Final project

Mehrab Atighi

7/6/2021

1)Library Data

for library Data we need to install kernlab packages.

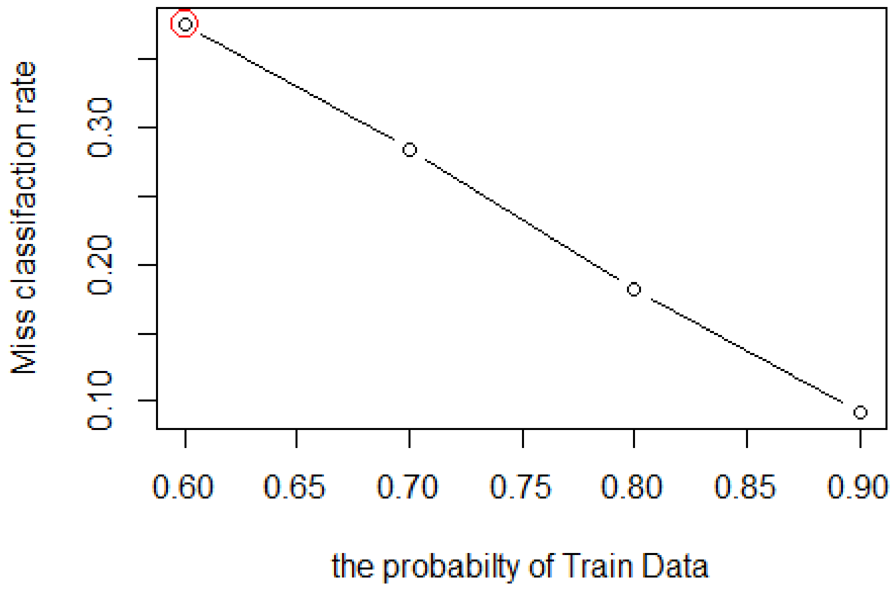
```
memory.limit(size=9999999999999)

## [1] 1e+13

#install.packages("kernlab")
library(kernlab)
data("spam")
Data<-spam
#View(Data)
attach(Data)
Data$type=ifelse(Data$type=="spam" , 1 ,0)
```

2) Make Model with diffrents methods

A) Validation set approch method

at the first we test it with Validation set approch method it means that we should get for example 60% and 40% of Data for Train and Test Data, after that we make model with train data and calculate the miss classification rate and other Criterion. Note: here we are calculate for 0.6,0.7,0.8,0.9 probabiltes for train data.

```
set.seed(2)
prob=c(0.6,0.7,0.8,0.9)
Miss.A.a<-c()
j=1
for(i in prob){
sample<-sample(c(TRUE , FALSE ) , nrow(Data) , replace = T , prob=c(i,1-i))
train.a<-Data[sample,]
test.a<-Data[!sample,]

#Multiclass logstic regression on validation set approch method
fit.A.a<-glm(type~. ,data = train.a , family = binomial)
predict.A.a<-ifelse(predict.glm(fit.A.a ,newdata = test.a , type=
"response")>0.5,1,0)
Miss.A.a[j]<-sum(as.numeric(predict.A.a==test.a$type))/nrow(Data)
j=j+1
}
Miss.A.a

## [1] 0.37600522 0.28428602 0.18169963 0.09193654

plot(prob,Miss.A.a,type="b" ,xlab="the probabilty of Train Data",ylab="Miss
classifaction rate")
points(prob[which.max(Miss.A.a)] ,max(Miss.A.a) ,col= "red" , cex=2)
```



According to this plot we can say that the Maximum of Miss classification rate for the probability of Train Data equal to 0.6 and the miss classification rate is equal to 0.3760 .

