# Shuting Chen

```
> library(faraway)
> library(carData)
> library(car)
```

> #1

```
> d0=data.frame(hsb)
> dim(d0)
[1] 200  11
> head(d0)
   id gender  race    ses schtyp     prog read write math science socst
1  70   male white    low public  general  57    52   41      47    57
2 121 female white middle public  vocation 68    59   53      63    61
3  86   male white   high public  general  44    33   54      58    31
4 141   male white   high public  vocation 63    44   47      53    56
5 172   male white middle public academic  47    52   57      53    61
6 113   male white middle public academic  44    52   51      63    61
> set.seed(1)
> bag1=randomForest(prog~math+science,d0,mtry=2,importance=T)
> par(mfrow=c(1,2))
> ypred=predict(bag1,newdata = d0)
> labels=c("academic","general","vocation")
> colors = c("black","green","red")
> plot(math~science,d0,col=d0$prog,pch=18,main="observed")
> legend("topleft",legend=labels,bty="n",text.col = colors,cex = 0.7)
> grid()
> plot(math~science,d0,col=ypred,pch=18,main="predicted")
> legend("topleft",legend=labels,bty="n",text.col = colors,cex = 0.7)
> grid()
> par(mfrow=c(1,1))
> 1-sum(diag(prop.table(table(d0$prog,ypred))))
[1] 0.09
> #error rate is 0.09
```

observed   predicted
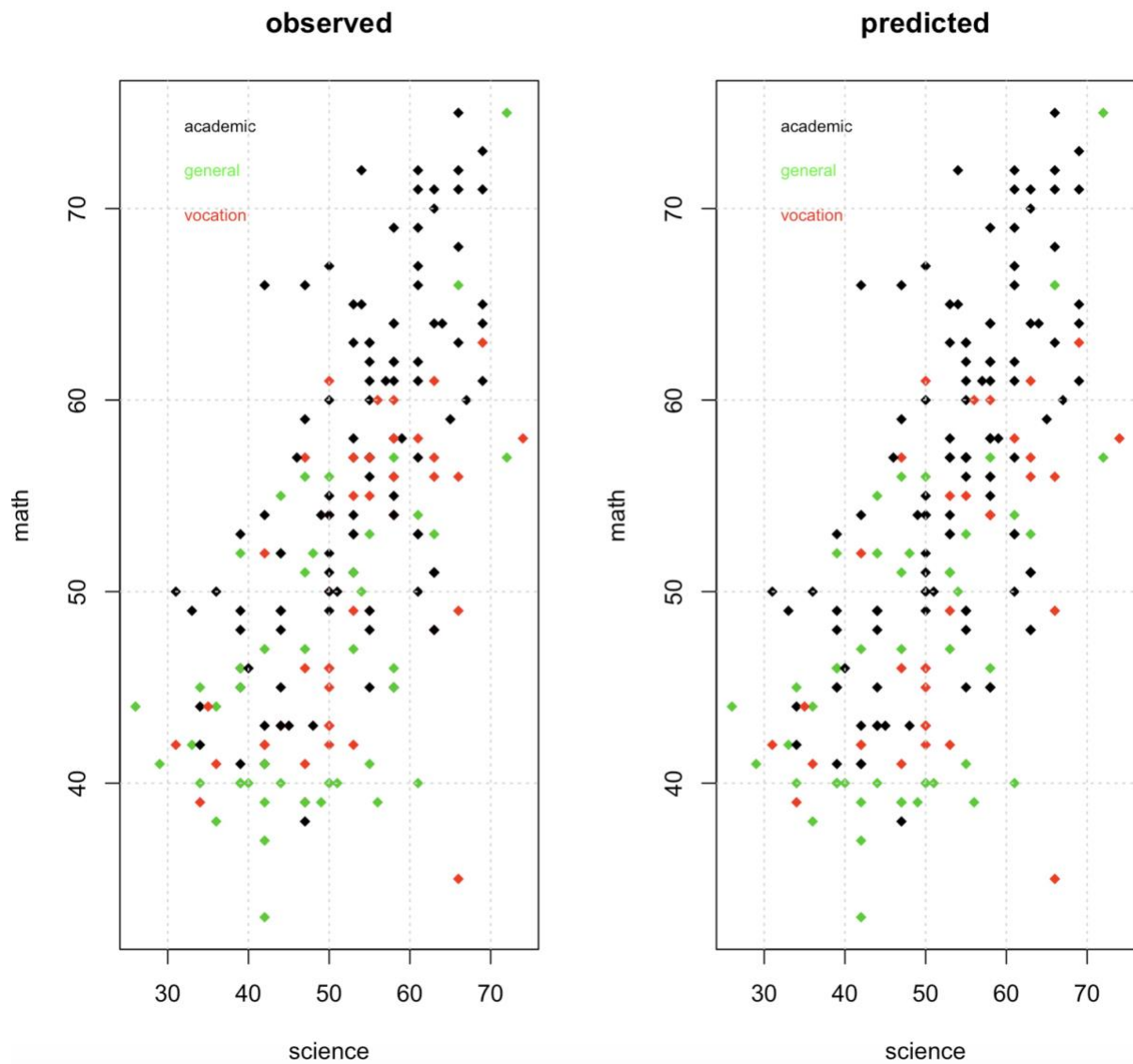
> **#2**
> **#a)**
> library(ISLR)
> library(randomForest)
> d0=data.frame(Caravan)
> sapply(Caravan,table)
$MOSTYPE

```
 1   2   3   4   5   6   7   8   9  10  11  12  13  15  16  17  18  19  20
124  82 249  52  45 119  44 339 278 165 153 111 179   5  16   9  19   3  25
 21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39
```

```
 15  98 251 180  82  48  50  25  86 118 205 141 810 182 214 225 132 339 328
 40  41
 71 205


$MAANTHUI

   1    2    3    4    5    6    7    8   10
5267  505   39    2    1    1    5    1    1

$MGEMOMV

   1    2    3    4    5
 284 2131 2646  693   68

$MGEMLEEF

   1    2    3    4    5    6
  74 1452 3000 1073  193   30

$MOSHOOFD

   1    2    3    4    5    6    7    8    9   10
 552  502  886   52  569  205  550 1563  667  276

……

Does not show all the result
……


> dim(d0)
[1] 5822   86
> d0=d0[,-which(names(d0)%in% c("PVRAAUT","AVRAAUT"))]
> train=1:1000
> d0train=d0[train,]
> d0test=d0[-train]
> set.seed(1)
> forest1=randomForest(Purchase~.,d0train,mtry=83,importance=T)
> importance(forest1)
               No        Yes MeanDecreaseAccuracy MeanDecreaseGini
MOSTYPE  9.36748604 -3.36760792          9.25053009     4.517191004
MAANTHUI 1.95373123 -0.06535802          1.91469335     1.151085123
MGEMOMV  7.83167926 -2.72448820          7.33494542     1.140477450
MGEMLEEF 4.10784489 -0.12529425          4.14172311     0.775692035
MOSHOOFD 3.77920421  0.51099817          3.97966332     1.451814974
MGODRK   4.08285089 -2.46532927          3.59848175     1.461014416
```

