CI-project-chemical

2022-12-08

Install packages & library

```
#install.packages("remotes")
#remotes::install_github("ygeunkim/propensityml")
#install.packages("dplyr")
#install.packages("sas7bdat")

library(propensityml)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(sas7bdat)
```

Chemical Dataset

Data reading

```
df<-read.table('/Users/kyungseonlee/Desktop/causal/poisox.txt')
colnames(df)=c('age','sex','prior','poisox','subseq','mortal')
df%>% head()
```

```
##   age sex prior poisox subseq mortal
## 1  31   0   126      1    140      0
## 2  35   1   115      1    124      1
## 3  27   1    92      0    101      0
## 4  66   0   118      0    112      1
## 5  54   1    89      1     90      0
## 6  35   1    93      1    104      0
```

```
df3=df
df3$blood_diff=df3$subseq-df3$prior
df3=df3[c(1,2,4,6,7)]
df3 %>% head()
```

```
##   age sex poisox mortal blood_diff
## 1  31   0      1      0         14
## 2  35   1      1      1          9
## 3  27   1      0      0          9
## 4  66   0      0      1         -6
## 5  54   1      1      0          1
## 6  35   1      1      0         11
```

```
df0=df[c(1,2,4,6)]
df0 %>% head()
```

```
##   age sex poisox mortal
## 1  31   0      1      0
## 2  35   1      1      1
## 3  27   1      0      0
## 4  66   0      0      1
## 5  54   1      1      0
## 6  35   1      1      0
```

Propensity score estimation- logistic regression, random Forest, CART

1. Logistic Regression

1-a. LR-ps estimation

```
log_reg_ps=glm(poisox~ age+sex+blood_diff, family = "binomial", data=df3)
logit_e_hat=predict(log_reg_ps)
print(logit_e_hat %>% head())
```

```
##         1         2         3         4         5         6
## 1.3632071 1.3822489 0.1119096 1.3075317 2.1539408 1.9435899
```

```
#exp(logit_e_hat)
e_hat=exp(logit_e_hat)/(1+exp(logit_e_hat))
print(e_hat %>% head())
```

```
##         1         2         3         4         5         6
## 0.7962804 0.7993519 0.5279482 0.7870998 0.8960365 0.8747460
```

```
df_lo=df3

df_lo[,"ps"]=e_hat
df_lo %>% head()
```

```
##   age sex poisox mortal blood_diff        ps
## 1  31   0      1      0         14 0.7962804
## 2  35   1      1      1          9 0.7993519
## 3  27   1      0      0          9 0.5279482
## 4  66   0      0      1         -6 0.7870998
## 5  54   1      1      0          1 0.8960365
## 6  35   1      1      0         11 0.8747460
```

1-b. LR-weighting

```
zi=df_lo$poisox
yi=df_lo$mortal
e=df_lo$ps

df_lo["ipw_wt"]= zi/e-(1-zi)/(1-e)
df_lo %>% head()
```

```
##   age sex poisox mortal blood_diff        ps    ipw_wt
## 1  31   0      1      0         14 0.7962804  1.255839
## 2  35   1      1      1          9 0.7993519  1.251013
## 3  27   1      0      0          9 0.5279482 -2.118412
## 4  66   0      0      1         -6 0.7870998 -4.697037
## 5  54   1      1      0          1 0.8960365  1.116026
## 6  35   1      1      0         11 0.8747460  1.143189
```

```
ATE_ipw_log=mean(zi*yi/e)-mean((1-zi)*yi/(1-e))
ATE_ipw_log
```

```
## [1] -0.129414
```

```
ATE_sipw_log=sum(zi*yi/e)/sum(zi/e)-sum((1-zi)*yi/(1-e))/sum((1-zi)/(1-e))
ATE_sipw_log
```

```
## [1] -0.005641556
```

