Data 1030 Project – Stroke Prediction

Yiwei Sang
Data Science Initiative, Brown University

https://github.com/yiweisang97/data1030_project

1 Introduction

1.1 Motivation

Stroke is a serious illness. It is a leading cause of death and long-term disability. It would be nice if we can predict whether an individual is likely to have a stroke and thus prevent stroke.

1.2 Dataset

I obtain the dataset from Kaggle. The dataset contains 43,400 observations and 12 columns. This is a classification problem. The target variable is 'stroke', a binary feature indicating whether an individual suffers from stroke or not. The dataset is imbalanced, only about 1.8 percent of the data is with label 1. The first column is 'id', which identifies unique individual. The other ten columns are ten features:

1. gender: Female, Male, Other
2. hypertension: 1 if the individual suffers from hypertension, 0 if the individual does not
3. heart_disease: 1 if the individual suffers from heart disease, 0 if the individual does not
4. ever_married: Yes, No
5. work_type: Private, Self-employed, Govt-job, children, Never_worked
6. Residence_type: Urban, Rural
7. smoking_status: never smoked, formerly smoked, smokes
8. age: age of patient in years

9. avg_glucose_level: average glucose level measured after meal, measured in mg/dL

10. bmi: body mass index, measured in kg/m^2

The first seven features are categorical and the last three are numerical.

2 EDA

I prepared stacked bar plot for each categorical feature, and prepared histogram and violin plot for each numerical feature. I prepared boxplot for 'age', grouped by the target variable 'stroke' and another categorical feature of interest, including 'hypertension', 'heart_disease', 'smoking_status' and 'ever_married'. All the graphs are in the figures section in the GitHub repository. In the report, I choose to demonstrate figures that I find interesting.

Figure 1 is a normalized histogram of feature 'bmi'. I marked the normal range of BMI with red lines. There is a peak around BMI of 30 for people who suffer from stroke. The normalized histogram of people who do not suffer from stroke and of people who do overlap as BMI becomes larger. For smaller values of BMI, the differences is relatively obvious: the proportion of people with normal range of BMI in 'not stroke' category is higher than that in 'stroke' category.

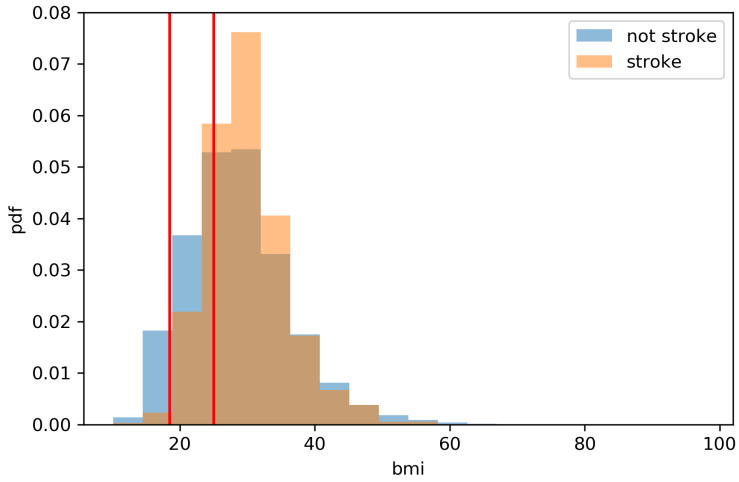

Figure 1: bmi - Histogram

Figure 2 is a normalized histogram of feature 'avg_glucose_level'. I marked the normal range of average glucose level with red lines. There are a great proportion of people with high average glucose level who suffer from stroke, remarkably higher than the proportion of people with high average glucose level

who do not suffer from stroke, as shown in the right part of the figure. While in the meantime, there are still a great proportion of people with normal average glucose level who suffer from stroke, though it is lower than the proportion of people with normal average glucose level who do not suffer from stroke.

