STAT 534: Homework 5

Fall 2021

Due: Friday, November 5

1. Load "titanic_data.csv" data downloaded from the Canvas. Use the following codes to preprocess the data:

```
library(mdsr)
library(tidyverse)
library(tidymodels)
library(yardstick)

filename <- "titanic_data.csv"
titanic <- read_csv(filename) %>%
dplyr::select(-c(home.dest, cabin, name, x, ticket))%>%
filter(embarked != "?")%>%
mutate(pclass = factor(pclass, levels = c(1, 2, 3), labels = c('Upper', 'Middle', 'Lower')),
       survived = factor(survived, levels = c(0, 1), labels = c('No', 'Yes')),
       age = as.numeric(age),
       sex = factor(sex),
       embarked = factor(embarked))%>%
na.omit()
```

   (a) Separate the training and testing datasets using 80%/20% splitting. (5 points)

   (b) Fit a logistic regression model to predict `survived` using all other available variables. Here are the formula:

```
survived ~ pclass + sex + age + sibsp + parch + fare + embarked
```

   Please show the confusion matrix for the training data and calculate the training accuracy rate. (10 points)

   (c) If we want to fit a random forests model using the same formula, how to set the parameter `mtry` (the number of candidates for splitting varibles)? (5 points)

   (d) Fit a random forests model and show the confusion matrix for the training data and calculate the training accuracy rate. You can set the number of `trees` as 201. (10 points).

   (e) Find the most important explanatory variable of the random forests model. (5 points)

   (f) Find the testing accuracy rate of the two models. (5 points)

   (g) Which accuracy should we use to compare the two models? What's your conclusion? (5 points)

   (h) Calculate the sensitivity and specificity of the two models for the testing data. (5 points)

   (i) (Bonus) Draw the ROC curves of the two models using the testing data. (Hint: when you modify the demo codes, remember to replace `` `.pred_>50K` `` with `` `.pred_Yes` ``.) (5 points)