Stroke Classification and Model Selection

jck2183 & lmt2203

1. Introduction

According to the World Health Organization (WHO), stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. This dataset is used to predict whether a patient is likely to get a stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

Data Source: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

All the features we had:

- 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

Note: "Unknown" in smoking_status means that the information is unavailable for this patient

The data is essentially clean and contain 12 variables. Excluding ID, we have 10 predictor variables and one binary outcome variable `stroke`. Categorical variables are converted from character to factor class in order for them to be included in the model and for the purpose of analysis. Binary predictor variables such as `hypertension`, `heart_disease`, and `ever_married` are recoded so that `0` means No and `1` means Yes. The `work_type` variable which has 5 levels (children, govt_job, never_worked, private, self_employed) are recoded to lower snake case.

We are mostly interested in training and comparing the performance of several predictive models on classification to see which one has the highest accuracy. Furthermore, we will also

look at variable importance and see which variables should be included in the final model. The models included in this report include logistic regression, linear discriminant analysis, classification trees, and random forests.

2. Exploratory Data Analysis

After importing data, we found that the imported dataset has 5110 observations in total. Excluding the id, we only have ten features and one binary outcome variable-stroke (0:no stroke, 1:stroke). We found that the stroke outcome distribution is highly imbalanced, with 4861 individuals had no stroke while 249 individuals suffered a stroke, which is only 4.87% of the total observations. (**Figure 1**)

We also looked at missing values across the dataset. There are 201 missing values for `bmi` variable. Among these 201 missing values in BMI, 40 observations have a stroke while 161 observations do not have a stroke. The distribution of BMI is right-skewed. To deal with the missing values, we then utilized preprocess imputation, knn Imputation, specifically, in the caret train function to address the imputation problem. We also have 1544 unknown values in smoking status and will treat those who answered unknown as a variable, thus there is no need to impute them.

In **Figure 2**, we show the proportion of strokes within each level of the factors. I also plot the distributions of the continuous predictors in **Figure 3**. From these plots, we can see that formerly smokers are more prone to having a stroke than smokers, this could be due to the fact former smokers quit smoking after having a stroke or after acquiring health conditions that heightened their risk of having a stroke. Self-employed individuals are also more likely to suffer a stroke than people with private or government jobs, maybe due to higher stress as a result of being self-employed. Urban residents, males and people with hypertension or heart disease are more prone to stroke than their counterparts. Age also seems to be an important factor in deciding whether one has a stroke or not, with higher age comes higher chance of having a stroke. Finally, people who had stroke tend to have higher glucose level than people who do not have a stroke.

To run our models, first, we have to convert character variables into factors to add them into our model and proceed with the analysis. Plus, we will also examine if there is any correlation among features. Meanwhile, we also found there is an observation who identified their gender as “Other.” We decide to omit this single subject so that we can proceed with our analysis.

Next, the characteristics of features will help us determine which model would be proper. As the outcome is binary, and the features are mixtures of continuous and categorical variables. We also have to decide how to partition the train and test data, which cross-validation method to use. Evaluation metrics should be used and set up an appropriate tuning grid corresponding to the tuning parameter.

We then used the imputed data to fit in the model to see which algorithm classifies the best. ROC or Kappa would be better evaluation metrics in this case because of the highly imbalance nature of the dataset.

3. Models

3.2 Logistic regression

Our logistic regression model assign class labels (either stroke or no stroke) based on estimated conditional class probabilities given the predictor variables. We also tried linear discriminant analysis (LDA) in the midterm project, however, the model's performance is not good as compared to logistic regression model and thus I decide to leave LDA out of the final report. The linear discriminant model would be more stable than the logistic regression model if the distribution of the predictors is approximately normal, which is not the case in this example. In addition, LDA performs better when we have more than two response classes.

After fitting a logistic regression model with train function in the `caret` package, we get a model with the most significant predictor (smallest p-value) is age, with a coefficient of 1.6449, followed by average glucose level with a coefficient of 0.14815. The positive coefficient suggests a positive association between age and average glucose level with higher chance of having a stroke. We used 0.5 as a simple predicted probability threshold for the classifier, meaning if the conditional probability is higher than 0.5, the observation is assigned to class label `1` or `stroke`.

K-nearest neighbor was also performed, which predicts the class label by identifying the observation that are nearest to it. For KNN, scaling of the variables matter, otherwise variables on a larger scale will have a larger effect on the distance between the observation. KNN assumes features are independent in each class. The accuracy of KNN model is slightly lower than our logistic model.