1-c. LR-Evaluate

```
cov_balance_lo=data.frame(rep(0),row.names = "Logistic Regression")

df3 %>% head()
```

```
##   age sex poisox mortal blood_diff
## 1  31   0      1      0         14
## 2  35   1      1      1          9
## 3  27   1      0      0          9
## 4  66   0      0      1         -6
## 5  54   1      1      0          1
## 6  35   1      1      0         11
```

```
for(i in colnames(df3)){
  if(i!="poisox" & i!="mortal"){
#    print(df1[i])
    t_weighted_mean=mean((df3[i]*df_lo$ipw_wt)[df3$poisox==1,])
    c_weighted_mean=mean((df3[i]*df_lo$ipw_wt)[df3$poisox==0,])
    weighted_mean_diff=abs(t_weighted_mean-c_weighted_mean)
    asam=weighted_mean_diff/sd((df3[i]*df_lo$ipw_wt)[df3$poisox==1,])
    cov_balance_lo[i]=asam
  }
}
cov_balance_lo=cov_balance_lo[,-1]
cov_balance_lo
```

```
##                          age      sex blood_diff
## Logistic Regression 14.1128 4.642901   2.377404
```

```
cov_balance_lo$ASAM=apply(cov_balance_lo,1,mean)
cov_balance_lo["ASAM"]
```

```
##                          ASAM
## Logistic Regression 7.044369
```

2. Random Forest

2-a. RF-ps estimation

```
#install.packages("randomForest")
library(randomForest)
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
rf=randomForest(poisox~ age+sex+blood_diff, data=df3)
```

```
## Warning in randomForest.default(m, y, ...): The response has five or fewer
## unique values. Are you sure you want to do regression?
```

```
rf_ps=predict(rf)
print(rf_ps %>% head())
```

```
##         1         2         3         4         5         6
## 0.7839251 0.9024729 0.5825203 0.1624645 0.9364468 0.9047788
```

```
df_rf=df3

for (i in 1:nrow(df_rf) ){
  df_rf[i,"ps"]=rf_ps[i]
}
df_rf %>% head()
```

```
##   age sex poisox mortal blood_diff        ps
## 1  31   0      1      0         14 0.7839251
## 2  35   1      1      1          9 0.9024729
## 3  27   1      0      0          9 0.5825203
## 4  66   0      0      1         -6 0.1624645
## 5  54   1      1      0          1 0.9364468
## 6  35   1      1      0         11 0.9047788
```

2-b. RF-weighting

```
zi=df_rf$poisox
yi=df_rf$mortal
e=df_rf$ps

df_rf["ipw_wt"]= zi/e-(1-zi)/(1-e)
df_rf %>% head()
```

```
##   age sex poisox mortal blood_diff        ps    ipw_wt
## 1  31   0      1      0         14 0.7839251  1.275632
## 2  35   1      1      1          9 0.9024729  1.108066
## 3  27   1      0      0          9 0.5825203 -2.395326
## 4  66   0      0      1         -6 0.1624645 -1.193979
## 5  54   1      1      0          1 0.9364468  1.067866
## 6  35   1      1      0         11 0.9047788  1.105242
```

```
ATE_ipw_rf=mean(zi*yi/e)-mean((1-zi)*yi/(1-e))
ATE_ipw_rf
```

```
## [1] 0.2138551
```

```
ATE_sipw_rf=sum(zi*yi/e)/sum(zi/e)-sum((1-zi)*yi/(1-e))/sum((1-zi)/(1-e))
ATE_sipw_rf
```

```
## [1] 0.067703
```

2-c. RF-Evaluation

```
cov_balance_rf=data.frame(rep(0),row.names = "Random Forest")

df3 %>% head()
```

```
##   age sex poisox mortal blood_diff
## 1  31   0      1      0         14
## 2  35   1      1      1          9
## 3  27   1      0      0          9
## 4  66   0      0      1         -6
## 5  54   1      1      0          1
## 6  35   1      1      0         11
```

```
for(i in colnames(df3)){
  if(i!="poisox" & i!="mortal"){
#    print(df1[i])
    t_weighted_mean=mean((df3[i]*df_rf$ipw_wt)[df3$poisox==1,])
    c_weighted_mean=mean((df3[i]*df_rf$ipw_wt)[df3$poisox==0,])
    weighted_mean_diff=abs(t_weighted_mean-c_weighted_mean)
    asam=weighted_mean_diff/sd((df3[i]*df_rf$ipw_wt)[df3$poisox==1,])
    cov_balance_rf[i]=asam
  }
}
cov_balance_rf=cov_balance_rf[,-1]
cov_balance_rf
```

```
##                     age      sex blood_diff
## Random Forest 7.432783 2.161662   1.068851
```

```
cov_balance_rf$ASAM=apply(cov_balance_rf,1,mean)
cov_balance_rf["ASAM"]
```

```
##                    ASAM
## Random Forest 3.554432
```

