Random Forest

Carrie Cheng

2023-04-30

```
library(dplyr)
dat <- read.csv('brfss_final.csv')
outcome <- data.frame(dat$X,dat$MICHD,dat$CVDINFR4,dat$CVDCRHD4)
outcome %>% group_by(dat.MICHD) %>% summarise(count=n())
```

```
## # A tibble: 2 x 2
##   dat.MICHD count
##       <int> <int>
## 1         1 14580
## 2         2 14580
```

```
outcome %>% group_by(dat.CVDINFR4) %>% summarise(count=n())
```

```
## # A tibble: 4 x 2
##   dat.CVDINFR4 count
##          <int> <int>
## 1            1  9188
## 2            2 19802
## 3            7   160
## 4            9    10
```

```
outcome %>% group_by(dat.CVDCRHD4) %>% summarise(count=n())
```

```
## # A tibble: 4 x 2
##   dat.CVDCRHD4 count
##          <int> <int>
## 1            1  9729
## 2            2 18874
## 3            7   550
## 4            9     7
```

```
## remove the ones that responded don't know & not sure in CVDINFR4 & CVDCRHD4
dat <- dat[-which(dat$CVDINFR4 == 7 | dat$CVDINFR4 == 9),]
dat <- dat[-which(dat$CVDCRHD4 == 7 | dat$CVDCRHD4 == 9),]
# remove columns that has only 1 value for all rows
dat <- dat[ , -which(names(dat) %in% c("MEDSHEPB","TOLDCFS", "HAVECFS", "WORKCFS"))]
```

