Regression-tree.R

rocka

2023-12-10

```
#Regression Tree
```

```
library(rpart)
library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 4.3.2
```

```
library(DMwR2)
```

```
## Warning: package 'DMwR2' was built under R version 4.3.2
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
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
```

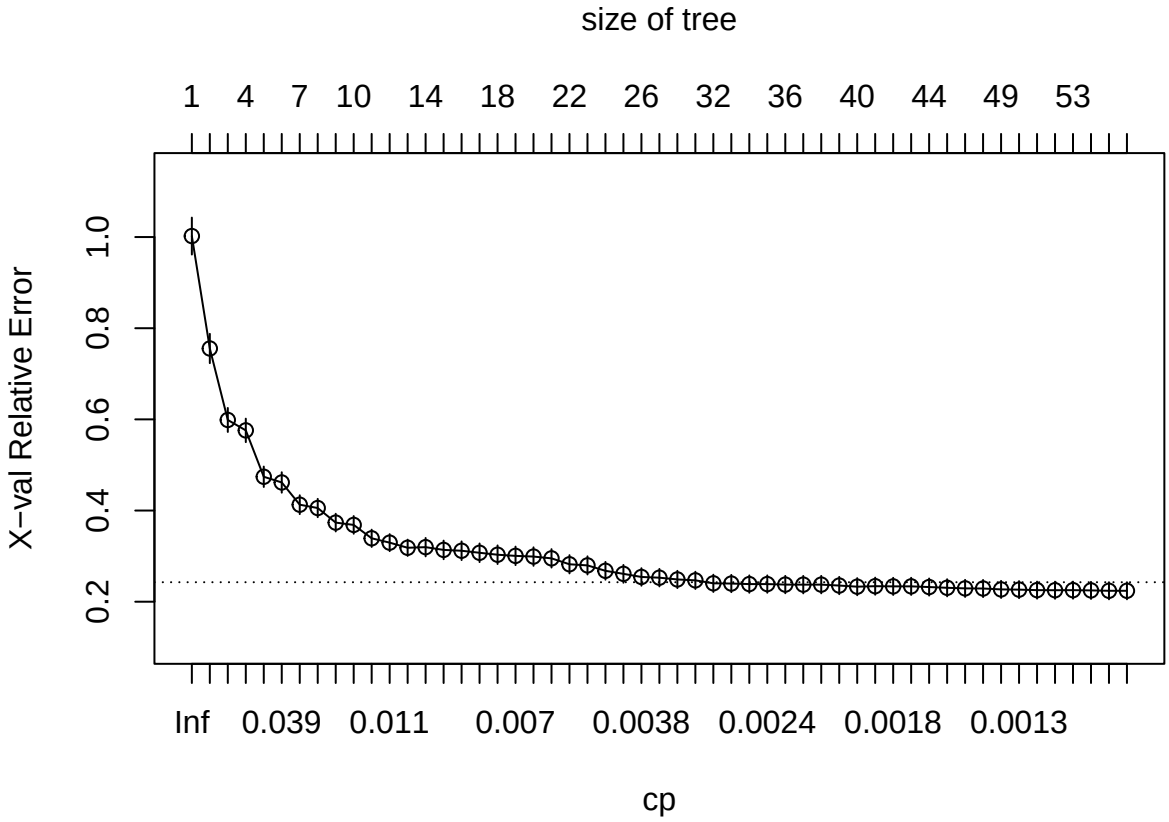
```
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.4
## v ggplot2   3.4.4     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
```