B) Leave one out cross validation method

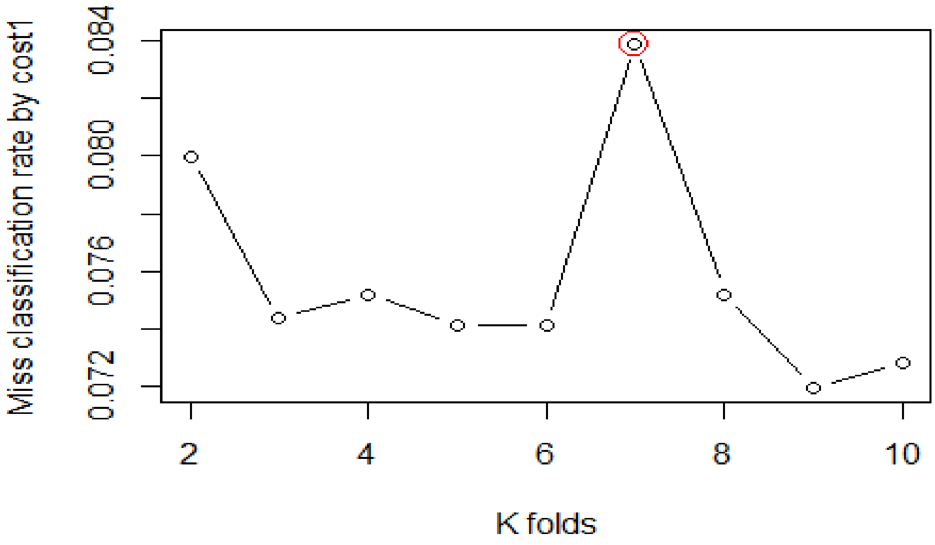
Now we want to select one of the (response , X(vectors)) paired for test Data and select another Data for Train Data after that we use these on muliclass logstic regression. after that we want to calculate the Miss classification rate and cost function and another rate for Criterion of this method. Note: the third method working with rock curve.

```
#Now here we want to test it with two cost function
set.seed(1)
library(boot)
fit.B.a<-glm(type~. ,data = Data ,family = binomial)
#first Criterion function:
cost1<- function(r, pi = 0) {mean(abs(r - pi) > 0.5)}
#second Criterion function:
cost2<- function(labels,pred){
mean(labels==ifelse(pred > 0.5, 1, 0))}

predict.B.a<-ifelse(predict.glm(fit.B.a , newdata = test.a , type =
"response")>0.5 ,1 ,0)
#Now we want to calculate these Criterion.
Miss.B.a<-cv.glm(Data , fit.B.a ,cost1)
Miss.B.b<-cv.glm(Data , fit.B.a ,cost2)



#third Criterion function:
#install.packages("pROC")
library(pROC)

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

LOOCV.AUC<-function(Dataset,Responsecolumn,Formula="type~."){

  #first create two vectors to hold the predictions and the responses and a
matrix to hold the final results
  rows<-nrow(Dataset)
  predictions<-rep(NA,rows)
  responses<-rep(NA,rows)
  results<-matrix(nrow=rows,ncol=3)
  #Now run the model, each time omitting the ith row of Dataset, then predict
on the ith row of Dataset
  for (i in 1:rows){
    model<-glm(Formula,data=Dataset[-i,],family=binomial)
    predictions[i]<-predict(model,Dataset[i,],type="response")
    responses[i]<-Dataset[i,Responsecolumn]
    }
  roc(responses,predictions)
  }
#Now we want to calculate third Criterion and compare it with another
Criterions
Loocv.B.a<-LOOCV.AUC(Dataset = Data , Responsecolumn = 58 )

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

Miss.B.a$delta[1]

## [1] 0.07281026

Miss.B.b$delta[1]

## [1] 0.9271897
```

C) K Fold Cross validation method

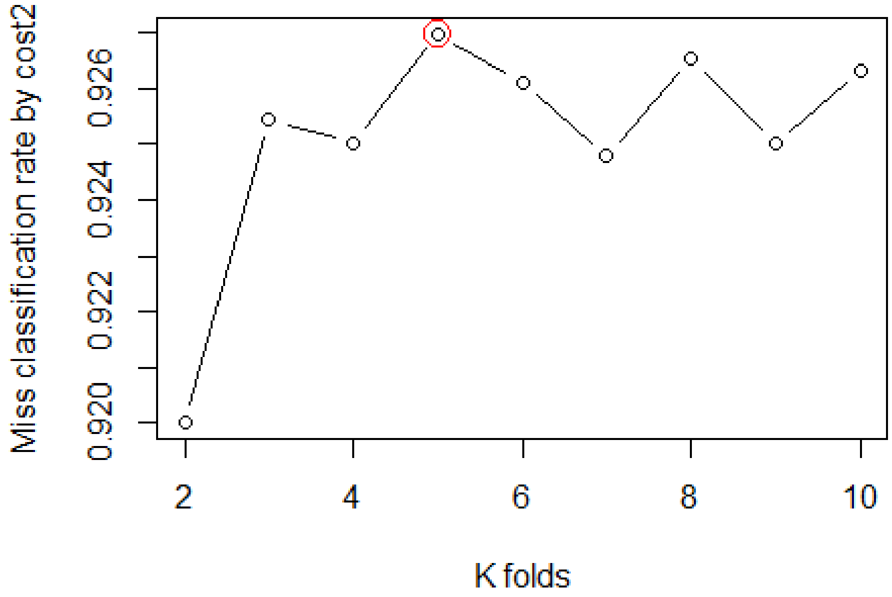
Now we want to select k(5 to 10 folds) subsets of Data and at the first, first sets is test data and another are train data in the next step we choose the second subset for test data and another train data and ... . Note: here we are do this method for k=2,3,...,10 after that we calculate and cost1 and cost2 Criterions values and compare these in plots.

```
set.seed(2)
fit.C.a<-glm(type~. ,data=Data,family = binomial)
q=1
Miss.C.a<-c()
Miss.C.b<-c()
cost1<- function(r, pi = 0) {mean(abs(r - pi) > 0.5)}
cost2<- function(labels,pred){
mean(labels==ifelse(pred > 0.5, 1, 0))}
for(i in c(2:10)){
Miss.C.a[i-1]<-cv.glm(Data,fit.C.a ,cost1, K= i)$delta[1]}
for(i in c(2:10)){
Miss.C.b[i-1]<-cv.glm(Data,fit.C.a ,cost2, K= i)$delta[1]}
Miss.C.a

## [1] 0.07998261 0.07433167 0.07520104 0.07411432 0.07411432 0.08389481
0.07520104
## [8] 0.07194088 0.07281026

plot(2:10 , Miss.C.a , xlab="K folds" , ylab="Miss classification rate by
cost1" , type="b")
points(which.max(Miss.C.a)+1 , max(Miss.C.a) , cex=2 , col="red")
```



according to this plot we can say that the k=7 have maximum of Miss classification rate by cost1 and its equal to 0.8389 .

```
plot(2:10 , Miss.C.b , xlab="K folds" , ylab="Miss classification rate by
cost2" , type="b")
points(which.max(Miss.C.b)+1 , max(Miss.C.b) , cex=2 , col="red")
```



```
Miss.C.b

## [1] 0.9200174 0.9254510 0.9250163 0.9269724 0.9261030 0.9247990 0.9265377
## [8] 0.9250163 0.9263204
```

according to this plot we can say that the k=5 have maximum of Miss classification rate by cost2 and its equal to 0.9269 .

