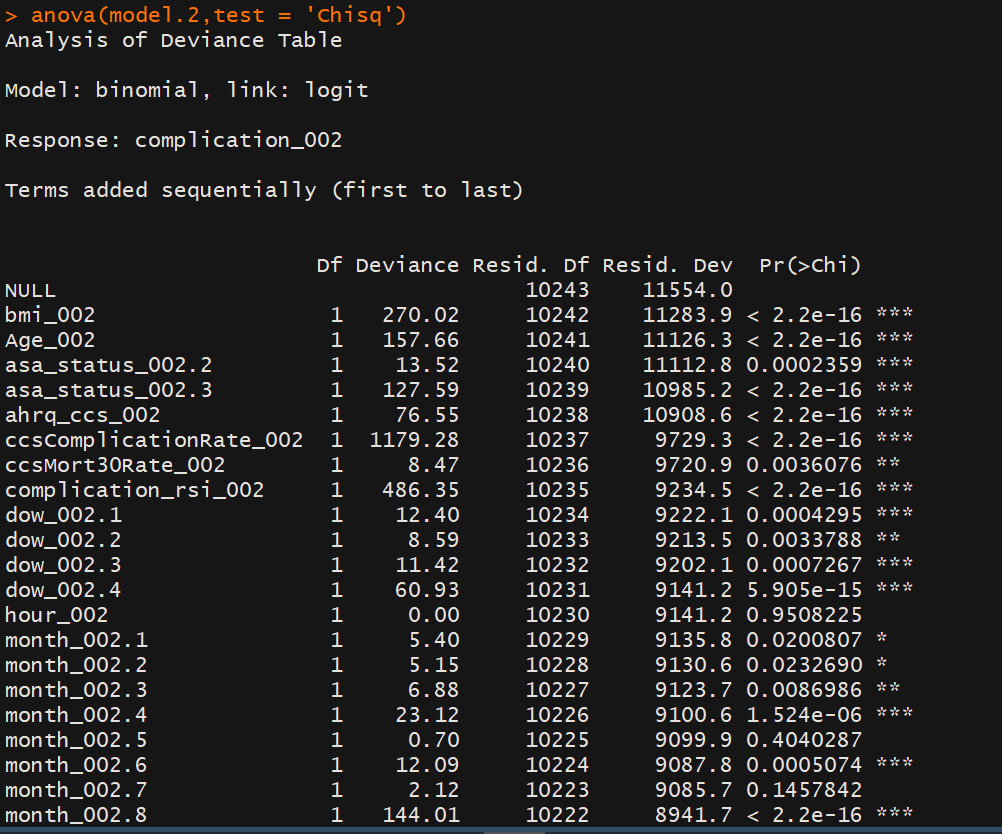
with(model.2,pchisq(null.deviance-deviance,df.null-df.residual,lower.tail = F))

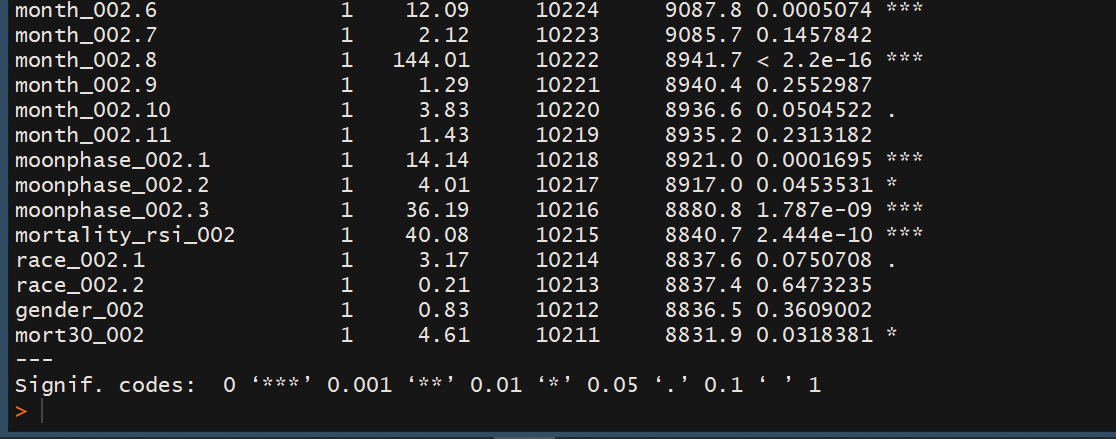


Comment: Above I have built a logistic regression model to predict the probability of having complication in surgery using all available predictors.

