# Stroke Factors: Classification & Predictive Analytics

Danyang Liu (500936348). Supervisor: Dr. Ceni Babaoglu

3/7/2022

```
#install.packages("RCurl")
#install.packages("MASS")
#install.packages("leaps")

library(RCurl)
```

```
## Warning: package 'RCurl' was built under R version 4.1.2
```

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.1.2
```

```
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 4.1.2
```

**Dataset: removed NAs, keep outliers**

**Converted cat variables into num**

**Using feature selection to narrow down variables**

**Read dataset**

```
stroke <- read.csv(file="stroke_1_raw.csv",header=T, sep=",")
```

**Exploratory Analytics and Data Cleaning**

```
# Descriptive analysis
str(stroke)
```

```
## 'data.frame':    5110 obs. of  12 variables:
##  $ id               : int  9046 51676 31112 60182 1665 56669 53882 10434 27419 60491 ...
##  $ gender           : chr  "Male" "Female" "Male" "Female" ...
##  $ age              : num  67 61 80 49 79 81 74 69 59 78 ...
```

```
##  $ hypertension     : int  0 0 0 0 1 0 1 0 0 0 ...
##  $ heart_disease    : int  1 0 1 0 0 0 1 0 0 0 ...
##  $ ever_married     : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ work_type        : chr  "Private" "Self-employed" "Private" "Private" ...
##  $ Residence_type   : chr  "Urban" "Rural" "Rural" "Urban" ...
##  $ avg_glucose_level: num  229 202 106 171 174 ...
##  $ bmi              : chr  "36.6" "N/A" "32.5" "34.4" ...
##  $ smoking_status   : chr  "formerly smoked" "never smoked" "never smoked" "smokes" ...
##  $ stroke           : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
summary(stroke)
```

```
##        id            gender               age         hypertension
##  Min.   :   67   Length:5110        Min.   : 0.08   Min.   :0.00000
##  1st Qu.:17741   Class :character   1st Qu.:25.00   1st Qu.:0.00000
##  Median :36932   Mode  :character   Median :45.00   Median :0.00000
##  Mean   :36518                      Mean   :43.23   Mean   :0.09746
##  3rd Qu.:54682                      3rd Qu.:61.00   3rd Qu.:0.00000
##  Max.   :72940                      Max.   :82.00   Max.   :1.00000
##  heart_disease     ever_married        work_type         Residence_type
##  Min.   :0.00000   Length:5110        Length:5110        Length:5110
##  1st Qu.:0.00000   Class :character   Class :character   Class :character
##  Median :0.00000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :0.05401
##  3rd Qu.:0.00000
##  Max.   :1.00000
##  avg_glucose_level     bmi            smoking_status         stroke
##  Min.   : 55.12    Length:5110        Length:5110        Min.   :0.00000
##  1st Qu.: 77.25    Class :character   Class :character   1st Qu.:0.00000
##  Median : 91.89    Mode  :character   Mode  :character   Median :0.00000
##  Mean   :106.15                                          Mean   :0.04873
##  3rd Qu.:114.09                                          3rd Qu.:0.00000
##  Max.   :271.74                                          Max.   :1.00000
```

```
# Convert 'N/A's (strings) in dataset to NA
is.na(stroke) <- stroke == "N/A"
# Count number of NAs in dataset
sum(is.na(stroke))
```

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

```
# Count number of NAs in all columns
colSums(is.na(stroke))
```

```
##                id            gender               age      hypertension
##                 0                 0                 0                 0
##     heart_disease      ever_married         work_type    Residence_type
##                 0                 0                 0                 0
## avg_glucose_level               bmi    smoking_status            stroke
##                 0               201                 0                 0
```

```r
# Count number of 'Unknown's in all columns
colSums(stroke == "Unknown")
```

```
##                id            gender               age      hypertension
##                 0                 0                 0                 0
##     heart_disease      ever_married         work_type    Residence_type
##                 0                 0                 0                 0
## avg_glucose_level               bmi    smoking_status            stroke
##                 0                NA              1544                 0
```

```r
# Remove first column 'id'; irrelevant to data analysis
stroke <- stroke[2:12]

# Check attribute levels and convert data types to numeric
# For binary "Yes"/"No" values, "Yes" = 1 and "No" = 2
str(stroke)
```

```
## 'data.frame':    5110 obs. of  11 variables:
##  $ gender           : chr  "Male" "Female" "Male" "Female" ...
##  $ age              : num  67 61 80 49 79 81 74 69 59 78 ...
##  $ hypertension     : int  0 0 0 0 1 0 1 0 0 0 ...
##  $ heart_disease    : int  1 0 1 0 0 0 1 0 0 0 ...
##  $ ever_married     : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ work_type        : chr  "Private" "Self-employed" "Private" "Private" ...
##  $ Residence_type   : chr  "Urban" "Rural" "Rural" "Urban" ...
##  $ avg_glucose_level: num  229 202 106 171 174 ...
##  $ bmi              : chr  "36.6" NA "32.5" "34.4" ...
##  $ smoking_status   : chr  "formerly smoked" "never smoked" "never smoked" "smokes" ...
##  $ stroke           : int  1 1 1 1 1 1 1 1 1 1 ...
```