D) backward subset selection

According to this method we have 3 step :

step 1)

Make a full(58 predictors) model with M0 name.

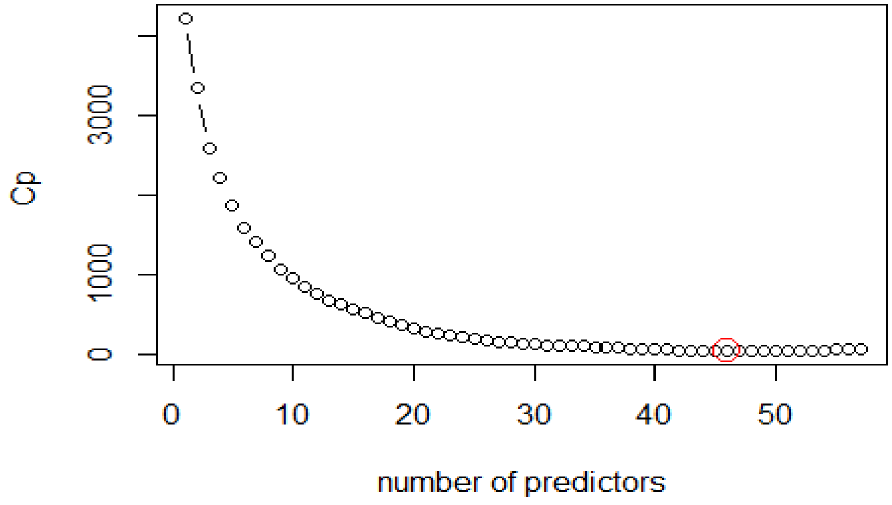
step 2)

we delet a predictor from M0 that have lowest correlation and relationship with our response and the best model after remove is M1 with 57 predictors. and do this until M58 that is empty model(with out any predictor).

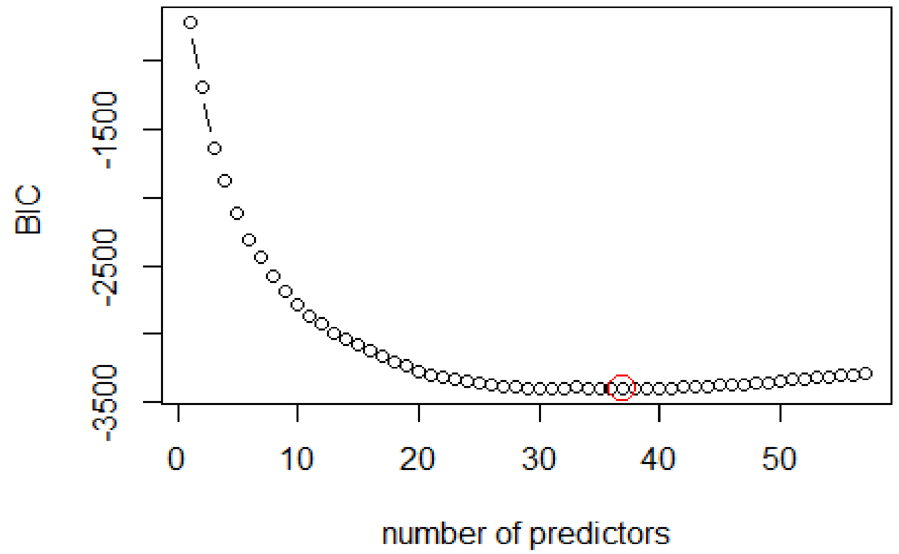
step 3)

According to our Criterions we choose the best model. Note: Criterions are including BIC , Cp , Adjust R squred for each Criterion we choose the best method.