The model is overall significant, but there are many variables that are not significant

1. Find out and report which variables are statistically significant in the logistic regression model built in (i).





Interpretation: Above, we can see the significance of all the variables used. All the variables with p value of chi square less than .05 are significant predictors. These include:

"bmi\_002" "Age\_002" "asa\_status\_002.2"

"asa\_status\_002.3" "ahrq\_ccs\_002" "ccsComplicationRate\_002"

"ccsMort30Rate\_002" "complication\_rsi\_002" "dow\_002.1"

"dow\_002.2" "dow\_002.3" "dow\_002.4"

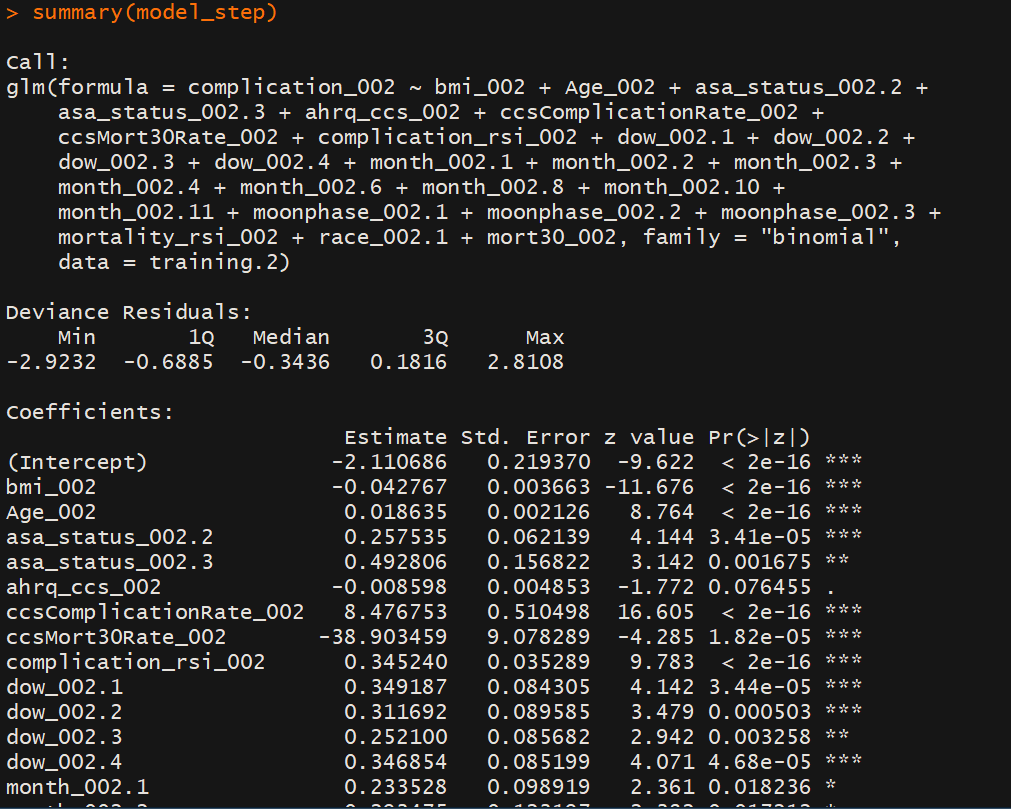
"month\_002.1" "month\_002.2" "month\_002.3" "month\_002.4" "month\_002.6" "month\_002.8" "month\_002.10" "moonphase\_002.1" "moonphase\_002.2" "moonphase\_002.3" “mortality\_rsi\_002" "race\_002.1"