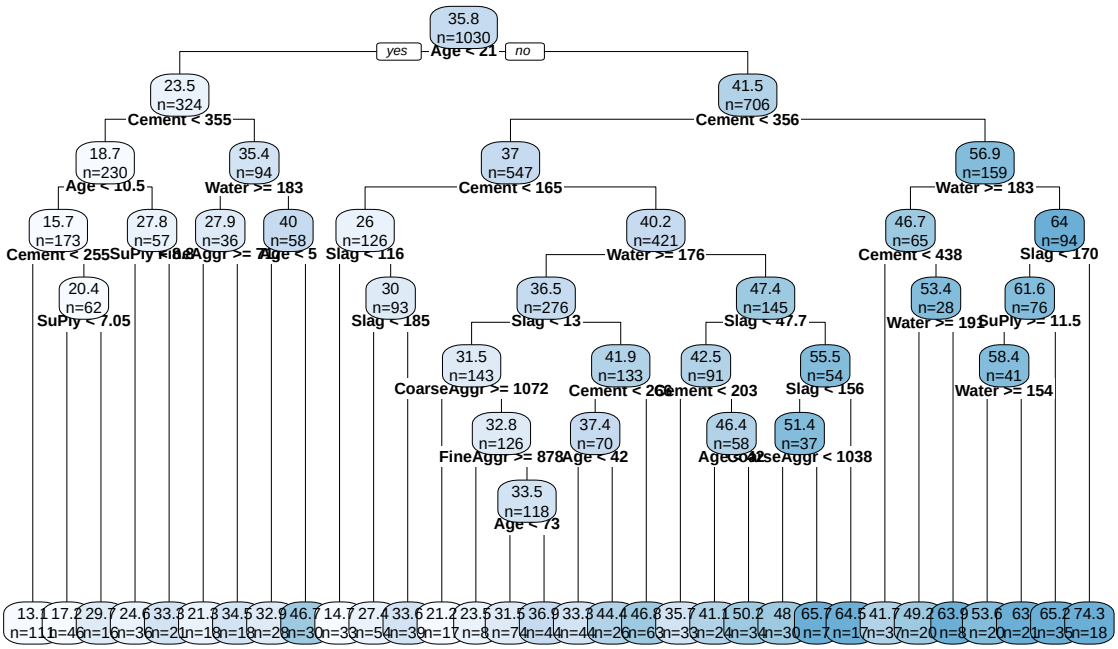
```
## -- Conflicts --------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
df<-read.csv("C:/Users/rocka/OneDrive/Documents/output_file.csv")


df <- rename(df, response=CCS)

SEED<-1234

set.seed(SEED)
rpart.modela<-rpart(response ~ .,data=df,method="anova", cp=0.001)

plotcp(rpart.modela)
```




```
rpart.modela.1se <- rt.prune(rpart.modela, se=1)

rpart.plot(rpart.modela.1se , extra=1, roundint=FALSE,
           digits=3, main="1se regression tree for concrete compressive strength (in MPa)",
           cex=0.5)
```

1se regression tree for concrete compressive strength (in MPa)




```
var(df$response)
```

```
## [1] 279.0818
```

```
printcp(rpart.modela.1se)
```

```
##
## Regression tree:
## rpart(formula = response ~ ., data = df, method = "anova", cp = 0.001)
##
## Variables actually used in tree construction:
## [1] Age        Cement     CoarseAggr FineAggr   Slag       SuPly      Water
##
## Root node error: 287175/1030 = 278.81
##
## n= 1030
##
##           CP nsplit rel error  xerror     xstd
## 1  0.2480823      0   1.00000 1.00217 0.040518
## 2  0.1714067      1   0.75192 0.75551 0.032159
## 3  0.0683911      2   0.58051 0.59872 0.026329
## 4  0.0645348      3   0.51212 0.57584 0.025762
## 5  0.0398308      4   0.44759 0.47410 0.022379
## 6  0.0391064      5   0.40775 0.46161 0.022415
```

```
## 7  0.0260850      6   0.36865 0.41282 0.020684
## 8  0.0217650      7   0.34256 0.40551 0.020089
## 9  0.0198351      8   0.32080 0.37354 0.019389
## 10 0.0198292      9   0.30096 0.36847 0.019389
## 11 0.0112837     10   0.28113 0.33909 0.018947
## 12 0.0101590     11   0.26985 0.32956 0.018673
## 13 0.0096024     12   0.25969 0.31831 0.017869
## 14 0.0085087     13   0.25009 0.31966 0.020159
## 15 0.0082445     14   0.24158 0.31362 0.020397
## 16 0.0076775     15   0.23334 0.31180 0.020444
## 17 0.0075075     16   0.22566 0.30733 0.020398
## 18 0.0071104     17   0.21815 0.30307 0.020384
## 19 0.0069893     18   0.21104 0.30057 0.020275
## 20 0.0069440     19   0.20405 0.29900 0.020276
## 21 0.0063950     20   0.19711 0.29509 0.020173
## 22 0.0061580     21   0.19071 0.28211 0.020024
## 23 0.0054207     22   0.18455 0.27966 0.019892
## 24 0.0043104     23   0.17913 0.26777 0.019487
## 25 0.0040058     24   0.17482 0.26109 0.019494
## 26 0.0035609     25   0.17082 0.25427 0.019208
## 27 0.0030721     26   0.16726 0.25271 0.019170
## 28 0.0030072     28   0.16111 0.24880 0.018947
## 29 0.0026963     29   0.15810 0.24685 0.018898
## 30 0.0026865     31   0.15271 0.24033 0.018937
```

```
(MSE<-278.81*0.24033)
```

```
## [1] 67.00641
```

```
(rmse<-sqrt(MSE))
```

```
## [1] 8.185744
```

```
# Use a validation set to estimate test-set MSE
set.seed(7)
train = sample(1:nrow(df), nrow(df)/2)

set.seed(SEED)
rpart.concrete.train <- rpart(response~
                                ., data = df[train,],
                              method="anova", cp=0)
rpart.concrete.train <- rt.prune(rpart.concrete.train, se=1)

Yhat <- predict(rpart.concrete.train, newdata = df[-train,])
concrete.test <- df[-train,"response"]

(MSE <- mean((Yhat-concrete.test)^2))
```

```
## [1] 68.21776
```

```r
(RMSE <- sqrt(MSE))
```

```
## [1] 8.259404
```

