Stroke Prediction

Ashleigh Colquhoun and Tong Chen

December 9, 2021

Abstract

Stroke is a leading cause of death, as stated by the World Health Organization. We analyze 11 variables such as gender, age, whether a person has hypertension or not, presence of heart disease, whether a patient has ever been married, the type of work they do, where patients reside, average glucose levels, BMI, smoking status to predict whether the patient had a stroke or not. In this paper, we identify factors that lead to stroke and choose the best predictive models based on both test errors and area under the curve of ROC among different models such as Bagging, Gradient Boosting Machine, Logistic Regression, Support Vector Machine(SVM), and Logistic Ridge Regression.

1. Introduction

Stroke is one of the leading causes of death and disability in the United States. Based on the Centers for Disease Control and Prevention, more than 795,000 people have a stroke in the United States, and about 610,000 of these are first or new strokes. A stroke is a medical emergency, and prompt treatment is crucial. Hence, early awareness would be helpful.

Motivated by Songhee Cheon's early research in 2019, his team used the Principal component analysis featuring quantile scaling to extract relevant background features from medical records to predict the stroke. Their study concluded that the DNN approach with an AUC of 83.48% was the best model that both patients and doctors could use to prescreen for possible stroke (Cheon et al. 2019). However, we are curious about whether we could use other models to predict stroke in this paper.

This paper aims to find the factors contributing to the higher probability of getting a stroke and the best model to predict the stroke. In order to find the optimal model, we identify factors that lead to stroke and derive a predictive model based on the results of the ROC curve by using different models such as bagging, random forest, logistic regression, support vector machine(SVM), and ridge regression. By calculating the areas under the curve of these models, we could compare their area, which indicates the optimal prediction model for our following analysis.

2. Materials and Methods

2.1. Subjects

We used data from Kaggle, centered on stroke data, updated in February 2021. The subjects were 5,111 people whose general statistical information was collected to predict the potential variables that cause the stroke.

Of these groups of people, 59% were female, and 41% were male. The mean age of these participants is 43.2 years old. Furthermore, most participants do not have hypertension or a history of heart disease based on this dataset. In addition, most are not married and work in private industries and live in urban areas.

2.2. Principal Variables

The dependent variable was the event of a stroke. And we have 11 Independent variables: gender, age, whether a person has hypertension or not, presence of heart disease, whether a patient has ever been married, the type of work they do, where patients reside, average glucose levels, BMI, and smoking status. Among our dataset, we have 4 binary variables, 4 categorical variables and 3 numerical variables. For this dataset, we are particularly interested in numerical variables such as BMI, age, and average glucose level.

2.3. Methods

This paper mainly uses two compassion methods to find the optimal models; we first use the test error to compare the selected models. Then we picked the models with relatively small test errors and a control group of one model with the highest test errors to work with the ROC curve. By comparing their AUC, we find the optimal model of SVM that fits our data best.

(1) *Bagging*

In the beginning, we used the bootstrap in our dataset since we have a relatively large sample size, and the bagging improves the performance of classification or regression methods.

(2) *Random forest*

In bagging, we find that the bootstrapped prediction rules are highly correlated, contributing to its relatively high test errors. Hence, we use the random forest to de-correlate, and it reduces test errors by 1.92995%, which is approximately 2%.

(3) *Gradient Boosting Machine*

The common ensemble techniques like random forests rely on simple averaging of models in the ensemble and the main idea of boosting is to add new models to the ensemble sequentially. The Gradient boosting machines established a connection with the statistical framework and a gradient-descent based formulation of boosting methods was derived (Natekin et al. 2013). Although the boosting process has the highest test error, it also gives us relative importance variables such as avg_glucose_level and age.

	var <chr>	**rel.inf** <dbl>
avg_glucose_level	avg_glucose_level	30.759126
age	age	29.341546
bmi	bmi	19.682147
smoking_status	smoking_status	5.496070
work_type	work_type	4.426816
hypertension	hypertension	2.955895
heart_disease	heart_disease	2.385362
gender	gender	1.878311
Residence_type	Residence_type	1.541541
ever_married	ever_married	1.533185

Table 1: the Gradient Boosting Machine

(4) *Logistic regression*

We used the logistic regression model to understand the relationship between the dependent variable and more independent variables by estimating probabilities. There is a summary of logistic regression we have:

```
Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -2.038e+01  5.911e+02  -0.034  0.97249
genderMale                 -1.766e-02  2.119e-01  -0.083  0.93361
age                         7.634e-02  8.762e-03   8.713  < 2e-16
hypertension1               4.977e-01  2.377e-01   2.094  0.03628
heart_disease1              4.239e-01  2.763e-01   1.534  0.12502
ever_marriedYes            -3.655e-01  3.008e-01  -1.215  0.22426
work_typeGovt_job           1.238e+01  5.911e+02   0.021  0.98328
work_typeNever_worked      -1.028e+00  3.009e+03   0.000  0.99973
work_typePrivate            1.235e+01  5.911e+02   0.021  0.98333
work_typeSelf-employed      1.174e+01  5.911e+02   0.020  0.98416
Residence_typeUrban         9.844e-02  2.094e-01   0.470  0.63825
avg_glucose_level           5.596e-03  1.778e-03   3.147  0.00165
bmi                         7.966e-03  1.670e-02   0.477  0.63340
smoking_statusnever smoked  1.489e-02  2.532e-01   0.059  0.95309
smoking_statussmokes        8.999e-02  3.301e-01   0.273  0.78512
smoking_statusUnknown      -5.994e-01  3.769e-01  -1.591  0.11172
```

Table 2: the logistic regression model

Based on the p-value each variable has, we found hypertension(hypertension1), age, and the average level of glucose (ave_glucose_level) are statistically significant at a 5% significance level, which indicates the potential variables that will contribute to the stroke.