```r
unique(stroke$gender)
```

```
## [1] "Male"   "Female" "Other"
```

```r
stroke$gender <- gsub("Male", 1, stroke$gender)
stroke$gender <- gsub("Female", 2, stroke$gender)
stroke$gender <- gsub("Other", 3, stroke$gender)
stroke$gender <- as.numeric(stroke$gender)
unique(stroke$gender)
```

```
## [1] 1 2 3
```

```r
unique(stroke$ever_married)
```

```
## [1] "Yes" "No"
```

```r
stroke$ever_married <- gsub("Yes", 1, stroke$ever_married)
stroke$ever_married <- gsub("No", 0, stroke$ever_married)
stroke$ever_married <- as.numeric(stroke$ever_married)
unique(stroke$ever_married)
```

```
## [1] 1 0
```

```
unique(stroke$work_type)
```

```
## [1] "Private"       "Self-employed" "Govt_job"      "children"
## [5] "Never_worked"
```

```
stroke$work_type <- gsub("Private", 1, stroke$work_type)
stroke$work_type <- gsub("Self-employed", 2, stroke$work_type)
stroke$work_type <- gsub("Govt_job", 3, stroke$work_type)
stroke$work_type <- gsub("children", 4, stroke$work_type)
stroke$work_type <- gsub("Never_worked", 5, stroke$work_type)
stroke$work_type <- as.numeric(stroke$work_type)
unique(stroke$work_type)
```

```
## [1] 1 2 3 4 5
```

```
unique(stroke$Residence_type)
```

```
## [1] "Urban" "Rural"
```

```
stroke$Residence_type <- gsub("Urban", 1, stroke$Residence_type)
stroke$Residence_type <- gsub("Rural", 2, stroke$Residence_type)
stroke$Residence_type <- as.numeric(stroke$Residence_type)
unique(stroke$Residence_type)
```

```
## [1] 1 2
```

```
stroke$bmi <- as.numeric(stroke$bmi)

unique(stroke$smoking_status)
```

```
## [1] "formerly smoked" "never smoked"    "smokes"          "Unknown"
```

```
stroke$smoking_status <- gsub("formerly smoked", 1, stroke$smoking_status)
stroke$smoking_status <- gsub("never smoked", 2, stroke$smoking_status)
stroke$smoking_status <- gsub("smokes", 3, stroke$smoking_status)
stroke$smoking_status <- gsub("Unknown", 4, stroke$smoking_status)
stroke$smoking_status <- as.numeric(stroke$smoking_status)
unique(stroke$smoking_status)
```

```
## [1] 1 2 3 4
```

```
# Check that all attributes are now numeric data types
str(stroke)
```

```
## 'data.frame':    5110 obs. of  11 variables:
##  $ gender           : num  1 2 1 2 2 1 1 2 2 2 ...
##  $ age              : num  67 61 80 49 79 81 74 69 59 78 ...
##  $ hypertension     : int  0 0 0 0 1 0 1 0 0 0 ...
##  $ heart_disease    : int  1 0 1 0 0 0 1 0 0 0 ...
##  $ ever_married     : num  1 1 1 1 1 1 1 0 1 1 ...
##  $ work_type        : num  1 2 1 1 2 1 1 1 1 1 ...
##  $ Residence_type   : num  1 2 2 1 2 1 2 1 2 1 ...
##  $ avg_glucose_level: num  229 202 106 171 174 ...
##  $ bmi              : num  36.6 NA 32.5 34.4 24 29 27.4 22.8 NA 24.2 ...
##  $ smoking_status   : num  1 2 2 3 2 1 2 2 4 4 ...
##  $ stroke           : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
# Deal with NAs
# Method 1: remove NAs
stroke_noNAs <- stroke[complete.cases(stroke), ]