Drop columns with more than 5% data missing, impute the rest using KNN

```r
# convert outcome variables
dat$MICHD <- factor(2-dat$MICHD)
dat$CVDINFR4 <- factor(2-dat$CVDINFR4)
dat$CVDCRHD4 <- factor(2-dat$CVDCRHD4)
# i believe X is the index column, not needed
# remove weights
dat <- dat[, !colnames(dat) %in% c('X', 'LLCPWT2', 'LLCPWT', 'CLLCPWT','STRWT','WT2RAKE')]
dat <- dat[, !colnames(dat) %in% c('QSTVER', 'STSTR','RAWRAKE')] # remove based on knowledge
threshold <- .1
ncol(dat) # 190
```

```
## [1] 187
```

```r
dat <- dat[, colMeans(is.na(dat)) <= threshold]
ncol(dat) # 52 columns left
```

```
## [1] 55
```

```r
columns_to_impute <- colnames(dat)[colSums(is.na(dat)) > 0]
columns_to_impute
```

```
##  [1] "TOLDHI3"  "CHOLMED3" "CPDEMO1B" "VETERAN3" "EMPLOY1"  "INCOME3"
##  [7] "DEAF"     "BLIND"    "DECIDE"   "DIFFWALK" "DIFFDRES" "DIFFALON"
## [13] "USENOW3"  "METSTAT"  "URBSTAT"  "MSCODE"   "DRDXAR3"  "BMI5"
## [19] "AIDTST4"  "FTJUDA2_" "GRENDA1_"
```

```r
str(dat[,columns_to_impute])
```

```
## 'data.frame':    28433 obs. of  21 variables:
##  $ TOLDHI3 : int  1 2 1 1 2 1 1 1 1 1 ...
##  $ CHOLMED3: int  1 2 1 1 2 2 1 1 1 1 ...
##  $ CPDEMO1B: int  1 1 8 1 1 8 8 1 1 2 ...
##  $ VETERAN3: int  2 2 2 2 1 2 1 2 2 2 ...
##  $ EMPLOY1 : int  8 7 2 7 7 7 7 8 7 7 ...
##  $ INCOME3 : int  77 3 99 77 7 99 5 77 5 10 ...
##  $ DEAF    : int  2 2 2 2 2 2 1 2 2 2 ...
##  $ BLIND   : int  1 2 2 2 2 2 2 2 2 2 ...
##  $ DECIDE  : int  1 2 1 2 1 2 2 2 2 2 ...
##  $ DIFFWALK: int  1 2 2 2 2 1 1 1 2 2 ...
##  $ DIFFDRES: int  2 2 2 2 2 1 2 2 2 2 ...
##  $ DIFFALON: int  1 2 2 2 2 1 1 2 2 2 ...
##  $ USENOW3 : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ METSTAT : int  1 1 1 1 1 2 1 2 1 1 ...
##  $ URBSTAT : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ MSCODE  : int  2 1 3 1 3 2 2 5 2 3 ...
##  $ DRDXAR3 : int  1 2 1 1 2 1 1 2 1 1 ...
##  $ BMI5    : int  NA 2829 2274 2923 2392 3543 NA NA 2224 3497 ...
##  $ AIDTST4 : int  2 2 2 2 2 1 2 2 1 2 ...
##  $ FTJUDA2_: int  0 0 0 29 7 NA 100 0 0 3 ...
##  $ GRENDA1_: int  14 0 29 71 29 14 3 10 14 17 ...
```

```r
complete_columns <- colnames(dat)[colSums(is.na(dat)) == 0 &
                                  !colnames(dat) %in% c('MICHD', 'CVDINFR4','CVDCRHD4')]
miss_names <- paste0("miss_", columns_to_impute)
dat[, miss_names] <- NA
for (i in 1:nrow(dat)){
  for (j in 1:length(miss_names)) {
    dat[i, miss_names[j]] <- as.numeric(any(is.na(dat[i, columns_to_impute[j]])))
  }
}
for (c in columns_to_impute) {
    col <- dat[[c]]
    scaled <- scale(dat[, complete_columns])
    knn <- knn(
        train = scaled[!is.na(col), complete_columns],
        test  = scaled[is.na(col), complete_columns],
        cl    = dat[!is.na(col), c]
        )

    dat[is.na(col), c] = knn
}
colSums(is.na(dat))
```

```
##       GENHLTH      PHYSHLTH      MENTHLTH      PRIMINSR      PERSDOC3
##             0             0             0             0             0
##      MEDCOST1      CHECKUP1       TOLDHI3      CHOLMED3      CVDINFR4
##             0             0             0             0             0
##      CVDCRHD4      CVDSTRK3      CHCSCNCR      CHCOCNCR      CHCCOPD3
##             0             0             0             0             0
##      ADDEPEV3      CHCKDNY2      DIABETE4       MARITAL      RENTHOM1
##             0             0             0             0             0
##      NUMHHOL3      CPDEMO1B      VETERAN3       EMPLOY1       INCOME3
##             0             0             0             0             0
##          DEAF         BLIND        DECIDE      DIFFWALK      DIFFDRES
##             0             0             0             0             0
##      DIFFALON       USENOW3       QSTLANG       METSTAT       URBSTAT
##             0             0             0             0             0
##        MSCODE       DUALUSE       TOTINDA       RFHYPE6       CHOLCH3
##             0             0             0             0             0
##         MICHD       ASTHMS1       DRDXAR3          RACE           SEX
##             0             0             0             0             0
##         AGE80          BMI5       CHLDCNT        EDUCAG       SMOKER3
##             0             0             0             0             0
##       CURECI1      DROCDY3_       AIDTST4      FTJUDA2_      GRENDA1_
##             0             0             0             0             0
##  miss_TOLDHI3 miss_CHOLMED3 miss_CPDEMO1B miss_VETERAN3  miss_EMPLOY1
##             0             0             0             0             0
##  miss_INCOME3     miss_DEAF    miss_BLIND   miss_DECIDE miss_DIFFWALK
##             0             0             0             0             0
## miss_DIFFDRES miss_DIFFALON  miss_USENOW3  miss_METSTAT  miss_URBSTAT
##             0             0             0             0             0
##   miss_MSCODE  miss_DRDXAR3     miss_BMI5  miss_AIDTST4 miss_FTJUDA2_
##             0             0             0             0             0
## miss_GRENDA1_
```

```
##             0
```

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
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:ggplot2':
##
##     margin
```

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

