

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 relevant 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)
stroke_df$stroke = as.factor(stroke_df$stroke)
```

```
stroke_df = stroke_df %>%
  mutate(stroke = recode(stroke,
    '0' = "no stroke",
    '1' = "stroke")) %>%

  na.omit() %>%
  select(-id)

summary(stroke_df)
```

```
##      gender      age      hypertension      heart_disease
## Length:5110    Min.   : 0.08    Min.   :0.00000    Min.   :0.00000
## Class :character 1st Qu.:25.00    1st Qu.:0.00000    1st Qu.:0.00000
## Mode  :character Median :45.00    Median :0.00000    Median :0.00000
##                Mean  :43.23    Mean  :0.09746    Mean  :0.05401
##                3rd Qu.:61.00    3rd Qu.:0.00000    3rd Qu.:0.00000
##                Max.   :82.00    Max.   :1.00000    Max.   :1.00000
## ever_married    work_type      Residence_type      avg_glucose_level
## Length:5110    Length:5110    Length:5110    Min.   : 55.12
## Class :character Class :character Class :character 1st Qu.: 77.25
## Mode  :character Mode  :character Mode  :character Median : 91.89
##                                     Mean  :106.15
##                                     3rd Qu.:114.09
##                                     Max.   :271.74
##      bmi      smoking_status      stroke
## Length:5110    Length:5110    no stroke:4861
## Class :character Class :character stroke   : 249
## Mode  :character Mode  :character
##
##
##
```

```
sapply("N/A", grepl, x=stroke_df)
```

```
##      N/A
## [1,] FALSE
## [2,] FALSE
## [3,] FALSE
## [4,] FALSE
## [5,] FALSE
## [6,] FALSE
## [7,] FALSE
## [8,] FALSE
## [9,] TRUE
## [10,] FALSE
## [11,] FALSE
```

```
bmi.index = stroke_df %>% filter(stroke_df$bmi == "N/A")
dim(bmi.index)
```

```
## [1] 201 11
```

```
bmi.index %>%  
  filter(bmi.index$stroke == "stroke") %>%  
  dim()
```

```
## [1] 40 11
```

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.

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.