# Deal with outliers
# Did not remove outliers

# Examine correlations between all Independent Variables
cor(stroke_noNAs[1:10])
```

```
##                        gender         age hypertension heart_disease
## gender             1.000000000  0.02981661 -0.021978158  -0.083013859
## age                0.029816612  1.00000000  0.274424873   0.257122776
## hypertension      -0.021978158  0.27442487  1.000000000   0.115990991
## heart_disease     -0.083013859  0.25712278  0.115990991   1.000000000
## ever_married       0.035542943  0.68078165  0.162406260   0.111245121
## work_type         -0.071262910 -0.41534434 -0.073404033  -0.054926544
## Residence_type    -0.003755064 -0.01094811  0.001074146   0.002361744
## avg_glucose_level -0.052612931  0.23583816  0.180542699   0.154525119
## bmi                0.025657719  0.33339800  0.167810584   0.041357443
## smoking_status    -0.040065223 -0.38667582 -0.132831660  -0.071396924
##                   ever_married   work_type Residence_type avg_glucose_level
## gender             0.035542943 -0.07126291  -0.0037550644      -0.052612931
## age                0.680781652 -0.41534434  -0.0109481144       0.235838155
## hypertension       0.162406260 -0.07340403   0.0010741462       0.180542699
## heart_disease      0.111245121 -0.05492654   0.0023617439       0.154525119
## ever_married       1.000000000 -0.37780605  -0.0049891711       0.151377377
## work_type         -0.377806049  1.00000000  -0.0130835508      -0.063151561
## Residence_type    -0.004989171 -0.01308355   1.0000000000       0.007616542
## avg_glucose_level  0.151377377 -0.06315156   0.0076165420       1.000000000
## bmi                0.341694652 -0.34724139   0.0001224412       0.175502176
## smoking_status    -0.310702330  0.31330828  -0.0027191093      -0.108983692
##                            bmi smoking_status
## gender              0.0256577189   -0.040065223
## age                 0.3333979952   -0.386675819
## hypertension        0.1678105844   -0.132831660
## heart_disease       0.0413574429   -0.071396924
## ever_married        0.3416946516   -0.310702330
## work_type          -0.3472413855    0.313308284
## Residence_type      0.0001224412   -0.002719109
## avg_glucose_level   0.1755021761   -0.108983692
```

```
## bmi                1.0000000000   -0.235739765
## smoking_status    -0.2357397646    1.000000000
```

### Dimensionality Reduction

```
# Feature selection - see best combo of attributes
subsets <- regsubsets(stroke~gender+age+hypertension+heart_disease+ever_married+work_type+Residence_type
sub.sum <- summary(subsets)
as.data.frame(sub.sum$outmat)
```

```
##          gender age hypertension heart_disease ever_married work_type
## 1  ( 1 )          *
## 2  ( 1 )          *
## 3  ( 1 )          *                          *
## 4  ( 1 )          *            *             *
## 5  ( 1 )          *            *             *           *
## 6  ( 1 )          *            *             *           *
## 7  ( 1 )          *            *             *           *         *
## 8  ( 1 )          *            *             *           *         *
##          Residence_type avg_glucose_level bmi smoking_status
## 1  ( 1 )
## 2  ( 1 )                                 *
## 3  ( 1 )                                 *
## 4  ( 1 )                                 *
## 5  ( 1 )                                 *
## 6  ( 1 )                                 *   *
## 7  ( 1 )                                 *   *
## 8  ( 1 )                                 *   *             *
```

```
# In order of importance:
# age (8x*), avg_glucose_level (7x*), heart_disease (6x*), hypertension(5x*), ever_married (4x*), bmi (
```

```
# Normalize continuous numeric variables
# Such as age, avg_blood_glucose, and bmi
# Using z-score methods
stroke_noNAs$age <- (stroke_noNAs$age - mean(stroke_noNAs$age))/sd(stroke_noNAs$age)
stroke_noNAs$avg_glucose_level <- (stroke_noNAs$avg_glucose_level - mean(stroke_noNAs$avg_glucose_level
stroke_noNAs$bmi <- (stroke_noNAs$bmi - mean(stroke_noNAs$bmi))/sd(stroke_noNAs$bmi)
```

### Classification

### Predictive Analytics: Logistic Regression

```
# Split dataset into 70% training, 30% testing sets
stroke_index1 <- sample(1:nrow(stroke_noNAs), 0.7 * nrow(stroke_noNAs))

# Assign selected sample as training set
# Assign leftover dataset as test set
train.set1 <- stroke_noNAs[stroke_index1,]
```

```r
test.set1 <- stroke_noNAs[-stroke_index1,]

# Logistic regression model for prediction
# Using only the top 4 features based on feature selection: age, avg_glucose_level, heart_disease, hype
glm_model1 <- glm(formula = stroke~age+avg_glucose_level+heart_disease+hypertension, data = train.set1,
summary(glm_model1)
```

```
##
## Call:
## glm(formula = stroke ~ age + avg_glucose_level + heart_disease +
##     hypertension, family = "binomial", data = train.set1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1426  -0.2981  -0.1622  -0.0787   3.5772
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -4.22842    0.17650 -23.957  < 2e-16 ***
## age                1.52506    0.14816  10.293  < 2e-16 ***
## avg_glucose_level  0.20291    0.06618   3.066  0.00217 **
## heart_disease      0.54195    0.23457   2.310  0.02087 *
## hypertension       0.49880    0.20803   2.398  0.01649 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1238.95  on 3435  degrees of freedom
## Residual deviance:  980.37  on 3431  degrees of freedom
## AIC: 990.37
##
## Number of Fisher Scoring iterations: 7
```

**Evaluation Metrics**

```r
predicted1 <- predict(glm_model1, test.set1, type = "response")
# Setting 0.5 as threshold - binary prediction
predicted_class1 <- ifelse(predicted1 >= 0.5, "Stroke", "No Stroke")
ConfusionMatrix1 <- table(actual = test.set1$stroke, predicted = predicted_class1)
ConfusionMatrix1
```

```
##       predicted
## actual No Stroke
##      0      1415
##      1        58
```

**Abysmal predictions using only feature selection and logistic regression applied to dataset with NAs removed (from BMI column) and outliers retained. No strokes are predicted at all**