Stroke Factors: Classification & Predictive Analytics

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

3/7/2022

Dataset: removed NAs, kept outliers.

Did not convert cat variables into num

No PCA analysis

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
```

```
# 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
```

```
# 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 ...
```

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

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

```
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
```

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

# Assign "No Stroke" and "Stroke" labels for Stroke attribute
stroke$stroke <- ifelse(stroke$stroke == 0, "No Stroke", "Stroke")
# Assign Stroke values as factor levels
stroke$stroke <- as.factor(stroke$stroke)

# 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     : 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              : num  36.6 NA 32.5 34.4 24 29 27.4 22.8 NA 24.2 ...
##  $ smoking_status   : chr  "formerly smoked" "never smoked" "never smoked" "smokes" ...
##  $ stroke           : Factor w/ 2 levels "No Stroke","Stroke": 2 2 2 2 2 2 2 2 2 2 ...
```

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

```r
# Leave outliers as is

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

# 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

```r
# Split dataset into 70% training, 30% testing sets
stroke_index2 <- sample(1:nrow(stroke_noNAs), 0.7 * nrow(stroke_noNAs))

# Assign selected sample as training set
# Assign leftover dataset as test set
train.set2 <- stroke_noNAs[stroke_index2,]
test.set2 <- stroke_noNAs[-stroke_index2,]

# Logistic regression model for prediction
glm_model2 <- glm(formula = stroke~., data = train.set2, family = "binomial")
summary(glm_model2)
```

```
##
## Call:
## glm(formula = stroke ~ ., family = "binomial", data = train.set2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0542  -0.2934  -0.1593  -0.0807   3.3955
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -3.90492    1.10781  -3.525 0.000424 ***
## gender                       0.24901    0.18608   1.338 0.180831
## age                          1.63407    0.16741   9.761  < 2e-16 ***
## hypertension                 0.46706    0.21053   2.219 0.026519 *
## heart_disease                0.15721    0.25729   0.611 0.541186
## ever_marriedYes             -0.01371    0.31545  -0.043 0.965332
## work_typeGovt_job           -0.81113    1.14795  -0.707 0.479825
## work_typeNever_worked      -10.04788  401.98179  -0.025 0.980058
## work_typePrivate            -0.58081    1.12971  -0.514 0.607166
## work_typeSelf-employed      -1.08318    1.15564  -0.937 0.348602
## Residence_typeUrban         -0.09988    0.17754  -0.563 0.573720
## avg_glucose_level            0.25035    0.06902   3.627 0.000286 ***
## bmi                         -0.02696    0.11017  -0.245 0.806683
```

```
## smoking_statusnever smoked  -0.05137    0.22496  -0.228 0.819377
## smoking_statussmokes         0.19374    0.28212   0.687 0.492241
## smoking_statusUnknown        0.01032    0.27287   0.038 0.969829
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1226.60  on 3435  degrees of freedom
## Residual deviance:  974.16  on 3420  degrees of freedom
## AIC: 1006.2
##
## Number of Fisher Scoring iterations: 14
```

Evaluation Metrics

```
predicted2 <- predict(glm_model2, test.set2, type = "response")
# Setting 0.5 as threshold - binary prediction
predicted_class2 <- ifelse(predicted2 >= 0.5, "Stroke", "No Stroke")
ConfusionMatrix2 <- table(actual = test.set2$stroke, predicted = predicted_class2)
ConfusionMatrix2
```

```
##            predicted
## actual      No Stroke
##   No Stroke      1413
##   Stroke           60
```

```
str(predicted2)
```

```
##  Named num [1:1473] 0.188 0.0427 0.2438 0.1659 0.118 ...
##  - attr(*, "names")= chr [1:1473] "3" "4" "6" "7" ...
```

```
summary(predicted2)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0000001 0.0039105 0.0159452 0.0441050 0.0572282 0.4366183
```

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