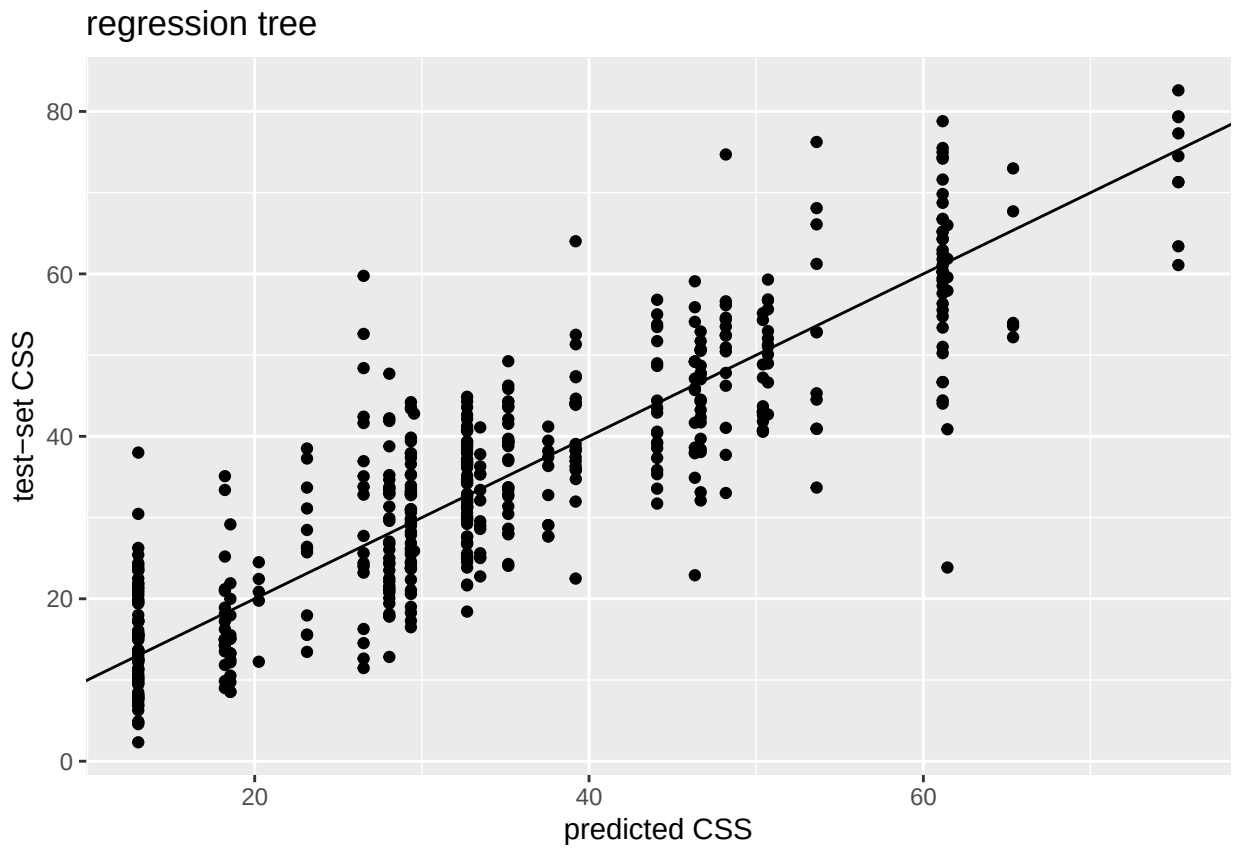
```r
#repeating with different seed for creating training set
set.seed(5)
train <- sample(1:nrow(df), nrow(df)/2)

set.seed(SEED)
rpart.concrete.train <- rpart(response~
                                .,data = df[train,],
                              method="anova", cp=0)
rpart.concrete.train <- rt.prune(rpart.concrete.train, se=1)

Yhat <- predict(rpart.concrete.train, newdata = df[-train,])
concrete.test <- df[-train,"response"]

(MSE <- mean((Yhat-concrete.test)^2))
```

```
## [1] 64.34217
```

```r
(RMSE <- sqrt(MSE))
```

```
## [1] 8.021357
```

```r
ggplot(data.frame(Yhat, concrete.test), aes(x=Yhat ,y=concrete.test)) +
  geom_point() +
  geom_abline(slope=1,intercept=0) +
  labs(x="predicted CSS",
       y="test-set CSS",
       title="regression tree")
```




```
#  Results:                                 MSE          RMSE
#     Cross-validated MSE                 67.00641     8.185744
#     Validation set MSE (set.seed(7))    68.21776     8.259404
#     Validation set MSE (set.seed(5))    64.34217     8.021357
```