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 burdens, including healthcare services and medications. Between 2012 and 2030, the total direct annual stroke-related medical costs are expected to increase from $71.55 billion to $183.13 billion (Ovbiagele, et al.). In fact, many risk factors can be modified to reduce the burdens of stroke in the population, such as smoking, physical inactivity, and hypertension (Boehme et al.). For these reasons, it is important and necessary to identify and study these modifiable risk factors for stroke. The purpose of this study is to analyze a dataset of patients and build an optimal model to predict the probability that a patient gets a 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 variables 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 factor level of 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. The cleaned dataset 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 having a 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 distributions of age, average glucose level, and bmi across two stroke groups suggest that older people with higher glucose levels and bmi tend to have a higher probability of experiencing a stroke (**Figure 1**). The very different density plots of response vs. age (**Figure 2**) also indicate that age is an informative variable in making predictions on stroke status. When checking collinearity, most predictors are not correlated with each other, except that marital status is positively correlated with age, which is expected.

Model Building

Since the response variable is binary, either positive or negative, it's appropriate to build classification models on this dataset and use AUC as the evaluation metrics. With the insights gained from 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 based on the training data are shown in **Table 1.** The penalized logistic regression model is found to have 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 or 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.

The best tuning parameters selected for the penalized logistic regression model using 5-fold cross-validation are 0.45 (alpha) and 0.00697 (lambda). The confusion matrix of the model based on the test data (**Table 2**) shows that the overall prediction accuracy is 0.9582 with a 95% CI (0.9437, 0.9698). However, the no-information rate is the same as the accuracy, meaning if we have no information and predict all observations to either positive or negative class, the accuracy would be 95.82%. In addition, the kappa coefficient is found to be 0, suggesting there is no agreement between classification and true values. 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 the positive predictive value (PPV) is N/A. The ROC curve (**Figure 3**) using the test data shows that the model's AUC is 0.833. Finally, hypertension, heart disease, age, and average glucose level are found to be significant predictors for stroke according to the variable importance plot (**Figure 4**).

Conclusions

By using 5-fold cross-validation, the penalized logistic regression model is selected to be the optimal model for predicting stroke because of its highest AUC. As we expected, in this model, whether a patient has hypertension or not, whether a patient has heart disease or not, age, and average glucose level are found to have statistically significant impacts on predicting the probability of suffering a stroke. Older people who have hypertension, heart disease, and higher glucose level are more likely to have a stroke. This finding brings significant implications regarding how to reduce the risk of stroke. People can diminish the risk factor, average glucose

level with a healthy dieting and exercise, especially for older people. Patients with hypertension or heart disease should particularly pay attention to guidelines about stroke prevention.

On the other hand, there are some problems with our final model. One of them is that it classifies all the test data into the negative class. Also, some significant risk factors from prior literature are not found to play important roles in this model, such as smoking status and bmi. These problems may be due to the limitations of this model. First, the “Unknown” smoking status is treated as an individual factor level while it means no information is available for this patient. Smoking status has a potential correlation with stroke, so accurate smoking status of these patients may help in building better models. Second, the prevalence of having a stroke in this dataset is only 4.3%. There is a large class imbalance between patients who had a stroke and those who didn’t, as a result, the models we used may be inaccurate in predicting a stroke. Including more patients who had a stroke or oversampling the data may help improve the model’s prediction performance to more accurately indicate the risk factors of a stroke.

Appendix

Table 1
ROC summaries of five models.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
glm	0.7459324	0.8261499	0.8469962	0.8460861	0.8726924	0.9200563	0
glmn	0.7504693	0.8329161	0.8541927	0.8507349	0.8751173	0.9247497	0
lda	0.7184397	0.8185553	0.8391084	0.8351982	0.8620314	0.8959473	0
gam	0.7413955	0.8257196	0.8466051	0.8458014	0.8726631	0.9203692	0
mars	0.7291927	0.8008350	0.8505945	0.8385021	0.8699827	0.9211514	0

Table 2
Confusion matrix of the penalized logistic regression model.

	neg	pos
neg	939	41
pos	0	0

Accuracy	0.9581633
Kappa	0.0000000
AccuracyLower	0.9436693
AccuracyUpper	0.9698128
AccuracyNull	0.9581633
AccuracyPValue	0.5414165
McnemarPValue	0.0000000

Sensitivity	0.0000000
Specificity	1.0000000
Pos Pred Value	NaN
Neg Pred Value	0.9581633
Precision	NA
Recall	0.0000000
F1	NA
Prevalence	0.0418367
Detection Rate	0.0000000
Detection Prevalence	0.0000000
Balanced Accuracy	0.5000000

Figure 1
Distributions of continuous variables across two stroke groups.

