Stony Brook University

Imbalanced Classification Techniques to Assess Stroke Risk

AMS 561 / DCS 521

Matthew Hureau, Amr Ibrahim

FAR BEYOND

Stony Brook University

Experience

Amr

Degree: Biomedical Engineering

Coding Languages: MATLAB, Python

Skills Gained: Visualization using plotly.express and dealing with both categorical and imbalanced data

Matt

Degree: Applied Mathematics and Statistics (Statistics Track)

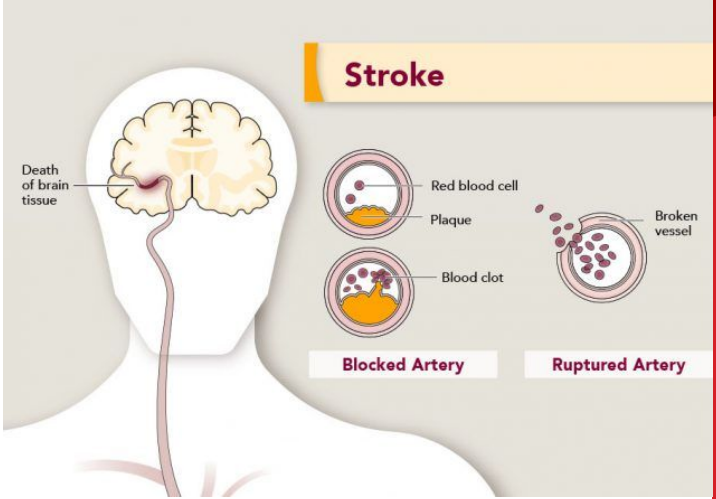
Coding Languages: R, SAS

Skills Gained: Visualization in Python, more complex machine learning procedures in R

Stony Brook University

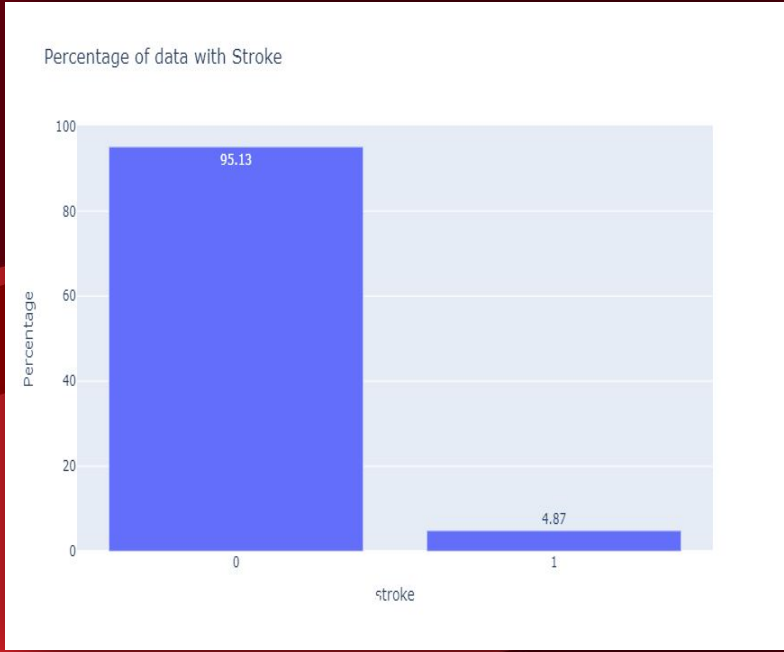
Motivation

- Stroke occurs when blood supply in the brain is blocked
- Stroke is a leading cause of death in the US, and roughly 80% of them are preventable
  - 800,000 strokes per year
  - 140,000 of which die
- Question: Which variables have significant predicting power for a stroke?
- Health related data relevant to areas of interest (Biomedical Engineering, Biostatistics)
- Becoming more well rounded data analysts (first project)



Stony Brook University

IMBALANCED CLASSIFICATION



What do imbalanced classes look like?

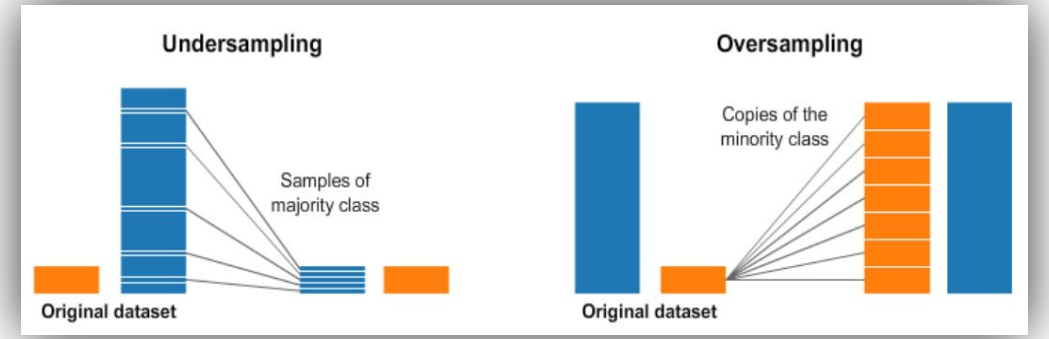
```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1210   66
         1    0    1

               Accuracy : 0.9483
                 95% CI : (0.9347, 0.9598)
    No Information Rate : 0.9475
    P-Value [Acc > NIR] : 0.4824

                  Kappa : 0.0279

 Mcnemar's Test P-Value : 1.235e-15

            Sensitivity : 0.0149254
            Specificity : 1.0000000
```

