RF

Yiru Gong, yg2832

2022-05-11

```
library(tidyverse)
library(summarytools)
library(corrplot)
library(caret)
library(MASS)
library(mlbench)
library(pROC) #ROCR
library(pdp)
library(vip)
library(AppliedPredictiveModeling) #for transparentTheme function
library(ISLR)
library(caret)
library(e1071)
library(kernlab)
library(keras)
library(tfruns)
library(ranger)
```

Data Input

```
data = read.csv('Covid19_vacc_predict_handout.csv')
data = data %>%
  na.omit() %>%
  dplyr::select(-id) %>%
  mutate(
    atlas_type_2015_mining_no = factor(atlas_type_2015_mining_no),
    covid_vaccination = factor(covid_vaccination),
    hum_region = factor(hum_region),
    sex_cd = factor(sex_cd),
    race_cd = factor(race_cd),
    lang_spoken_cd = factor(lang_spoken_cd),
    atlas_low_education_2015_update = factor(atlas_low_education_2015_update)
    )
dfSummary(data[,c(5,7,8,10,11,17,18)])
```

Data Frame Summary
Dimensions: 8308 x 7
Duplicates: 7802

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	atlas_type_2015_mining_no [factor]	1. 0 2. 1	8177 (98.4%) 131 (1.6%)	IIIIIIIIIIIIIIIIIII	8308 (100.0%)	0 (0.0%)
2	covid_vaccination [factor]	1. no_vacc 2. vacc	6682 (80.4%) 1626 (19.6%)	IIIIIIIIIIIIIIII III	8308 (100.0%)	0 (0.0%)
3	hum_region [factor]	1. CALIFORNIA/NEVADA 2. CENTRAL 3. CENTRAL WEST 4. EAST 5. EAST CENTRAL 6. FLORIDA 7. GREAT LAKES/CENTRAL NORTH 8. GULF STATES 9. INTERMOUNTAIN 10. MID-ATLANTIC/NORTH CAROLI [5 others]	299 (3.6%) 551 (6.6%) 238 (2.9%) 491 (5.9%) 1370 (16.5%) 607 (7.3%) 1111 (13.4%) 454 (5.5%) 220 (2.6%) 845 (10.2%) 2122 (25.5%)	I	8308 (100.0%)	0 (0.0%)
4	sex_cd [factor]	1. F 2. M	4527 (54.5%) 3781 (45.5%)	IIIIIIIIII IIIIIIIII	8308 (100.0%)	0 (0.0%)
5	lang_spoken_cd [factor]	1. * 2. CHI 3. CRE 4. ENG 5. KOR 6. OTH 7. SPA 8. VIE	10 (0.1%) 13 (0.2%) 4 (0.0%) 7957 (95.8%) 7 (0.1%) 34 (0.4%) 276 (3.3%) 7 (0.1%)		8308 (100.0%)	0 (0.0%)
6	atlas_low_education_2015_update [factor]	1. 0 2. 1	7769 (93.5%) 539 (6.5%)	IIIIIIIIIIIIIIIIII I	8308 (100.0%)	0 (0.0%)
7	race_cd [factor]	1. 0 2. 1 3. 2 4. 3 5. 4 6. 5 7. 6	160 (1.9%) 7317 (88.1%) 558 (6.7%) 80 (1.0%) 56 (0.7%) 129 (1.6%) 8 (0.1%)	IIIIIIIIIIIIIIIII I	8308 (100.0%)	0 (0.0%)

```r
data2 = model.matrix(covid_vaccination ~ ., data)[ ,-1]
```

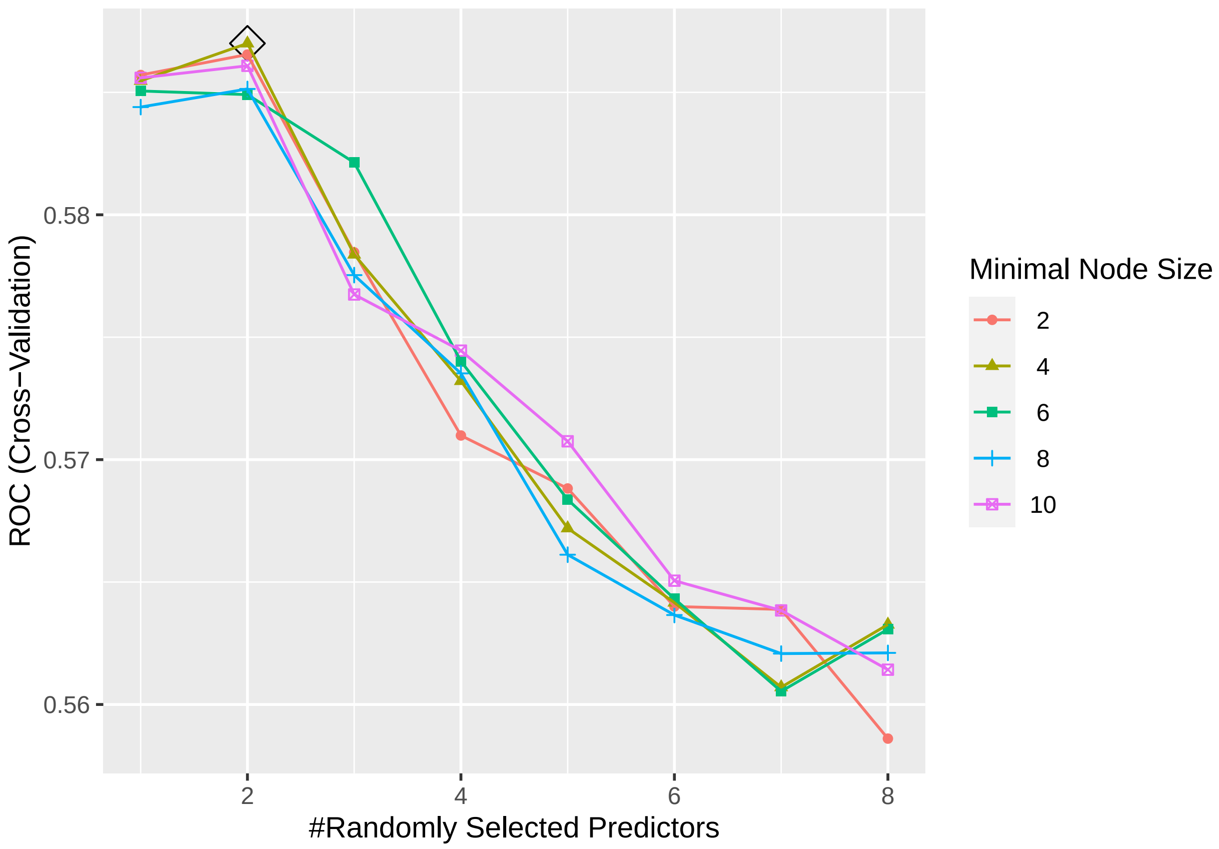
Data split

```r
set.seed(1)
rowTrain <- createDataPartition(y = data$covid_vaccination,
                                p = 0.7,
                                list = FALSE)
x = data2[rowTrain,]
```