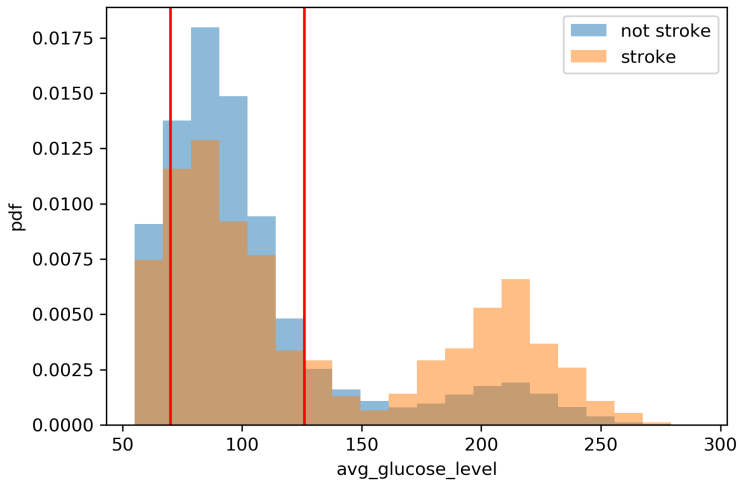

Figure 2: avg glucose level - Histogram

Figure 3 is a normalized histogram of feature 'age'. As age becomes larger, the proportion of people who suffer from stroke increase, the increase is sharp around age 78.

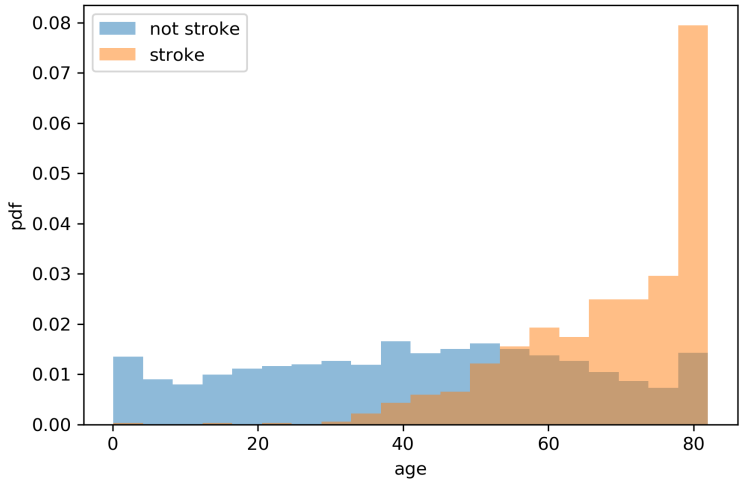

Figure 3: age - Histogram

Figure 4 is a boxplot of feature 'age', grouped by label - 'stroke' and 'smoking_status'. Among the group of people who suffer from stroke, the average age of people who smokes < the average age of people who formerly smoked < the average age of people who never smoked.

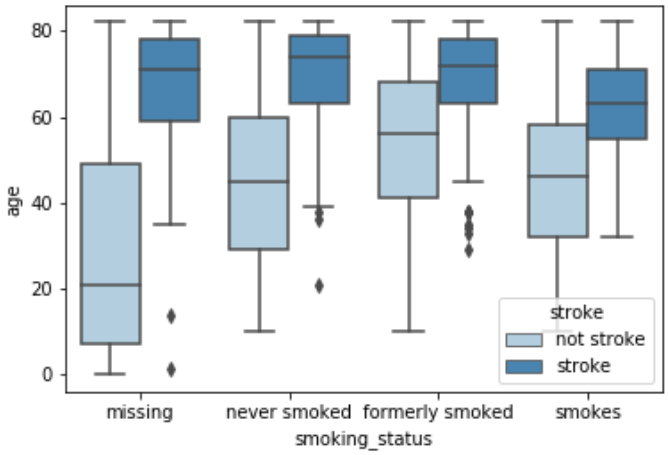

Figure 4: Interaction between smoking status and age with stroke

3 Methods

3.1 Handling missing values and Preprocessing

I first dropped column 'id' which identifies unique individual. It is not a contributing factor to stroke.

The dataset contains missing values in features 'bmi' and 'smoking_status'. I replace the missing values in 'smoking_status' with 'missing' as a separate category. I use IterativeImputer which uses RandomForestRegressor as estimator to impute the missing values in 'bmi'.

I use OneHotEncoder for all categorical features excluding 'hypertension' and 'heart_disease', which are already $0, 1$ valued. I use MinMaxScaler for all numerical features since they are all reasonably bounded. I did not preprocess the label, which is also already $0, 1$ valued.I did not preprocess the label, which is also $0, 1$ valued.

