**setwd("C:/STONY/Practice/R (No.5)/adoption/adoption")**

**##########**

**## No.1 ##**

**##########**

**### read the cluster data 20 times**

**cls <- list()**

**names\_read <- paste0('cluster',1:20,'\_edge','.csv')**

**for (i in 1:20){**

**cls[[i]] <- read.csv(names\_read[i],header = FALSE)**

**}**

**### check if we read the data successfully**

**dim(cls[[1]])**

**### compute the group size**

**size <- list()**

**for (i in 1:20){**

**a <- cls[[i]]**

**size[[i]] <- dim(a)[1]**

**}**

**size = size %>% as.numeric()**

**### compute the network density of every group**

**# define a function to calculate the density**

**cls\_dens <- function(x){**

**col.num <- c()**

**a <- dim(x)[2]**

**for (i in 1:a){**

**col.num[i] <- sum(x[,i])**

**}**

**num <- sum(col.num)**

**density <- num/a%\*%a**

**return(density)**

**}**

**# use a for loop to get the density of every group**

**density <- list()**

**for (i in 1:20){**

**a <- cls[[i]]**

**density[[i]] <- cls\_dens(a)**

**}**

**density = density %>% as.numeric()**

**### compute the network coeffient**

**library(igraph)**

**library(dplyr)**

**cls\_coeff <- list()**

**for (i in 1:20){**

**a <- cls[[i]] %>% as.matrix()**

**cls\_coeff[[i]] <- graph\_from\_adjacency\_matrix(a, mode = 'undirected') %>%**

**transitivity(type = 'undirected')**

**}**

**cls\_coeff = cls\_coeff %>% as.numeric()**

**### generate a new dataframe**

**group <- c(1:20)**

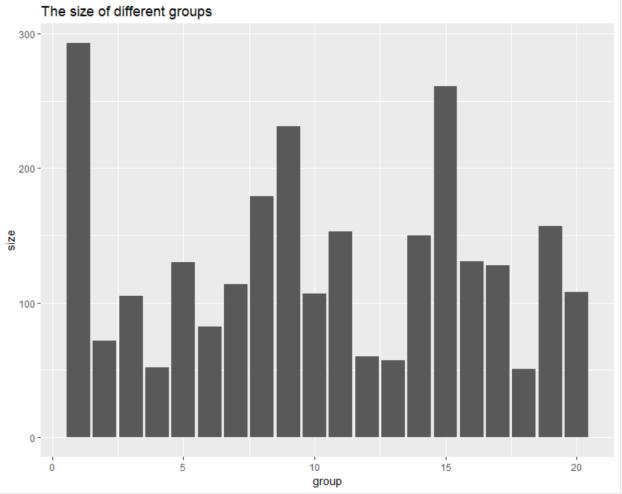
**all\_grp <- cbind(size,density,cls\_coeff,group) %>% as.data.frame()**

