Stroke Prediction: A Logistic Regression Analysis to Predict Stroke Occurrence

Background

A stroke is a health condition that results in cell death in the brain due to insufficient blood flow, leading to brain dysfunction. Stroke is the 2nd leading cause of death globally, according to the WHO, responsible for around 11% of total deaths.

Project Goal

The goal of this project was to build a model that would predict stroke occurrence and identify any significant attribute that contributes to someone's stroke occurrence.

Project Metrics

Since I worked with categorical data, I used the Classification method and ran the Logistic Regression Analysis. I looked at three metrics: accuracy rate, recall rate, and precision rate.

- **Accuracy rate**:
    - Depicts the percentage of total predicted values that match the true (actual) values (positives and negatives).
- **Precision rate**:
    - Shows the proportion of correctly identified positive values in the class (true positives) over the sum of true and false positives (the items incorrectly labeled as belonging to the class).
- **Recall rate:**
    - Represents the ratio of the true positives to the total number of actual positive instances, including correctly identified positives and those that were missed (false negatives).

The original dataset, sourced from Kaggle.com, contained 5000+ observations, with 11 predictors (1 prediction (Yes/No), 3 quantitative and 7 qualitative predictors). The model takes patients' conditions, such as age, gender, smoking or not, etc. The current data was imbalanced as most people were diagnosed with no stroke. Therefore, I would look at the recall rate to find how many true positives were missed.

Research Questions

Some of the questions we wish to answer by building the model and analyzing the datasets were

- What predictors/attributes have a significant influence on the occurrence of a stroke?
- Given a certain list of attributes, what is the probability of a person having a stroke vs not having a stroke?
- Which of our predictive models best predicts having a stroke?

Project Detail

1. Data Cleaning
    - Before we perform the analysis, we first need to clean the datasets. The data cleaning process involved:
        A. Imported the data to R and stored as a data frame
        B. Removed any null values in the BMI column and removed "Other" values in the Gender column
    - After cleaning the data, the current data sets showed that 95.7% of people will not have a stroke with underlying conditions.
2. Logistic Regression Analysis
    - In this project, I ran a Logistic Regression Analysis using selected predictors and two other resampling methods (Validation-Set and Leave-One-Out cross-validation (LOOCV)).
    - Logistic Regression Analysis by approaches:
        A. Logistic Regression with Selected Predictors
            - For the initial analysis, I used 8 predictors from the datasets, i.e., gender, age, hypertension, heart_disease, ever_married, residence_type, avg_glucose_level, smoking_status, and ran the model.
            - The result showed there were 3 significant predictors, which were:
                - age (p-value < 2e-16)
                - avg_glucose level (p-value 0.000137)
                - hypertension1 / yes for having a hypertension (p-value 0.002658)

```
> glm2<-glm(stroke~gender+age+hypertension+heart_disease+ever_married+Residence_type+avg_gl
ucose_level+smoking_status,data=df,family=binomial)
> summary(glm2)

Call:
glm(formula = stroke ~ gender + age + hypertension + heart_disease +
    ever_married + Residence_type + avg_glucose_level + smoking_status,
    family = binomial, data = df)

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             -7.643637   0.466109 -16.399  < 2e-16 ***
genderMale              -0.009296   0.154182  -0.060 0.951922
age                      0.069117   0.005774  11.971  < 2e-16 ***
hypertension1            0.522217   0.173796   3.005 0.002658 **
heart_disease1           0.366366   0.206451   1.775 0.075966 .
ever_marriedYes         -0.108958   0.242494  -0.449 0.653200
Residence_typeUrban      0.009024   0.149689   0.060 0.951928
avg_glucose_level        0.004805   0.001260   3.813 0.000137 ***
smoking_statusnever smoked -0.061386   0.188360  -0.326 0.744504
smoking_statussmokes     0.319684   0.228646   1.398 0.162064
smoking_statusUnknown   -0.261710   0.245503  -1.066 0.286418
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1728.3  on 4907  degrees of freedom
Residual deviance: 1369.3  on 4897  degrees of freedom
AIC: 1391.3

Number of Fisher Scoring iterations: 8
```

- Using the predict() function, we could predict the probability that the person will have a stroke or not. I used 0.25 as a cut-off value for the probability that the person will have a stroke.
- Next, I generated the confusion matrix to determine how many observations were correctly or incorrectly classified.

```
> table(glm.pred2,stroke)
         stroke
glm.pred2    0    1
        0 4620  186
        1   79   23
```

  - 4620 observations were true negatives (correctly predicted no stroke)
  - 23 observations were true positives (correctly predicted as will have a stroke).
- Through this approach, the model yielded:
  - 94.6% accuracy rate
  - 22.5% precision rate
  - 11.00% recall rate

```
> mean(glm.pred2==stroke)
[1] 0.9460065
> #training error rate
> 1-mean(glm.pred2==stroke)
[1] 0.05399348
> table(glm.pred2,stroke)[2,2]/(table(glm.pred2,stroke)[2,2]+table(glm.pred2,stroke)[2,1])
[1] 0.2254902
> #recall rate
> table(glm.pred2,stroke)[2,2]/(table(glm.pred2,stroke)[2,2]+table(glm.pred2,stroke)[1,2])
[1] 0.1100478
```

B. Validation-Set Approach (Selected Predictors)

- With the Validation-Set approach, I took half of the datasets as training data and the other half as test data determined by R at random.

```
> set.seed(11)
> train=sample(1:nrow(df), nrow(df)/2)
> df.test=df[-train,]
> dim(df.test)
[1] 2454   12
>
> #create the DV in the test data
> stroke.test=stroke[-train]
```

