Stroke Factors: Classification & Predictive Analytics

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

3/7/2022

Dataset: removed NAs, keep outliers

Converted cat variables into num

With PCA analysis and Normalization

Read dataset

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

Exploratory Analytics and Data Cleaning

```r
# 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 ...
```

```r
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
```

```r
# 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
```

```r
# 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 ...
```

```
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
```

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

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

```
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
```

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

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

```r
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
```

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

unique(stroke$smoking_status)
```

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

```r
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
```

```r
# 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     : 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 ...
```

```r
# Deal with NAs
# Method 1: remove NAs
stroke_noNAs <- stroke[complete.cases(stroke), ]
# Method 2: replace NAs with values using k-NN algorithm?

# 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
```

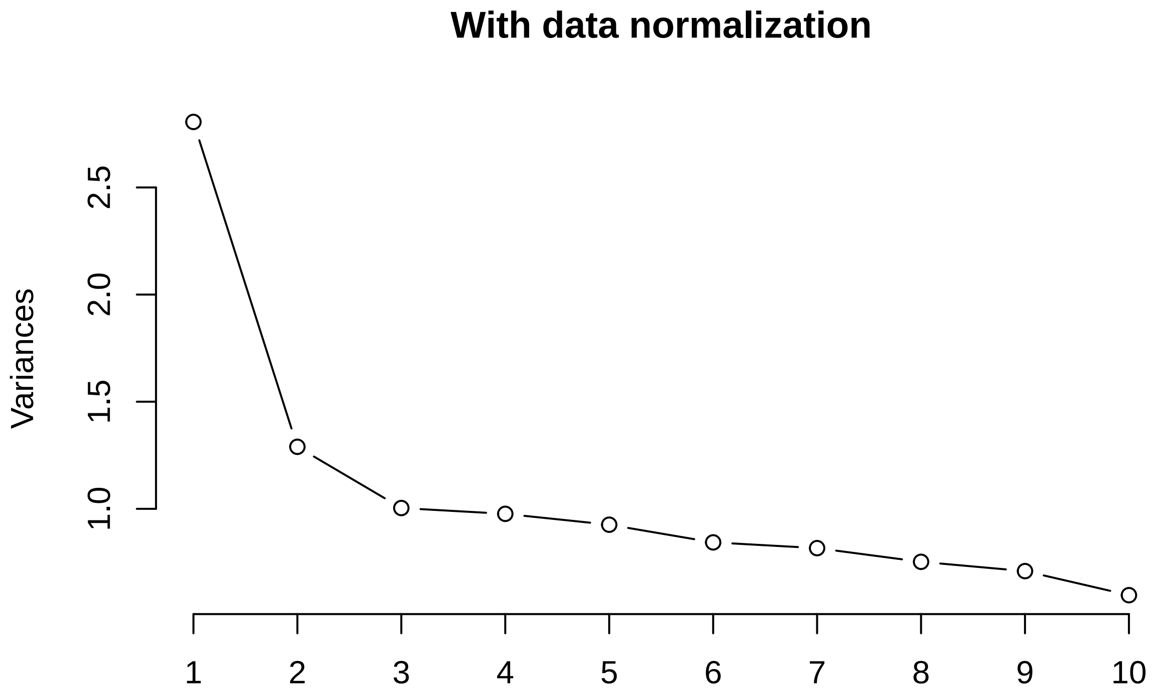
```r
# PCA and normalization
stroke.pca.normdata <- prcomp(stroke_noNAs, scale = TRUE, center = TRUE)
stroke.pca.normdata$rotation
```