```
library(ggplot2)
library(ROCR)
set.seed(263)
train_index <- createDataPartition(dat$MICHD, p = 0.8, list = FALSE)
train <- dat[train_index, ]
test <- dat[-train_index, ]
train$weights <- ifelse(as.numeric(train$MICHD) == 1,
                        1/mean(as.numeric(train$MICHD) == 1),
                        1/(1-mean(as.numeric(train$MICHD) == 1)))
train$weights <- as.numeric(train$weights)
test$weights <- ifelse(as.numeric(test$MICHD) == 1,
                        1/mean(as.numeric(test$MICHD) == 1),
                        1/(1-mean(as.numeric(test$MICHD) == 1)))
test$weights <- as.numeric(test$weights)
index_weight <- which(names(train) == "weights")
summary(train)
```

```
##     GENHLTH         PHYSHLTH        MENTHLTH        PRIMINSR
##  Min.   :1.000   Min.   : 1.00   Min.   : 1.00   Min.   : 1.000
##  1st Qu.:2.000   1st Qu.:20.00   1st Qu.:30.00   1st Qu.: 3.000
##  Median :3.000   Median :88.00   Median :88.00   Median : 3.000
##  Mean   :2.941   Mean   :59.02   Mean   :65.33   Mean   : 7.463
##  3rd Qu.:4.000   3rd Qu.:88.00   3rd Qu.:88.00   3rd Qu.: 3.000
##  Max.   :9.000   Max.   :99.00   Max.   :99.00   Max.   :99.000
##     PERSDOC3        MEDCOST1        CHECKUP1        TOLDHI3         CHOLMED3
```

```
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.0
##  1st Qu.:1.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.0
##  Median :1.000   Median :2.000   Median :1.000   Median :1.000   Median :1.0
##  Mean   :1.498   Mean   :1.972   Mean   :1.246   Mean   :1.498   Mean   :1.5
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:2.0
##  Max.   :9.000   Max.   :9.000   Max.   :9.000   Max.   :9.000   Max.   :9.0
##  CVDINFR4  CVDCRHD4     CVDSTRK3        CHCSCNCR        CHCOCNCR
##  0:15844   0:15091   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1: 6903   1: 7656   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##                      Median :2.000   Median :2.000   Median :2.000
##                      Mean   :1.921   Mean   :1.835   Mean   :1.837
##                      3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000
##                      Max.   :9.000   Max.   :9.000   Max.   :9.000
##     CHCCOPD3        ADDEPEV3        CHCKDNY2        DIABETE4
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:1.000
##  Median :2.000   Median :2.000   Median :2.000   Median :3.000
##  Mean   :1.867   Mean   :1.854   Mean   :1.938   Mean   :2.532
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:3.000
##  Max.   :9.000   Max.   :9.000   Max.   :9.000   Max.   :9.000
##     MARITAL         RENTHOM1        NUMHHOL3        CPDEMO1B
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000
##  Median :2.000   Median :1.000   Median :2.000   Median :1.000
##  Mean   :2.173   Mean   :1.248   Mean   :1.789   Mean   :2.345
##  3rd Qu.:3.000   3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :9.000   Max.   :9.000   Max.   :9.000   Max.   :9.000
##     VETERAN3        EMPLOY1         INCOME3           DEAF
##  Min.   :1.000   Min.   :1.000   Min.   : 1.00   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:5.000   1st Qu.: 5.00   1st Qu.:2.000
##  Median :2.000   Median :7.000   Median : 7.00   Median :2.000
##  Mean   :1.832   Mean   :5.701   Mean   :25.92   Mean   :1.878
##  3rd Qu.:2.000   3rd Qu.:7.000   3rd Qu.:11.00   3rd Qu.:2.000
##  Max.   :9.000   Max.   :9.000   Max.   :99.00   Max.   :9.000
##      BLIND           DECIDE         DIFFWALK        DIFFDRES
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000
##  Median :2.000   Median :2.000   Median :2.000   Median :2.000
##  Mean   :1.967   Mean   :1.957   Mean   :1.758   Mean   :1.971
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :9.000   Max.   :9.000   Max.   :9.000   Max.   :9.000
##     DIFFALON        USENOW3         QSTLANG         METSTAT
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:3.000   1st Qu.:1.000   1st Qu.:1.000
##  Median :2.000   Median :3.000   Median :1.000   Median :1.000
##  Mean   :1.934   Mean   :3.027   Mean   :1.007   Mean   :1.381
##  3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:1.000   3rd Qu.:2.000
##  Max.   :9.000   Max.   :9.000   Max.   :2.000   Max.   :2.000
##     URBSTAT         MSCODE         DUALUSE         TOTINDA
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000
##  Median :1.000   Median :3.000   Median :1.000   Median :1.000
##  Mean   :1.199   Mean   :3.031   Mean   :2.326   Mean   :1.362
##  3rd Qu.:1.000   3rd Qu.:5.000   3rd Qu.:1.000   3rd Qu.:2.000
```