Logistic Regression model fit to imbalanced data

FAR BEYOND

Techniques and Tools

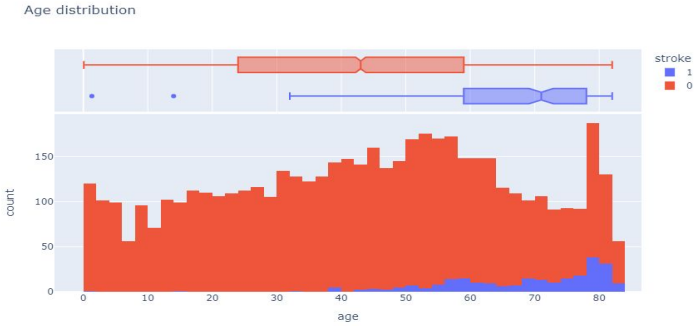
- Python and R
- Python Modules
  - pandas, plotly.express, sklearn
- R Packages
  - ROSE, ggplot, caret, rpart, and more
- **Undersampling**, **Oversampling**, **Under/Over Sampling,** and **Synthetic Data Generation** via ROSE

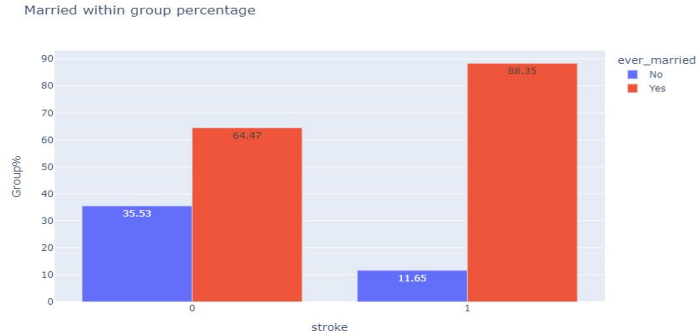


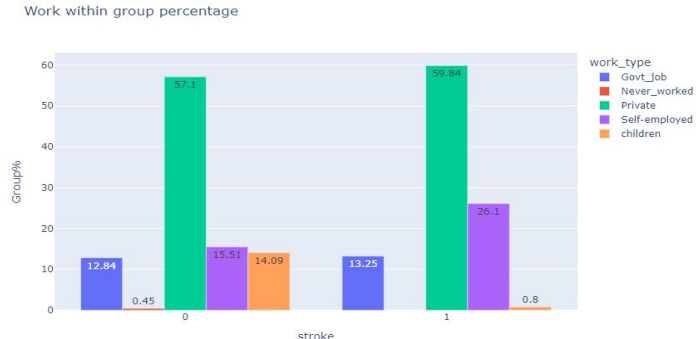
Stony Brook University

Data Visualization

- 5109 observations on 11 variables (10 predictors)
- Continuous predictors:
  - **Age**
  - BMI
  - Average glucose level
- Binary categorical predictors:
  - Health-related:
    - **Hypertension**
    - Heart disease
  - Lifestyle-related:
    - Gender
    - **Ever married**
    - Residence type
- Non-binary categorical predictors:
  - Work type
  - Smoking status







FAR BEYOND

Building the Best Model

- We wanted to treat this task as we would in a professional setting
- 10-Fold CV for models with “tunable” parameters
- Logistic Regression
  - Stepwise, Backward Selection, Forward Selection
- Ridge, Lasso, Elastic Net
- Neural Network, Random Forest, K-Nearest Neighbors
- Support Vector Machine
  - Used exclusively linear kernel for the three larger training sets
  - Polynomial Kernel was not considered for any of the training sets