We fit a generalized additive model (GAM), which allow non-linearity of each of the variables while maintaining additivity so that we can examine the effect of each predictor on the outcome individually while holding all of the other variables fixed. The disadvantage of GAM is that model is restricted to be additive and important interactions can be missed.

3.2. Classification tree

For classification tree, classification error rate could be used as a criterion in making the binary splits, however, it is not sensitive enough for tree growing, and thus Gini index – a measure for node purity - is a more preferable alternative.

Here we fit a decision tree (CART) and a more elaborated version of CART, which is conditional inference tree (CIT). Although CART and CIT are pretty similar and usually treated as the same thing, CIT provide more accurate variable importance measures. Inference trees are non-parametric and thus not rely on distributional assumptions.

The most important predictor of stroke seems to be age since it was used to differentiate the first branch, followed by heart disease and average glucose level. Our classification trees have lower ROC, which is somewhat expected as trees generally do not have the same predictive power as other classification methods.

3.2.1 Random Forest Ensemble

Random forest is an extension of bootstrap aggregation (bagging) of decision trees and can be used for classification and regression problems. Because models in the ensemble are unpruned, making them slightly overfit to the training dataset. This is desirable as it helps to make each tree more different and have less correlated predictions or prediction errors. We included all the predictors into the model plotting the variable importance. As for assumption check, we found that Random Forest usually does not make any assumptions about the underlying distribution of your data and can implicitly handle collinearity in features.

There are two main tuning parameters for random forest classifier, including mtry: number of variables available for splitting at each tree node and min.node.size: the minimum number of data points in a node required for the node to be split further. We will use grid search and see if a pair of best parameters resides in the given grid.

According to the variable importance plot, average glucose level is the most important variable in predicting whether someone has a stroke, followed by BMI and then age.

3.3 SVM

Since we tried to classify two groups of the outcome, SVMs, which can be defined as linear classifiers when the margin is large enough, can potentially do a good job in this case.

Although Random Forest Ensemble has indicated several predictors play important roles, we still decided to process SVMs with all predictors included in the model for fear of not comparing two models' performance with all conditions holding constant.
Given that SVM is quite tolerant of input data, especially the soft-margin version. The only assumptions of support vector machines are independent and identically distributed data.

C is also known as the cost parameter, is the tuning parameter in the SVM algorithm. We used the exponentiated sequence of numbers within the cross-validation training section to see any best parameters for this model. Plotting out the tuning will also help us have a better visualization on deciding hyperparameters.

We ran both linear and radial models with smote sampling the imbalanced data. The training accuracy rate for linear SVM is 0.7571 with a balanced accuracy rate equals 0.76116, while the test accuracy rate is 0.7546 with the balanced accuracy rate equals 0.76202. The AUC of linear SVM equals 0.762.

The training accuracy rate for radial SVM is 0.9041 with a balanced accuracy rate equals 0.93062, while the test accuracy rate is 0.8401 with the balanced accuracy rate equals 0.51191. The AUC of radial SVM equals 0.5119101.

We can conclude that the linear SVM outperforms the radial SVM by comparing their AUC. It seems that the radial SVM model is too flexible, even though it gave a low training error rate

and an unbalanced test error rate. The AUC and balanced test accuracy rate suggested that radial SVM do a bad job in classifying two categories of the outcome.

4. Conclusion

Since there are only 10 predictors, we included all of them. If the data set was more complex and had more predictors, we would have used stepwise or different models or dimension reduction to narrow down the number of features included.

We evaluate the models' performance on the test data using `predict` function. It seems that the logistic regression models perform better than the rest with the highest ROC. This is probably due to the fact that we only have two classes of outcome and ten predictors, mostly categorical in the model. (**Figure 4)**

Although all the models have pretty similar ROC except for the classification tree model. Our models have low sensitivity, which is understandable given the highly imbalanced data. We could fix this by oversampling, however, for this particular analysis, I opted to evaluate normal sampling data to avoid biased prediction results. The imbalance nature of the data also affects the validity of accuracy levels as a method of evaluation. We did run in issue when trying to compare the models, that is the SVM models cannot be included because they were not trained using the caret package, and thus do not produce compatible results.