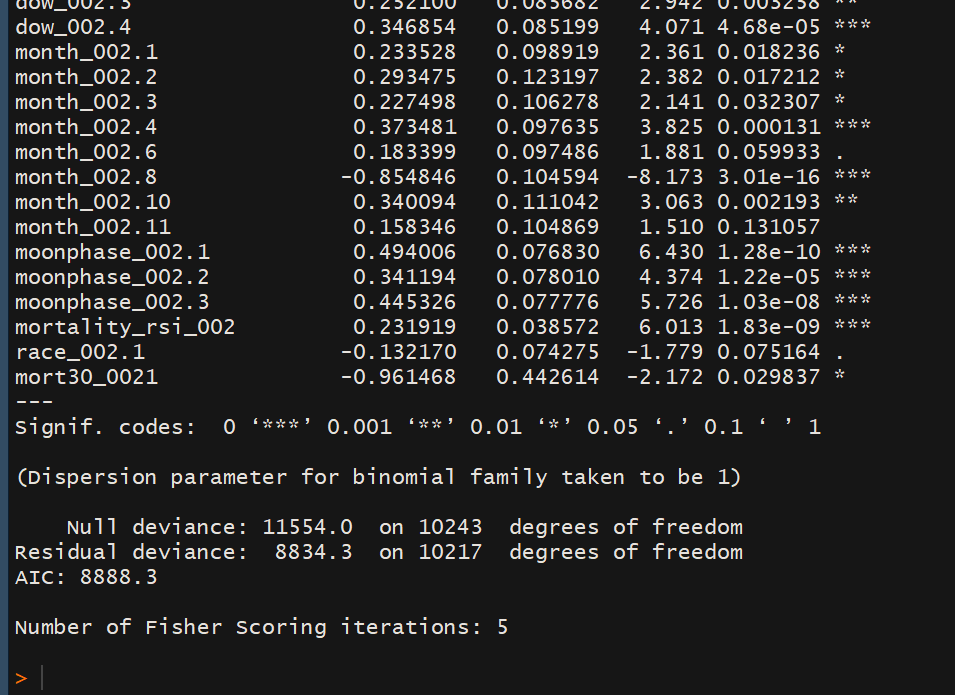
"mort30\_002" "complication\_002"

1. Build a new logistic regression model using only significant features.Report the model diagnostics followed to build this model.

Ans: model\_step = step(glm(complication\_002~.,data = dat\_transformed,family = 'binomial'))

summary(model\_step)





The model diagonostic used for this is AIC.

1. Write an estimated logistic regression model obtained in (iii)

Ans: The estimated logistic regression model is:

(similarly all the variables will be added)

1. Use Youden’s index to find the most optimal cut-off probability value for the best model chosen in (iii).

p = predict(model\_step,type = 'response')

p

actual = training.2$complication\_002

#function to calculate accuracy metrics and youden index

acc <- function(mod, pp, p,actual) {

out = c()

## Classification table

pred <- ifelse(p<pp,0,1)

# pred\_test <- ifelse(p\_test<pp,0,1)

tab<- table(pred,actual = actual)

out$sumtab<- addmargins(tab,FUN=sum)

TAP <- sum(tab[,2]) #Total actual positives

TAN <- sum(tab[,1]) # Total actual negatives

TP <- out$sumtab[2,2]

TN <- out$sumtab[1,1]

FP <- out$sumtab[2,1]

FN <- out$sumtab[1,2]

out$TPR = TP/TAP # Sensitivity or recall ## ability to correctly classify

out$FPR = FP/TAN

out$TNR = TN/TAN # Specificity

out$FNR = FN/TAP

out$accuracy = (TP+TN)/(TAN+TAP)

out$miss\_classification\_error = 1-out$accuracy

out$precision = TP/(TP+FP)

# conditional probability of being positive when predicted positive

out$specificity <- TN/TAN

out$f\_score = TP/(TP+0.5\*(FP+FN))

out$cut\_off = pp

out$youden = out$TPR+out$TNR-1

return(out)

