ams580\_hw5

Mustafa Yigit Isik

2023-03-08

#**Question 1**

#Install Packs  
if (!requireNamespace("tidyverse")) install.packages('tidyverse')

## Loading required namespace: tidyverse

if (!requireNamespace("caret")) install.packages('caret')

## Loading required namespace: caret

if (!requireNamespace("randomForest")) install.packages('randomForest')

## Loading required namespace: randomForest

if (!requireNamespace("party")) install.packages('party')

## Loading required namespace: party

library(tidyverse)

## ── Attaching packages  
## ───────────────────────────────────────  
## tidyverse 1.3.2 ──

## ✔ ggplot2 3.4.0 ✔ purrr 1.0.1   
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.3.0 ✔ stringr 1.5.0   
## ✔ readr 2.1.4 ✔ forcats 1.0.0   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(caret)

## Loading required package: lattice  
##   
## Attaching package: 'caret'  
##   
## The following object is masked from 'package:purrr':  
##   
## lift

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

library(party)

## Loading required package: grid  
## Loading required package: mvtnorm  
## Loading required package: modeltools  
## Loading required package: stats4  
## Loading required package: strucchange  
## Loading required package: zoo  
##   
## Attaching package: 'zoo'  
##   
## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric  
##   
## Loading required package: sandwich  
##   
## Attaching package: 'strucchange'  
##   
## The following object is masked from 'package:stringr':  
##   
## boundary

#Read/Clean Data  
data <- read.csv('/Users/mustafayigitisik/Desktop/stuff/semesters/spring 2023/ams 580/hw5\_rf\_gini/Titanic1.csv')  
#remove the instructed 3 columns  
data <- subset(data, select = -c(Name, Ticket, Cabin))  
#remove rows where age is empty   
data <- data[!(is.na(data$Age) | data$Age==""), ]  
cat('There are', nrow(data), 'observations left.')

## There are 714 observations left.

data$Survived <- as.factor(data$Survived)  
data$Sex <- ifelse(data$Sex=="male",1,0)  
str(data)

## 'data.frame': 714 obs. of 9 variables:  
## $ PassengerId: int 1 2 3 4 5 7 8 9 10 11 ...  
## $ Survived : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 2 2 2 ...  
## $ Pclass : int 3 1 3 1 3 1 3 3 2 3 ...  
## $ Sex : num 1 0 0 0 1 1 1 0 0 0 ...  
## $ Age : num 22 38 26 35 35 54 2 27 14 4 ...  
## $ SibSp : int 1 1 0 1 0 0 3 0 1 1 ...  
## $ Parch : int 0 0 0 0 0 0 1 2 0 1 ...  
## $ Fare : num 7.25 71.28 7.92 53.1 8.05 ...  
## $ Embarked : chr "S" "C" "S" "S" ...

#Split Data  
set.seed(123)  
training.samples <- data$Survived %>%   
 createDataPartition(p = 0.75, list = FALSE)  
train.data <- data[training.samples, ]  
test.data <- data[-training.samples, ]

# Question 2

set.seed(123)  
model <- train(  
 Survived ~., data = train.data, method = "rf",  
 trControl = trainControl("cv", number = 10),  
 importance = TRUE  
 )  
model$bestTune

## mtry  
## 1 2

model$finalModel

##   
## Call:  
## randomForest(x = x, y = y, mtry = param$mtry, importance = TRUE)   
## Type of random forest: classification  
## Number of trees: 500  
## No. of variables tried at each split: 2  
##   
## OOB estimate of error rate: 19.96%  
## Confusion matrix:  
## 0 1 class.error  
## 0 281 37 0.1163522  
## 1 70 148 0.3211009

## Sensitivity  
148/(148+70)

## [1] 0.6788991

## Specificity  
281/(281+37)

## [1] 0.8836478

## Accuracy  
(148+281)/(148+37+281+70)

## [1] 0.8003731

# Question 3

pred <- model %>% predict(test.data)  
table(pred,test.data$Survived)

##   
## pred 0 1  
## 0 90 18  
## 1 16 54

## Sensitivity  
54/(54+16)

## [1] 0.7714286

## Specificity  
90/(90+18)

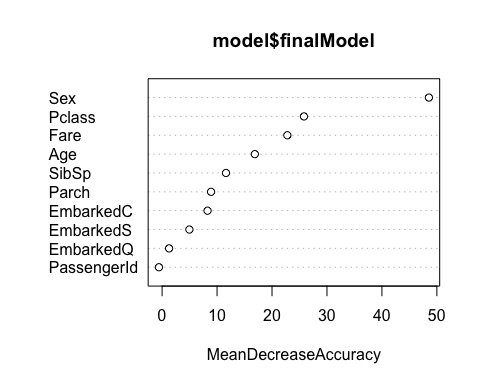
## [1] 0.8333333

## Accuracy  
(54+90)/(54+16+90+18)

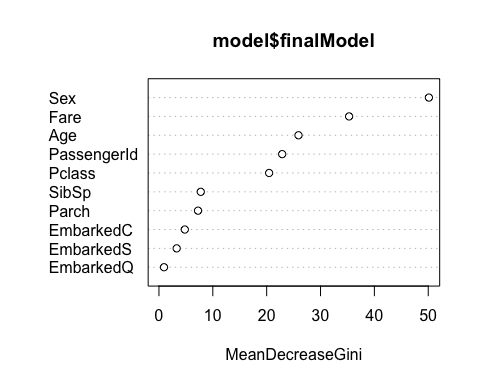
## [1] 0.8089888

# Question 4

# Plot MeanDecreaseAccuracy  
varImpPlot(model$finalModel, type = 1)



# Plot MeanDecreaseGini  
varImpPlot(model$finalModel, type = 2)

 # Question 5

varImp(model, type = 1)

## rf variable importance  
##   
## Overall  
## Sex 100.000  
## Pclass 53.729  
## Fare 47.559  
## Age 35.528  
## SibSp 24.836  
## Parch 19.302  
## EmbarkedC 17.994  
## EmbarkedS 11.278  
## EmbarkedQ 3.761  
## PassengerId 0.000

##just for experimenting, I wanted to try based on meanDecreaseGini as well  
varImp(model, type = 2)

## rf variable importance  
##   
## Overall  
## Sex 100.000  
## Fare 69.882  
## Age 50.775  
## PassengerId 44.617  
## Pclass 39.683  
## SibSp 13.879  
## Parch 12.845  
## EmbarkedC 7.846  
## EmbarkedS 4.781  
## EmbarkedQ 0.000

# Question 6

The number of variables

sqrt(36)

## [1] 6