MGODPR   7.95704145 2.42704345   8.21955518   3.546813354
MGODOV   2.12706237 1.56932854   2.46568801   1.310703255
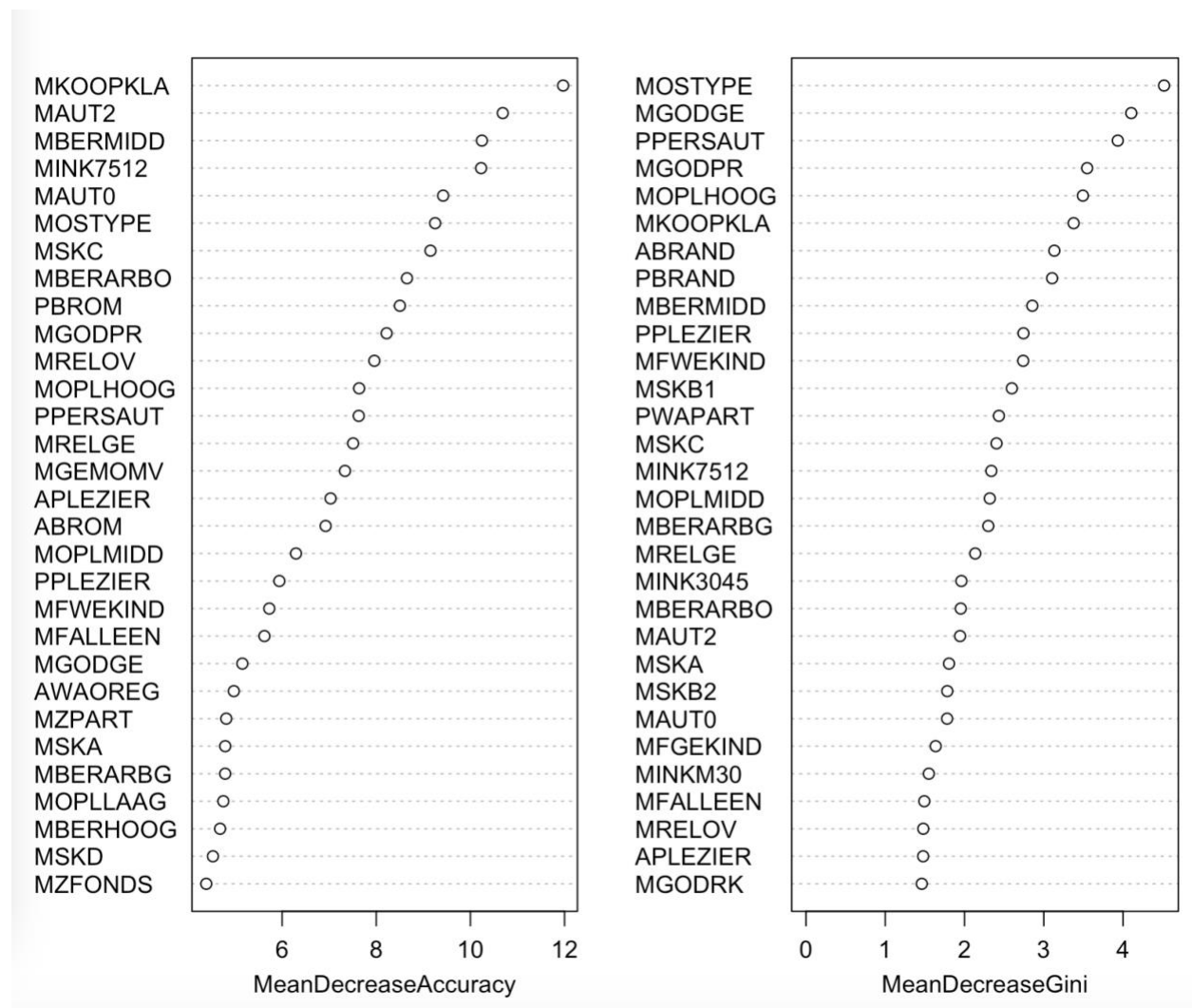
……