3. CART

3-a. CART-ps estimation

```
library(rpart)

#control = rpart.control(minbucket = 2)
df1_c=df3
df1_c %>% head(5)
```

```
##   age sex poisox mortal blood_diff
## 1  31   0      1      0         14
## 2  35   1      1      1          9
## 3  27   1      0      0          9
## 4  66   0      0      1         -6
## 5  54   1      1      0          1
```

```
cart2=rpart(poisox~ age+sex+blood_diff+blood_diff, data=df1_c ,method='poisson',contr
ol = rpart.control(maxdepth = 3))
#summary(cart2)
cart2$cptable
```

```
##           CP nsplit rel error    xerror       xstd
## 1 0.46629674      0 1.0000000 1.0001353 0.01808422
## 2 0.07551949      1 0.5337033 0.5340874 0.01884688
## 3 0.01873951      3 0.3826643 0.3840556 0.01625246
## 4 0.01000000      4 0.3639248 0.3658785 0.01459797
```

```
cart_ps2=predict(cart2)
cart_ps2 %>% head()
```

```
##           1           2           3           4           5           6
## 0.734080091 0.734080091 0.734080091 0.002835322 0.978666503 0.734080091
```

```
df1_c[,"ps2"]=cart_ps2

df1_c %>% head()
```

```
##   age sex poisox mortal blood_diff         ps2
## 1  31   0      1      0         14 0.734080091
## 2  35   1      1      1          9 0.734080091
## 3  27   1      0      0          9 0.734080091
## 4  66   0      0      1         -6 0.002835322
## 5  54   1      1      0          1 0.978666503
## 6  35   1      1      0         11 0.734080091
```

