Irene Chen

Ming Qiu

Fred You

Prof. Eran Mukamel

COGS 109

June 4th 2021

Predicting Stroke Based on Patient Lifestyle and Health Information

Introduction

A stroke is when blood flow in the brain is blocked, leading to life-threatening brain cell death[1]. The occurrence of a stroke may further lead to other medical conditions, such as dementia[2]. In the United States alone, stroke is the leading cause of death, making it a very relevant topic to our health; thankfully, it is preventable. The Center for Disease and Control suggest a variety of healthy living habits to help prevent the chance of developing a stroke, implying the fact that there are certain factors which may play a part in the potential occurrence of a stroke [3]. As such, our goal is to investigate and predict whether an individual will have a stroke based on their gender, age, life style, and medical history.

Our intended dataset is one found on Kaggle [4]; it details the chance of an individual having a stroke, given certain features (eg. gender, age, BMI, etc). The dataset has 11 features and 5110 observations. The features are that concerning an individual's lifestyle and health: gender, age, whether the patient has underlying diseases (such as hypertension and heart disease), marital status, work status, residence, average glucose level, BMI, and smoking status.

---

[1]Stroke. (2021)
[2]Stroke: A global response is needed. (2017)
[3] Preventing Stroke: Healthy Living Habits. (2020)
[4] Here is the dataset: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

All of the features are categorical, except for average glucose level, age and BMI which are continuous.

Additionally, according to John Hopkins Medicine, risk factors for stroke include: stress, diabetes, heart disease, smoking, obesity and more[5]. Moreover, our hypothesis is as follows:

The occurrence of stroke can be better predicted by a combination of cardiovascular disease (heart diseases and blood pressure), BMI and smoking status.

Methods

We are focusing on prediction of stroke based on the features given in the dataset. We are not looking for an inference. In the dataset, the chance of stroke is classified as 1 or 0. As such, there is only the outcome of "Yes, a stroke has occurred" or "No, a stroke has not occurred". In addition, the majority of the predictors are categorical and binary. Hence, we have decided to use logistic regression over other classifiers like Linear Discriminant Analysis. Logistic regression is a statistical method used to classify data points. It functions as an extension of linear regression. In the case of our dataset, we will implement logistic regression to predict whether a stroke occurs (classified as 1), or not (classified as 0). Our decision boundary is one that satisfies as such $P(y = 1|x) = P(y = 0|x) = 0.5$. We performed feature selection using backward stepwise selection, and finally cross validation using k-fold cross validation. Detailed explanations on how we implemented both methods can be found below in *Model Selection*.

Results

Data Cleaning

The dataset contains 11 features, and 5110 observations. In our dataset, [chance of stroke] is classified as either 1 (the occurrence of a stroke) or 0 (no occurrence of a stroke). Due to the occurrence of 'NaN's in our dataset, we have decided to remove them during data cleaning. After

[5]Risk Factors for Stroke. (n.d.).

we removed all the rows containing at least one null value, we had a dataset with 4909 observations. Moreover, most of the features are categorical; as such, we decided to change them to numerical values. For example, in the original data, the feature 'Residence_type' contained values 'Urban' and 'Rural'. This column was broken down into two columns, 'Residence_type_Urban' and 'Residence_type_Rural', which is occupied with 0s for False (Not Urban/Not Rural) and 1s for (Urban/Rural). Only one column was kept as having both would be redundant. Additionally, some columns were renamed as the original dataset had naming conventions that were not uniform.

Model Selection

In order to prevent overfitting, we have implemented feature selection. We have specifically chosen to perform backward stepwise selection. After doing so, we performed cross validation on the models in an attempt to pick the best model out of them.

Backward Stepwise Selection

Due to the number of features being significantly less than the number of observations, backward stepwise selection is favored over forward stepwise selection[6]. By implementing backward stepwise selection, we can start off considering all of the features. In this particular implementation of backward selection, we selected features through the examination of p-values. A logistic regression model is computed with the given formula, and then p-values are observed. The feature with the highest p-value is removed from the formula, and the method is repeated. Once the largest p-value of the remaining features is less than the threshold value (chosen to be 0.05), backward stepwise selection is halted.

After performing backward stepwise selection, these were the models that were produced:

[6]Choueiry, G. (n.d.).