**### figure out the size**

**ggplot(all\_grp,aes(group,size)) +**

**geom\_bar(stat="identity",position="dodge") +**

**ggtitle("The size of different groups")**

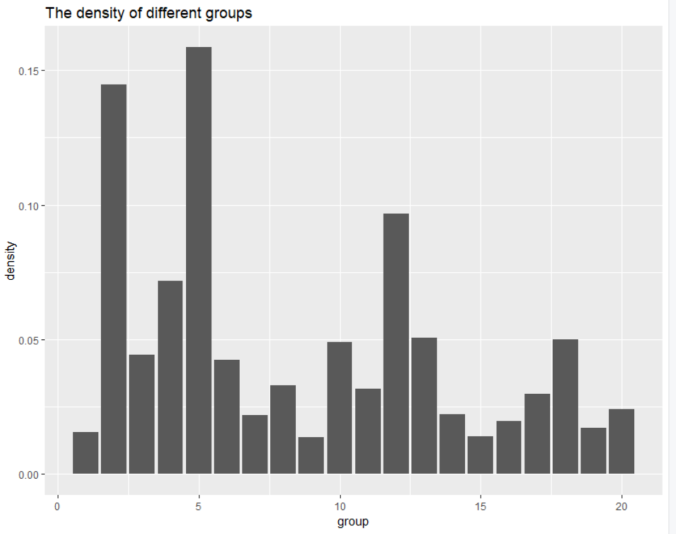
**# output: **

**### group 5 hsa the highest density**

**ggplot(all\_grp,aes(group,density)) +**

**geom\_bar(stat="identity",position="dodge") +**

**ggtitle("The density of different groups")**

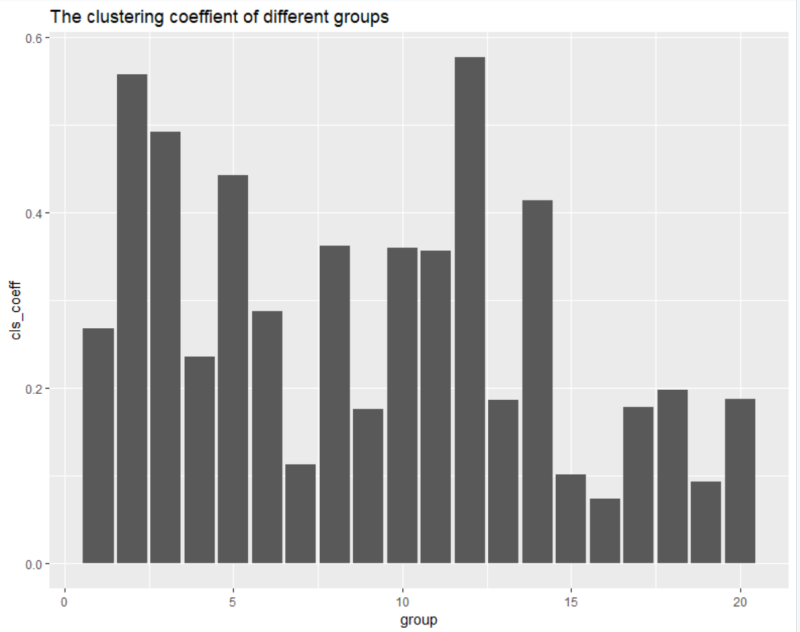
**# output:** ****

**### group 12 has the highest clustering coefficient**

**ggplot(all\_grp,aes(group,cls\_coeff)) +**

**geom\_bar(stat="identity",position="dodge") +**

**ggtitle("The clustering coeffient of different groups")**

**# output:** ****

**##########**

**## No.2 ##**

**##########**

**### read the sample data 20 times**

**sample <- list()**

**names\_read1 <- paste0('sample',1:20,'.csv')**

**for (i in 1:20){**

**sample[[i]] <- read.csv(names\_read1[i],header = TRUE)**

**}**

**dim(sample[[1]])**

**adoption <- list()**

**for (i in 1:20){**

**a <- sample[[i]]**

**adoption[[i]] <- sum(a[,1])**

**}**

**adoption <- adoption %>% as.numeric()**

**### generate the