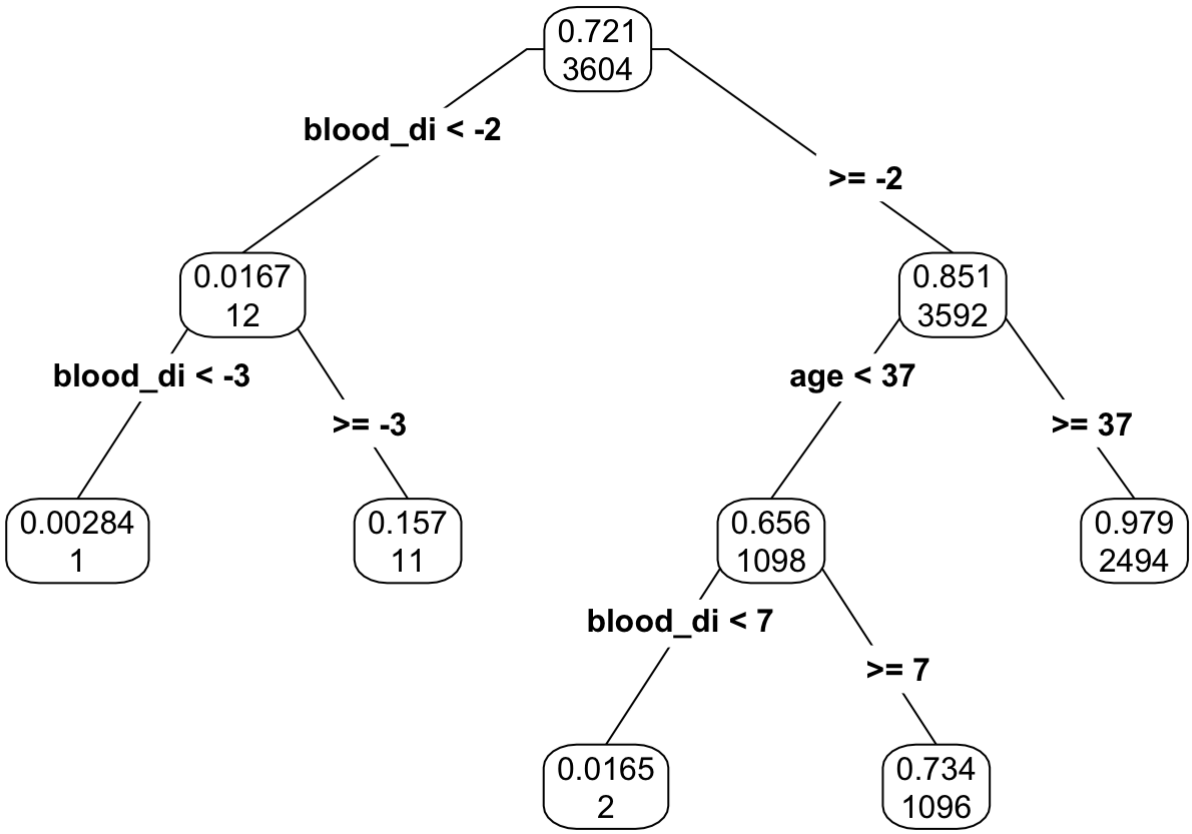
```
sum(df1_c$ps2==0|df1_c$ps2==1)
```

```
## [1] 0
```

```
#install.packages("rpart.plot")

library(rpart.plot)

prp(cart2, type=4, extra=2, digits=3)
```



3-b. CART-weighting

```
zi=df1_c$poisox
yi=df1_c$mortal
e=df1_c$ps2


df1_c["ipw_wt"]= zi/e-(1-zi)/(1-e)
df1_c %>% head()
```

```
##   age sex poisox mortal blood_diff         ps2    ipw_wt
## 1  31   0      1      0         14 0.734080091  1.362249
## 2  35   1      1      1          9 0.734080091  1.362249
## 3  27   1      0      0          9 0.734080091 -3.760531
## 4  66   0      0      1         -6 0.002835322 -1.002843
## 5  54   1      1      0          1 0.978666503  1.021799
## 6  35   1      1      0         11 0.734080091  1.362249
```

```
ATE_ipw_cart=mean(zi*yi/e)-mean((1-zi)*yi/(1-e))
ATE_ipw_cart
```

```
## [1] -0.003643648
```

```
ATE_sipw_cart=sum(zi*yi/e)/sum(zi/e)-sum((1-zi)*yi/(1-e))/sum((1-zi)/(1-e))
ATE_sipw_cart
```

```
## [1] 0.03710139
```

3-c. CART-evaluation

```
cov_balance_c=data.frame(rep(0),row.names = "CART")

df3%>% head()
```

```
##   age sex poisox mortal blood_diff
## 1  31   0      1      0         14
## 2  35   1      1      1          9
## 3  27   1      0      0          9
## 4  66   0      0      1         -6
## 5  54   1      1      0          1
## 6  35   1      1      0         11
```

```
for(i in colnames(df3)){
  if(i!="poisox" & i!="mortal"){
#    print(df1[i])
    t_weighted_mean=mean((df3[i]*df1_c$ipw_wt)[df3$poisox==1,])
    c_weighted_mean=mean((df3[i]*df1_c$ipw_wt)[df3$poisox==0,])
    weighted_mean_diff=abs(t_weighted_mean-c_weighted_mean)
    asam=weighted_mean_diff/sd((df3[i]*df1_c$ipw_wt)[df3$poisox==1,])
    cov_balance_c[i]=asam
  }
}
cov_balance_c=cov_balance_c[,-1]
cov_balance_c
```

```
##            age       sex blood_diff
## CART 0.6550871 0.3402262  0.6428177
```

```
cov_balance_c$ASAM=apply(cov_balance_c,1,mean)
cov_balance_c["ASAM"]
```

```
##          ASAM
## CART 0.5460436
```

