Stroke Project

Yuechen Liu, Yanhao Li, Mufeng Xu

Group 22

Introduction

Stroke is a serious life-threatening medical condition.[1] According to the World Health Organization, stroke is the second leading cause of death globally.[2] To better understand which factors correlate to the stroke event, our group found a stroke prediction dataset. This dataset contains twelve columns. The first column labels the unique identifier of the patient. The last column records the occurrence of stroke by 1 (Yes) or 0 (No). The other ten columns contain the observations of predictors and demographic variables.[2]

For exploratory analysis, we modified the raw data and created the first data frame. We removed the demographic variables. Then, we transformed the “bmi” predictor to numeric class. In addition, we transformed outcome and all categorical predictors to factor class. For modeling, we modified the first data frame and created the second data frame. Firstly, we cleaned the predictor names and removed rows that contain “NA” in “bmi” and “Other” in “gender”. Secondly, we replaced phrases in categorical predictors by numbers and numbers in outcome by phrases. Finally, we rearranged the variables to place three continuous variables in the beginning.

Exploratory Data Analysis

A summary table (Table 1) was created to show the descriptive statistics of all predictors with and without experiencing strokes. Frequency and percentage are shown for categorical predictors. Mean, standard deviation, and range are shown for continuous predictors.

According to the percentages of hypertension, there may be an association between hypertension and stroke events. Percentage of hypertension among people who

experienced stroke is almost three times the percentage of hypertension among people without stroke experience. Similar to hypertension, the percentage of heart disease among people who experienced stroke is almost four times the percentage of heart disease among people without stroke experience. There may be an association between heart disease and stroke events.

According to the percentages of gender, there may be no association between gender and stroke events. Percentage of female and male among people who experienced stroke is almost the same as the percentage of female and male among people without stroke experience.

	0 (N=4861)	1 (N=249)	Total (N=5110)	p value
gender				0.790
- Female	2853 (58.7%)	141 (56.6%)	2994 (58.6%)	
- Male	2007 (41.3%)	108 (43.4%)	2115 (41.4%)	
- Other	1 (0.0%)	0 (0.0%)	1 (0.0%)	
age				< 0.001
- Mean (SD)	41.972 (22.292)	67.728 (12.727)	43.227 (22.613)	
- Range	0.080 - 82.000	1.320 - 82.000	0.080 - 82.000	
hypertension				< 0.001
- 0	4429 (91.1%)	183 (73.5%)	4612 (90.3%)	
- 1	432 (8.9%)	66 (26.5%)	498 (9.7%)	
heart_disease				< 0.001
- 0	4632 (95.3%)	202 (81.1%)	4834 (94.6%)	
- 1	229 (4.7%)	47 (18.9%)	276 (5.4%)	
avg_glucose_level				< 0.001
- Mean (SD)	104.796 (43.846)	132.545 (61.921)	106.148 (45.284)	
- Range	55.120 - 267.760	56.110 - 271.740	55.120 - 271.740	
bmi				0.003
- N-Miss	161	40	201	
- Mean (SD)	28.823 (7.908)	30.471 (6.329)	28.893 (7.854)	
- Range	10.300 - 97.600	16.900 - 56.600	10.300 - 97.600	

Table 1: summary statistics of all predictors

We performed the density plots (Figure 1) for the three continuous variables. According to the density plot for age, there may be an association between age and stroke event. There is a clearly left-skewed curve for patients that experienced stroke. Compared to that, the curve for patients without stroke experience is close to bell shape.

According to the density plot for bmi, there may be no association between bmi and stroke events. Two curves are quite similar.

According to the density plot for average glucose level, it is hard to predict association. Two curves have similarities. However, there are still some differences on the magnitudes of peaks.

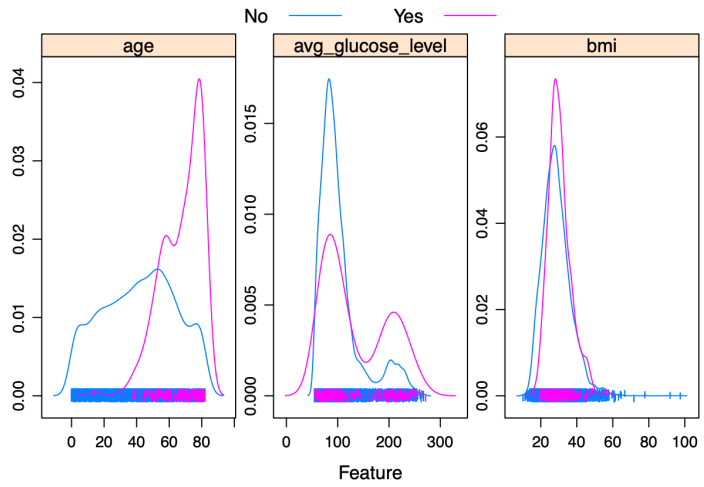

Figure 1: density plots of 3 continuous variables(age, average glucose level and bmi) without stroke(blue), and with stroke(pink)

Models

Included Predictors

In the modeling, we included all six predictors. They are “gender”, “hypertension”, “heart disease”, “age”, “average glucose level”, and “bmi”. The first three are categorical predictors. The last three are continuous predictors. In the previous exploratory analysis, we made hypotheses that “hypertension”, “heart disease”, and “age” would play relatively important roles in modeling. In contrast, we hypothesized that “gender” and “bmi” would not play important roles in modeling.

Used Technique

To start the modeling, we created our training data. Each row has a 80% chance to get in training data. The held-out data is test data. By this method, my training data has 3928 rows. My test data has 980 rows.

We used several models to fit the dataset: GLM, GLMN, MARS, GAM, LDA, Random Forest, gbmA. Svml and Svmr. This dataset met our assumptions: variables need to be independent and sample size should be large.