Model Number	Models
1	$stroke \sim \sim 1 + age + avg\_glucose\_level + bmi + gender\_Female + gender\_Male + gender\_Other + hypertension\_1 + heart\_disease\_1 + ever\_married\_Yes + work\_type\_Govt\_job + work\_type\_Never\_worked + work\_type\_Private + work\_type\_Self\_employed + work\_type\_children + Residence\_type\_Rural + smoking\_status\_Unknown + smoking\_status\_formerly\_smoked + smoking\_status\_never\_smoked + smoking\_status\_smokes$
2	$stroke \sim \sim 1 + avg\_glucose\_level + bmi + gender\_Female + gender\_Male + gender\_Other + hypertension\_1 + heart\_disease\_1 + ever\_married\_Yes + work\_type\_Govt\_job + work\_type\_Never\_worked + work\_type\_Private + work\_type\_Self\_employed + work\_type\_children + Residence\_type\_Rural + smoking\_status\_Unknown + smoking\_status\_formerly\_smoked + smoking\_status\_never\_smoked + smoking\_status\_smokes$
3	$stroke \sim \sim 1 + bmi + gender\_Female + gender\_Male + gender\_Other + hypertension\_1 + heart\_disease\_1 + ever\_married\_Yes + work\_type\_Govt\_job + work\_type\_Never\_worked + work\_type\_Private + work\_type\_Self\_employed + work\_type\_children + Residence\_type\_Rural + smoking\_status\_Unknown + smoking\_status\_formerly\_smoked + smoking\_status\_never\_smoked + smoking\_status\_smokes$
4	$stroke \sim \sim 1 + bmi + gender\_Female + gender\_Male + gender\_Other + hypertension\_1 + heart\_disease\_1 + ever\_married\_Yes + work\_type\_Govt\_job + work\_type\_Never\_worked + work\_type\_Private + work\_type\_Self\_employed + work\_type\_children + smoking\_status\_Unknown + smoking\_status\_formerly\_smoked + smoking\_status\_never\_smoked + smoking\_status\_smokes$
5	$stroke \sim \sim 1 + gender\_Female + gender\_Male + gender\_Other + hypertension\_1 + heart\_disease\_1 + ever\_married\_Yes + work\_type\_Govt\_job + work\_type\_Never\_worked + work\_type\_Private + work\_type\_Self\_employed + work\_type\_children + smoking\_status\_Unknown + smoking\_status\_formerly\_smoked + smoking\_status\_never\_smoked + smoking\_status\_smokes$
6	$stroke \sim \sim 1 + gender\_Male + gender\_Other + hypertension\_1 + heart\_disease\_1 + ever\_married\_Yes + work\_type\_Govt\_job + work\_type\_Never\_worked + work\_type\_Private + work\_type\_Self\_employed + work\_type\_children + smoking\_status\_Unknown + smoking\_status\_formerly\_smoked + smoking\_status\_never\_smoked + smoking\_status\_smokes$
7	$stroke \sim \sim 1 + gender\_Other + hypertension\_1 + heart\_disease\_1 + ever\_married\_Yes + work\_type\_Govt\_job + work\_type\_Never\_worked + work\_type\_Private + work\_type\_Self\_employed + work\_type\_children + smoking\_status\_Unknown + smoking\_status\_formerly\_smoked + smoking\_status\_never\_smoked + smoking\_status\_smokes$
8	$stroke \sim \sim 1 + hypertension\_1 + heart\_disease\_1 + ever\_married\_Yes + work\_type\_Govt\_job + work\_type\_Never\_worked + work\_type\_Private + work\_type\_Self\_employed + work\_type\_children + smoking\_status\_Unknown + smoking\_status\_formerly\_smoked + smoking\_status\_never\_smoked + smoking\_status\_smokes$

The features age, average glucose level, BMI, gender (female, male, other) as well as place of residence (Rural or Urban) were removed,according to their p-values. According to the CDC, cigarette smoking increases the likelihood of developing a stroke[7], aligning with the inclusion of smoking related features in our model.

While backward stepwise selection does not guarantee the best possible model, it does provide us with a valuable option to consider when trying to prevent overfitting. As there is no full search, this implementation gives the added advantage of being much computationally cost efficient in comparison to some other feature selection methods like best subset selection. We are only going through $O(p^2)$ models. Moreover, as we are not going through all possible models and doing a full search like in best subset selection, it is possible that backward stepwise selection will yield a model that is less likely to be influenced by noise (which leads to overfitting).