new dataframe**

**grp\_ado\_net <-cbind(group,size,density,cls\_coeff,adoption) %>% as.data.frame()**

**str(grp\_ado\_net)**

**### build the model to see if density and network coefficient have impacts on adoption**

**### here we can see the size and density are more important than coefficient through the output below.**

**model <- {adoption~size+density+cls\_coeff}**

**fit <- lm(model,data=grp\_ado\_net)**

**summary(fit)**

**anova(fit,test="Chisq")**

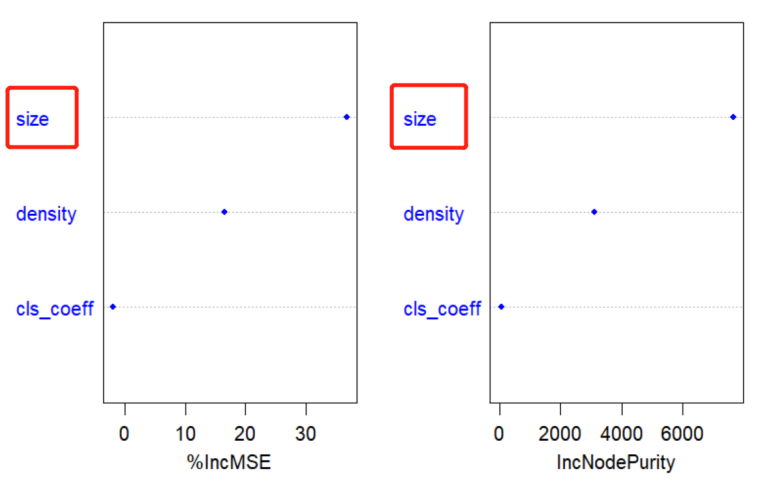
**step(fit,direction="both")**

**set.seed(1234)**

**library(randomForest)**

**bk.rf.fit <- randomForest(model,data=grp\_ado\_net,mtry=3,ntree=1000,importance=TRUE)**

**varImpPlot(bk.rf.fit,color="blue",pch=20,cex=1.25,main="")**

**# output:** ****

The

**##########**

**## No.3 ##**

**##########**

**### here we choose the data of group 3**

**cls\_3 <- cls[[3]] %>% as.matrix()**

**sample\_3 <- sample[[3]] %>% as.matrix()**

**### calculate the centrality of every group**

**net\_3 <- list()**

**for (i in 1:105){**

**net\_3[[i]] <- sum(cls\_3[i,])**

**}**

**net\_3 <- net\_3 %>% as.numeric()**

**### generate a new data