Machine learning decision tree

Tejasvini Mavuleti

2022-08-03

```
library(rpart)
library(caret)

## Warning: package 'caret' was built under R version 4.2.1

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.2.1

## Loading required package: lattice

## Warning: package 'lattice' was built under R version 4.2.1

library(e1071)

## Warning: package 'e1071' was built under R version 4.2.1

library(DMwR2)

## Warning: package 'DMwR2' was built under R version 4.2.1

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

set.seed(100)
options(warm=-1)
library(lubridate)

##
## Attaching package: 'lubridate'

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

library(ggplot2)
library(randomForest)

## Warning: package 'randomForest' was built under R version 4.2.1

## randomForest 4.7-1.1

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

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

## The following object is masked from 'package:ggplot2':
##
##     margin

library(dplyr)

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:randomForest':
##
##     combine

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

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

library(rpart)
library(caret)
library(caretEnsemble)

## Warning: package 'caretEnsemble' was built under R version 4.2.1

##
## Attaching package: 'caretEnsemble'

## The following object is masked from 'package:ggplot2':
##
##     autoplot

library(e1071)
library(corrplot)

## corrplot 0.92 loaded

library(mlbench)

## Warning: package 'mlbench' was built under R version 4.2.1

library(pROC)

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

##
## Attaching package: 'pROC'
```

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

```
clean_dataset <- function() {
  datasetloc = "C:/Users/mavul/OneDrive/Desktop/Health care data.csv"
  if (file.exists(datasetloc)) {
    data <- read.csv(file=datasetloc, header = T)
  }
  return(data)
}
```

```
age <- function(dob, age.day = today(), units = "years", floor = TRUE) {
  calc.age = interval(dob, age.day) / duration(num = 1, units = units)
  if (floor) return(as.integer(floor(calc.age)))

  return(calc.age)
}
```

```
age_group <- function(ag) {
  ifelse(ag<25,25, ifelse(ag<40, 40, ifelse(ag<50,50,65)))
}
```

```
e_europe <- c('Ukraine','Russia','Poland','Czech Republic','Hungary')
w_europe <- c('Austria','Belgium','France','Germany','Italy','Netherlands','P
ortugal','Spain','Switzerland')
n_europe <- c('Sweden', 'Finland', 'Denmark')
c_europe <- c('England','Scotland','Ireland')
```

```
ethnic_group <- function(countryname) {
  ifelse((countryname %in% e_europe), 'e_europe',
         ifelse((countryname %in% w_europe) ,'w_europe',
                ifelse((countryname %in% n_europe), 'n_europe',
                       ifelse((countryname %in% c_europe), 'c_europe',
                              countryname))))
}
```

```
patients <- clean_dataset()
```

```
#Removing the patient IDs from the data set
```

```
patients <- patients[,-1]
str(patients)
```

```
## 'data.frame':    2000 obs. of  13 variables:
##  $ gender            : chr  "female" "female" "male" "male" ...
##  $ dob               : chr  "1944-03-09" "1966-07-02" "1981-05-31" "1945-0
2-13" ...
##  $ zipcode           : int  89136 94105 89127 44101 89136 94105 60612 4322
1 89127 43210 ...
```

```
##  $ employment_status : chr  "retired" "employed" "employed" "retired" ...
##  $ education         : chr  "bachelors" "phd/md" "masters" "bachelors" ...
##  $ marital_status    : chr  "married" "married" "married" "married" ...
##  $ children          : int  1 4 2 2 3 2 0 2 2 7 ...
##  $ ancestry          : chr  "Portugal" "Sweden" "Germany" "Denmark" ...
##  $ avg_commute       : num  13.4 15.2 23.6 19.6 36.5 ...
##  $ daily_internet_use: num  2.53 6.77 3.63 5 7.75 3.34 6.75 3.01 4.12 3.15
...