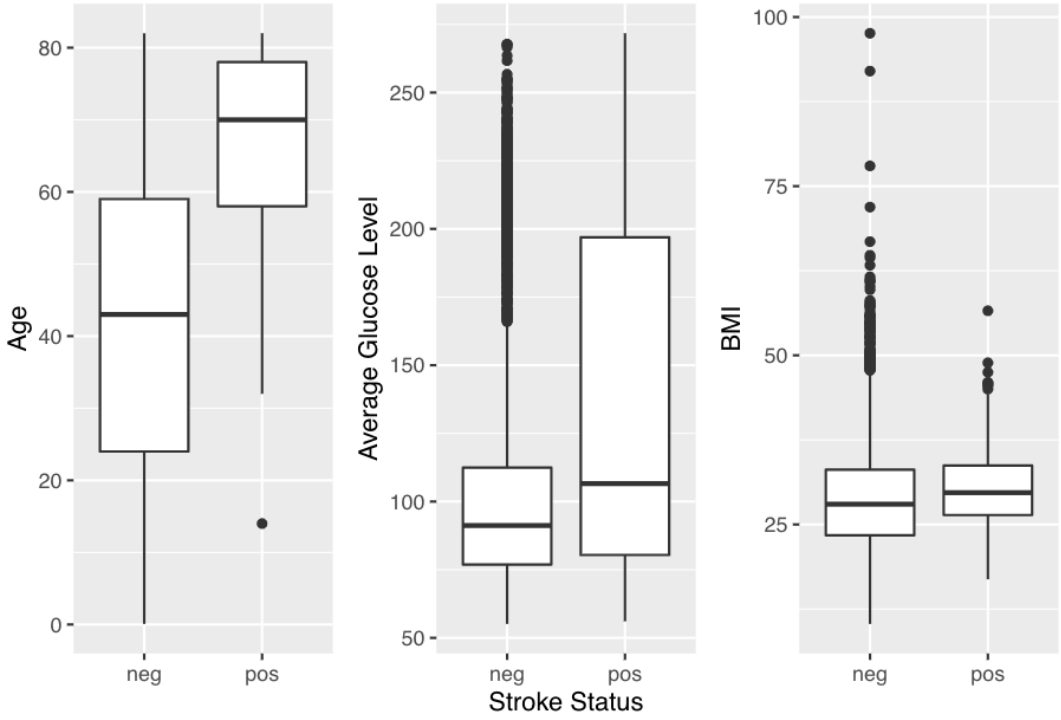


Figure 2
Density plots of stroke vs. continuous variables.

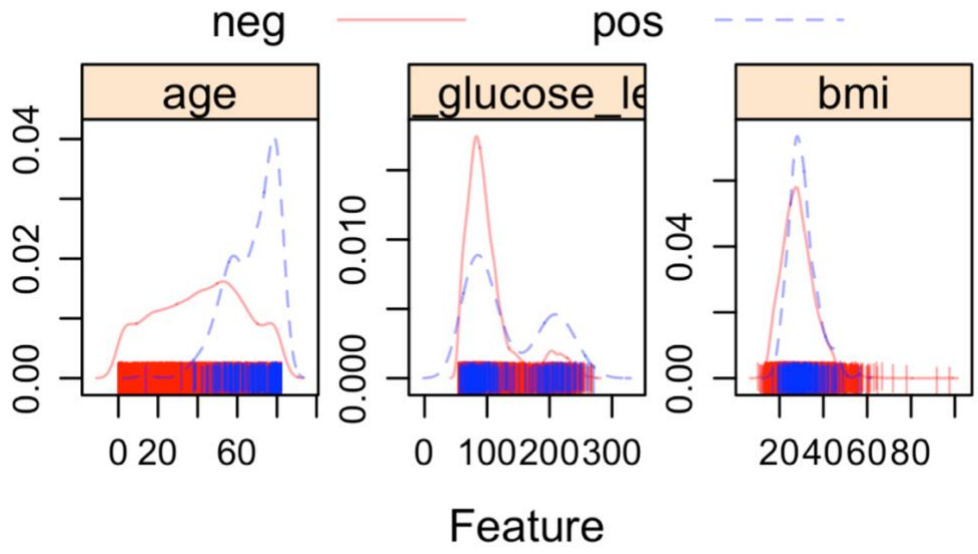


Figure 3
The ROC curve of the penalized logistic regression model.

