Healthcare Issue: Patients' Stroke Prediction

Objective

Using a specific characteristic, forecast whether the patient will experience a stroke or not. The AUC-ROC Score is the evaluation metric.

Data Understanding

The data's column definitions are listed below:

1. Patient ID, id
2. gender-Gender Patients
3. Age of the Patient
4. hypertension-0 means there is no hypertension, while 1 means you have hypertension.
5. heart disease-0 means there is no heart disease, and 1 means you have heart disease.
6. ever married-Yes/No
7. occupation type: Work Type
8. Residential type-Area Residential type (Urban/Rural)
9. Average Glucose Level (avg glucose level) (measured after meal)
10. Body mass index (BMI)
11. patient's smoking status (smoking status)
12. stroke-0: no stroke; stroke-1: stroke occurred.

Loading the Train data and the Final Test data on which I filed the Results will be the first step.

```
##        id            gender            age          hypertension
##  Min.   :    1   Female:25665   Min.   : 0.08   Min.   :0.00000
##  1st Qu.:18038   Male  :17724   1st Qu.:24.00   1st Qu.:0.00000
##  Median :36352   Other :   11   Median :44.00   Median :0.00000
##  Mean   :36326                  Mean   :42.22   Mean   :0.09357
##  3rd Qu.:54514                  3rd Qu.:60.00   3rd Qu.:0.00000
##  Max.   :72943                  Max.   :82.00   Max.   :1.00000
##
##  heart_disease     ever_married         work_type     Residence_type
##  Min.   :0.00000   No :15462    children     : 6156   Rural:21644
##  1st Qu.:0.00000   Yes:27938    Govt_job     : 5440   Urban:21756
##  Median :0.00000                Never_worked :  177
##  Mean   :0.04751                Private      :24834
##  3rd Qu.:0.00000                Self-employed: 6793
##  Max.   :1.00000
##
##  avg_glucose_level      bmi                smoking_status
##  Min.   : 55.00    Min.   :10.10                  :13292
##  1st Qu.: 77.54    1st Qu.:23.20   formerly smoked: 7493
##  Median : 91.58    Median :27.70   never smoked   :16053
##  Mean   :104.48    Mean   :28.61   smokes         : 6562
##  3rd Qu.:112.07    3rd Qu.:32.90
##  Max.   :291.05    Max.   :97.60
##                    NA's   :1462
##      stroke
##  Min.   :0.00000
##  1st Qu.:0.00000
##  Median :0.00000
##  Mean   :0.01804
##  3rd Qu.:0.00000
##  Max.   :1.00000
##
```

```
##        id            gender            age          hypertension
##  Min.   :    2   Female:10957   Min.   : 0.08   Min.   :0.00000
##  1st Qu.:18542   Male  : 7642   1st Qu.:24.00   1st Qu.:0.00000
##  Median :36717   Other :    2   Median :43.00   Median :0.00000
##  Mean   :36747                  Mean   :42.86   Mean   :0.09317
##  3rd Qu.:55114                  3rd Qu.:60.00   3rd Qu.:0.00000
##  Max.   :72942                  Max.   :82.00   Max.   :1.00000
##
##  heart_disease     ever_married         work_type     Residence_type
##  Min.   :0.00000   No : 6662    children     : 2613   Rural:9291
##  1st Qu.:0.00000   Yes:11939    Govt_job     : 2302   Urban:9310
##  Median :0.00000                Never_worked :   75
##  Mean   :0.04806                Private      :10750
##  3rd Qu.:0.00000                Self-employed: 2861
##  Max.   :1.00000
##
##  avg_glucose_level      bmi                smoking_status
##  Min.   : 55.00    Min.   :10.20                  :5751
##  1st Qu.: 77.55    1st Qu.:23.30   formerly smoked:3260
##  Median : 91.83    Median :27.70   never smoked   :6833
##  Mean   :104.39    Mean   :28.55   smokes         :2757
##  3rd Qu.:112.31    3rd Qu.:32.80
##  Max.   :275.72    Max.   :88.30
##                    NA's   :591
```

We must forecast for 18601 patients despite having data for 43400 patients. We can see that for a small number of patients, BMI and smoking status are absent.

Let's link the two tables to see if we can find a suitable replacement for them.

