# P8106_Midterm_jck2183

jck2183_Chia-wen Kao

2021/3/23

```
library(tidyverse)
library(caret)
library(glmnet)
library(mlbench)
library(pROC) #generate ROC curve and calculate AUC
library(pdp) #partial dependent plot
library(vip) #variable importance plot: global impact on different predictor
library(AppliedPredictiveModeling) # for visualization purpose
```

## Introduction:

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relavant information about the patient.

Data Source: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

All the features we had:

- id: unique identifier
- gender: "Male", "Female" or "Other"
- age: age of the patient
- hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
- heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
- ever_married: "No" or "Yes"
- work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
- Residence_type: "Rural" or "Urban"
- avg_glucose_level: average glucose level in blood
- bmi: body mass index
- smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*
- stroke: 1 if the patient had a stroke or 0 if not *Note: "Unknown" in smoking_status means that the information is unavailable for this patient

### Import Data

```
stroke_df = read.csv("./data/healthcare-dataset-stroke-data.csv")
# head(stroke_df)
```

```r
stroke_df$stroke = as.factor(stroke_df$stroke)
stroke_df$gender = as.factor(stroke_df$gender)
stroke_df$ever_married = as.factor(stroke_df$ever_married)
stroke_df$work_type = as.factor(stroke_df$work_type)
stroke_df$Residence_type = as.factor(stroke_df$Residence_type)
stroke_df$smoking_status = as.factor(stroke_df$smoking_status)
stroke_df$heart_disease = as.factor(stroke_df$heart_disease)
stroke_df$hypertension = as.factor(stroke_df$hypertension)
stroke_df$work_type = as.factor(stroke_df$work_type)
stroke_df$bmi = as.numeric(stroke_df$bmi)
```

```
## Warning: NAs introduced by coercion
```

```r
stroke_df = stroke_df %>%
    mutate(stroke = recode(stroke,
                           `0` = "no stroke",
                           `1` = "stroke")) %>%
    select(-id)

summary(stroke_df)
```

```
##     gender          age        hypertension heart_disease ever_married
##  Female:2994   Min.   : 0.08   0:4612       0:4834        No :1757
##  Male  :2115   1st Qu.:25.00   1: 498       1: 276        Yes:3353
##  Other :   1   Median :45.00
##                Mean   :43.23
##                3rd Qu.:61.00
##                Max.   :82.00
##
##          work_type    Residence_type avg_glucose_level      bmi
##  children     : 687   Rural:2514     Min.   : 55.12    Min.   :10.30
##  Govt_job     : 657   Urban:2596     1st Qu.: 77.25    1st Qu.:23.50
##  Never_worked :  22                  Median : 91.89    Median :28.10
##  Private      :2925                  Mean   :106.15    Mean   :28.89
##  Self-employed: 819                  3rd Qu.:114.09    3rd Qu.:33.10
##                                      Max.   :271.74    Max.   :97.60
##                                                        NA's   :201
##          smoking_status       stroke
##  formerly smoked: 885   no stroke:4861
##  never smoked   :1892   stroke   : 249
##  smokes         : 789
##  Unknown        :1544
##
##
##
```

The imported dataset has 5110 observations in total. Excluding the id, we only gave 10 features and one binary outcome variable-stroke (0:no stroke, 1:stroke). We found that the stroke outcome distribution is imbalanced with 4861 observations have no stroke while 249 observations have a stroke.

We find out there are 201 observations with missing values in BMI. Among these missing values, 40 observations have a stroke while 161 observations without stroke.

Our main task for this project is to find out the appropriate models that have a better performance on prediction by comparing several models' performance.

First, we have to convert character variables into factors, so that we can add them into our model and proceed with the analysis. We also have to figure out how to deal with the missing values, since the missing data reside only in the BMI variable, and it is a continuous variable. We can discuss how to impute these values, such as taking an average BMI for the stroke group and the non-stroke group respectively. Plus, we will also examine if there is any correlation among features. Meanwhile, we also found there are some unreasonable values that appear in the data such as the minimum age is 0.08, minimum BMI is 10, and so on. There is also an observation who identified their gender as "Other". We need to discuss if we need to remove them or how to deal with them.

Second, we have to handle the imbalanced distribution among stroke and non-stroke groups. The imbalance distribution problem can be solved via the sampling method in cross-validation.

Third, the characteristics of features will help us determine which model would be proper. As the outcome is binary, and the features are mixtures of continuous and categorical variables. In the meanwhile, we also have to decide how to partition the train and test data, which cross-validation method to use, which evaluation metrics should be used, and set up a reasonable tuning grid corresponding to the tuning parameter.