}

acc(model\_step,0.5, p,actual)$youden

acc(model\_step,0.4, p,actual)$youden

acc(model\_step,0.3, p,actual)$youden

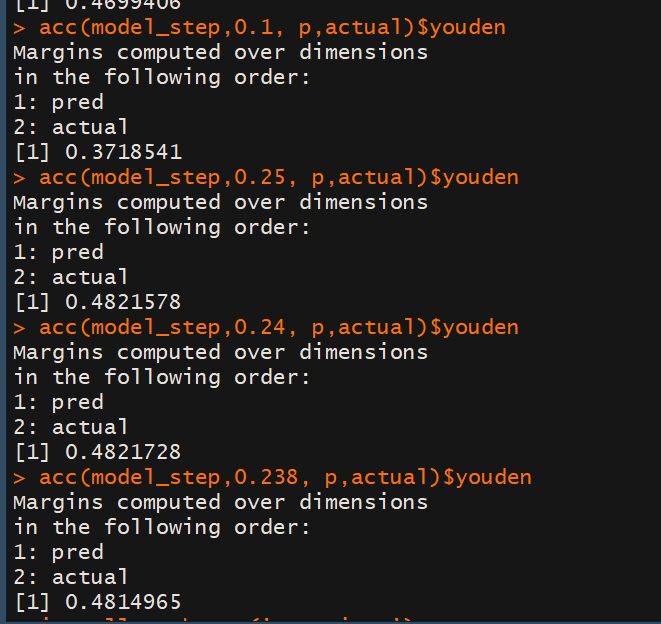
acc(model\_step,0.2, p,actual)$youden

acc(model\_step,0.1, p,actual)$youden

acc(model\_step,0.25, p,actual)$youden

acc(model\_step,0.24, p,actual)$youden

acc(model\_step,0.238, p,actual)$youden



The optimal youden index was found to be at cut off 0.24, with its max value 0.4831728

(c)Decision tree classifier model

# normalizing the variables

attach(data\_dummy.2)

Age\_002 = (Age\_002 - mean(Age\_002))/sd(Age\_002)

bmi\_002= (bmi\_002 - mean(bmi\_002))/sd(bmi\_002)

ccsComplicationRate\_002= ( ccsComplicationRate\_002- mean(ccsComplicationRate\_002))/sd(ccsComplicationRate\_002)

ccsMort30Rate\_002= (ccsMort30Rate\_002 - mean(ccsMort30Rate\_002))/sd(ccsMort30Rate\_002)

complication\_rsi\_002= ( complication\_rsi\_002- mean(complication\_rsi\_002))/sd(complication\_rsi\_002)

detach(data\_dummy.2)

# Test Train split

smp\_size.2<-floor(0.7\*nrow(data\_dummy.2))

set.seed(1024)

trainingdata.2 <- sample(seq\_len(nrow(data\_dummy.2)),size=smp\_size.2)

training.2<-data\_dummy.2[trainingdata.2,]

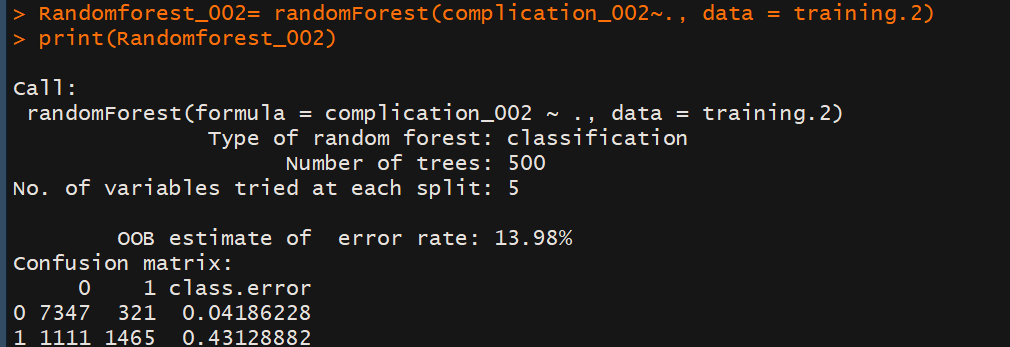
testing.2<-data\_dummy.2[-trainingdata.2,]

1. Build Random Forest decision tree and clearly identify and report predictor which is classifying the patient having complications of post-surgery.