Regarding the variable importance based on the logistic model, age seems to be the most important predictor indicating whether someone has a stroke or not, followed by the average glucose level and hypertension.

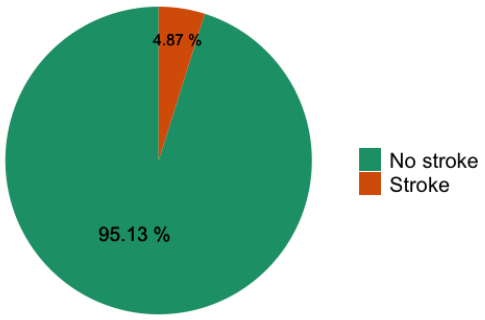


Figure 1. Distribution of the stroke outcome

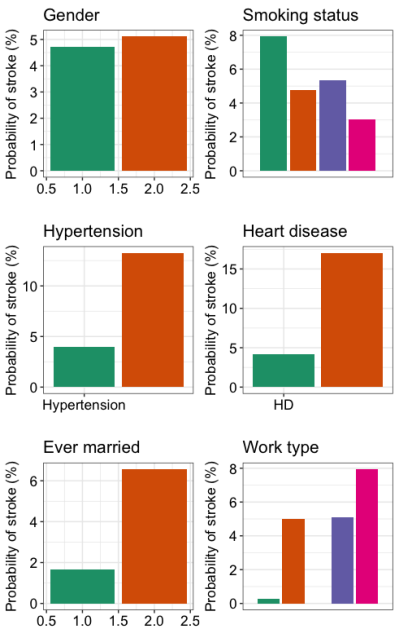


Figure 2. Probability of having a stroke in each category