Total ATE table

```
ATE_table= rbind(ATE_ipw_log,
      ATE_sipw_log,
      ATE_ipw_rf,
      ATE_sipw_rf,
      ATE_ipw_cart,
      ATE_sipw_cart)

colnames(ATE_table)="ATE table in Chemical Dataset"
knitr :: kable(ATE_table,"simple")
```

	ATE table in Chemical Dataset
ATE_ipw_log	-0.1294140
ATE_sipw_log	-0.0056416
ATE_ipw_rf	0.2138551
ATE_sipw_rf	0.0677030
ATE_ipw_cart	-0.0036436
ATE_sipw_cart	0.0371014

Evaluation visualization - Chemical dataset

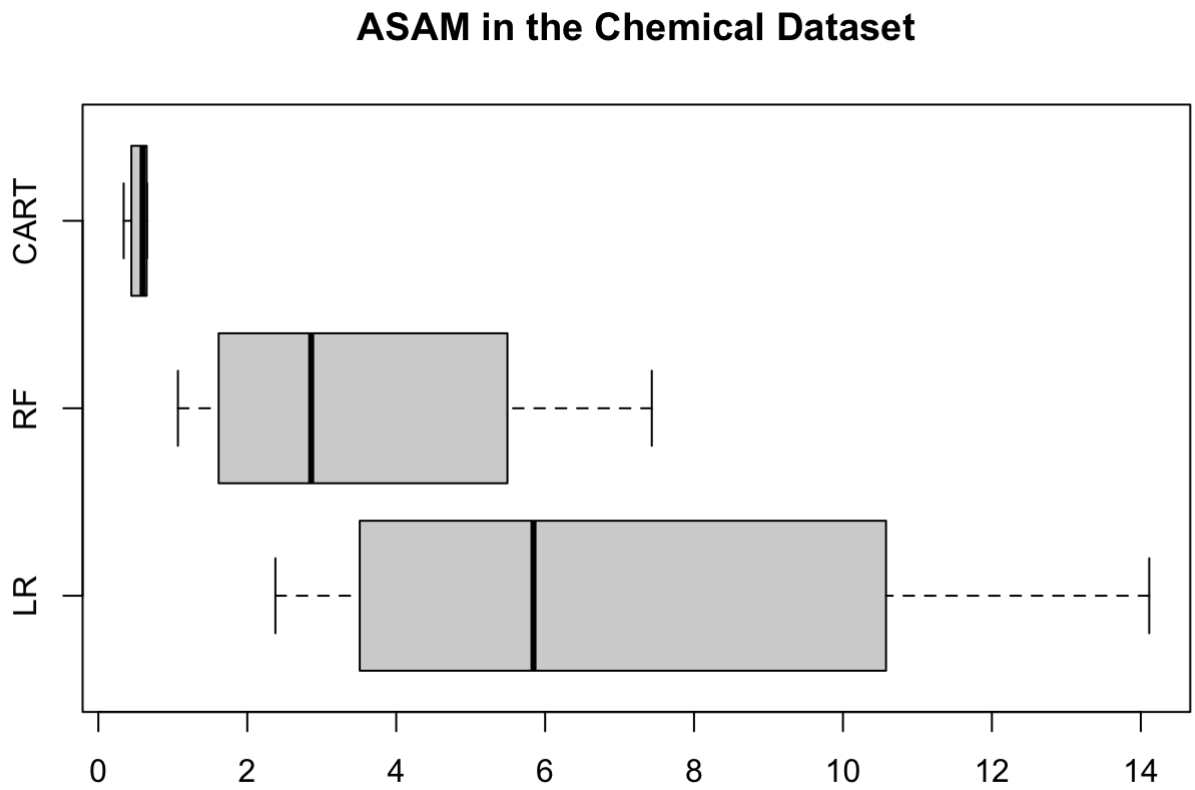
a. ASAM table

```
ASAM_table1=rbind(cov_balance_lo["ASAM"] ,cov_balance_rf["ASAM"] ,cov_balance_c["ASA
M"])
colnames(ASAM_table1)="ASAM in Chemical Dataset"
knitr::kable(ASAM_table1,"simple")
```