After preprocessing, the dataset still contains 43,400 observations and have 21 features.

3.2 Model Selection and Parameter Tuning

I used Logistic Regression with L1(Lasso) regularization, Random Forest Classfier and Support Vector Classification(SVC).

a. Logistic Regression: I tune the parameter alpha. I tried 20 values ranging from 10^{-5} to 10^{2}, both ends included, evenly spaced on log scale.

b. Random Forest Classfier: I tune the parameters max_depth and min_sample_split. The values I tried for max_depth range from 1 to 10, both ends included. The values I tried for min_sample_split range from 102 to 498, both ends included, incremented 3 each time.

c. SVC: I tune the parameters C and gamma. I tried 20 values for C ranging from 10^{-3} to 10^{5}, both ends included, evenly spaced on log scale. I tried 20 values for gamma ranging from 10^{-10} to 10^{3}, both ends included, evenly spaced on log scale.

I split 20 percent of data points into test set and then use stratified k-fold cross validation on the other 80 percent of the data points to tune parameters. Most best parameters tuned are not near the boundary of the values I tried.

3.3 Evaluation Metric

The candidates for evaluation metrics includes accuracy, precision, recall and f1-score. I use f1-score as evaluation metric.

Since the dataset is imbalanced, accuracy will not be an informative evaluation metric. Precision should be used when we do not factor false positive and recall should be used when we do not favor false negative. I think both situations are not favored. False positive: people who are wrongly diagnosed with stroke may need to receive unnecessary yet expensive treatment; false negative: people who are not correctly diagnosed as stroke may miss the best timing for stroke treatment. f1-score is the harmonic mean of precision and recall, which seeks balance between precision and recall. Thus I think f1-score would be a suitable evaluation metric.

3.4 Uncertainties

The project use random_state in data splitting, Random Forest Classifier and Iterative Imputer which uses Random Forest Regressor as estimator, which may introduce uncertainties to the models. Thus for each model, I use 10 random states ranging from 42^{0} to 42^{9}, both ends included, evenly spaced, and calculate the average and standard deviation of f1-score under different random states.

4 Results

4.1 Comparison to baseline model

The baseline model is one that predicts 1(stroke) for all data points. The f1-score of the baseline model is approximately 0.03544.

a. Logistic Regression: The average f1-score is 0.036 with standard deviation rather small. Logistic Regression improves about 1.6 percent on average, compared to the baseline model. The low standard deviation shows that the results are relatively stable.

b. Random Forest Classifier: The average f1-score is 0.038 with standard deviation 0.0014. Random Forest Classifier improves about 7.2 percent compared to the baseline model.

c. SVC: The average f1-score is 0.038 with standard deviation 0.0013. SVC improves about 7.2 percent on average, compared to the baseline model.

There is some improvement in f1-score compared to the baseline model.

4.2 Global Feature Importance

4.2.1 Logistic Regression

I choose random_state=294, and use the best parameters found in the process of k-folds cross validation (k-folds=5). I standardized the preprocessed data and plot the feature coefficients of logistic regression in descending order of absolute values. The figures are similar. All figures are in the figures section in GitHub repository. In the report, I present two figures.