```
##  Max.   :2.000   Max.   :5.000   Max.   :9.000   Max.   :9.000
##     RFHYPE6         CHOLCH3      MICHD         ASTHMS1        DRDXAR3
##  Min.   :1.000   Min.   :1.000   0:11664   Min.   :1.00   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:1.000   1:11083   1st Qu.:3.00   1st Qu.:1.000
##  Median :2.000   Median :1.000             Median :3.00   Median :1.000
##  Mean   :1.665   Mean   :1.429             Mean   :2.81   Mean   :1.476
##  3rd Qu.:2.000   3rd Qu.:1.000             3rd Qu.:3.00   3rd Qu.:2.000
##  Max.   :9.000   Max.   :9.000             Max.   :9.00   Max.   :2.000
##       RACE            SEX            AGE80            BMI5         CHLDCNT
##  Min.   :1.000   Min.   :1.000   Min.   :18.00   Min.   :   6   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:65.00   1st Qu.:2296   1st Qu.:1.000
##  Median :1.000   Median :2.000   Median :72.00   Median :2663   Median :1.000
##  Mean   :1.632   Mean   :1.571   Mean   :69.94   Mean   :2647   Mean   :1.272
##  3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:80.00   3rd Qu.:3099   3rd Qu.:1.000
##  Max.   :9.000   Max.   :2.000   Max.   :80.00   Max.   :9684   Max.   :9.000
##      EDUCAG        SMOKER3        CURECI1          DROCDY3_         AIDTST4
##  Min.   :1.00   Min.   :1.00   Min.   :1.000   Min.   :  0.00   Min.   :1.000
##  1st Qu.:2.00   1st Qu.:3.00   1st Qu.:1.000   1st Qu.:  0.00   1st Qu.:2.000
##  Median :3.00   Median :4.00   Median :1.000   Median :  0.00   Median :2.000
##  Mean   :2.99   Mean   :3.63   Mean   :1.399   Mean   : 69.66   Mean   :2.272
##  3rd Qu.:4.00   3rd Qu.:4.00   3rd Qu.:1.000   3rd Qu.: 17.00   3rd Qu.:2.000
##  Max.   :9.00   Max.   :9.00   Max.   :9.000   Max.   :900.00   Max.   :9.000
##     FTJUDA2_          GRENDA1_        miss_TOLDHI3     miss_CHOLMED3
##  Min.   :   0.00   Min.   :   0.00   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:   0.00   1st Qu.:  14.00   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :   3.00   Median :  36.00   Median :0.00000   Median :0.00000
##  Mean   :  42.71   Mean   :  68.57   Mean   :0.07008   Mean   :0.07148
##  3rd Qu.:  43.00   3rd Qu.:  71.00   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :9900.00   Max.   :9900.00   Max.   :1.00000   Max.   :1.00000
##  miss_CPDEMO1B      miss_VETERAN3       miss_EMPLOY1       miss_INCOME3
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000   Min.   :0.00000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.00000
##  Median :0.000000   Median :0.000000   Median :0.000000   Median :0.00000
##  Mean   :0.001583   Mean   :0.004352   Mean   :0.008045   Mean   :0.01706
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.00000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.000000   Max.   :1.00000
##    miss_DEAF        miss_BLIND       miss_DECIDE      miss_DIFFWALK
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.0000   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.0284   Mean   :0.03047   Mean   :0.03319   Mean   :0.03592
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##  miss_DIFFDRES     miss_DIFFALON      miss_USENOW3      miss_METSTAT
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.000000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.000000
##  Mean   :0.03768   Mean   :0.03983   Mean   :0.04401   Mean   :0.008309
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.000000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.000000
##   miss_URBSTAT       miss_MSCODE       miss_DRDXAR3        miss_BMI5
##  Min.   :0.000000   Min.   :0.00000   Min.   :0.000000   Min.   :0.00000
##  1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.00000
##  Median :0.000000   Median :0.00000   Median :0.000000   Median :0.00000
```