frame which contains the sample and centrality**

**data\_3 <- cbind(sample\_3,net\_3) %>% as.data.frame()**

**## plot the network distribution #**

**## generate the adjacency matirx**

**m=as.matrix(cls\_3)**

**igraph=graph\_from\_adjacency\_matrix(m,mode="undirected",weighted=NULL,diag=FALSE)**

**## define the color**

**color<-sample[[3]][,1]**

**b <- data.frame(ado=color)**

**V(igraph)$name <- c(1:105)**

**V(igraph)$ado=as.character(b[V(igraph)$name,])**

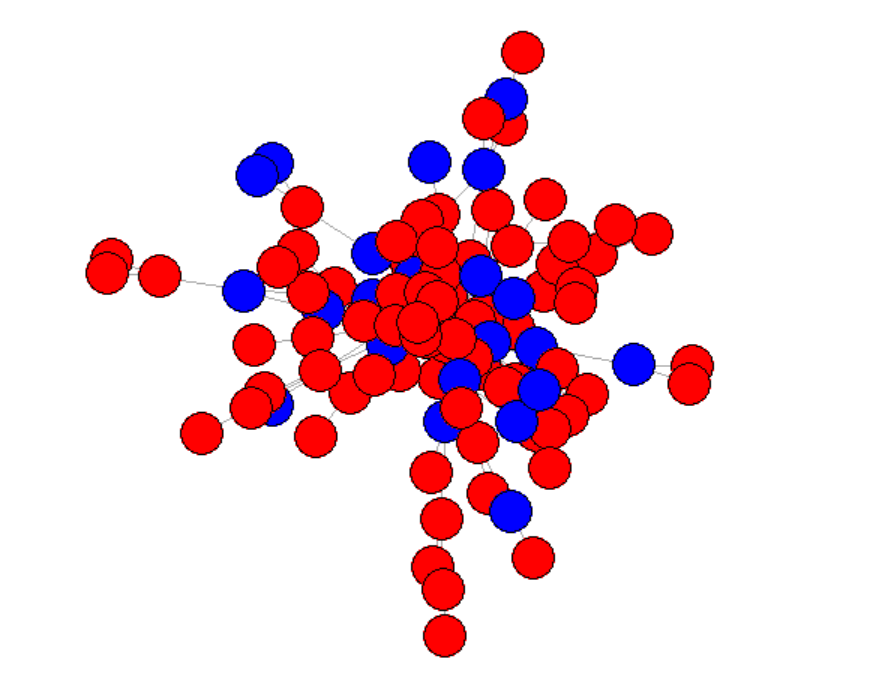
**head(V(igraph)$ado)**

**V(igraph)$color=V(igraph)$ado**

**V(igraph)$color=gsub("0","red",V(igraph)$color) #0 will be red**

**V(igraph)$color=gsub("1","blue",V(igraph)$color) #1 will be blue**

**plot.igraph(igraph,vertex.label=NA,layout=layout.fruchterman.reingold)**

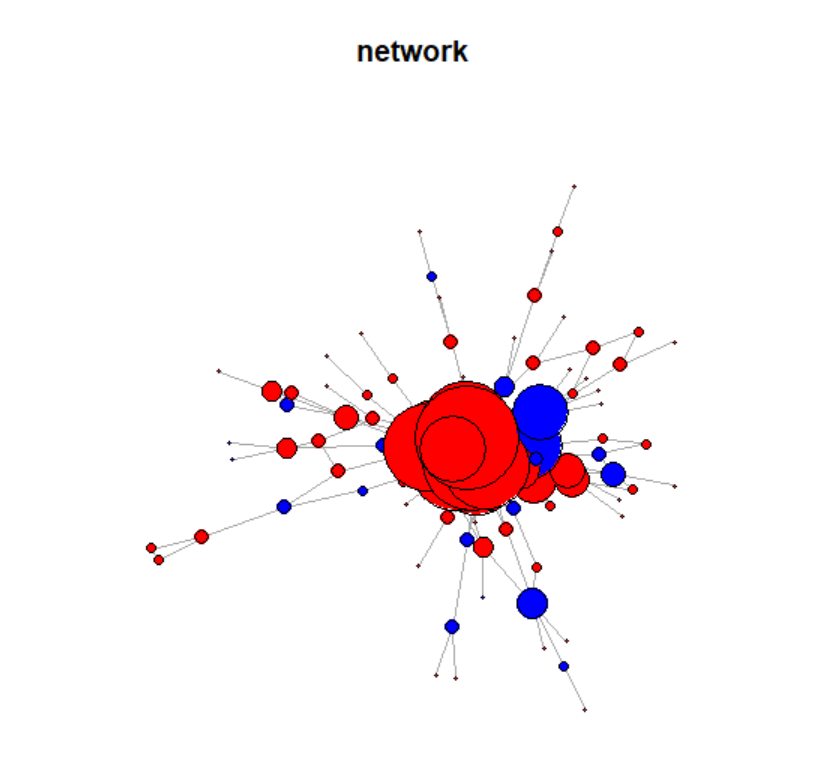
****

**## define the size**

**set.seed(123)**

**V(igraph)$size=degree(igraph)\*2**

**plot.igraph(igraph,main="network",vertex.label=NA,layout=layout.fruchterman.reingold)**

****

**### out put : we have changed the color and size of nodes, however, it seems the nodes in the center have large centrality.**