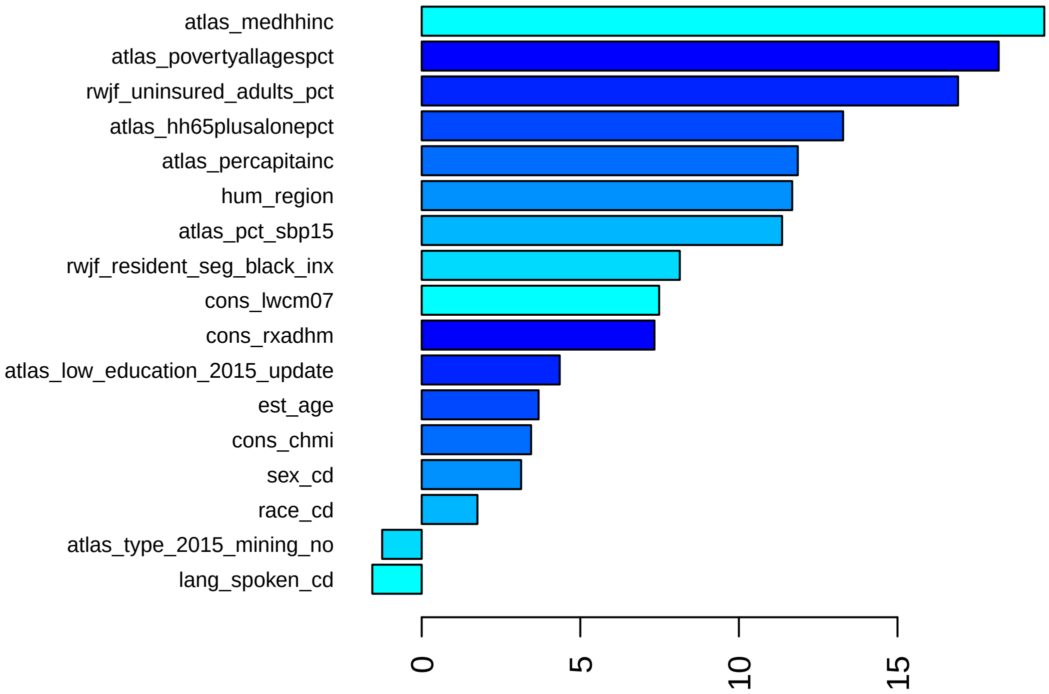
```
y = data$covid_vaccination[rowTrain]
x2 = data2[-rowTrain,]
y2 = data$covid_vaccination[-rowTrain]

save(x,y,x2,y2,file = "split_data.Rdata")
```

Random Forest

```
ctrl <- trainControl(method = "cv",
                     summaryFunction = twoClassSummary,
                     classProbs = TRUE)

rf.grid <- expand.grid(mtry = 1:8,
                       splitrule = "gini",
                       min.node.size = seq(from = 2, to = 10,
                                           by = 2))

set.seed(1)
rf.fit <- train(covid_vaccination ~ . ,
                data,
                subset = rowTrain,
                method = "ranger",
                tuneGrid = rf.grid,
                metric = "ROC",
                trControl = ctrl)

ggplot(rf.fit, highlight = TRUE)
```



```r
# variable importance
set.seed(1)
rf2.final.per <- ranger(covid_vaccination ~ . ,
                data[rowTrain,],
                mtry = rf.fit$bestTune[[1]],
                min.node.size = rf.fit$bestTune[[3]],
                splitrule = "gini",
                importance = "permutation",
                scale.permutation.importance = TRUE)

par(mar = c(3,12,3,3))
barplot(sort(ranger::importance(rf2.final.per), decreasing = FALSE),
        las = 2, horiz = TRUE, cex.names = 0.7,
        col = colorRampPalette(colors = c("cyan","blue"))(8))
```


atlas_medhhinc
atlas_povertyallagespct
rwjf_uninsured_adults_pct
atlas_hh65plusalonepct
atlas_percapitainc
hum_region
atlas_pct_sbp15
rwjf_resident_seg_black_inx
cons_lwcm07
cons_rxadhm
atlas_low_education_2015_update
est_age
cons_chmi
sex_cd
race_cd
atlas_type_2015_mining_no
lang_spoken_cd
0
5
10
15