```
##  Mean   :0.008309   Mean   :0.01376   Mean   :0.007166   Mean   :0.09592
##  3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.00000
##  Max.   :1.000000   Max.   :1.00000   Max.   :1.000000   Max.   :1.00000
##   miss_AIDTST4     miss_FTJUDA2_     miss_GRENDA1_        weights
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :1.950
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:1.950
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :1.950
##  Mean   :0.05873   Mean   :0.09148   Mean   :0.09184   Mean   :2.000
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:2.052
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :2.052
```

Parameter Tuning

Let's tune number of trees ntrees and number of features selected to place split mtry. In the following, let's use 10-fold cross-validation.

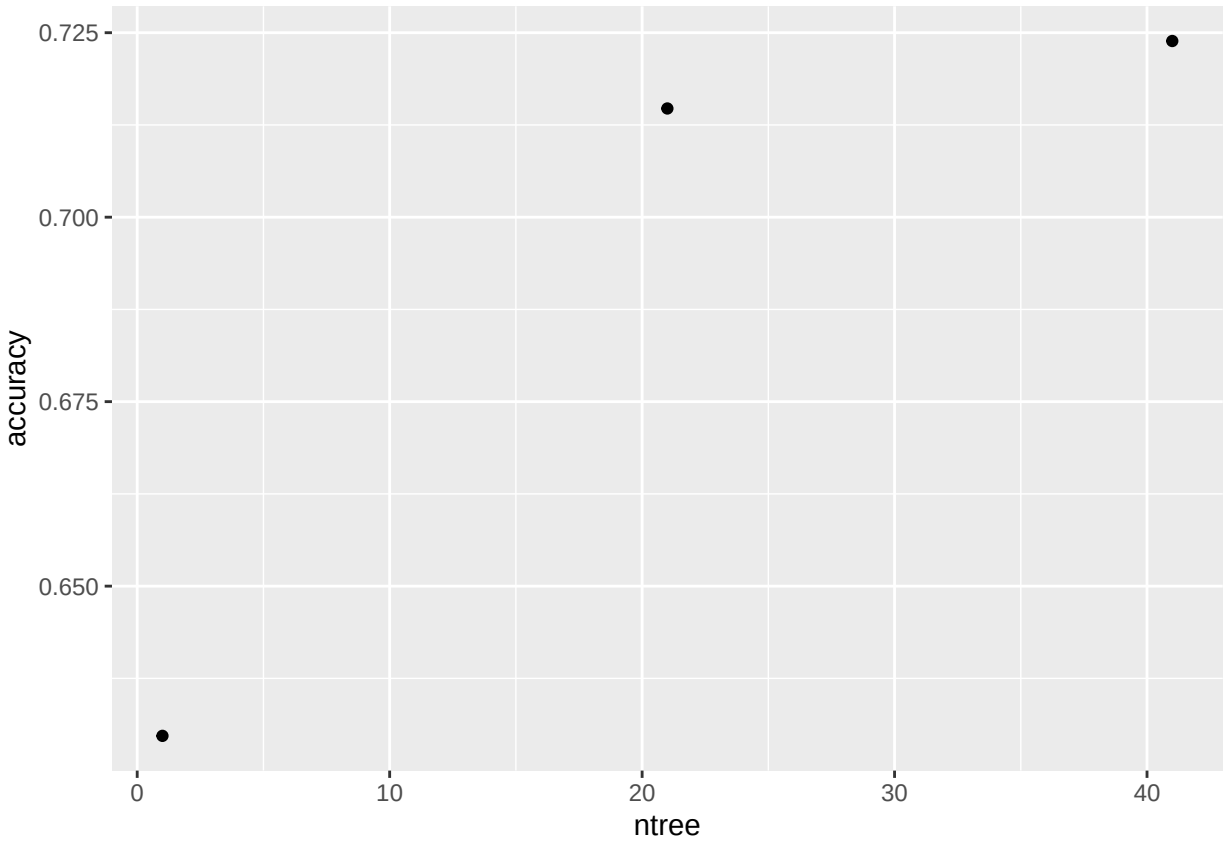
```
## get index of the other two outcomes

index_michd <- which(names(train) == "MICHD")
index_infr <- which(names(train) == "CVDINFR4")
index_crhd <- which(names(train) == "CVDCRHD4")
```

Tune number of trees

Let's set mtry = 10.