**##########**

**## No.4 ##**

**##########**

**### degree centrality ###**

**### here we choose the group 3 and combine the adoption and centrality as a new dataframe**

**data\_3 <- cbind(sample\_3,net\_3) %>% as.data.frame()**

**### here we use the logistic regression to figure out if the centrality has an impact on adoption**

**model <- {adoption~net\_3+smart+gender}**

**fit <- glm(model,family=binomial,data=data\_3)**

**summary(fit)**

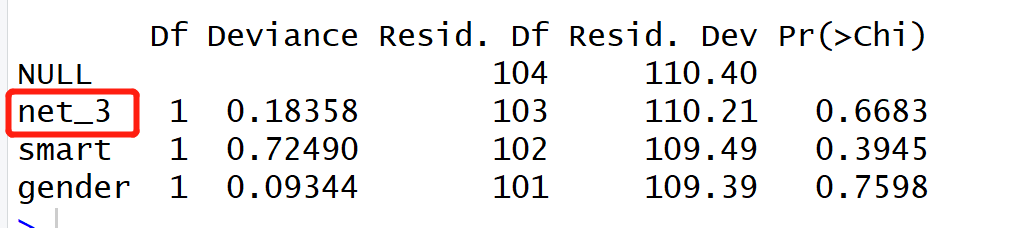
**anova(fit,test="Chisq")**

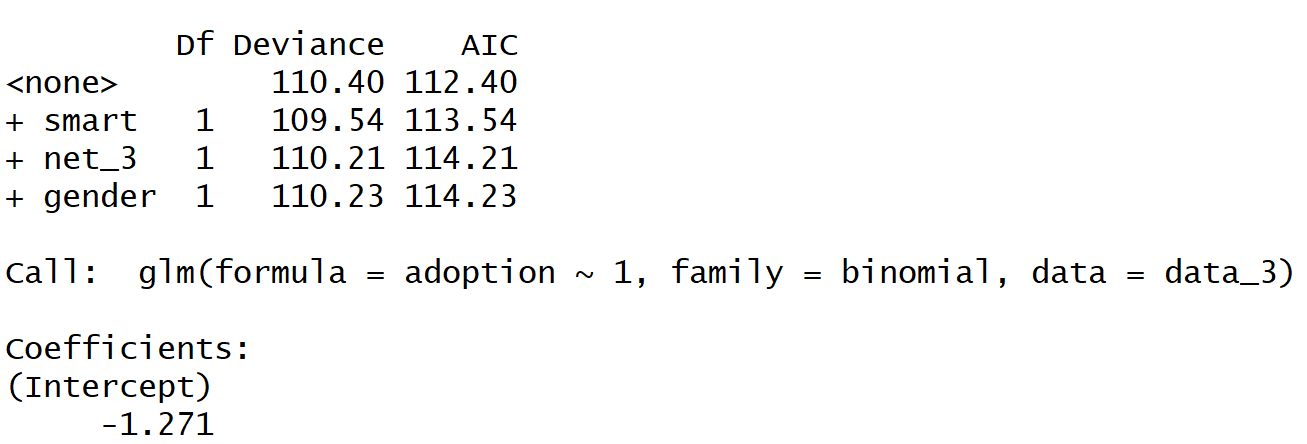
**step(fit,direction="both")**

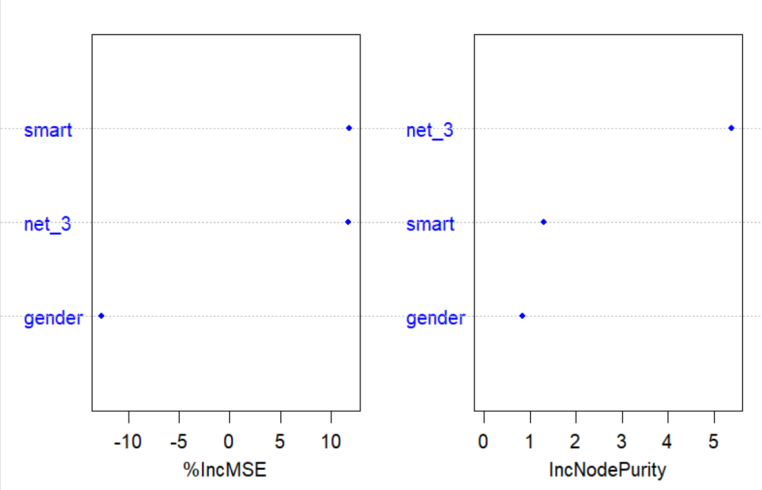
**set.seed(1234)**

**bk.rf.fit <- randomForest(model,data=data\_3,mtry=3,ntree=1000,importance=TRUE)**

**varImpPlot(bk.rf.fit,color="blue",pch=20,cex=1.25,main="")**

****

****

****

**### out put : our result obviously shows that the centrality doesn’t have an important impact on the adoption in Samsung Note II. However, through the random forest model, we see that net\_3 is better than gender but is not better than smart phone usage to some degree.**