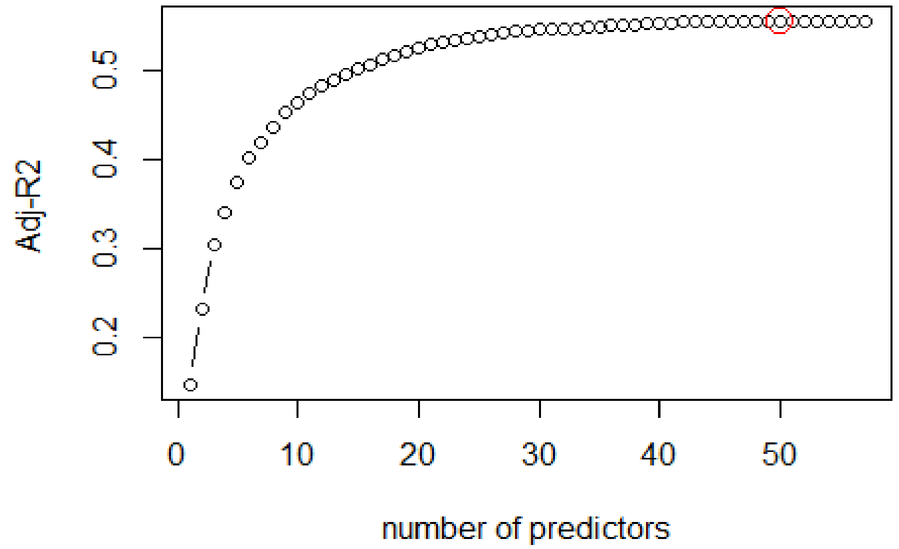
```
set.seed(1)
library(leaps)
backward.subset<-regsubsets(type~ . ,data = Data ,nvmax=58 , method =
"backward")
result<-summary(backward.subset)
# as there bottem line have alot of outputs i make is as comment it show that
in each step how it choose the best subset.
#result
#Now we want to plot these models with diffrents criterion.
par(mfrow=c(1,1))
plot(1:length(result$cp) ,result$cp , xlab="number of predictors" , ylab="Cp
"   , type = "b")
points(which.min(result$cp) , min(result$cp) ,col="red" , cex=2)
```



```
plot(1:length(result$bic) ,result$bic , xlab="number of predictors" ,
ylab="BIC "  , type = "b")
points(which.min(result$bic) , min(result$bic) ,col="red" , cex=2)
```



```
plot(1:length(result$adjr2) ,result$adjr2 , xlab="number of predictors" ,
ylab="Adj-R2 "  , type = "b")
points(which.max(result$adjr2) , max(result$adjr2) ,col="red" , cex=2)
```



we know that the minimum of BIC, Cp cirterions are the better model and the maximum of the adjust R squred is better model too. Now here we can see that the backward model selection choose the Model with 50 predictors as the better(with adjust R squred corterion.) bottem we can see this predictor names and Coefficients values.

```
coef(backward.subset , id=which.max(result$adjr2))
```

```
##       (Intercept)              make           address               all 
##      2.011350e-01     -5.012417e-02     -1.210617e-02      3.996102e-02 
##             num3d               our              over            remove 
##      1.186138e-02      8.437590e-02      1.204162e-01      2.124195e-01 
##          internet             order              mail           receive 
##      9.399767e-02      7.487099e-02      1.614349e-02      5.746932e-02 
##              will              free          business             email 
##     -2.841598e-02      7.475768e-02      5.137680e-02      5.771563e-02 
##               you            credit              your              font 
##      1.425568e-02      6.127913e-02      5.251263e-02      4.468358e-02 
##            num000             money                hp               hpl 
##      1.791273e-01      9.110257e-02     -2.315356e-02     -2.144791e-02 
##            george              labs            telnet              data 
##     -1.221953e-02     -5.286056e-02     -2.385966e-02     -4.208898e-02 
##            num415             num85        technology           num1999 
##      5.315424e-02     -3.047991e-02      2.692394e-02     -3.410158e-02 
##             parts                pm            direct           meeting 
##     -5.233530e-02     -1.931751e-02      4.217186e-02     -3.946366e-02 
##          original           project                re               edu 
##     -6.217463e-02     -3.227688e-02     -3.528565e-02     -3.889380e-02 
##             table        conference     charSemicolon  charRoundbracket 
##     -1.949352e-01     -5.842266e-02     -1.411539e-01     -6.100067e-02 
## charSquarebracket   charExclamation        charDollar          charHash 
##     -6.071287e-02      6.785124e-02      2.355408e-01      2.724968e-02 
##        capitalAve       capitalLong      capitalTotal 
##      2.126329e-04      7.156337e-05      8.055017e-05
```

```
which.max(result$adjr2)
```

```
## [1] 50
```

Now we want to chek the backward method with leave one out and k fold cross validation ways and again calculate the cirterions and compare them.

```
fit.D.a<-
glm(type~our+over+remove+internet+free+credit+your+font+num000+money+hp+georg
e+meeting+re+
       +edu+charSemicolon+charExclamation+charDollar+capitalTotal,data = Data
, family = binomial)
#cost 2 function:
cost2<- function(labels,pred){
mean(labels==ifelse(pred > 0.5, 1, 0))}
```

in the Leave one out method we say the bottem function so just now we use it. and we are compare the cost 2 corterion valuse buy cost 2 in leave one out cross validation with out backward model selection and leave one out cross validation method with backward model selection.

```
(LOOCV.D.a<-cv.glm(Data,fit.D.a,cost2)$delta[1])

## [1] 0.9171919

(Miss.B.b$delta[1])

## [1] 0.9271897
```

according to this values we can see that the backward selection and with out any model selection cost 2 dont have signifact diffrent becuse we delet about 9 predictors and it just has 1 lower values and its great.

Now we want to compare the miss classification rate corterion valuse in leave one out cross validation with out backward model selection and leave one out cross validation method with backward model selection.