```
ntree <- seq(1, 51, by = 20)
accuracy <- sapply(ntree, function(n){
  train(as.factor(MICHD) ~ ., method = "rf",
        data = train[, -c(index_infr, index_crhd, index_weight)],
        weights = train$weights,
        tuneGrid = data.frame(mtry = 10),
        ntree = n, trControl = trainControl(method = "cv", number = 10))$results$Accuracy
})

qplot(ntree, accuracy)
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

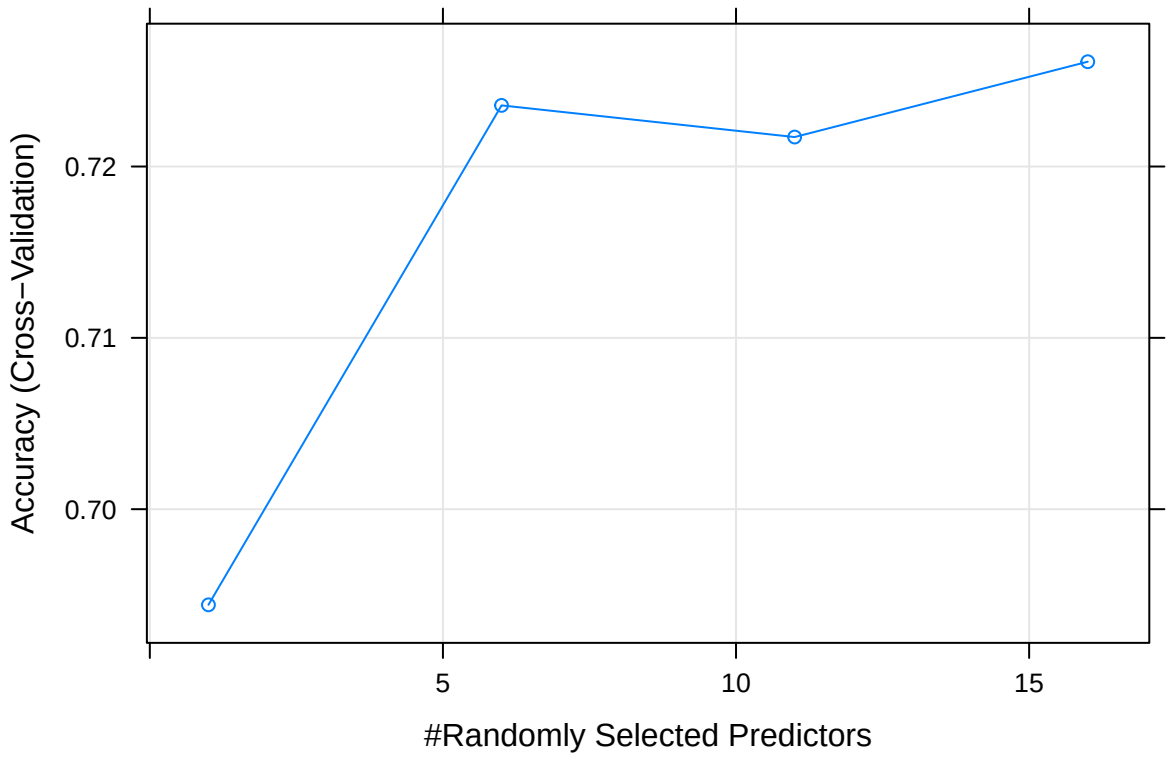



```
best_ntree <- ntree[which(accuracy == max(accuracy))]
best_ntree <- min(best_ntree)
print(paste("The best ntree is", best_ntree))
```

```
## [1] "The best ntree is 41"
```