Does not show all the result
……

AFIETS   -0.56967023 -2.62350658   -1.20496853   0.793364260
AINBOED   0.05537678 0.00000000   0.05606031   0.031434525
ABYSTAND 1.09725797 -1.09627420   0.86464848   0.248707510
> varImpPlot(forest1,main="")
> #the best predictor is MOSTYPE

> #b)

```
> library(gbm)
> set.seed(1)
> n=nrow(d0)
> y=rep(0,n)
> y[d0$Purchase=="Yes"]=1
> d0$Purchase=y
> train=1:1000
> d0train=d0[train,]
> d0test=d0[-train,]
> boost1 = gbm(Purchase~.,data=d0[train,],distribution="bernoulli",
n.trees=1000,shrinkage=0.01)
> summary(boost1)
            var     rel.inf
PPERSAUT PPERSAUT 14.63504779
MKOOPKLA MKOOPKLA  9.47091649
MOPLHOOG MOPLHOOG  7.31457416
MBERMIDD MBERMIDD  6.08651965
PBRAND    PBRAND  4.66766122
MGODGE    MGODGE  4.49463264
ABRAND    ABRAND  4.32427755
MINK3045 MINK3045  4.17590619
MOSTYPE   MOSTYPE  2.86402583
PWAPART   PWAPART  2.78191075
MAUT1      MAUT1  2.61929152
MBERARBG MBERARBG  2.10480508
MSKA        MSKA  2.10185152
MAUT2      MAUT2  2.02172510
......
Does not show all the result
......
AGEZONG   AGEZONG  0.00000000
AWAOREG   AWAOREG  0.00000000
AZEILPL   AZEILPL  0.00000000
APLEZIER APLEZIER  0.00000000
AFIETS     AFIETS  0.00000000
AINBOED   AINBOED  0.00000000
ABYSTAND ABYSTAND  0.00000000
> grid()
>
> #the most important predictor is PPERSAUT
```