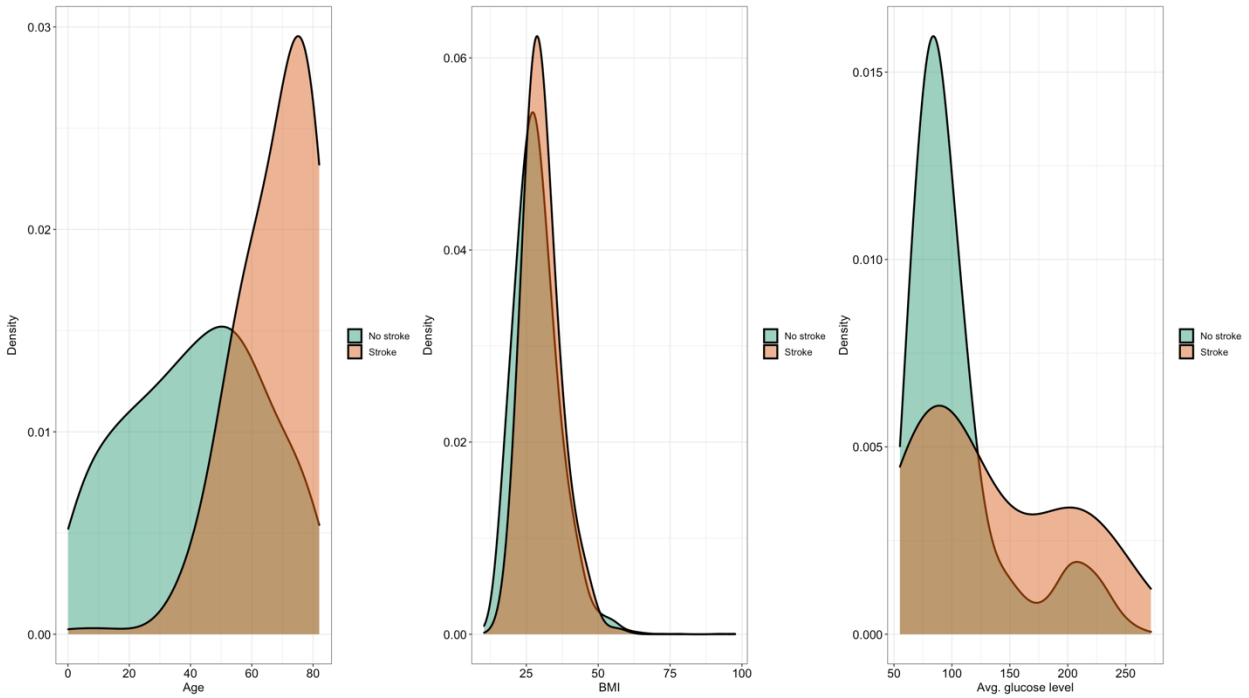


Figure 3. Distribution of continuous outcomes

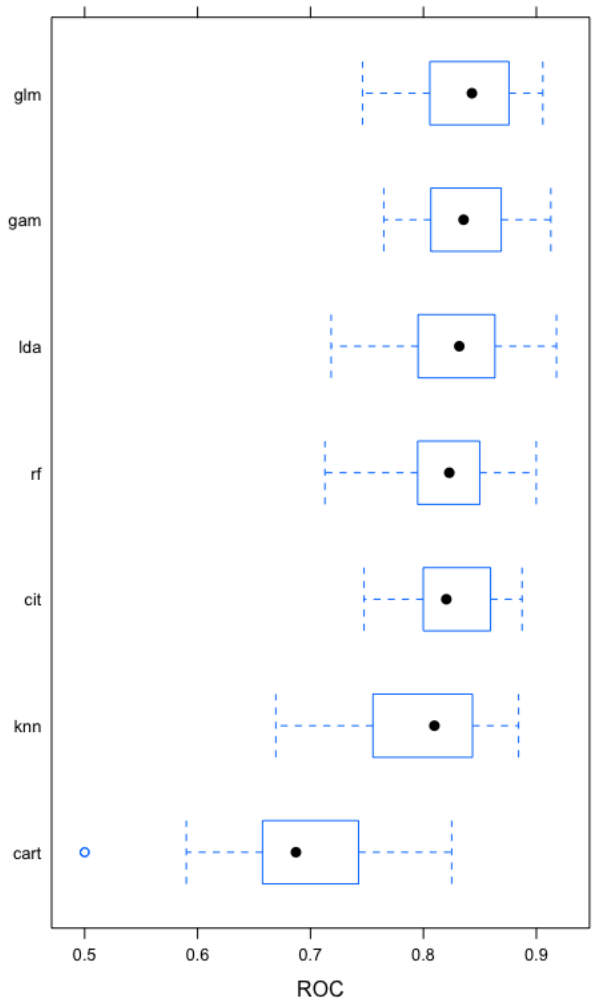
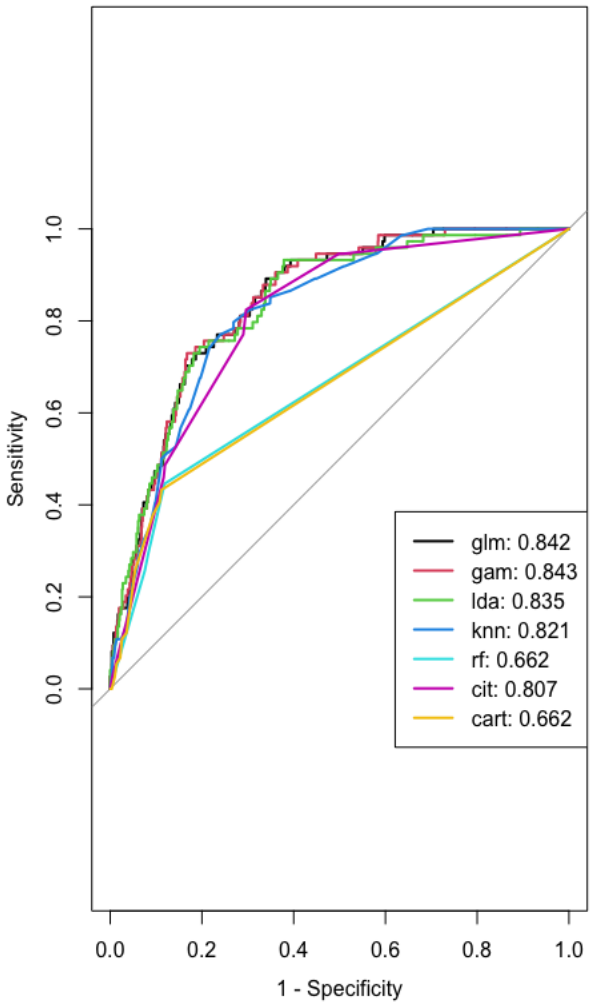


Figure 4. ROC


Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1 - Specificity
0.0
0.2
0.4
0.6
0.8
1.0
glm: 0.842
gam: 0.843
lda: 0.835
knn: 0.821
rf: 0.662
cit: 0.807
cart: 0.662