Furthermore, a model that includes most or all of the features is likely to have a lower train MSE than a model with less features. More complex models (the ones with more features) are less likely to have high training set MSE, but they are likely to lead to overfitting.

To select our final model and model parameters, we perform cross validation. Here we will be using K-Fold cross validation.

K-Fold Cross Validation

In model selection, there are several key concepts to consider; one of them is the idea of training error vs. testing error. If the testing error is much greater than the training error, then overfitting occurs. Additionally, one must also consider the tradeoffs between bias and variance. Higher variance will also lead to overfitting. To help prevent such a case, we perform cross validation. To compare the models we implemented with different flexibilities, we decided to use

[7] Preventing Stroke: Healthy Living Habits. (2020)

k-fold cross validation with a common k value of 5. k= 5 is common in practice which we believe to have a good balance between bias and variance. In comparison to LOOCV, another cross validation method, we are not only able to reduce bias by using most of the data in training, but also reduce variance as most of the data is used in testing.

We first randomize the order of each data point to ensure that each fold is not too biased, and split the data into 5 folds. While 4 (k-1) folds will be used to train the model and one remaining fold is withheld for the purpose of testing the model. We use the KFold package to implement this step by setting the parameter $k = 5$ (split into 5 subsets) and shuffle = True (randomize the order). The next step is to fit the model with the training set. In addition to the models obtained from backward stepwise selection, we will also create a model which depends on the features suggested in our hypothesis. In total, we are comparing 9 different models. For each fold, we fit the models with different feature combination and get the prediction based on the testing set; we then calculate the the test MSEs, which is the mean of $(actual\ test\ outcome - predicted\ outcome)^2$. The last step is to calculate and compare the mean of the 5 test MSEs.

```
mean of mMSE_1  0.04237103749239626
mean of mMSE_2  0.04237103749239626
mean of mMSE_3  0.04257470348017631
mean of mMSE_4  0.04257470348017631
mean of mMSE_5  0.04257470348017631
mean of mMSE_6  0.04257470348017631
mean of mMSE_7  0.04257470348017631
mean of mMSE_8  0.04257470348017631
mean of mMSE_9  0.04257470348017631
```

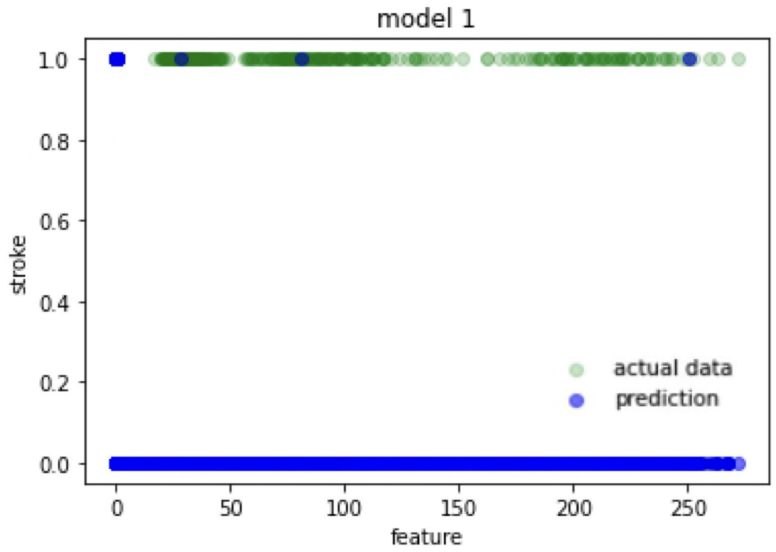
Fig. The mean test MSE of Models 1 to 9

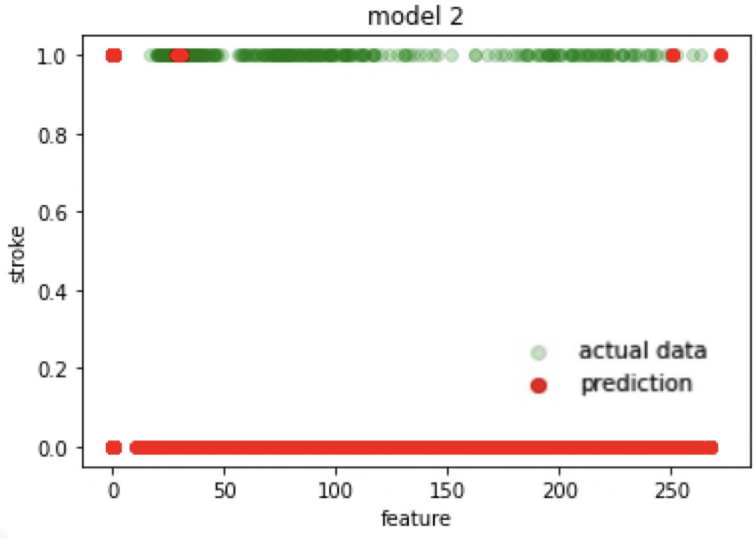
Note: Models 1 - 8 are models produced by backward stepwise selection, while model 9 is the hypothesis model