##  $ available_vehicles: int  2 2 1 3 1 0 2 3 1 1 ...
##  $ military_service  : chr  "no" "no" "no" "no" ...
##  $ disease           : chr  "hypertension" "endometriosis" "prostate cance
r" "multiple sclerosis" ...

summary(patients)

##     gender              dob                zipcode      employment_status
##  Length:2000        Length:2000        Min.   :10001   Length:2000
##  Class :character   Class :character   1st Qu.:43221   Class :character
##  Mode  :character   Mode  :character   Median :60612   Mode  :character
##                                        Mean   :63388
##                                        3rd Qu.:90008
##                                        Max.   :94110
##   education          marital_status        children       ancestry
##  Length:2000        Length:2000        Min.   :0.000   Length:2000
##  Class :character   Class :character   1st Qu.:1.000   Class :character
##  Mode  :character   Mode  :character   Median :2.000   Mode  :character
##                                        Mean   :2.267
##                                        3rd Qu.:3.000
##                                        Max.   :7.000
##   avg_commute    daily_internet_use available_vehicles military_service
##  Min.   :-2.47   Min.   :1.010      Min.   :0.000      Length:2000
##  1st Qu.:23.46   1st Qu.:4.020      1st Qu.:1.000      Class :character
##  Median :30.32   Median :5.010      Median :2.000      Mode  :character
##  Mean   :30.38   Mean   :4.993      Mean   :1.746
##  3rd Qu.:37.13   3rd Qu.:5.973      3rd Qu.:3.000
##  Max.   :63.73   Max.   :8.820      Max.   :4.000
##    disease
##  Length:2000
##  Class :character
##  Mode  :character
##
##
##

patients$education <- ifelse(patients$education == 'highscool', as.character(
'highschool'), as.character(patients$education))
patients$education <- ifelse(as.factor(patients$education) == 'phD/MD', as.ch
aracter('phd/md'), as.character(patients$education))
patients$education <- as.factor(patients$education)
```

```r
patients$ancestry <- as.factor(ethnic_group(patients$ancestry))

patients$age <- age(patients$dob)

binary_value <- function(value, compare_to) {
  ifelse(value==compare_to,1,0)
}

patients$prostate_cancer <- binary_value(patients$disease,'prostate cancer')
patients$skin_cancer <- binary_value(patients$disease,'skin cancer')
patients$breast_cancer <- binary_value(patients$disease,'breast cancer')
patients$hiv_aids <- binary_value(patients$disease,'HIV/AIDS')
patients$diabetes <- binary_value(patients$disease,'diabetes')
patients$heart_disease <- binary_value(patients$disease,'heart disease')
patients$hypertension <- binary_value(patients$disease,'hypertension')
patients$endometriosis <- binary_value(patients$disease,'endometriosis')
patients$multiple_sclerosis <- binary_value(patients$disease,'multiple sclero
sis')
patients$schizophrenia <- binary_value(patients$disease,'schizophrenia')
patients$kidney_disease <- binary_value(patients$disease,'kidney disease')
patients$gastritis <- binary_value(patients$disease,'gastritis')
patients$alzheimer <- binary_value(patients$disease,'Alzheimer disease')
str(patients)

## 'data.frame':    2000 obs. of  27 variables:
##  $ gender            : chr  "female" "female" "male" "male" ...
##  $ dob               : chr  "1944-03-09" "1966-07-02" "1981-05-31" "1945-0
2-13" ...
##  $ zipcode           : int  89136 94105 89127 44101 89136 94105 60612 4322
1 89127 43210 ...
##  $ employment_status : chr  "retired" "employed" "employed" "retired" ...
##  $ education         : Factor w/ 4 levels "bachelors","highschool",..: 1 4
3 1 3 2 4 1 3 2 ...
##  $ marital_status    : chr  "married" "married" "married" "married" ...
##  $ children          : int  1 4 2 2 3 2 0 2 2 7 ...
##  $ ancestry          : Factor w/ 4 levels "c_europe","e_europe",..: 4 3 4
3 4 4 2 1 4 2 ...
##  $ avg_commute       : num  13.4 15.2 23.6 19.6 36.5 ...
##  $ daily_internet_use: num  2.53 6.77 3.63 5 7.75 3.34 6.75 3.01 4.12 3.15
...
