Stroke Prediction

Lin Yang ly2565

**Introduction**

Stroke is a medical emergency and a brain attack that interrupts blood supply and oxygen to the brain. According to the World Health Organization (WHO), stroke is the second leading cause of death globally, and leads to approximately 11% of total deaths. An estimated 17.9 million people died from cardiovascular diseases in 2019, and 85% of these deaths were due to heart attack and stroke(WHO, 2021). The high stroke mortality has caused significant cost burden, including healthcare services and medications. In fact, there are many risk factors that can be modified to reduce the burden of stroke in the population, such as smoking, physical inactivity, and hypertension (). For these reasons, it is important and necessary to identify and study these modifiable risk factors for stroke. This report is an analysis of a dataset of patients for the purpose of predicting the probability that a patient gets stroke based on predictors like gender, age, hypertension, work type, and body mass index (BMI). Each row in the dataset provides relevant information about the patient.

All columns we have are shown below:

* id: unique identifier
* gender: "Male", "Female", or "Other"
* age: age of the patient
* hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
* heart\_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
* ever\_married: "No" or "Yes"
* work\_type: "children", "Govt\_jov", "Never\_worked", "Private", or "Self-employed"
* residence\_type: "Rural" or "Urban"
* avg\_glucose\_level: average glucose level in blood
* bmi: body mass index
* smoking\_status: "formerly smoked", "never smoked", "smokes", or "Unknown"
* stroke: 1 if the patient had a stroke or 0 if not

**Data Cleaning and Exploratory Data Analysis**

To build predictive models and extract insights from the data, data cleaning needs to be performed. First, an observation with the gender of “Other” and the ones with missing bmi values are removed. This approach is appropriate because the removed observations are a small proportion of the dataset. Then, all categorical variables are converted to factors and are assigned numeric values. Note that “Unknown” is treated as a level of the factor variable, smoking status, even it means information is unavailable for the patient. The response variable, stroke, is converted to a factor variable with “pos” representing the patient had a stroke or “neg” representing the other way. After data cleaning, the cleaned data contains 4908 observations of 11 variables. Three of them are continuous variables: age, average glucose level, and bmi, the remaining are categorical variables. The prevalence of stroke in this dataset is found to be 4.3%. Of all the patients, there are 2897 females and 2011 males, 3204 patients were ever married, 451 patients have hypertension, and 243 patients have heart disease. The distribution of age between two stroke groups suggests that order people tend to have higher probability of experiencing a stroke(Figure?). This can be verified by the very different density plots of response vs. age (Figure?), indicating that age is informative in making predictions on stroke status. A correlation plot of all predictors is also made, showing that most predictors are not correlated with each other, except that marital status is positively correlated with age, which is reasonable(Figure?).

**Model Building**

Since the response variable is binary, either positive or negative, classification models are appropriate in analyzing this dataset. With the insights gained form the exploratory data analysis step, we fitted multiple models to analyze risk factors for stroke: logistic regression, penalized logistic regression, linear discriminant analysis (LDA) model, generalized additive model (GAM), and multivariate adaptive regression splines (MARS) model. ROC summaries of the five models and boxplots of AUC scores are shown in Figure ?. It shows that the penalized logistic regression model has the highest AUC score, meaning it has the best performance at distinguishing between positive and negative stroke cases, it is thus selected to be the optimal model predicting whether a patient would get a stroke a not. Fitting a logistic regression model requires some assumptions. First, logistic regression requires the observations to be independent of each other. Second, there should be no or little correlation between predictors. Third, it assumes the log odds and independent variables to be linearly related. Based on the results in the EDA part, these assumptions are not violated.

Tuning parameter test data performance The confusion matrix and ROC curve of the penalized logistic regression model is shown in Figure ? The overall prediction accuracy, 0.9582 with a 95% CI (0.9437, 0.9698), is quite high. However, the no-information rate is the same as the accuracy, indicating if we have no information and predict all observations to either positive or negative class, the accuracy would be 50%. This can be explained by the fact that the penalized logistic regression model classifies all observations to the negative class. It’s also the reason why the specificity is 1 and positive predictive value (PPV) is N/A. The model’s AUC is 0.833.

Important variables

**Limitations**

Unknown Smoking status missing bmi values

The early expectations lead to the belief that Age, Hypertension, Heart Disease, and average glucose level are the most indicative risk factors for predicting a stroke based on this data. However, as only 4.87% of the dataset did experience a stroke, it is likely that the models will be inaccurate in being able to predict a stroke.

**References**

Boehme AK, Esenwa C, Elkind MS. Stroke Risk Factors, Genetics, and Prevention. Circ Res. 2017;120(3):472-495. doi:10.1161/CIRCRESAHA.116.308398