(5) Logistic Ridge regression

Furthermore, our team wants to analyze multiple regression data that suffer from multicollinearity. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors. It also proved by our results that ridge regression has relatively small test errors.

(6) *SVM*

The support vector machines are supervised learning methods used for classification, regression, and outlier detection. It's very effective in high-dimensional spaces, and it's one of the most robust and accurate algorithms among other classification algorithms. We used this method in order to improve our test error. Although random forest improved our test error by 2%, we are still trying to find a better model to improve our results.

(7) *ROC curve*

ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. In general, it indicates the trade-off between sensitivity and specificity. We used the ROC curve to compare the models we selected by test errors.

3. Results and discussion

3.1 Results

By calculating the test errors for the models we have, we found that the Gradient Boosting Machine has the highest test errors while SVM has the lowest. In general, the random forest and bagging have very similar results. Similarly, the logistic regression is very close to Logistic Ridge Regression but slightly different. Here is a table for the test errors of the different models in descending order:

Method	Test error
Gradient Boosting Machine	0.04768
Bagging	0.04197
Random forest	0.04116
Logistic regression	0.03912
Logistic Ridge regression	0.03871
SVM	0.02037

***Table 3**: the comparison of different methods with the test errors*

By comparing the test errors, we selected the models with relatively small test errors and a control group of Gradient Boosting Machine for our next evaluation step. Here is our output by using the ROC curve:

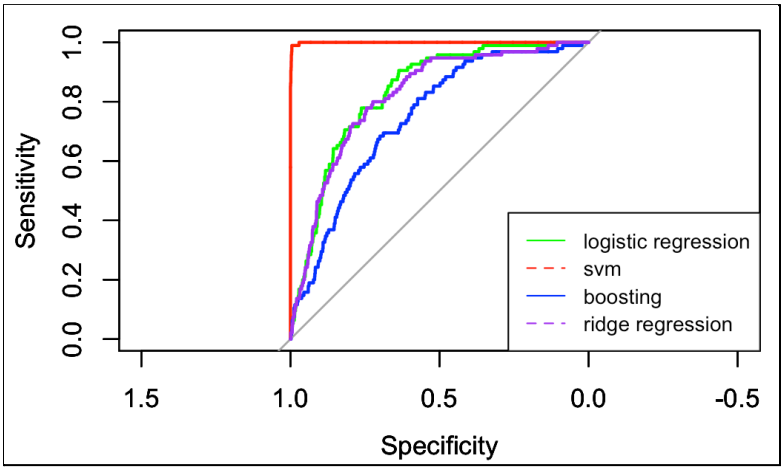

Figure 1: ROC curve

Methods	AUC
SVM	0.9994
Logistic regression	0.8333
Logistic Ridge regression	0.8202
Gradient Boosting Machine	0.7463

Table 4: the comparison of AUC

By comparing the area under the curve, we found the SVM has an overwhelming advantage over other models. At the same time, the Gradient Boosting Machine has the lowest AUC, and it's the same results as we got by comparing test errors in table 2.

3.2 Discussion

From Cheon's 2019 research paper, we find in their dataset, the SVM will be the worst method however in our project, the SVM behaves so well. The reasons for that could vary. Firstly, in their paper, they only have 4 variables as mean age, gender, mortality and stroke types which is relatively simple compared to the dataset we have. Secondly, ¾ of their variables are categorical and our dataset is the mixed of the numerical and categorical variables. Hence, we consider the different dataset may impact the results of selecting the optimal model. Furthermore, It's built on the test data and we used the radial function instead of the kernel function in the SVM.

4. Conclusion

Based on the logistic regression and Gradient Boosting Machine, age and average glucose level are the most critical factors contributing to a person having a stroke. Among these two factors, the effect of average glucose level is more apparent based on our Gradient Boosting Machine methods.

We found the optimal model as SVM to predict the stroke by combining the results of test errors and ROC curve.

References

1. "Stroke Prediction Dataset." n.d. Accessed December 1, 2021. https://kaggle.com/fedesoriano/stroke-prediction-dataset.
2. "Stroke Facts | Cdc.Gov." 2021. May 25, 2021. https://www.cdc.gov/stroke/facts.htm.
3. *Cheon,* Songhee, Jungyoon Kim, and Jihye Lim. 2019. "The Use of Deep Learning to Predict Stroke Patient Mortality." *International Journal of Environmental Research and Public Health* 16 (11): 1876. https://doi.org/10.3390/ijerph16111876.
4. Natekin, Alexey, and Alois Knoll. 2013. "Gradient Boosting Machines, a Tutorial." *Frontiers in Neurorobotics* 7: 21. https://doi.org/10.3389/fnbot.2013.00021.