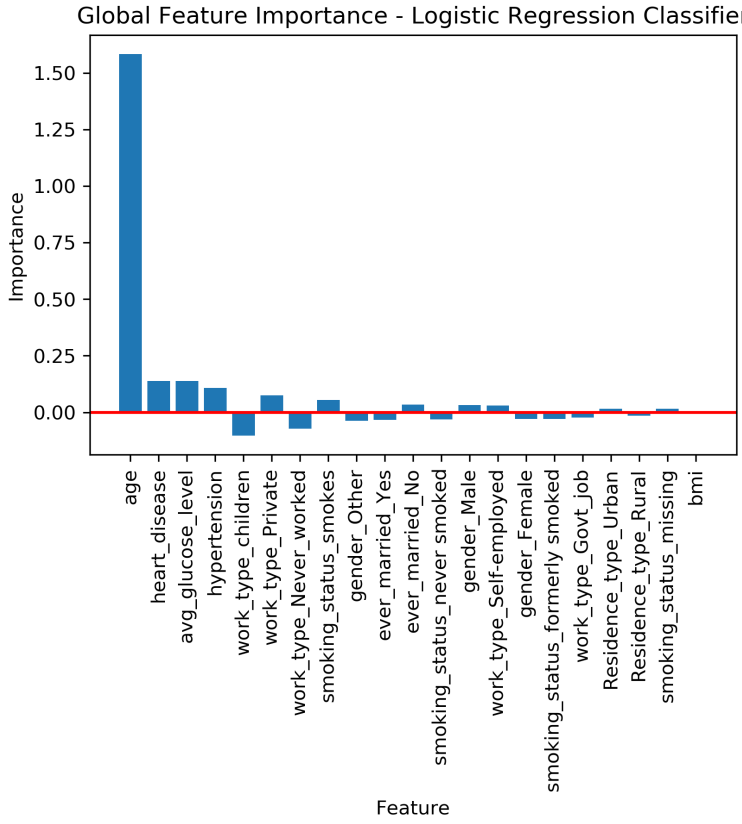

Figure 5: global feature importance - logistic regression 1

The coefficient of age is positive and its absolute value is remarkably higher than that of other features. Hypertension, heart_disease and average glucose level all have relatively high positive coefficients. work_type_children and work_type_never_worked have relatively high negative coefficients. Other than age, the internal rankings of absolute value of coefficients within the features mentioned above are slightly different.

a. Age is a rather important contributing factor to stroke. Young people are not likely to have stroke. As people get older, they are more likely to have stroke. This result makes sense since people's body functions decrease as their age increase. Elder people should be paid more attention as they are high-risk group for stroke.

b. Hypertension and heart_disease is a relatively important factor to stroke. People who suffer from hypertension or heart disease are more likely to have stroke. The result makes sense since it is a medically proven fact that hypertension and certain kind of heart disease may cause stroke.

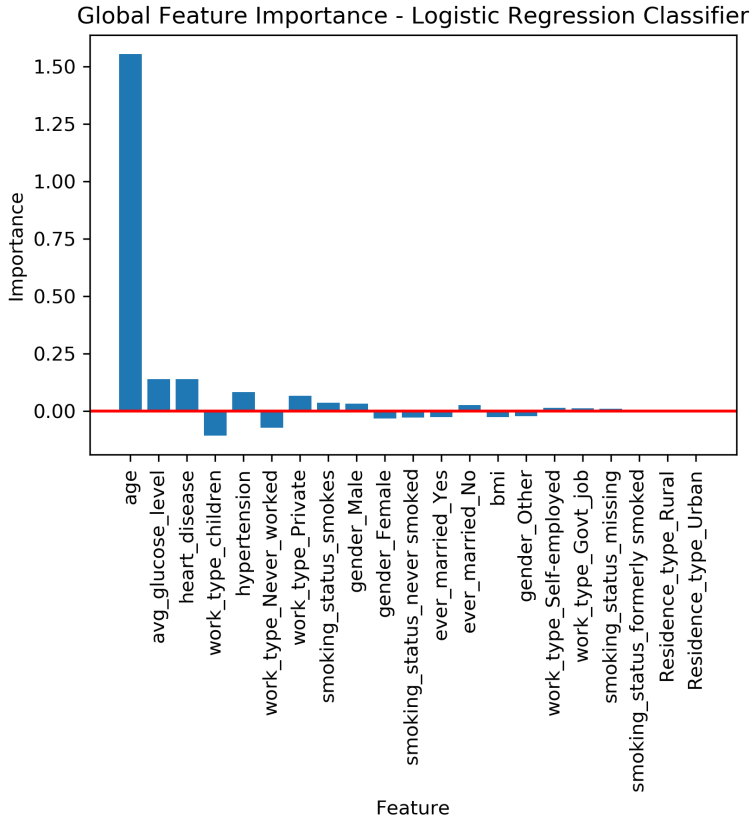

Figure 6: global feature importance - logistic regression 2

c. Average glucose level is a relatively important factor to stroke. People with high average glucose level are more likely to have stroke. The result makes sense since high level of average glucose normally indicates bad health. People can constantly monitor their average glucose level and take actions(for instance, see a doctor), trying to control their average glucose level in normal range.

d. work_type_children and work_type_never_worked are two relatively important negative factors to stroke, meaning when people's work type is children or never_worked, they are not likely to have stroke. I think such result is due to that children are all under age of 18, and people who never worked are usually younger. This result is a demonstration of the effect of age to stroke. In addition, another hypothesis to explain the result is that the stress level of children and people who never worked is normally low. However, in order to prove this hypothesis, further information about the stress levels of different work types is needed.