##  $ available_vehicles: int  2 2 1 3 1 0 2 3 1 1 ...
##  $ military_service  : chr  "no" "no" "no" "no" ...
##  $ disease           : chr  "hypertension" "endometriosis" "prostate cance
r" "multiple sclerosis" ...
##  $ age               : int  78 56 41 77 82 65 75 59 75 64 ...
##  $ prostate_cancer   : num  0 0 1 0 0 0 0 0 0 0 ...
##  $ skin_cancer       : num  0 0 0 0 1 0 0 0 0 0 ...
##  $ breast_cancer     : num  0 0 0 0 0 0 0 1 0 0 ...
##  $ hiv_aids          : num  0 0 0 0 0 0 0 0 0 1 ...
##  $ diabetes          : num  0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ heart_disease     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ hypertension      : num  1 0 0 0 0 0 0 0 0 0 ...
##  $ endometriosis     : num  0 1 0 0 0 0 0 0 0 0 ...
##  $ multiple_sclerosis: num  0 0 0 1 0 0 0 0 0 0 ...
##  $ schizophrenia     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ kidney_disease    : num  0 0 0 0 0 0 1 0 0 0 ...
##  $ gastritis         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ alzheimer         : num  0 0 0 0 0 0 0 0 0 0 ...
```

```
os_alzheimer <- select(patients, age, employment_status, education, marital_s
tatus, ancestry, available_vehicles, avg_commute,zipcode, children,daily_inte
rnet_use,military_service, alzheimer)
train <- sample(nrow(os_alzheimer), 0.7*nrow(os_alzheimer), replace = FALSE)
TrainSet <- os_alzheimer[train,]
TestSet <- os_alzheimer[-train,]
summary(TrainSet)
```

```
##       age        employment_status       education   marital_status
##  Min.   :24.00   Length:1400        bachelors :750   Length:1400
##  1st Qu.:58.00   Class :character   highschool:314   Class :character
##  Median :69.00   Mode  :character   masters   :199   Mode  :character
##  Mean   :68.02                      phd/md    :137
##  3rd Qu.:79.00
##  Max.   :98.00
##      ancestry   available_vehicles  avg_commute       zipcode
##  c_europe:217   Min.   :0.000      Min.   :-2.47   Min.   :10001
##  e_europe:318   1st Qu.:1.000      1st Qu.:23.61   1st Qu.:43221
##  n_europe:211   Median :2.000      Median :30.39   Median :60612
##  w_europe:654   Mean   :1.746      Mean   :30.43   Mean   :62877
##                 3rd Qu.:3.000      3rd Qu.:37.18   3rd Qu.:90008
##                 Max.   :4.000      Max.   :63.73   Max.   :94110
##     children     daily_internet_use military_service     alzheimer
##  Min.   :0.000   Min.   :1.010      Length:1400        Min.   :0
##  1st Qu.:1.000   1st Qu.:4.070      Class :character   1st Qu.:0
##  Median :2.000   Median :5.020      Mode  :character   Median :0
##  Mean   :2.227   Mean   :5.009                         Mean   :0
##  3rd Qu.:3.000   3rd Qu.:5.945                         3rd Qu.:0
##  Max.   :7.000   Max.   :8.640                         Max.   :0
```

```
summary(TestSet)
```

```
##       age        employment_status       education   marital_status
##  Min.   :30.00   Length:600         bachelors :326   Length:600
##  1st Qu.:59.00   Class :character   highschool:149   Class :character
##  Median :69.00   Mode  :character   masters   : 81   Mode  :character
##  Mean   :67.88                      phd/md    : 44
##  3rd Qu.:77.00
##  Max.   :98.00
##      ancestry   available_vehicles  avg_commute       zipcode
##  c_europe: 87   Min.   :0.000      Min.   : 4.63   Min.   :10001
##  e_europe:151   1st Qu.:1.000      1st Qu.:23.30   1st Qu.:43221
```

```