for leave one out cross validation

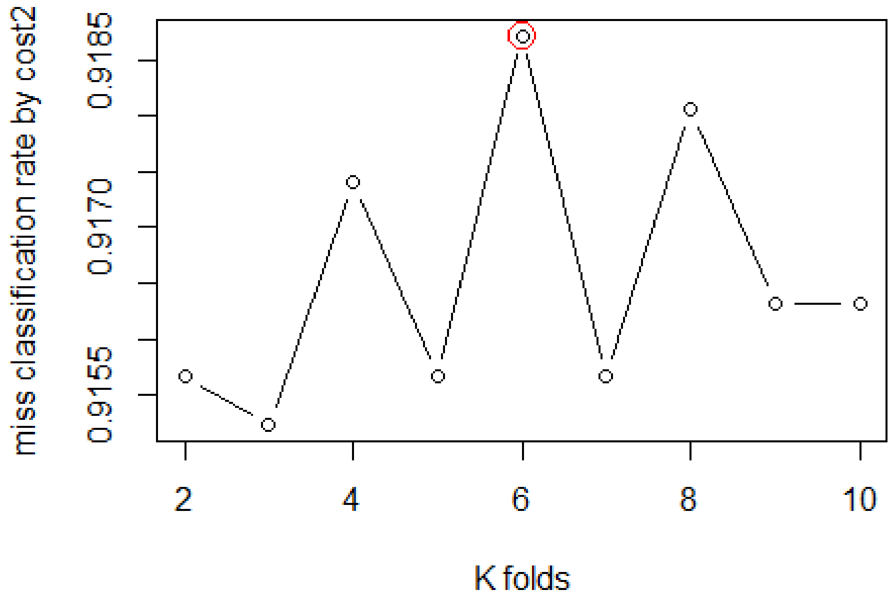
```
LOOCV.AUC(Data , Responsecolumn = 58 ,
          Formula
=type~our+over+remove+internet+free+credit+your+font+num000+money+hp+george+m
eeting+re+
       +edu+charSemicolon+charExclamation+charDollar+capitalTotal)

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

##
## Call:
## roc.default(response = responses, predictor = predictions)
##
## Data: predictions in 2788 controls (responses 0) < 1813 cases (responses
1).
## Area under the curve: 0.9683

Loocv.B.a

##
## Call:
## roc.default(response = responses, predictor = predictions)
##
## Data: predictions in 2788 controls (responses 0) < 1813 cases (responses
1).
## Area under the curve: 0.9715
```

again according to this values we can see that the backward selection and with out any model selection miss classification rate dont have signifact diffrent becuse we delet about 9 predictors and it just have small diffrent.

for k-fold cross validation

Now we want to just calculate cost 2 cirterion values for model with backward selection and compare it with the values with out any model selection.

```
Miss.D.a<-c()
for(i in c(2:10)){
Miss.D.a[i-1]<-cv.glm(Data,fit.D.a ,cost2, K= i)$delta[1]}
#the bottem outputs are for each k k=2:9
Miss.D.a

## [1] 0.9156705 0.9152358 0.9174093 0.9156705 0.9187133 0.9156705 0.9180613
## [8] 0.9163225 0.9163225

Miss.C.b

## [1] 0.9200174 0.9254510 0.9250163 0.9269724 0.9261030 0.9247990 0.9265377
## [8] 0.9250163 0.9263204

par(mfrow=c(1,1))
plot(2:10 , Miss.D.a , xlab="K folds" , ylab="miss classification rate by
cost2" , type="b")
points(which.max(Miss.D.a)+1 , max(Miss.D.a) , cex=2 , col="red")
points(c(2:10),Miss.C.b , col="orange")
```

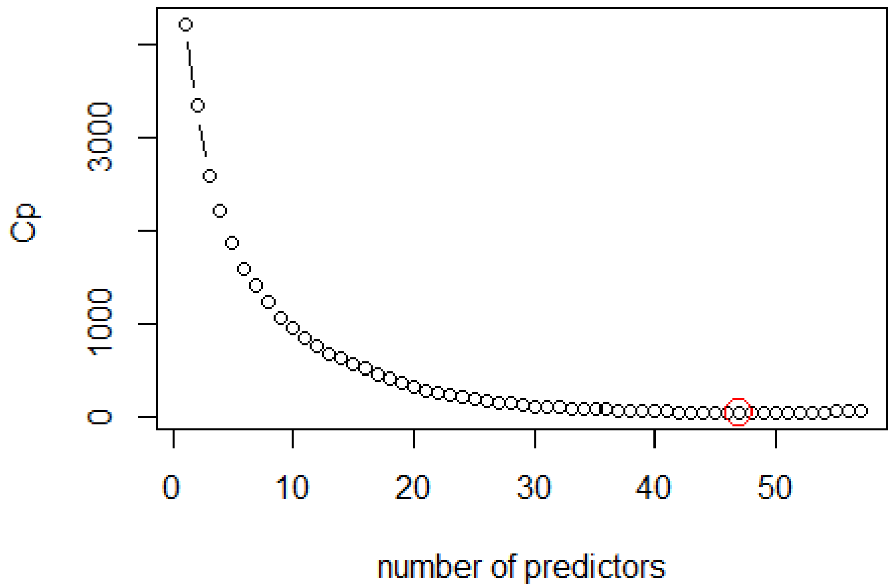


again according to this values we can see that the backward selection and with out any model selection miss classification rate buy cost2 dont have signifact diffrent becuse we delet about 9 predictors and it just have small diffrent.