4.2.2 Random Forest Classification

I choose random_state=294, and use the best parameters found in the process of k-folds cross validation (k-folds=5). I plot the attribute feature_importances_ of Random Forest Classifier. The figures are similar. All figures are in the figures section in GitHub repository. In the report, I present one figure.

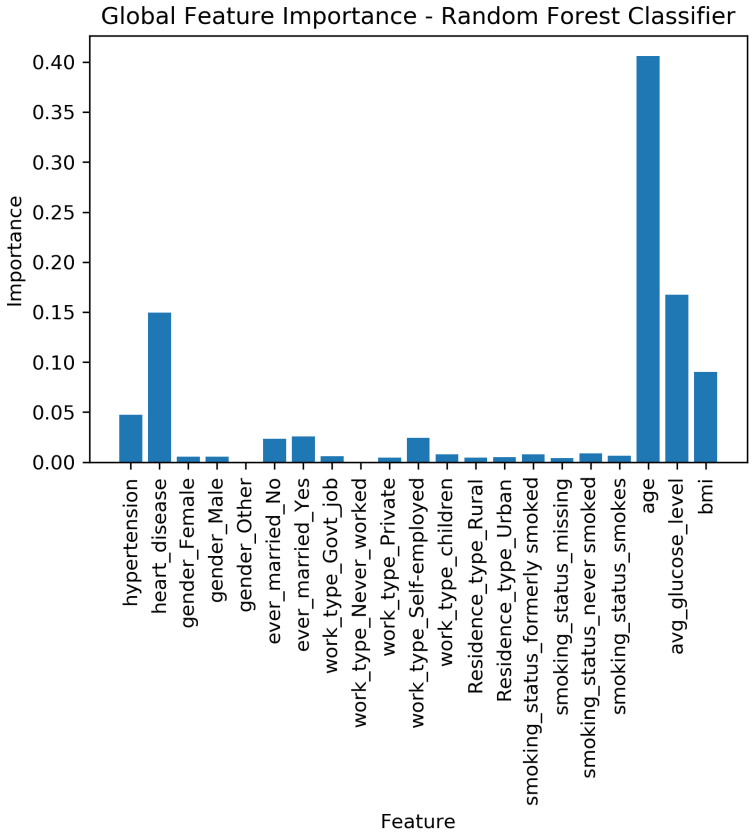

Figure 7: global feature importance - Random Forest Classifier

The feature importance of age is remarkably higher than that of other features. Other relatively important features are avg_glucose_level, heart_disease, bmi, hypertension(in descending order of feature importance, the order is the same for the five figures). The results are similar to the feature importance analysis for Logistic Regression except bmi, work_type_children and work_type_never_worked.

a. The feature importance of bmi is relatively high. The result makes sense to some extent. High BMI normally indicates bad health. However the feature importance of bmi in Logistic Regression is rather small. I think it might be that the difference between the normalized histogram of BMI

of two categories is relatively large for BMI in normal and below-normal range, and Random Forest Classifiers capture such difference.

b. The feature importances of work_type_children and work_type_never_worked are low. I think it might be that Random Forest Classifier is able to capture that these two features mostly reflect the importance of age, and Random Forest Classifier capture such relations.

5 Outlooks

One potential way that I may improve my model is to tune cutoff values for the three classification models.
I think the weakness of my model is that the number of features available are too small. Thus I think another potential way to improve my model is to collect more feature data. For instance, I want to collect data on number of years people have smoked, which should be more informative than 'smoking_status'. I also want to collect data on the type of heart disease of patients if the patients suffer from heart disease and other features about hearts(such as weight and cardiac diameters). I also want to collect data on stress levels of people.

6 References

a https://www.kaggle.com/asaumya/healthcare-dataset-stroke-data

b https://www.kaggle.com/asaumya/healthcare-problem-prediction-stroke-patients

c https://www.kaggle.com/tbourton/stroke-analysis-with-decision-trees