##  n_europe: 91   Median :2.000      Median :29.91   Median :60612
##  w_europe:271   Mean   :1.747      Mean   :30.26   Mean   :64579
##                 3rd Qu.:3.000      3rd Qu.:37.09   3rd Qu.:90008
##                 Max.   :4.000      Max.   :61.66   Max.   :94110
##     children     daily_internet_use military_service     alzheimer
##  Min.   :0.000   Min.   :1.250      Length:600         Min.   :0
##  1st Qu.:1.000   1st Qu.:3.938      Class :character   1st Qu.:0
##  Median :2.000   Median :4.930      Mode  :character   Median :0
##  Mean   :2.358   Mean   :4.958                         Mean   :0
##  3rd Qu.:3.000   3rd Qu.:5.990                         3rd Qu.:0
##  Max.   :7.000   Max.   :8.820                         Max.   :0
```

Compare model of Random Forest with Decision Tree model

```
ctrl <- trainControl(method = "repeatedcv",
                     number = 10,
                     repeats = 10,
                     verboseIter = FALSE,
                     sampling = "smote")

set.seed(42)
patients <- read.csv("C:/Users/mavul/OneDrive/Desktop/Health care data.csv")
patients <- patients[ ,14]
str(patients)

ubSMOTE <- function(X= input, Y=response, perc=40, method="percPos"){

}

data <- ubSMOTE(X= input, Y=response, perc=40, method="percPos")
us_dataset <- cbind(data$X, class=data$Y)

model_rf_smote <- caret::train(disease ~ .,data = patients,
     method = "rf",preProcess = c("scale", "center"),trControl = ctrl)

   ##  chr [1:2000] "hypertension" "endometriosis" "prostate cancer" ...

response <- as.factor(TrainSet$os_alzheimer)
input <- select(TrainSet, age, employment_status, education, marital_status,
ancestry)

ubUnder <- function(X= input, Y=response, perc=40, method="percPos"){
}

data <- ubUnder(X=input, Y=response, perc=40, method="percPos")
us_alzheimer <- cbind(data$X, class=data$Y)

ubOver <- function(X= input, Y=response, perc=40, method="percPos"){
}
```

```r
data <- ubOver(X=input, Y=response)
os_alzheimer <- cbind(data$X, class=data$Y)

ubSMOTE <- function(X= input, Y=response, perc=40, method="percPos"){
}

data <- ubSMOTE(X=input, Y=response)
smote_alzheimer <- cbind(data$X, class=data$Y)

train_control <- trainControl(method = "repeatedcv", number = 10, repeats=3,
savePredictions = TRUE)
ubUnder <- function(X= input, Y=response, perc=40, method="percPos"){
}

summary(clean_dataset())
```

```
##       id                 gender              dob                zipcode
##  Length:2000        Length:2000        Length:2000        Min.   :10001
##  Class :character   Class :character   Class :character   1st Qu.:43221
##  Mode  :character   Mode  :character   Mode  :character   Median :60612
##                                                           Mean   :63388
##                                                           3rd Qu.:90008
##                                                           Max.   :94110
##  employment_status   education          marital_status        children
##  Length:2000        Length:2000        Length:2000        Min.   :0.000
##  Class :character   Class :character   Class :character   1st Qu.:1.000
##  Mode  :character   Mode  :character   Mode  :character   Median :2.000
##                                                           Mean   :2.267
##                                                           3rd Qu.:3.000
##                                                           Max.   :7.000
##    ancestry           avg_commute     daily_internet_use available_vehicles
##  Length:2000        Min.   :-2.47   Min.   :1.010      Min.   :0.000
##  Class :character   1st Qu.:23.46   1st Qu.:4.020      1st Qu.:1.000
##  Mode  :character   Median :30.32   Median :5.010      Median :2.000
##                     Mean   :30.38   Mean   :4.993      Mean   :1.746
##                     3rd Qu.:37.13   3rd Qu.:5.973      3rd Qu.:3.000
##                     Max.   :63.73   Max.   :8.820      Max.   :4.000
##  military_service     disease
##  Length:2000        Length:2000
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```