**##########**

**## No.5 ##**

**##########**

**##### LOGISTIC REGRESSION #####**

**### get the train data**

**## get the centrality of whole 20 groups**

**a <- list()**

**net\_20 <- list()**

**for (i in 1:20){**

**a[[i]] <- cls[[i]] %>% as.matrix()**

**b <- list()**

**for (j in 1:dim(a[[i]])[1]){**

**b[[j]] <- sum(a[[i]][j,])**

**b <- b %>% as.data.frame()**

**}**

**net\_20[[i]] <- b**

**#net\_20[[i]] <- net\_20[[i]] %>% as.data.frame()**

**net\_20[[i]] <- t(net\_20[[i]]) %>% as.numeric() %>% as.matrix()**

**}**

**### put the result into numeric format**

**for (i in 1:20){**

**net\_20[[i]] <- net\_20[[i]] %>% as.numeric()**

**}**

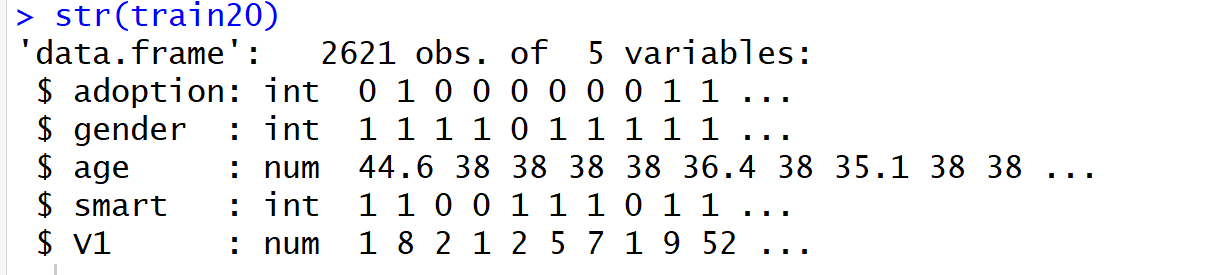
**train <- unlist(net\_20) %>% as.matrix() %>% as.data.frame()**

**### get the train data by combining the 20 groups**

**a <- bind\_rows(sample)**

**train20 <- cbind(a,train)**

**str(train20)**

**### out put : **

**### get the logit model of training set**

**traindata <- train20[1:1747,]**

**traindata$adoption %>% as.factor()**

**model <- {adoption ~ gender+age+smart+V1}**

**traindata.lr.fit <- glm(model,family=binomial,data=traindata)**

**### in order to see the accuracy of model, we will plot the ROC curve**

**# area under ROC curve for TRAINING data**

**library(lattice) # lattice plot**

**library(vcd) # mosaic plots**

**library(ROCR) # ROC curve objects for binary classification**

**traindata$lr.predprob <- predict(traindata.lr.fit,type="response")**

**traindata.lr.pred <- prediction(traindata$lr.predprob,traindata$adoption)**

**traindata.lr.auc <- as.numeric(performance(traindata.lr.pred,"auc")@y.values)**

**# area under ROC curve for TEST data**

**testdata <- train20[1748:2621,]**

**testdata$lr.predprob <- as.numeric(predict(traindata.lr.fit,**

**testdata,type="response"))**

**testdata.lr.pred <- prediction(testdata$lr.predprob,testdata$adoption)**

**testdata.lr.auc <- as.numeric(performance(testdata.lr.pred,"auc")@y.values)**

**# genertae a function to plot the ROC**

**plot.roc <- function(train.roc,train.auc,test.roc,test.auc) {**

**plot(train.roc,col="blue",lty="solid",main="",lwd=2,**

**xlab="False Positive Rate",ylab="True Positive Rate")**

**plot(test.roc,col="red",lty="dashed",lwd=2,add=TRUE)**

**abline(c(0,1))**

**train.legend <- paste("Training AUC = ",round(train.auc,digits=3))**

**test.legend <- paste("Test AUC = ",round(test.auc,digits=3))**

**legend("bottomright",legend=c(train.legend,test.legend),**

**lty=c("solid","dashed"),lwd=2,col=c("blue","red"))**

**}**

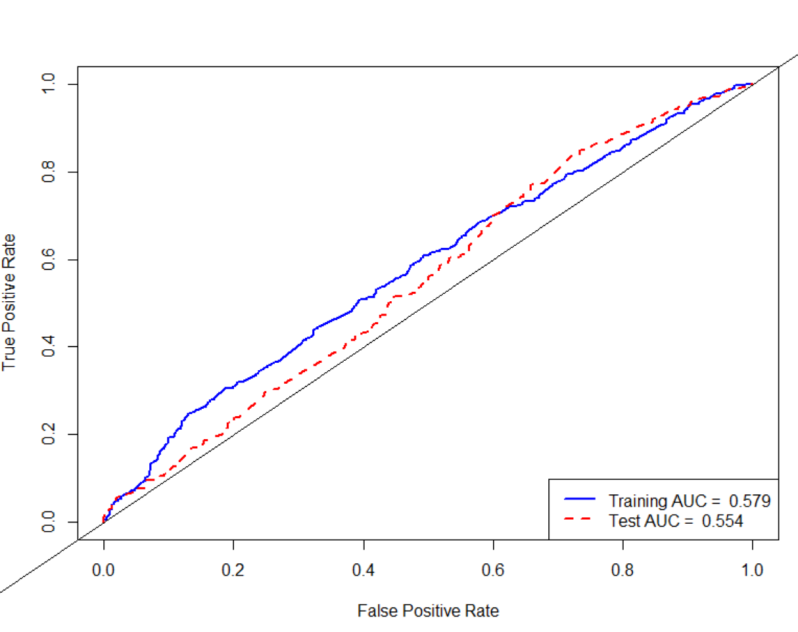
**# ROC for logistic regression to know the accuracy**