Ranfor= randomForest(complication\_002~., data = training.2)

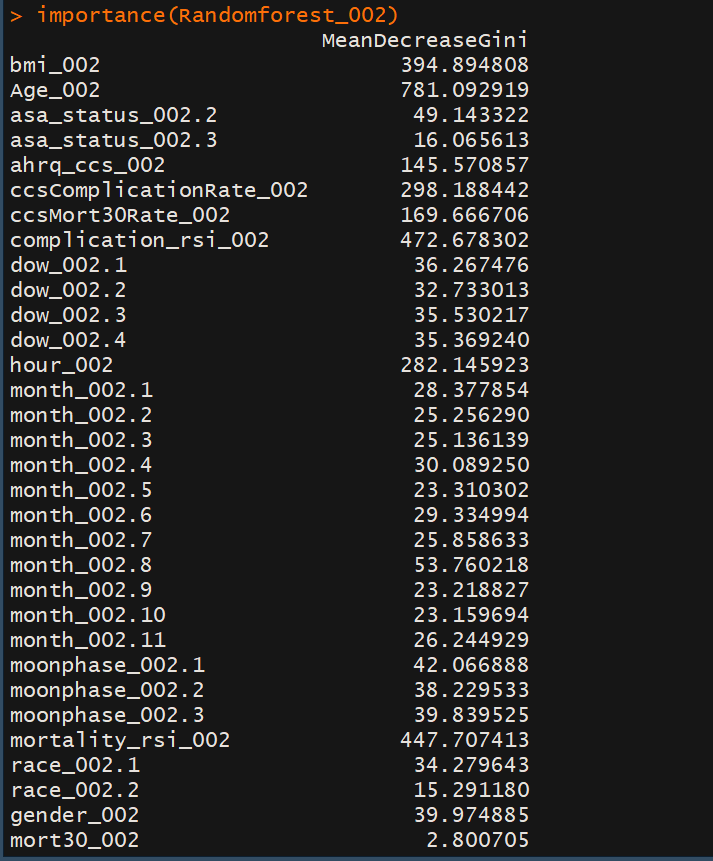
summary(Ranfor)

print(Ranfor)



Above is the random forest model.

Below are the variables used for the decision tree



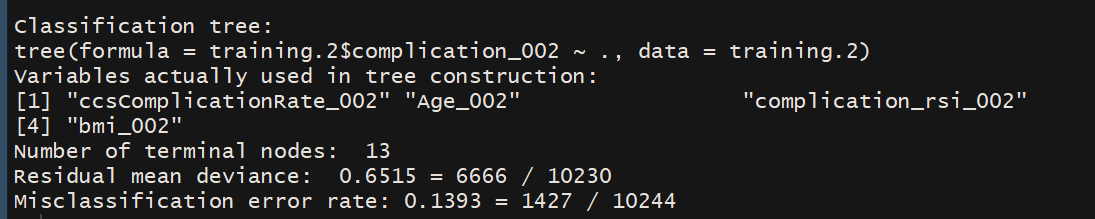
1. Also plot that decision tree and report the most important splitting criteria (rule).

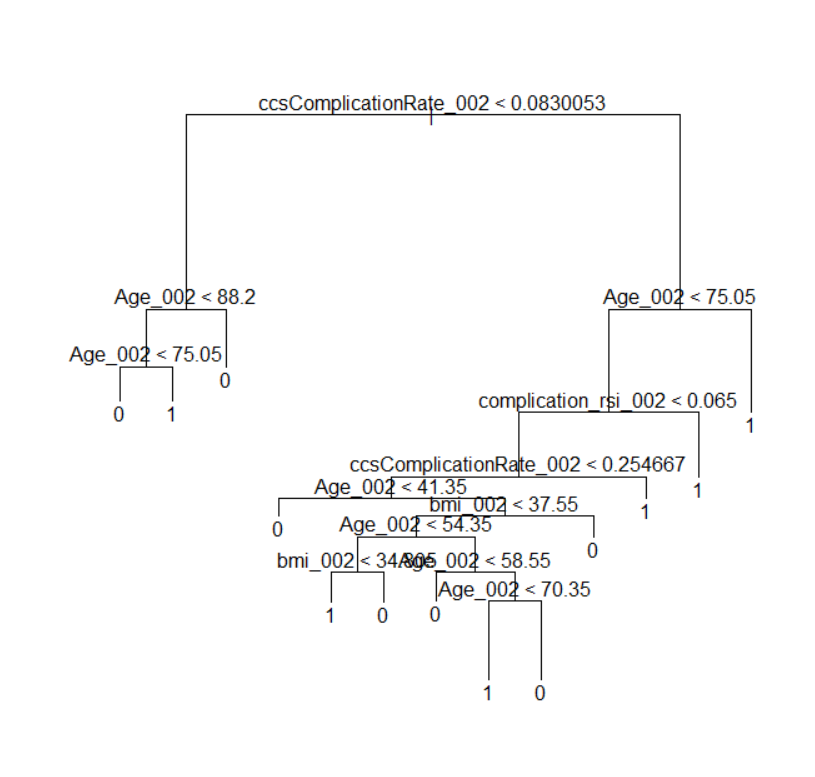
treemod.2= tree(training.2$complication\_002~., data=training.2)