- Once I had the training data, I ran the Logistic Regression using the training data as the subset and 3 significant predictors (age, avg_glucose level, and hypertension1). Compared to the previous analysis, age was the most significant predictor while hypertension1 and avg_glucose_level were less significant than before.

```
> #run logistic regression in the training dataset
> glm.fits=glm(stroke~.,data=df,family=binomial, subset=train)
> summary(glm.fits)

Call:
glm(formula = stroke ~ ., family = binomial, data = df, subset = train)

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)               -2.070e+01  5.941e+02  -0.035  0.97220
id                         9.439e-07  5.155e-06   0.183  0.85471
genderMale                 1.908e-01  2.183e-01   0.874  0.38215
age                        7.294e-02  9.144e-03   7.977 1.49e-15 ***
hypertension1              7.805e-01  2.372e-01   3.290  0.00100 **
heart_disease1             4.641e-01  2.891e-01   1.605  0.10847
ever_marriedYes           -1.101e-01  3.699e-01  -0.298  0.76598
work_typeGovt_job          1.203e+01  5.941e+02   0.020  0.98384
work_typeNever_worked     -7.640e-01  3.432e+03   0.000  0.99982
work_typePrivate           1.268e+01  5.941e+02   0.021  0.98297
work_typeSelf-employed     1.181e+01  5.941e+02   0.020  0.98414
Residence_typeUrban       -2.649e-01  2.150e-01  -1.232  0.21792
avg_glucose_level          5.873e-03  1.810e-03   3.245  0.00117 **
bmi                        4.114e-03  1.695e-02   0.243  0.80821
smoking_statusnever smoked 1.161e-01  2.709e-01   0.429  0.66828
smoking_statussmokes       1.899e-01  3.380e-01   0.562  0.57431
smoking_statusUnknown     -6.295e-02  3.422e-01  -0.184  0.85404
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 873.47  on 2453  degrees of freedom
Residual deviance: 663.44  on 2437  degrees of freedom
AIC: 697.44

Number of Fisher Scoring iterations: 18
```

- Next, I reran the model to predict the probability of having a stroke with the test dataset, predicted probabilities of having a stroke with the test dataset with significant predictors only, and generated the confusion matrix to compare the predicted values and decision variables.

```
> glm.fits=glm(stroke~gender+age+hypertension+heart_disease+ever_married+Residence_type+avg
_glucose_level+smoking_status,data=df,family=binomial, subset=train)
> summary(glm.fits)

Call:
glm(formula = stroke ~ gender + age + hypertension + heart_disease +
    ever_married + Residence_type + avg_glucose_level + smoking_status,
    family = binomial, data = df, subset = train)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -7.907246   0.680710 -11.616  < 2e-16 ***
genderMale                0.219228   0.217127   1.010 0.312650
age                       0.067961   0.008475   8.019 1.07e-15 ***
hypertension1             0.790082   0.234107   3.375 0.000738 ***
heart_disease1            0.454000   0.287026   1.582 0.113709
ever_marriedYes          -0.049221   0.366745  -0.134 0.893236
Residence_typeUrban      -0.236290   0.212701  -1.111 0.266610
avg_glucose_level         0.005818   0.001744   3.336 0.000850 ***
smoking_statusnever smoked  0.123419   0.268984   0.459 0.646353
smoking_statussmokes      0.204733   0.333980   0.613 0.539870
smoking_statusUnknown    -0.096672   0.341329  -0.283 0.777006
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 873.47  on 2453  degrees of freedom
Residual deviance: 677.00  on 2443  degrees of freedom
AIC: 699

Number of Fisher Scoring iterations: 8
```

- From the 2454 observations in the test data, 2299 observations were correctly predicted as no stroke (true negatives), and 14 were correctly predicted as true positives.

```
> table(glm.pred,stroke.test)
        stroke.test
glm.pred    0    1
       0 2299   89
       1   52   14
```

- Through the Validation-Set Approach, the model yielded higher precision and recall rates:
  - 94.25% accuracy rate
  - 21.21% precision rate
  - 13.59% recall rate

C. Leave-One-Out Cross Validation (LOOCV)

- I used 1 dataset for the training and used the remainder as test data for the model.
- With the LOOCV, I looked at the accuracy rate of the model, and after performing the analysis, the model yielded a 96.24% accuracy rate and 3.75% cross-validation error.

```
> cv.err$delta #cross validation error
[1] 0.03753966 0.03753961
>
> #accuracy rate for LOOCV
> 1-cv.err$delta[1]
[1] 0.9624603
```

- Built a persona and ran the Logistic Regression
  - I wanted to test the model to predict whether someone will have a stroke or not based on the data I provided.
  - I built a persona using the relative frequency and median of the dataset for each predictor.
  - Supposedly, I wanted to predict the probability of having a stroke for a 28-year-old single male who never smokes, doesn't have heart disease or hypertension, and lives in an urban area.
  - I created a data frame with the following data for each predictor:
    - gender: male
    - age: 28
    - hypertension: 0 (does not have hypertension)
    - heart_diesase: 0, (does not have heart disease)
    - ever_married: no
    - residence_type: urban
    - glucose_level: 99
    - smoking_status: never smoked
  - After rerunning the analysis, the model gave a probability of 0.485% that the person will have a stroke. With this result, I concluded that the model predicted the person is less likely to have a stroke.

Project Report

- Through different resampling methods, I found that the Logistic Regression with Selected Predictors yielded the highest precision and recall rate.
- The most significant attributes to determine whether someone will have a stroke or not are age, average glucose levels, and whether the person has hypertension.
- There is a 0.485% probability that someone will have a stroke given the underlying condition the individual has and based on what the model predicted.
- To further improve this project, obtaining more samples of stroke occurrence would help mitigate the effects of the imbalance data and allow for performing other Classification methods, such as K-NN and Random Forest.