**traindata.lr.roc <- performance(traindata.lr.pred,"tpr","fpr")**

**testdata.lr.roc <- performance(testdata.lr.pred,"tpr","fpr")**

**plot.roc(train.roc=traindata.lr.roc,train.auc=traindata.lr.auc,**

**test.roc=testdata.lr.roc,test.auc=testdata.lr.auc)**

**### out put : **

**### get the prediction data**

**## get the cluster data of 20 groups**

**setwd("C:/STONY/Practice/R (No.5)/prediction/prediction")**

**### read the prediction data from group 21 to 40**

**pred <- list()**

**names\_read2 <- paste0('prediction',21:40,'.csv')**

**for (i in 1:20){**

**pred[[i]] <- read.csv(names\_read2[i],header = TRUE)**

**}**

**str(pred[[3]])**

**dim(pred\_cls[[3]])**

**### read the clustering data from group 21 to 40**

**pred\_cls <- list()**

**names\_read3 <- paste0('cluster',21:40,'\_edge','.csv')**

**names\_read3[1]**

**for (i in 1:20){**

**pred\_cls[[i]] <- read.csv(names\_read3[i],header = FALSE)**

**}**

**## get the centrality of group 21-40**

**pre\_cen <- list()**

**net\_40 <- list()**

**for (i in 1:20){**

**pre\_cen[[i]] <- pred\_cls[[i]] %>% as.matrix()**

**b <- list()**

**for (j in 1:dim(pre\_cen[[i]])[1]){**

**b[[j]] <- sum(pre\_cen[[i]][j,])**

**b <- b %>% as.data.frame()**

**}**

**net\_40[[i]] <- b**

**net\_40[[i]] <- t(net\_40[[i]]) %>% as.numeric() %>% as.matrix()**

**}**

**for (i in 1:20){**

**net\_40[[i]] <- net\_40[[i]] %>% as.numeric()**

**}**

**### get the test data by combining the 20 groups**

**pred\_40\_c <- list() # to save the data from net\_40**

**pred\_40\_s <- list() # to save the data from pred**

**k <- list() # to save the result of combing net\_40 an pred**

**for (i in 1:20){**

**pred\_40\_c[[i]] <- net\_40[[i]]**

**pred\_40\_s[[i]] <- pred[[i]]**

**k[[i]] <- cbind(pred\_40\_s[[i]],V1=pred\_40\_c[[i]])**

**}**

**#### PREDICTION !**

**# get the prediction for adoption for every group. if the probability is less than 0.35, we say the adoption is 1, else we say it’s 0.**

**adop <- c()**

**h <- c()**

**for (i in 1:20){**

**for (j in 1:dim(k[[i]])[1]){**

**h[j] <- predict(traindata.lr.fit,newdata=k[[i]][j,],type="response")**

**if(h[j] >= 0.35){**

**h[j] = 1**

**}**

**else{**

**h[j] = 0**

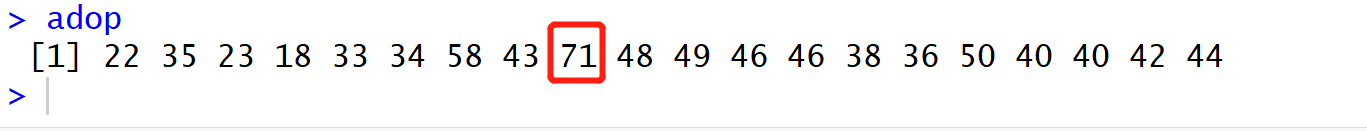
**}**

**}**

**adop[[i]] <- sum(h)**

**}**

**adop**

**### out put : **

**## we can see group 29 has the biggest number of adoption**

**## and the top 10 groups are 7-13,16,19-20**