Tune mtry

```
train_rf <- train(as.factor(MICHD) ~ ., method = "rf",
                  data = train[, -c(index_infr, index_crhd, index_weight)],
                  weights = train$weights,
                  tuneGrid = data.frame(mtry = seq(1, 20, by = 5)),
                  ntree = best_ntree,
                  nodesize = 10, trControl = trainControl(method = "cv", number = 10))

plot(train_rf)
```




```
best_mtry <- train_rf$bestTune

result_cv <- train_rf$results

print(paste("The best mtry is ", best_mtry))
```

```
## [1] "The best mtry is  16"
```

Use the best model to train random forest

The below is the confusion matrix on the test set.

```
rf_best <- randomForest(as.factor(MICHD) ~.,
                        data = train[, -c(index_infr, index_crhd, index_weight)],
                        mtry = best_mtry[[1]], ntree = best_ntree, nodesize = 10,
                        weights = train$weights)

pred_test <- predict(rf_best, test)
cm_test <- confusionMatrix(as.factor(pred_test), as.factor(test$MICHD))

cm_test
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 2060  655
```

```
##          1  856 2115
##
##                Accuracy : 0.7343
##                  95% CI : (0.7226, 0.7457)
##     No Information Rate : 0.5128
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4691
##
##  Mcnemar's Test P-Value : 2.673e-07
##
##             Sensitivity : 0.7064
##             Specificity : 0.7635
##          Pos Pred Value : 0.7587
##          Neg Pred Value : 0.7119
##              Prevalence : 0.5128
##          Detection Rate : 0.3623
##    Detection Prevalence : 0.4775
##       Balanced Accuracy : 0.7350
##
##        'Positive' Class : 0
##
```

```
metric_test <- c(cm_test$overall[["Accuracy"]],
                 cm_test$byClass[c("Sensitivity","Specificity")])

cat(paste("The overall accuracy using the best tuned random forest model is",
      metric_test[1], "\n",
      "Sensitivity is", metric_test[2], "\n",
      "Specificity is", metric_test[3]))
```

```
## The overall accuracy using the best tuned random forest model is 0.73425958494548
##  Sensitivity is 0.706447187928669
##  Specificity is 0.763537906137184
```