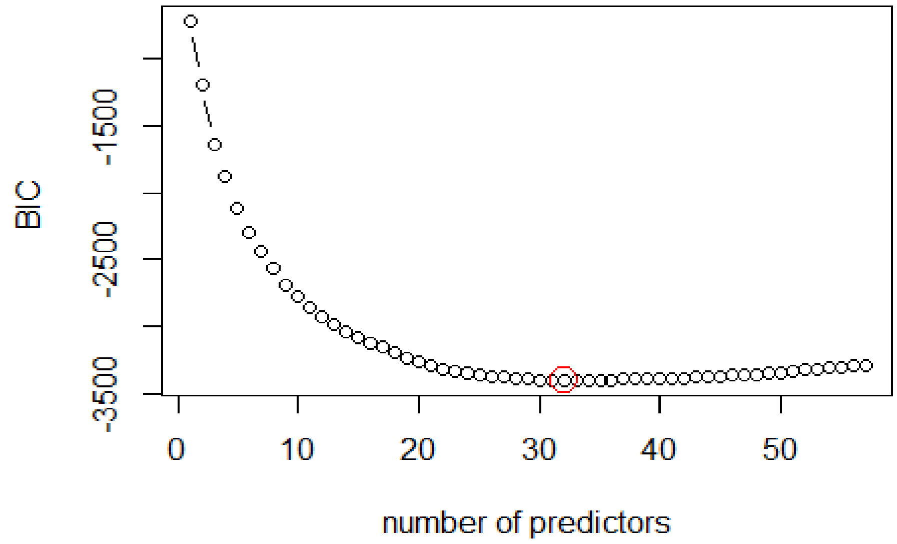
E) forward subseet selection

According to this method we have 3 step : ### step 1) Make an empty model with M0 name. ### step 2) we added a predictor to M0 that have best correlation and relationship with our response and the best model after adding is M1 with 1 predictor. and do this until M58 that is full model(with 58 predictors). ### step 3) According to our Criterions we choose the best model. Note: Criterions are including BIC , Cp , Adjust R squred for each Criterion we choose the best method.

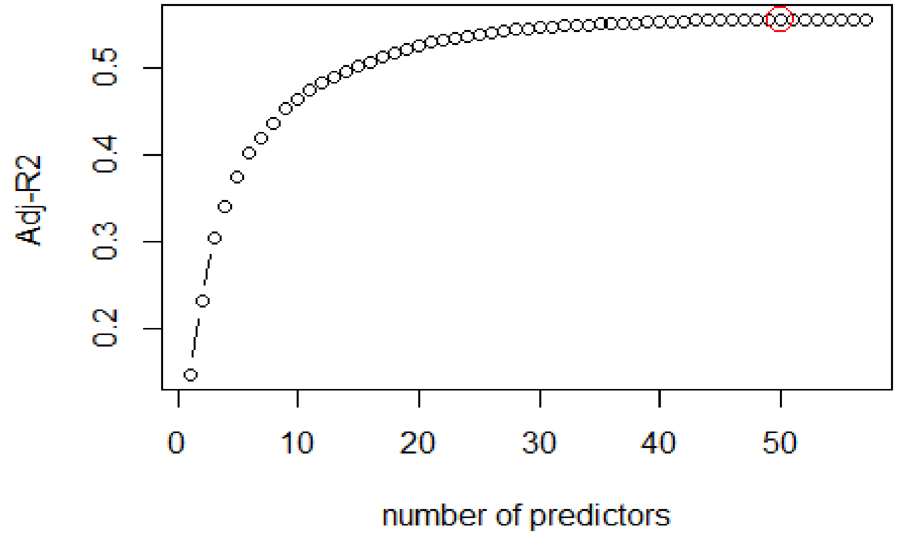
```
set.seed(1)
library(leaps)
forward.subset<-regsubsets(type~ . ,data = Data ,nvmax=58 , method =
"forward")
result<-summary(forward.subset)
# as there bottem line have alot of outputs i make is as comment it show that
in each step how it choose the best subset.
#result

par(mfrow=c(1,1))

plot(1:length(result$cp) ,result$cp , xlab="number of predictors" , ylab="Cp
"   , type = "b")
points(which.min(result$cp) , min(result$cp) ,col="red" , cex=2)
```



```
plot(1:length(result$bic) ,result$bic , xlab="number of predictors" ,
ylab="BIC "  , type = "b")
points(which.min(result$bic) , min(result$bic) ,col="red" , cex=2)
```



```
plot(1:length(result$adjr2) ,result$adjr2 , xlab="number of predictors" ,
ylab="Adj-R2 "  , type = "b")
points(which.max(result$adjr2) , max(result$adjr2) ,col="red" , cex=2)
```



we know that the minimum of BIC, Cp cirterions are the better model and the maximum of the adjust R squred is better model too. Now here we can see that the forward model selection choose the Model with 50 predictors as the better(with adjust R squred corterion.). bottem we can see this predictor names and Coefficients values.

```
coef(forward.subset , id=which.max(result$adjr2))
```

```
##       (Intercept)              make           address               all 
##      2.011350e-01     -5.012417e-02     -1.210617e-02      3.996102e-02 
##             num3d               our              over            remove 
##      1.186138e-02      8.437590e-02      1.204162e-01      2.124195e-01 
##          internet             order              mail           receive 
##      9.399767e-02      7.487099e-02      1.614349e-02      5.746932e-02 
##              will              free          business             email 
##     -2.841598e-02      7.475768e-02      5.137680e-02      5.771563e-02 
##               you            credit              your              font 
##      1.425568e-02      6.127913e-02      5.251263e-02      4.468358e-02 
##            num000             money                hp               hpl 
##      1.791273e-01      9.110257e-02     -2.315356e-02     -2.144791e-02 
##            george              labs            telnet              data 
##     -1.221953e-02     -5.286056e-02     -2.385966e-02     -4.208898e-02 
##            num415             num85        technology           num1999 
##      5.315424e-02     -3.047991e-02      2.692394e-02     -3.410158e-02 
##             parts                pm            direct           meeting 
##     -5.233530e-02     -1.931751e-02      4.217186e-02     -3.946366e-02 
##          original           project                re               edu 
##     -6.217463e-02     -3.227688e-02     -3.528565e-02     -3.889380e-02 
##             table        conference     charSemicolon  charRoundbracket 
##     -1.949352e-01     -5.842266e-02     -1.411539e-01     -6.100067e-02 
## charSquarebracket   charExclamation        charDollar          charHash 
##     -6.071287e-02      6.785124e-02      2.355408e-01      2.724968e-02 
##        capitalAve       capitalLong      capitalTotal 
##      2.126329e-04      7.156337e-05      8.055017e-05
```

```
which.max(result$adjr2)
```

```
## [1] 50
```