Tuning Parameters

Tuning parameters were used in the MARS model to minimize prediction error. The number of retained terms and degree of freedom were selected using cross-validation. The tuning parameters were set to be 8 retained terms and 1 degree of freedom.

Test Performance

For all the models, we applied a simple classifier with a cut-off of 0.6. Then, we evaluated the performance of models by the test data.

The no information rate is 0.9582 in all models. For GLM, GLMN, and MARS, they all have accuracies at 0.9592, which is slightly larger than no information rate. Even though these values are very close, 0.9592 is the highest accuracy we got. For GAM, Random Forest, gbmA, Svml, and Svmr, they all have accuracies at 0.9582, which equals no information rate. For LDA, it has accuracy at 0.9531, which is lower than no information rate. Consequently, we consider LDA as the worst model.

For GAM, Random Forest, gbmA, Svml, and Svmr, they all have sensitivities at 0, which is problematic. Sensitivity measures the proportion of positives that are correctly identified. When sensitivity equals 0, the model predicts all positives as negatives. In the circumstance of this dataset, these models predict all people who experienced stroke as people without stroke experience. As a result, we consider GAM, Random Forest, gbmA, Svml, and Svmr as bad models.

For GLM, GLMN, and MARS, they all have very low sensitivities, which is still problematic. However, we have to say that these three models are better than previous models.

Variable Importance

By generating an importance plot (Figure 2), we can see that the most important predictor in predicting stroke for people is “age”, followed by “hypertension” and “bmi”. All variables were kept for further prediction and analysis since there is no clue in how one variable is not associated with the outcome.

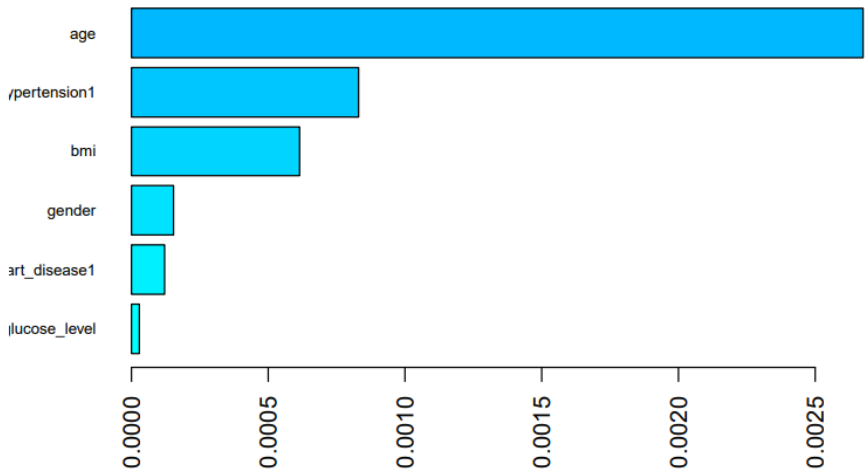

Figure 2: importance plot for all variables, deeper blue indicating more important

Final Model

To compare between models, ROC boxplots (Figure 3) were created. Higher ROC indicates that the model predicts the outcome better. Based on Figure , GLMN has the best ability in predicting whether a participant has experienced strokes. Bias would be created when there is a large number of collinear variables.

However, the assumption of normality does not hold all the time, we may have our categorical features such as gender, ever_married, etc. or we may have time series data which are not normally distributed.

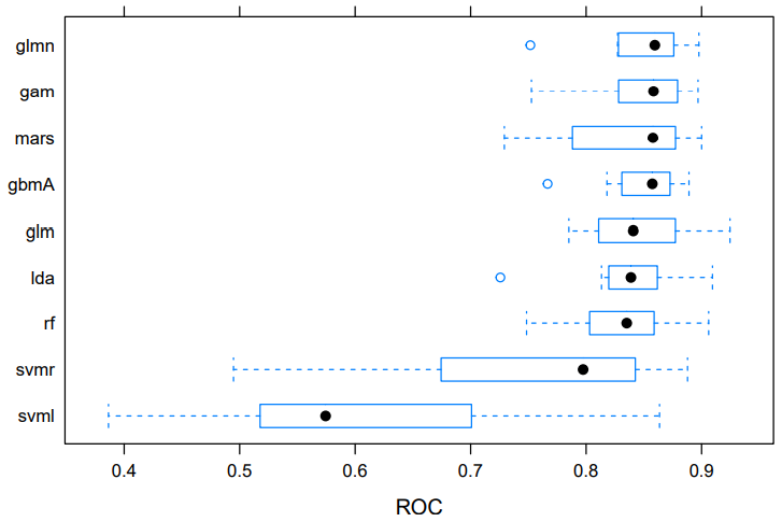

Figure 3: ROC boxplot for all models

Conclusion

According to the importance plot (Figure 2), we can see that the most important predictor in predicting stroke for people is “age”, followed by “hypertension” and “bmi”. Some predictors match our hypotheses found in exploratory analysis. For example, “age” and “hypertension” do appear in the first two places. However, some predictors do not match our hypotheses. For example, “heart disease” appears to be less important than “bmi”.

In conclusion, elder or fatter people, or people with hypertension are more associated with strokes. Females, people with heart disease or higher average glucose level are less associated with strokes.

However, when making predictions using all models above, we found that they did not have significant advantages or weaknesses and none of them worked well with the dataset. In comparison, GLMN has the best ability in predicting whether people will

have strokes using their information. Further analysis and more models should be implemented in order to have a higher predicting accuracy.

Citation

1. Stroke. *nhs.uk* https://www.nhs.uk/conditions/stroke/ (2017).
2. Stroke Prediction Dataset. https://kaggle.com/fedesoriano/stroke-prediction-dataset