```
##                             PC1          PC2        PC3          PC4
## gender             0.0259233572 -0.383197165 0.22647254  0.783583699
## age                0.5054853391  0.024182201 0.03883891  0.016166292
## hypertension       0.2369608561  0.321117657 0.02457679  0.163307429
## heart_disease      0.1788912023  0.469186383 -0.01005996 -0.056521570
## ever_married       0.4530676198 -0.146862148  0.01048988 -0.073641577
## work_type         -0.3662424678  0.342969301  0.05342298  0.091335526
## Residence_type    -0.0004500146 -0.003674695 -0.95169736  0.294293470
## avg_glucose_level  0.2227372897  0.396705545 -0.05857439  0.003802351
## bmi                0.3462318579 -0.189572614 -0.07265088 -0.169550552
## smoking_status    -0.3408295914  0.185892059  0.01933302  0.044595103
## stroke             0.1749366056  0.402795091  0.16947481  0.474074630
##                           PC5         PC6         PC7          PC8         PC9
## gender            -0.06772847 -0.42123786  0.05364083 -0.005462004 -0.02148917
## age                0.16530094  0.00622259  0.07387716 -0.059389034  0.34286770
## hypertension      -0.52425447  0.19820837  0.67143088 -0.049056126 -0.06175224
## heart_disease      0.50408051 -0.54919880  0.24541895 -0.090709343 -0.31740497
## ever_married       0.13105817  0.02072971  0.02774393 -0.182774917  0.56966539
## work_type         -0.13222204 -0.04754579  0.11038298  0.134263672  0.31359043
## Residence_type     0.06191834  0.03238549  0.01572021 -0.017392912  0.03365133
## avg_glucose_level -0.45729942 -0.39599194 -0.53528418  0.257542874  0.16563130
## bmi               -0.35759919 -0.04962683 -0.18219990 -0.396018211 -0.48514813
## smoking_status    -0.08362968 -0.12086745 -0.09810147 -0.838212797  0.23912035
## stroke             0.23803758  0.55135610 -0.37156941 -0.091175962 -0.18289258
##                          PC10         PC11
## gender            -0.04038217  0.012904548
## age               -0.03671119 -0.766078324
## hypertension       0.18403730  0.091768057
## heart_disease     -0.05942138  0.126941895
## ever_married      -0.17098964  0.601738850
## work_type         -0.76333035 -0.077946990
## Residence_type    -0.03157301 -0.009390111
## avg_glucose_level  0.20269042  0.047543638
## bmi               -0.50263049 -0.042270955
## smoking_status     0.22096432 -0.084176084
## stroke            -0.06455502  0.095055029
```

```
# Values of normalized data after transformation
head(stroke.pca.normdata$x)
```

```
##        PC1      PC2        PC3       PC4        PC5        PC6        PC7
## 1 4.106707 4.634341  1.1650141 0.3926441  2.3963837 -0.2970315 -2.4863540
## 3 3.288214 3.814086 -0.4984660 1.1094590  3.9892364  0.7807132 -0.9279121
## 4 1.920969 1.554716  1.7720993 2.3566927  0.3414241  1.6781614 -3.0042256
## 5 2.955541 3.102574  0.1276187 3.7972126 -0.7268401  2.5458755 -0.1555969
## 6 3.047806 2.290637  1.3618064 0.8236563  0.9586390  2.6651581 -2.8836456
## 7 3.569566 4.722408 -0.3293020 1.7775113  2.7314833  1.8168057  1.9267721
##          PC8        PC9       PC10        PC11
## 1  0.4565377 -2.0617585 -0.3885461  0.86198365
## 3 -0.8863863 -1.7819676 -0.5679247  0.21512247
## 4 -0.8490067 -0.5555427  0.1166312  0.71031713
## 5  0.2968865  0.4668517  0.4387088  0.05720661
## 6  0.9747550 -0.0747238  0.1551344 -0.20324665
## 7 -0.9909720 -1.9055035  0.2418273  0.72568769
```

```
# Visualize PCA to see most important principal components
plot(stroke.pca.normdata, type = "l", main = "With data normalization")
```



```
# Elbow point occurs around PC3 (if 1.0 as threshold), so PC1 and PC2 explain most of the variance in t
```

```
# Check correlations of original vs normalized transformed data
# Original
cor(stroke_noNAs)
```

```
##                         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
## stroke            -0.007020754  0.23233086  0.142514606   0.137937788
##                   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
## stroke             0.105089144 -0.05753360  -0.0060314265       0.138935862
##                            bmi smoking_status       stroke
## gender             0.0256577189   -0.040065223 -0.007020754
## age                0.3333979952   -0.386675819  0.232330856
## hypertension       0.1678105844   -0.132831660  0.142514606
## heart_disease      0.0413574429   -0.071396924  0.137937788
## ever_married       0.3416946516   -0.310702330  0.105089144
## work_type         -0.3472413855    0.313308284 -0.057533605
## Residence_type     0.0001224412   -0.002719109 -0.006031426
## avg_glucose_level  0.1755021761   -0.108983692  0.138935862
## bmi                1.0000000000   -0.235739765  0.042373661
## smoking_status    -0.2357397646    1.000000000 -0.075919784
## stroke             0.0423736611   -0.075919784  1.000000000
```