Smoking Status data was absent in 30.71% of cases, and 3.31% of cases lacked BMI data. We'll go ahead and replace it because the BMI is lower; the mean and median figures are similarly near. I'm replacing that with mean.

It is not suggested to fill in the missing data for smoking (30.71%) with mean or median. I'm going ahead and creating two models, one with available smoking data and the other without.

Handling Imbalanced Data

Let's now examine the number of positive and negative stroke data cases that we have.

Here are some smoke data:

```
##
##     0     1
## 29470   638
```

And here is for Non smoke data

```
##
##     0     1
## 13147   145
```

In both instances, it is clear that the data set is unbalanced, and if we proceed, there is a good chance that the machine learning system will correctly predict that no strokes will occur for all of the data. Therefore, we must balance the data.

In order to address this and balance the data, I am utilizing the ROSE approach, which creates fake data to balance the set.

Applying Model

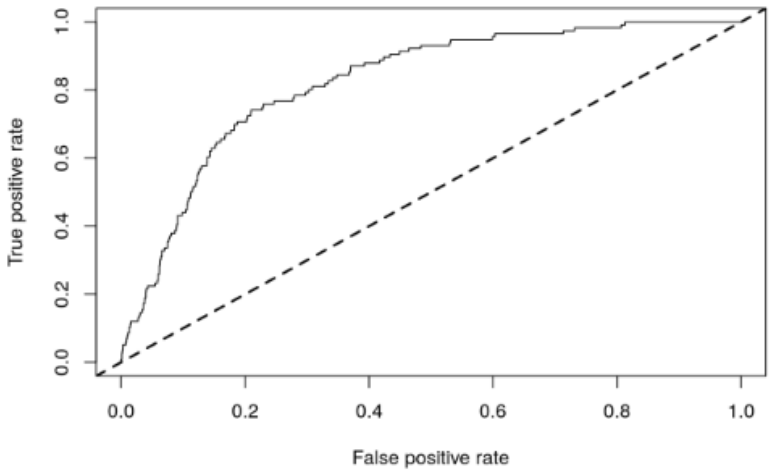
using Logistic regression, dividing the data into Train and Validate (80:20), and

Here is the model after being applied on a dataset with Smoke.

```
##
## Call:
## glm(formula = stroke ~ ., family = "binomial", data = train_filter_With_Smoke_Data_R
ose)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.73328  -0.79937  -0.00051   0.81476   2.76735
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -1.633e+01  8.581e+01  -0.192 0.847680
## id                           1.316e-06  6.995e-07   1.881 0.059917 .
## genderMale                   6.224e-02  3.264e-02   1.907 0.056514 .
## genderOther                 -1.345e+01  7.679e+02  -0.018 0.986029
## age                          6.634e-02  1.137e-03  58.355  < 2e-16 ***
## hypertension1                5.689e-01  3.984e-02  14.280  < 2e-16 ***
## heart_disease1               6.744e-01  5.158e-02  13.074  < 2e-16 ***
## ever_marriedYes             -6.312e-02  4.958e-02  -1.273 0.202960
## work_typeGovt_job            1.219e+01  8.581e+01   0.143 0.886993
## work_typeNever_worked       -1.578e-01  2.255e+02  -0.001 0.999442
## work_typePrivate             1.220e+01  8.581e+01   0.144 0.885892
## work_typeSelf-employed       1.233e+01  8.581e+01   0.145 0.884648
## Residence_typeUrban          1.218e-01  3.147e-02   3.871 0.000108 ***
## avg_glucose_level            3.651e-03  2.804e-04  13.024  < 2e-16 ***
## bmi                         -1.575e-02  2.438e-03  -6.463 1.03e-10 ***
## smoking_statusnever smoked -2.211e-01  3.732e-02  -5.924 3.13e-09 ***
## smoking_statussmokes         1.795e-01  4.464e-02   4.021 5.79e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 33389  on 24085  degrees of freedom
## Residual deviance: 24544  on 24069  degrees of freedom
## AIC: 24578
##
## Number of Fisher Scoring iterations: 14
```

So, as can be seen, significant variables include age, hypertension, heart disease, type of residence, average glucose level, BMI, and smoking status. Some of these are also intuitive, but others, including gender, marital status, and employment position, can be disregarded.