summary(treemod.2)

plot (treemod.2)

text(treemod.2,pretty=0)

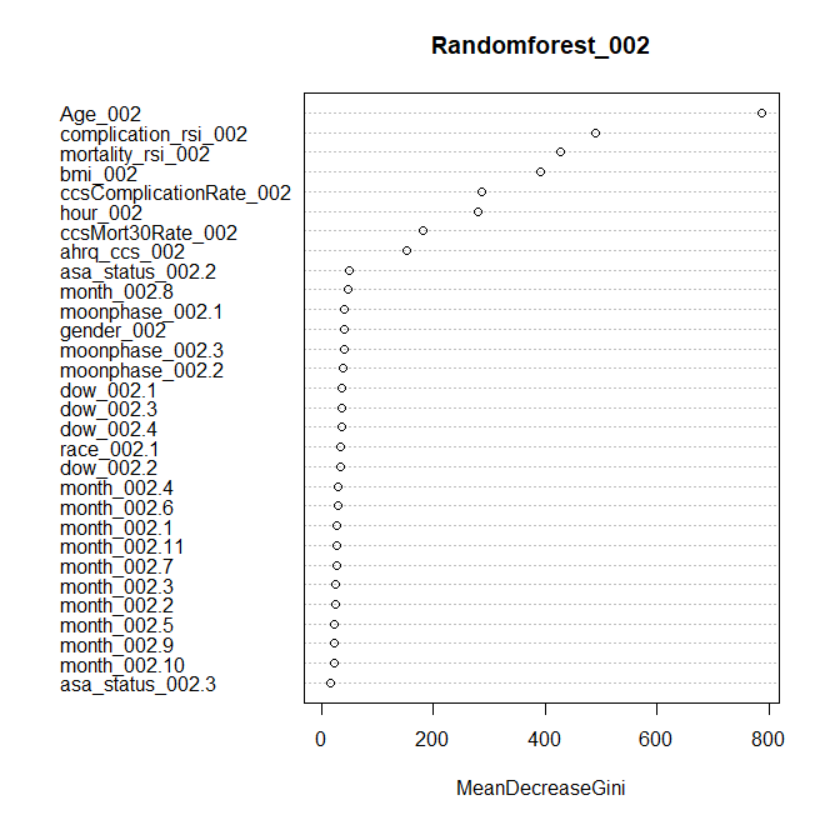




The most important splitting criteria is whether the cssComplilation rate is <0.083003 or not.

1. Use variance importance plot, report the variables which are classifying the surgery complication.

varImpPlot(Randomforest\_002)



The important variables for classifying are age, complication\_rsi, mortality\_rsi, bmi, ccsComplicationRate, hour, cssMort30Rate and ahrq\_ccs.

(d) Compare the logistic regression model and decision tree classifier performance using confusion matrix with specific accuracy measures or ROC and AUC?

acc <- function(mod, pp, p,actual) {

out = c()

## Classification table

pred <- ifelse(p<pp,0,1)

# pred\_test <- ifelse(p\_test<pp,0,1)

tab<- table(pred,actual = actual)

out$sumtab<- addmargins(tab,FUN=sum)

TAP <- sum(tab[,2]) #Total actual positives

TAN <- sum(tab[,1]) # Total actual negatives

TP <- out$sumtab[2,2]

TN <- out$sumtab[1,1]

FP <- out$sumtab[2,1]