To prevent overfitting, we also calculated and stored the train MSEs, and the result is slightly smaller than the test MSEs, while the two models with the lowest test MSEs have lowest train MSEs as well. Given that the training and testing MSes are not too far apart, it implies that there is not a high chance of overfitting occurring in our models.

Model Estimation

From the mean MSEs across the 5-fold cross validation, we see that the models that have the lowest MSE of 0.04237 are the two shown in *Conclusion and Discussion* (model 1 and 2).





Hence, the best models out of the 9 are models 1 and 2. That is,

Model 1

$$stroke \sim \sim 1 + age + avg\_glucose\_level + bmi + gender\_Female + gender\_Male + gender\_Other + hypertension\_1 + heart\_disease\_1 + ever\_married\_Yes + work\_type\_Govt\_job + work\_type\_Never\_worked + work\_type\_Private + work\_type\_Self\_employed + work\_type\_children + Residence\_type\_Rural + smoking\_status\_Unknown + smoking\_status\_formerly\_smoked + smoking\_status\_never\_smoked + smoking\_status\_smokes$$

Model 2

$$stroke \sim \sim 1 + avg\_glucose\_level + bmi + gender\_Female + gender\_Male + gende hypertension\_1 + heart\_disease\_1 + ever\_married\_Yes + work\_type\_Govt\_job + work\_type\_Never\_worked + work\_type\_Private + work\_type\_Self\_employed + work\_type\_children + Residence\_type\_Rural + smoking\_status\_Unknown + smoking\_status\_formerly\_smoked + smoking\_status\_never\_smoked + smoking\_status\_smokes$$

Conclusion and Discussion

Through model selection, we can conclude that our best model is either the first or second model. Although heart disease, BMI and smoking status are both features that contribute to model 1 and model 2, from cross validation, our hypothesis was based on only the combination of these three types of features. Both model 1 and model2 had the lowest test MSE after performing k-folds cross validation with test MSE ~ 0.04237. It is interesting that our best models are the ones with most (or all) of the features. While it does lead to a low training MSE, these two models surprisingly also lead to the lowest testing MSE out of the 9 models. This suggests that perhaps there is less of an effect of overfitting. Moreover, the hypothesis model (model 9) was not the one with the lowest test MSE. As such, we can conclude that our hypothesis is not supported.

As for potential implications and next steps, one must acknowledge that our dataset is not very large. If future researchers were interested in the same topic, the collection of more data is

recommended. In doing so, we will be able to reduce variance and increase complexity of models without leading to overfitting. Furthermore, it is acknowledged that backward stepwise selection may not have led to the best model (or models in this case). Other feature selection methods, such as best subset selection or regularization methods, could be applied and compared using cross validation. Future researchers may also be interested in applying other classifiers such as Linear Discriminant Analysis or KNN Classification. Future researchers may also be interested in combining this data set with another one concerning other diseases attributed to stroke, like Dementia. Life-threatening diseases will always benefit from further study, and such research could help create advances in the healthcare industry.

Works Cited

Dataset: Retrieved from https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

Choueiry, G. (n.d.). Stepwise Selection. Retrieved from https://quantifyinghealth.com/stepwise-selection/

Preventing Stroke: Healthy Living Habits. (2020, January 31). Retrieved from https://www.cdc.gov/stroke/healthy_living.htm

Risk Factors for Stroke. (n.d.). Retrieved from https://www.hopkinsmedicine.org/health/conditions-and-diseases/stroke/risk-factors-for-stroke

Stroke. (2021, February 09). Retrieved from https://www.mayoclinic.org/diseases-conditions/stroke/symptoms-causes/syc-20350113

Stroke: A global response is needed. (2017, December 08). Retrieved from https://www.who.int/bulletin/volumes/94/9/16-181636/en/