ROC curve

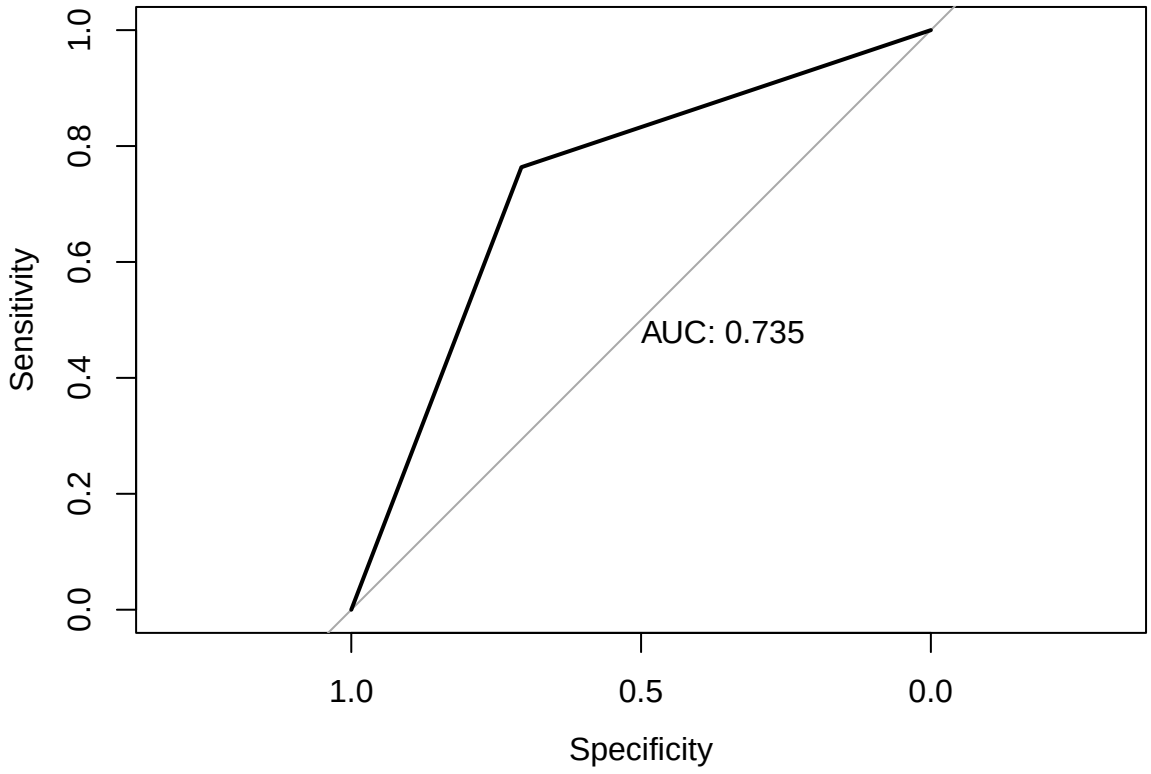
```
library(pROC)
```

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

##
## Attaching package: 'pROC'

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

```
roc_rf <- roc(as.numeric(test$MICHD) ~ as.numeric(pred_test),
              plot = TRUE, print.auc = TRUE)
```

```
## Setting levels: control = 1, case = 2
```

```
## Setting direction: controls < cases
```




```
print(paste("AUC is", as.numeric(roc_rf$auc)))
```

```
## [1] "AUC is 0.734992547032927"
```

Importance Features

```
importance(rf_best)
```

```
##               MeanDecreaseGini
## GENHLTH             711.862637
## PHYSHLTH            226.248796
## MENTHLTH            162.379993
## PRIMINSR            194.612805
## PERSDOC3            156.118201
## MEDCOST1             21.074964
## CHECKUP1             66.919519
## TOLDHI3             130.041153
```

```
## CHOLMED3           831.827109
## CVDSTRK3           175.570332
## CHCSCNCR            60.878042
## CHCOCNCR            59.803548
## CHCCOPD3           137.168431
## ADDEPEV3            48.680383
## CHCKDNY2            67.959545
## DIABETE4           150.956374
## MARITAL            140.603848
## RENTHOM1            63.952908
## NUMHHOL3            68.655144
## CPDEMO1B            91.664466
## VETERAN3            85.280955
## EMPLOY1            222.730660
## INCOME3            268.645292
## DEAF                64.241666
## BLIND               45.225568
## DECIDE              51.402183
## DIFFWALK           217.148434
## DIFFDRES            29.247107
## DIFFALON            62.481659
## USENOW3             26.324198
## QSTLANG              4.940709
## METSTAT             41.278064
## URBSTAT             42.644058
## MSCODE             116.152190
## DUALUSE             32.801411
## TOTINDA             63.744204
## RFHYPE6            206.316823
## CHOLCH3             39.530379
## ASTHMS1             73.892691
## DRDXAR3             71.853098
## RACE               105.902927
## SEX                216.899257
## AGE80              630.905354
## BMI5               512.703215
## CHLDCNT             39.686864
## EDUCAG             128.522533
## SMOKER3            137.302086
## CURECI1             17.913457
## DROCDY3_           201.076068
## AIDTST4             85.085988
## FTJUDA2_           235.275564
## GRENDA1_           319.942699
## miss_TOLDHI3        20.752127
## miss_CHOLMED3       25.872908
## miss_CPDEMO1B        1.307365
## miss_VETERAN3        2.226573
## miss_EMPLOY1         2.335442
## miss_INCOME3         4.162709
## miss_DEAF            3.905504
## miss_BLIND           4.291004
## miss_DECIDE          4.847821
## miss_DIFFWALK        5.017279
```

```
## miss_DIFFDRES         4.075043
## miss_DIFFALON         4.461494
## miss_USENOW3          5.820074
## miss_METSTAT          4.855854
## miss_URBSTAT          5.522441
## miss_MSCODE          10.521266
## miss_DRDXAR3          8.753314
## miss_BMI5            20.971275
## miss_AIDTST4         11.670755
## miss_FTJUDA2_        25.047514
## miss_GRENDA1_        24.284334
```

```
varImpPlot(rf_best)
```