I have a dataset containing family information of married couples, which have around 10 variables & 600+ observations.

Independent variables are ~ gender, age, years married, children, religion etc.

I have one response variable which is number of extra marital affairs.

Now, I want to know what all factor influence the chances of extra marital affair.

Since extra marital affair is a binary variable (either a person will have or not),

so we can fit logistic regression model here to predict the probability of extra marital affair.

install.packages('AER')

data(Affairs,package="AER")

#importing the data

data <- read.csv(file.choose())

View(data)

str(data)

##Here we have two variable which are factor must changed to numeric

data$gender <- factor(data$gender,levels = c('male','female'),labels = c(1,0))

data$children <- factor(data$children,levels = c('yes','no'),labels = c(1,0))

View(data)

str(data)

attach(data)

##Here the dependent variable affairs and it has some categorical data we should make it binomial data

affairs1<- ifelse(affairs>0,1,0)

data <- cbind(affairs1,data)

View(data)

##Model building ##

table(affairs1)

model <- glm(affairs1~gender+children+religiousness+rating+education+yearsmarried+age+occupation,data = data,family = "binomial")

summary(model)

exp(coef(model))

pre <- predict(model,data,type = "response")

pre

# Confusion matrix and considering the threshold value as 0.5

confusion<-table(pre>0.5,data$affairs1)

confusion

# Model Accuracy

Accuracy<-sum(diag(confusion)/sum(confusion))

Accuracy #0. 765391

# Creating empty vectors to store predicted classes based on threshold value

pred\_values <- NULL

yes\_no <- NULL

pred\_values <- ifelse(pre>=0.5,1,0)

yes\_no <- ifelse(pre>=0.5,"yes","no")

# Creating new column to store the above values

data[,"pre"] <- pre

data[,"pred\_values"] <- pred\_values

data[,"yes\_no"] <- yes\_no

View(data[,c(1,11:13)])

table(data$affairs1,data$pred\_values)

##precision=0.75

##recall=0.16

##True positive Rate=25/150=0.16

##False positive Rate=16/451=

# Calculate the below metrics

# ROC Curve => used to evaluate the betterness of the logistic model

# more area under ROC curve better is the model

# We will use ROC curve for any classification technique not only for logistic

library(ROCR)

rocrpred<-prediction(pre,data$affairs1)

rocrperf<-performance(rocrpred,'tpr','fpr')

str(rocrperf)

plot(rocrperf,colorize=T,text.adj=c(-0.2,1.7))

# More area under the ROC Curve better is the logistic regression model obtained

## Getting cutt off or threshold value along with true positive and false positive rates in a data frame

str(rocrperf)

rocr\_cutoff <- data.frame(cut\_off = rocrperf@alpha.values[[1]],fpr=rocrperf@x.values,tpr=rocrperf@y.values)

colnames(rocr\_cutoff) <- c("cut\_off","FPR","TPR")

View(rocr\_cutoff)

library(dplyr)

rocr\_cutoff$cut\_off <- round(rocr\_cutoff$cut\_off,6)

# Sorting data frame with respect to tpr in decreasing order

rocr\_cutoff <- arrange(rocr\_cutoff,desc(TPR))

View(rocr\_cutoff)

