```
###  Offshore distribution of yelkouan shearwaters in the north-western
Adriatic Sea: insight from machine learning
###  authors: Silvia Bonizzoni, Blair D. Sterba-Boatwright, Sarah Piwetz,
Giovanni Bearzi


library(ROSE)
library(randomForest)
#  library(AUC) # compute area under receiver operating curve
#  library(generalhoslem) # goodness of fit function for logistic

example("ROSE-package")

data <- read.csv("~/Desktop/myNewDataForInterpret.csv")
str(data)
data$presence <- factor(data$presence)
data.ROSE <- ROSE(presence ~ ., data = data)

rf <- randomForest(presence ~ ., data = data.ROSE$data, do.trace = 50)
rf.predict <- predict(rf, newdata = data, type = "prob")[,2]
rf_roc <- roc.curve(data$presence, rf.predict)
rf_roc$auc

#  Cross-validation similar to EBM methods
gp.no <- (1:nrow(data)) %% 5
gp.no <- sample(gp.no, length(gp.no))
accum <- data.frame(presence = data$presence, pred = 0)
for (k in 0:4) {
  data.k.train <- data[gp.no != k, ]
  data.k.test <- data[gp.no == k, ]
  data.k.ROSE <- ROSE(presence ~ ., data = data.k.train)
  rf.k <- randomForest(presence ~ ., data = data.k.ROSE$data, do.trace =
50)
  rf.predict.k <- predict(rf.k, newdata = data.k.test, type = "prob")[,
2]
  rf.roc.k <- roc.curve(data.k.test$presence, rf.predict.k)
  print(rf.roc.k$auc)
  accum$pred[gp.no == k] <- rf.predict.k
}
rf.all <- roc.curve(accum$presence, accum$pred)
rf.all$auc

aucv <- numeric(10)
for (i in 1:10) {
  gp.no <- sample(gp.no, length(gp.no))
  accum <- data.frame(presence = data$presence, pred = 0)
  for (k in 0:4) {
    data.k.train <- data[gp.no != k, ]
    data.k.test <- data[gp.no == k, ]
    data.k.ROSE <- ROSE(presence ~ ., data = data.k.train)
    rf.k <- randomForest(presence ~ ., data = data.k.ROSE$data)
    rf.predict.k <- predict(rf.k, newdata = data.k.test, type = "prob")[,
2]
    rf.roc.k <- roc.curve(data.k.test$presence, rf.predict.k)
    print(rf.roc.k$auc)
    accum$pred[gp.no == k] <- rf.predict.k
    print(paste("k =", k))
  }
```

```
  rf.all <- roc.curve(accum$presence, accum$pred)
  aucv[i] <- rf.all$auc
  print(paste("i =", i))
}
aucv
#  [1] 0.7920859 0.7805680 0.7883114 0.7884030 0.7727770 0.7774747
#  [7] 0.7937165 0.7854725 0.7876285 0.7725924
mean(aucv) # 0.784
sd(aucv) # 0.0077

roc_sw <- roc(accum$pred, accum$presence)
auc(roc_sw)
plot(roc_sw)
#  sensitivity + specificity - 1 => y + (1 - x) - 1 = c => y = x + c
abline(0, 1, col = "red")
length(unique(accum$pred))
accum$x <- 0; accum$y <- 0
accum <- arrange(accum, pred)
whack <- sensitivity(accum$pred, accum$presence)
thud <- specificity(accum$pred, accum$presence)
lines(1 - whack$cutoffs, whack$measure, col = "red")
max_c <- 0; max_sens <- 0; max_spec <- 0
for (i in 1:length(whack$cutoffs)) {
  this_c <- whack$measure[i] + whack$cutoffs[i] - 1
  if (this_c > max_c) {
    max_c <- this_c; max_sens <- whack$measure[i]; max_spec <-
whack$cutoffs[i]
  }
}
max_c # 0.427976
max_sens # 0.68
max_spec # 0.747976
```