```
# Normalized transformed
cor(stroke.pca.normdata$x)
```

```
##                PC1           PC2           PC3           PC4           PC5
## PC1   1.000000e+00  2.336483e-15  9.352877e-16  2.024261e-15  3.283038e-15
## PC2   2.336483e-15  1.000000e+00 -2.274753e-15 -8.908095e-15  2.822281e-15
## PC3   9.352877e-16 -2.274753e-15  1.000000e+00  3.580131e-15 -2.002969e-15
## PC4   2.024261e-15 -8.908095e-15  3.580131e-15  1.000000e+00 -6.255659e-15
## PC5   3.283038e-15  2.822281e-15 -2.002969e-15 -6.255659e-15  1.000000e+00
## PC6   3.690637e-15 -6.168064e-15  2.799895e-15  1.014889e-14  1.451579e-15
## PC7  -3.139860e-15  5.919405e-15 -2.007565e-15 -6.981174e-15 -1.306442e-15
## PC8  -2.275190e-15 -9.488412e-16 -1.292954e-16 -5.315636e-16 -6.358075e-16
## PC9   3.738978e-15 -1.101285e-15  1.609169e-15  4.000332e-15  1.811088e-15
## PC10 -6.718098e-16  5.288342e-16  4.199467e-16  1.697325e-15 -1.089210e-15
## PC11  1.365207e-14  2.285802e-15 -5.306541e-16 -2.423821e-15  1.930271e-15
##                PC6           PC7           PC8           PC9          PC10
## PC1   3.690637e-15 -3.139860e-15 -2.275190e-15  3.738978e-15 -6.718098e-16
## PC2  -6.168064e-15  5.919405e-15 -9.488412e-16 -1.101285e-15  5.288342e-16
## PC3   2.799895e-15 -2.007565e-15 -1.292954e-16  1.609169e-15  4.199467e-16
## PC4   1.014889e-14 -6.981174e-15 -5.315636e-16  4.000332e-15  1.697325e-15
## PC5   1.451579e-15 -1.306442e-15 -6.358075e-16  1.811088e-15 -1.089210e-15
## PC6   1.000000e+00  7.612556e-15 -7.589438e-16 -2.412575e-16 -1.892336e-15
## PC7   7.612556e-15  1.000000e+00 -3.593953e-16 -1.033947e-15  1.623512e-15
## PC8  -7.589438e-16 -3.593953e-16  1.000000e+00 -1.939783e-15  3.251519e-16
## PC9  -2.412575e-16 -1.033947e-15 -1.939783e-15  1.000000e+00 -2.496878e-16
## PC10 -1.892336e-15  1.623512e-15  3.251519e-16 -2.496878e-16  1.000000e+00
## PC11 -5.135392e-16  2.984933e-15 -2.749967e-16  3.286801e-15 -3.725563e-16
##               PC11
## PC1   1.365207e-14
## PC2   2.285802e-15
## PC3  -5.306541e-16
## PC4  -2.423821e-15
## PC5   1.930271e-15
## PC6  -5.135392e-16
## PC7   2.984933e-15
## PC8  -2.749967e-16
## PC9   3.286801e-15
## PC10 -3.725563e-16
## PC11  1.000000e+00
```

```
# Correlations between PCs in normalized transformed data are almost 0 - these PCs are now orthogonal
```

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

# Logistic regression model for prediction
glm_model1 <- glm(formula = stroke~., data = train.set1, family = "binomial")
summary(glm_model1)
```

```
##
## Call:
## glm(formula = stroke ~ ., family = "binomial", data = train.set1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0911  -0.2965  -0.1575  -0.0789   3.5598
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -3.925199   0.577648  -6.795 1.08e-11 ***
## gender            -0.003442   0.182008  -0.019  0.98491
## age                1.513812   0.154018   9.829  < 2e-16 ***
## hypertension       0.595272   0.204749   2.907  0.00365 **
## heart_disease      0.548366   0.234024   2.343  0.01912 *
## ever_married      -0.103646   0.283948  -0.365  0.71510
## work_type          0.062088   0.120293   0.516  0.60576
## Residence_type    -0.158369   0.178967  -0.885  0.37621
## avg_glucose_level  0.150758   0.068875   2.189  0.02861 *
## bmi                0.024427   0.110186   0.222  0.82456
## smoking_status    -0.060481   0.088497  -0.683  0.49434
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1220.41  on 3435  degrees of freedom
## Residual deviance:  968.21  on 3425  degrees of freedom
## AIC: 990.21
##
## Number of Fisher Scoring iterations: 8
```

Evaluation Metrics

```
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      1412
##      1        61
```

Abysmal predictions using only logistic regression applied to dataset with NAs removed (from BMI column) and outliers retained. Hardly any strokes are predicted at all