	ASAM in Chemical Dataset
Logistic Regression	7.0443691
Random Forest	3.5544320
CART	0.5460436

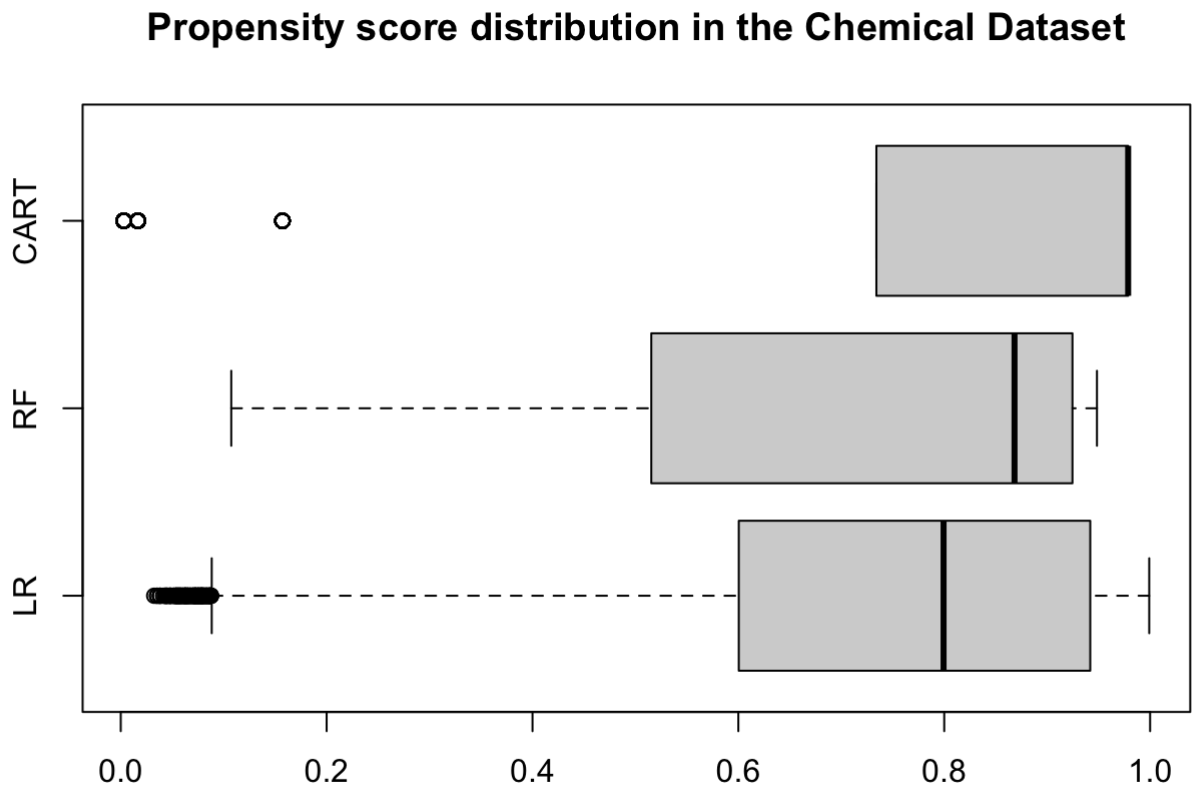
b. ASAM box plot

```
a=cbind(t(cov_balance_lo) ,t(cov_balance_rf) ,t(cov_balance_c))
colnames(a)=c("LR","RF","CART")
boxplot(a, main="ASAM in the Chemical Dataset",horizontal = TRUE)
```