FN <- out$sumtab[1,2]

out$TPR = TP/TAP # Sensitivity or recall ## ability to correctly classify

out$FPR = FP/TAN

out$TNR = TN/TAN # Specificity

out$FNR = FN/TAP

out$accuracy = (TP+TN)/(TAN+TAP)

out$miss\_classification\_error = 1-out$accuracy

out$precision = TP/(TP+FP)

# conditional probability of being positive when predicted positive

out$specificity <- TN/TAN

out$f\_score = TP/(TP+0.5\*(FP+FN))

out$cut\_off = pp

out$youden = out$TPR+out$TNR-1

return(out)

}

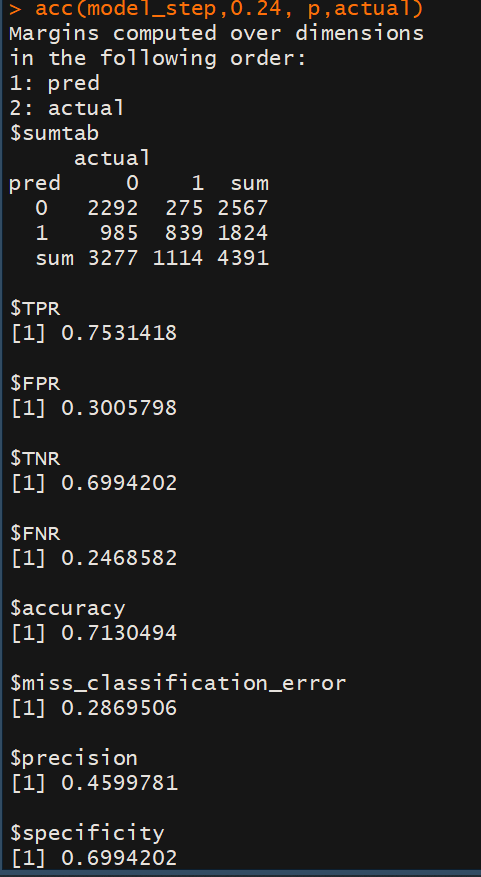
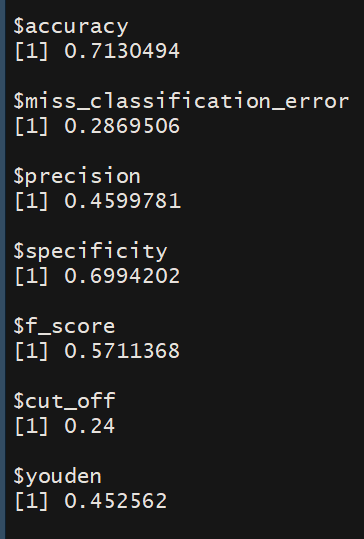
# Accuracy of logistic regression, with cut off probability = 0.24

acc(model\_step,0.24, p,actual)

p = predict(model\_step,type = 'response',newdata = testing.2)

p

actual = testing.2$complication\_002

# accuracy of random forest

p\_random = predict(Randomforest\_002,type = 'response',newdata = testing.2)

actual = testing.2$complication\_002

acc <- function(mod, p,actual) {

out = c()

## Classification table

#pred <- ifelse(p<pp,0,1)

pred = p

# pred\_test <- ifelse(p\_test<pp,0,1)

tab<- table(pred,actual = actual)

out$sumtab<- addmargins(tab,FUN=sum)

TAP <- sum(tab[,2]) #Total actual positives

TAN <- sum(tab[,1]) # Total actual negatives

TP <- out$sumtab[2,2]

TN <- out$sumtab[1,1]

FP <- out$sumtab[2,1]

FN <- out$sumtab[1,2]

out$TPR = TP/TAP # Sensitivity or recall ## ability to correctly classify

out$FPR = FP/TAN

out$TNR = TN/TAN # Specificity

out$FNR = FN/TAP

out$accuracy = (TP+TN)/(TAN+TAP)

out$miss\_classification\_error = 1-out$accuracy

out$precision = TP/(TP+FP)

# conditional probability of being positive when predicted positive

out$specificity <- TN/TAN

out$f\_score = TP/(TP+0.5\*(FP+FN))

#out$cut\_off = pp

out$youden = out$TPR+out$TNR-1

return(out)

}