45 machine learning models fit across the four imbalanced classification methods

(around 800 lines of R code)

```
##
## Call:
## glm(formula = stroke ~ ., family = "binomial", data = training.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0310  -0.7888   0.1778   0.8243   2.5175
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -2.380772   0.967823  -2.460   0.0139 *
## genderMale                    -0.160190   0.265378  -0.604   0.5461
## age                            0.071396   0.010100   7.069 1.56e-12 ***
## hypertension                   0.242258   0.321646   0.753   0.4513
## heart_disease                 -0.084672   0.373546  -0.227   0.8207
## ever_marriedYes               -0.020911   0.417553  -0.050   0.9601
## work_typeGovt_job             -1.991576   1.049762  -1.897   0.0578 .
## work_typeNever_worked        -12.604695 882.743924  -0.014   0.9886
## work_typePrivate              -1.231746   1.013303  -1.216   0.2241
## work_typeSelf-employed        -2.000617   1.080368  -1.852   0.0641 .
## Residence_typeUrban            0.027337   0.258127   0.106   0.9157
## avg_glucose_level              0.005341   0.002503   2.134   0.0329 *
## bmi                           -0.027488   0.021680  -1.268   0.2048
## smoking_statusnever smoked    -0.004514   0.332441  -0.014   0.9892
## smoking_statussmokes           0.194115   0.401774   0.483   0.6290
## smoking_statusUnknown         -0.470086   0.388051  -1.211   0.2257
## ---
```

Stony Brook University

```
##
## Call:
## glm(formula = stroke ~ ., family = "binomial", data = training.data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.0310   -0.7888    0.1778    0.8243    2.5175
##
## Coefficients:
##                             E
## (Intercept)                 -2.3807                        0.0139 *
## genderMale                  -0.160190   0.265378  -0.604   0.5461
## age                          0.071396   0.010100   7.069 1.56e-12 ***
## hypertension                 0.242258   0.321646   0.753   0.4513
## heart_disease               -0.084672   0.373546  -0.227   0.8207
## ever_marriedYes             -0.020911   0.417553  -0.050   0.9601
## work_typeGovt_job           -1.991576   1.049762  -1.897   0.0578 .
## work_typeNever_worked      -12.604695 882.743924  -0.014   0.9886
## work_typePrivate            -1.231746   1.013303  -1.216   0.2241
## work_typeSelf-employed      -2.000617   1.080368  -1.852   0.0641 .
## Residence_typeUrban          0.027337   0.258127   0.106   0.9157
## avg_glucose_level            0.005341   0.002503   2.134   0.0329 *
## bmi                         -0.027488   0.021680  -1.268   0.2048
## smoking_statusnever smoked  -0.004514   0.332441  -0.014   0.9892
## smoking_statussmokes         0.194115   0.401774   0.483   0.6290
## smoking_statusUnknown       -0.470086   0.388051  -1.211   0.2257
```

Other variables held constant:
1 unit (year) increase in age
= change in log odds of having a stroke by
0.071396 (odds are 7% higher)

FAR BEYOND

```
##
## Call:
## glm(formula = stroke ~ ., family = "binomial", data = training.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0310  -0.7888   0.1778   0.8243   2.5175
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -2.380772   0.967823  -2.460   0.0139 *
## genderMale                    -0.160190   0.265378  -0.604   0.5461
## age                            0.071396   0.010100   7.069 1.56e-12 ***
## hypertension                   0.242258   0.321646   0.753   0.4513
## heart_disease                 -0.084672   0.373546  -0.227   0.8207
## ever_marriedYes               -0.020911   0.417553  -0.050   0.9601
## work_typeGovt_job             -0.[illegible]
## work_typeNever_worked         -0.[illegible]  0.9886
## work_typePrivate              -1.231746   1.013303  -1.216   0.2241
## work_typeSelf-employed        -2.000617   1.080368  -1.852   0.0641 .
## Residence_typeUrban            0.027337   0.258127   0.106   0.9157
## avg_glucose_level              0.[illegible]  0.[illegible]  [illegible]  0.[illegible]329 *
## bmi                           -0.[illegible]  0.[illegible]  [illegible]  0.[illegible]48
## smoking_statusnever           -0.[illegible]  0.[illegible]  [illegible]  0.[illegible]92
## smoking_statussmokes          -0.[illegible]  0.[illegible]  [illegible]  0.[illegible]290
## smoking_statusUnknown         -0.4700[illegible]51  -1.211   0.2257
```

Negative coefficient: the event (stroke) is less likely for this class than the reference level "children"

The odds of having a stroke for someone who is self employed is exp(-2.000617) = 0.135 times that of a stay at home parent (odds are 7.4 times higher for a stay at home parent)

Results

- The oversampling method yielded the best results for our data, but we recommend being as exhaustive as possible
- “Best model” is context dependent
- Apply imbalanced techniques **AFTER** you split the data into training and testing
  - Overfitting

BEST MODEL: Stepwise/Backward/Forward Logistic Regression (Undersampled)

logit(p) = - 3.38 + 0.073*age - 1.90*work_type_govt - 12.77*work_type_never - 1.14*work_type_private - 1.88*work_type_selfemployed + 0.0043*avg_glucose

Note: “children” is the reference level for work_type

Overall Model Accuracy	Sensitivity	Specificity
0.7565	0.7910	0.7546