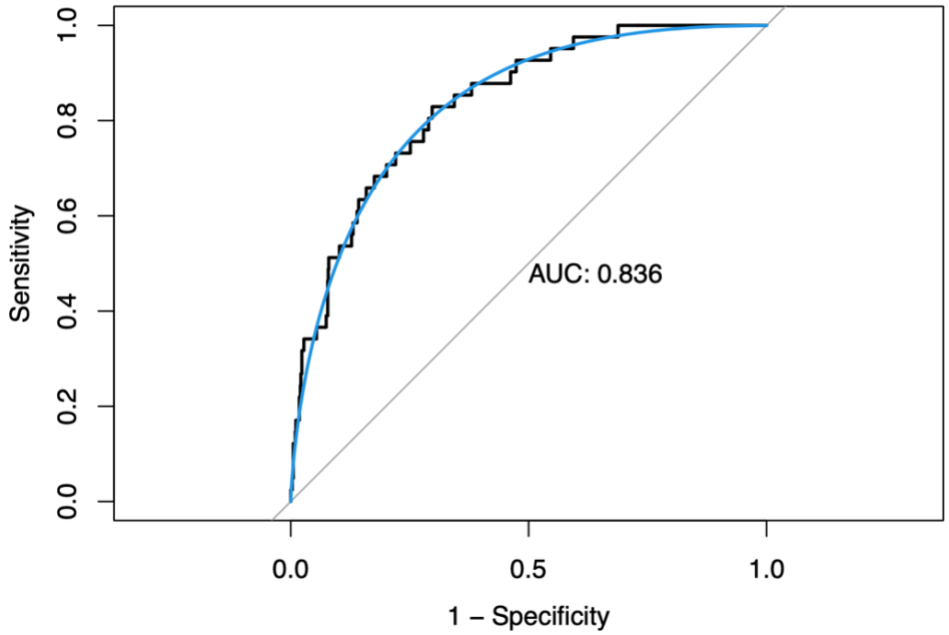
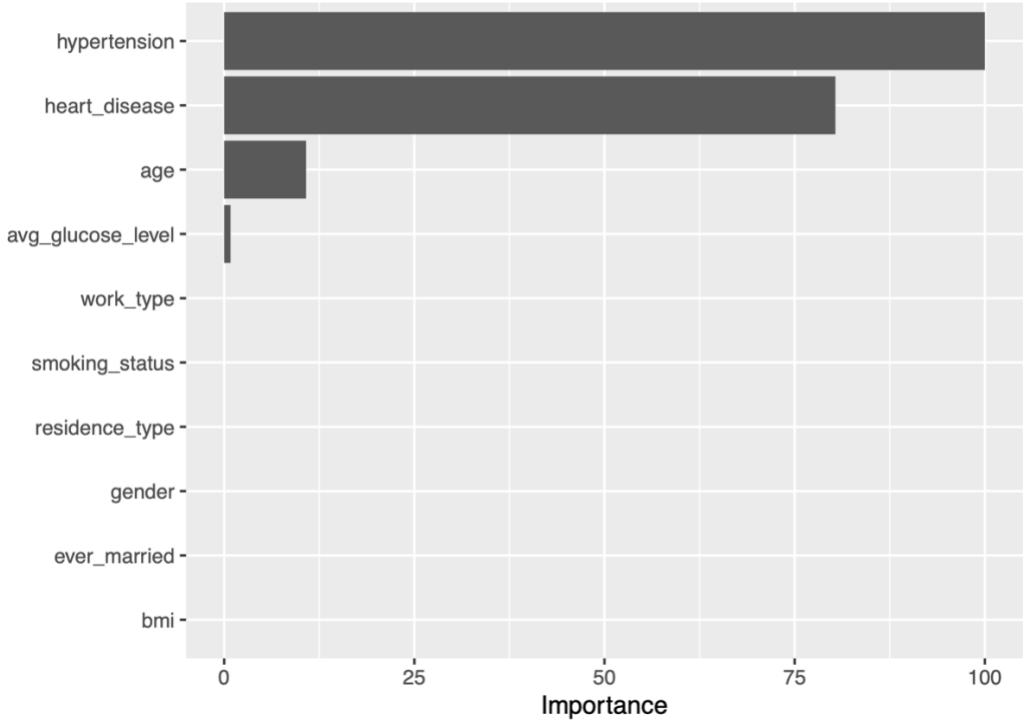


Figure 4
The variable importance plot of the penalized logistic regression model.



References

Ovbiagele B, Goldstein LB, Higashida RT, Howard VJ, Johnston SC, Khavjou OA, et al. Forecasting the future of stroke in the united states: A policy statement from the american heart association and american stroke association. Stroke; a journal of cerebral circulation. 2013;44:2361–2375.

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.