We forecast using this model for the validation set, plot the AUC-ROC, and determine the value to determine the accuracy of the validation set.



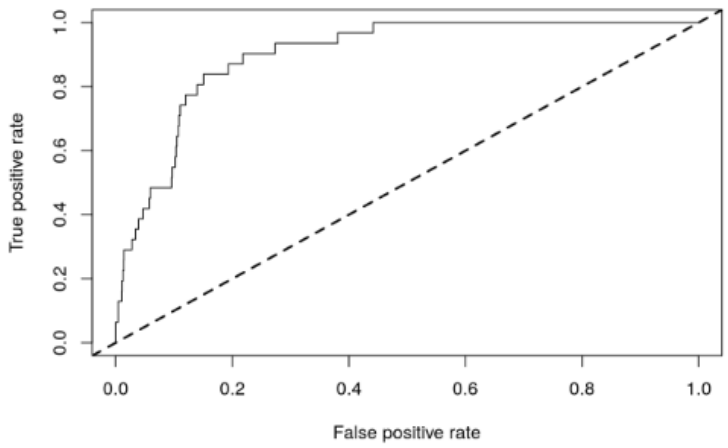
AUC over validation set for a dataset with Smoke data is 82.36%

Let's built the model for Dataset with non-smoke data

```
##
## Call:
## glm(formula = stroke ~ ., family = "binomial", data = train_filter_Without_Smoke_Dat
a_Rose)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7383  -0.4307  -0.1419   0.6647   3.0562
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -4.318e+00  1.559e-01 -27.694  < 2e-16 ***
## id                       1.754e-06  1.233e-06   1.423   0.1546
## genderMale               4.083e-01  5.505e-02   7.418 1.19e-13 ***
## age                      5.670e-02  1.723e-03  32.911  < 2e-16 ***
## hypertension1            1.726e-01  9.373e-02   1.842   0.0655 .
## heart_disease1           1.104e+00  1.050e-01  10.516  < 2e-16 ***
## ever_marriedYes          8.928e-01  8.829e-02  10.112  < 2e-16 ***
## work_typeGovt_job       -1.353e-01  1.702e-01  -0.795   0.4265
## work_typeNever_worked   -1.073e+01  1.522e+02  -0.070   0.9438
## work_typePrivate        -1.345e-01  1.539e-01  -0.874   0.3821
## work_typeSelf-employed  -8.095e-03  1.679e-01  -0.048   0.9615
## Residence_typeUrban      1.140e-01  5.423e-02   2.102   0.0356 *
## avg_glucose_level        3.752e-03  5.334e-04   7.033 2.02e-12 ***
## bmi                     -3.295e-04  3.924e-03  -0.084   0.9331
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14740.1  on 10632  degrees of freedom
## Residual deviance:  8756.4  on 10619  degrees of freedom
## AIC: 8784.4
##
## Number of Fisher Scoring iterations: 12
```

Here, we can see that features include gender, age, heart disease, marital status, and average blood sugar level. These diverge significantly from the prior design.

We forecast using this model for the validation set, plot the AUC-ROC, and determine the value to determine the accuracy on the validation set.



AUC over validation set for dataset with Smoke data is 90.05%

Prediction and Result

Post that I used the models to predict the value for the Test set given in for the competition and submitted the result. It gave 75% AUC-ROC value for the submitted result.

Have shared the final results as well

Conclusion

Overall, logistic regression was used to predict whether a patient would experience a stroke or not. We must deal with the unbalanced data that characterizes these healthcare issues. We might test several approaches to dealing with unbalanced data, such SMOTE, to improve the model.

Additionally, there were various ways we could have handled the missing smoke status data, for instance. Patients with ages under 10 or 15 may have been labeled as never smoked, etc.

The age distribution of the two data sets may be one factor explaining why the two models were so different. The median age of the datasets for smokers and non-smokers was 48 and 21, respectively.

These are some potential improvements to the logistic model.

If you have any suggestions for how this model could have been enhanced or if there is another ML method we could employ for such datasets, please let us know.

Reference :

https://www.kaggle.com/code/asaumya/healthcare-problem-prediction-stroke-patients