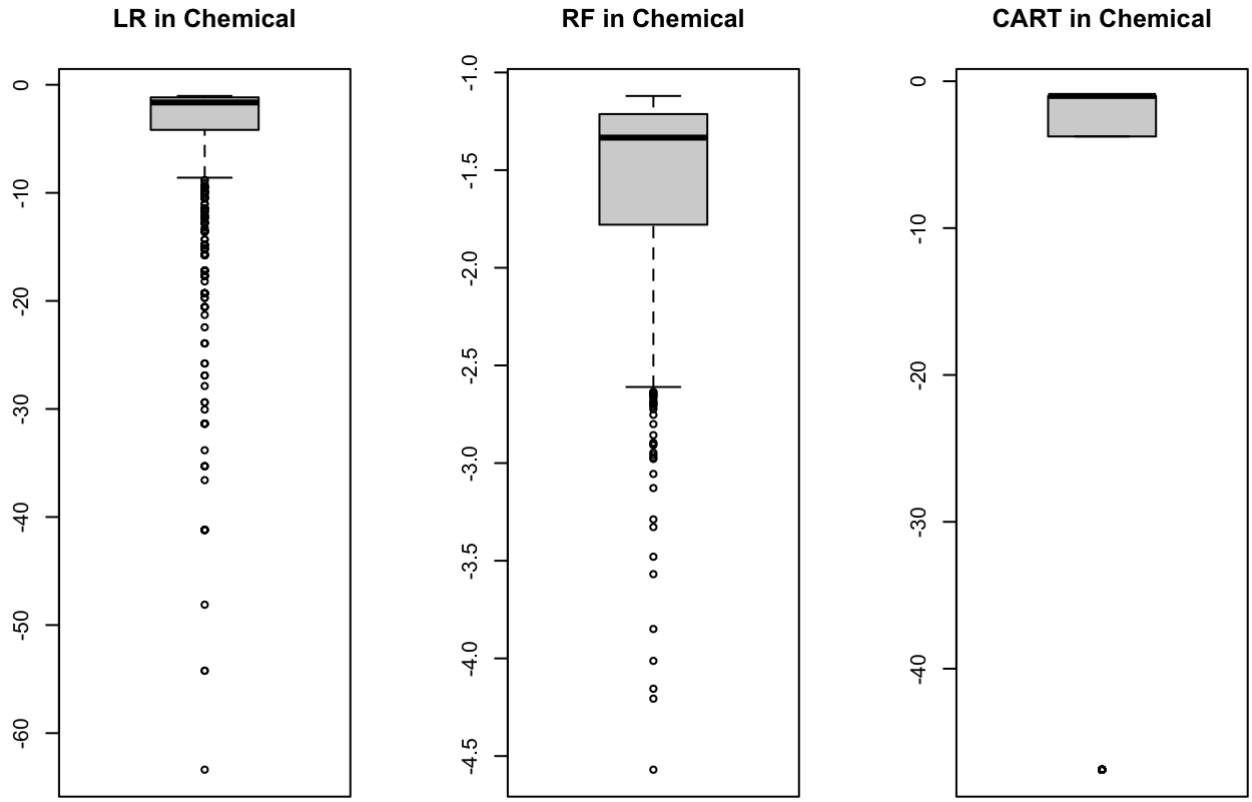
c. ps distribution

```
b=cbind(df_lo["ps"],df_rf["ps"],df1_c["ps2"])
colnames(b)=c("LR","RF","CART")
boxplot(b,horizontal = TRUE,main="Propensity score distribution in the Chemical Datas
et")
```



d. weight distribution

```
par(mfcol=c(1,3))
boxplot(df_lo[(df3$poisox==0),"ipw_wt"],main="LR in Chemical")
boxplot(df_rf[(df3$poisox==0),"ipw_wt"],main="RF in Chemical")
boxplot(df1_c[(df3$poisox==0),"ipw_wt"],main="CART in Chemical")
```


LR in Chemical
0
-10
-20
-30
-40
-50
-60
RF in Chemical
-1.0
-1.5
-2.0
-2.5
-3.0
-3.5
-4.0
-4.5
CART in Chemical
0
-10
-20
-30
-40