Now we want to chek the forward method with leave one out and k fold cross validation ways and again calculate the cirterions and compare them.

for leave one out cross validation

```
fit.E.a<-
glm(type~our+over+remove+internet+free+email+credit+your+font+num000+money+hp
+george+meeting
       +edu+charSemicolon+charExclamation+charDollar+capitalTotal,data = Data
, family = binomial)

cost2<- function(labels,pred){
mean(labels==ifelse(pred > 0.5, 1, 0))}
```

in the Leave one out method we say the bottem function so just now we use it. and we are compare the cost 2 corterion valuse buy cost 2 in K-fold cross validation with out backward model selection and K-fold cross validation method with backward model selection.

```
(LOOCV.E.a<-cv.glm(Data,fit.E.a,cost2)$delta[1])

## [1] 0.9161052

Miss.B.b$delta[1]

## [1] 0.9271897
```

according to this values we can see that the forward selection and with out any model selection cost2 dont have signifact diffrent becuse we delet about 39 predictors and it just have small diffrent.

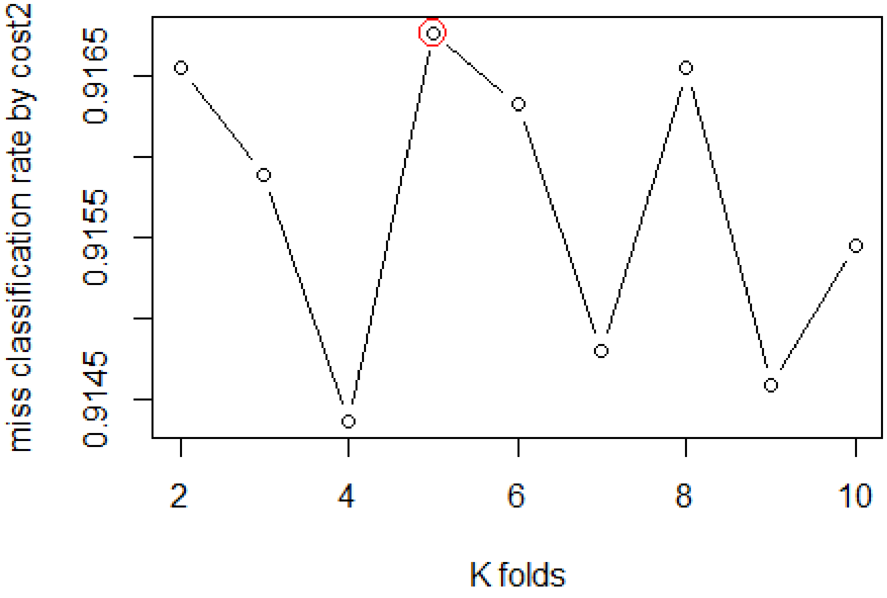
```
LOOCV.AUC(Data , Responsecolumn = 58 ,
          Formula
=type~our+over+remove+internet+free+credit+your+font+num000+money+hp+george+m
eeting+re+
       +edu+charSemicolon+charExclamation+charDollar+capitalTotal)

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

##
## Call:
## roc.default(response = responses, predictor = predictions)
##
## Data: predictions in 2788 controls (responses 0) < 1813 cases (responses
1).
## Area under the curve: 0.9683

Loocv.B.a

##
## Call:
## roc.default(response = responses, predictor = predictions)
##
## Data: predictions in 2788 controls (responses 0) < 1813 cases (responses
1).
## Area under the curve: 0.9715
```

again according to this values we can see that the forward selection and with out any model selection miss classification rate buy cost2 dont have signifact diffrent becuse we delet about 9 predictors and it just have small diffrent.

for K-fold cross validation

Now we want to just calculate cost 2 cirterion values for model with forward selection and compare it with the values with out any model selection.

```
Miss.E.a<-c()
for(i in c(2:10)){
Miss.E.a[i-1]<-cv.glm(Data,fit.E.a ,cost2, K= i)$delta[1]}
Miss.E.a

## [1] 0.9165399 0.9158879 0.9143664 0.9167572 0.9163225 0.9148011 0.9165399
## [8] 0.9145838 0.9154532

par(mfrow=c(1,1))
plot(2:10 , Miss.E.a , xlab="K folds" , ylab="miss classification rate by
cost2" , type="b")
points(which.max(Miss.E.a)+1 , max(Miss.E.a) , cex=2 , col="red")
points(c(2:10),Miss.C.b , col="orange")
```



again according to this values we can see that the forward selection and with out any model selection miss classification rate buy cost2 dont have signifact diffrent becuse we delet about 9 predictors and it just have small diffrent.

F) Boot straping

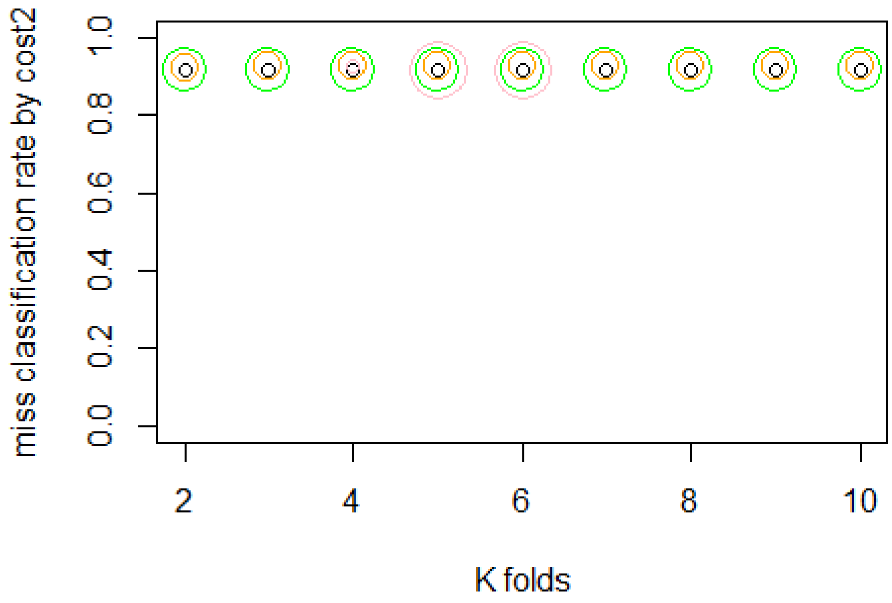
as there we have alot of Data so we dont use boot strapping method but i put the codes here:

```
#library(boot)
#stat<-function(data , index){
  #fit.F.a<-glm(type~. , data=Data ,subset = index ,family = binomial)
  #coef(fit.F.a)}
#stat(Data , 1:58)
#boot(Data , stat , 10000)
```

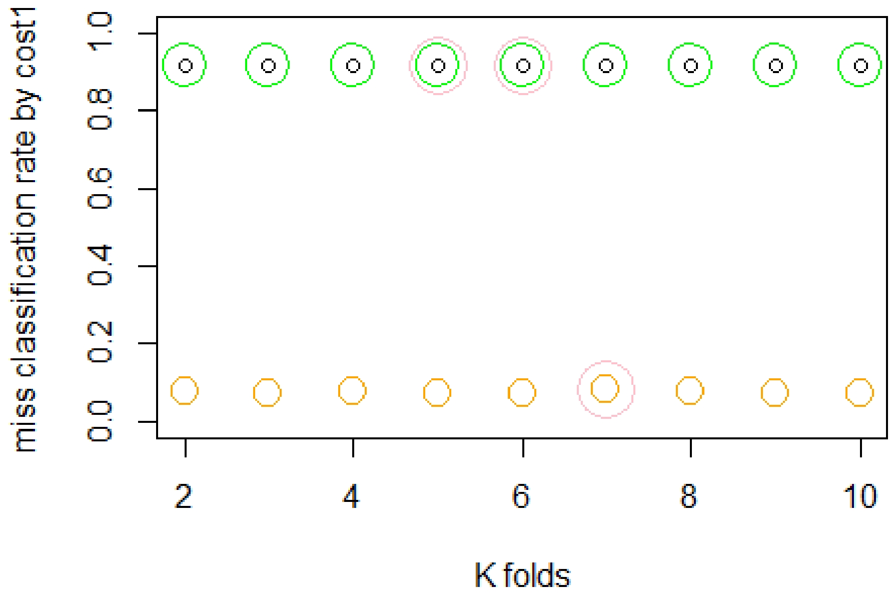
3)Conclusion

Now we want to compare the backward and forward, leave one out cross validation and K-fold cross validation here.

```
# for leave one out cross validation:

plot(2:10 , Miss.E.a , xlab="K folds" , ylab="miss classification rate by 
cost2" , type="p" ,col="black" ,ylim=c(0,1),cex=1)
points(c(2:10),Miss.C.b , col="orange",cex=2)
points(c(2:10) , Miss.D.a , col="green",cex=3)
points(which.max(Miss.C.b) , max(Miss.C.b) ,col="pink")
points(which.max(Miss.D.a)+1 , max(Miss.D.a) , cex=4 , col="pink")
points(which.max(Miss.E.a)+1 , max(Miss.E.a) , cex=4 , col="pink")
```



```
#for K-fold cross validation:

plot(2:10 , Miss.C.a , xlab="K folds" , ylab="miss classification rate by
cost1" , type="p" ,col="orange" ,ylim = c(0,1) ,cex=2)
points(which.max(Miss.C.a)+1 , max(Miss.C.a) , cex=4 , col="pink")
points(c(2:10) , Miss.D.a , col="green",cex=3)
points(which.max(Miss.D.a)+1 , max(Miss.D.a) , cex=4 , col="pink")
points(c(2:10), Miss.E.a ,col="black",cex=1)
points(which.max(Miss.E.a)+1 , max(Miss.E.a) , cex=4 , col="pink")
```



Produced by Mehrab Atighi. Thanks to Dr. Seyed Noorullah Mousavi.