COGS109 Final Project Report

Group 15

Meiwen (Cecilia) Liu

Wenqian Zhao

Muchan Li

University of California, San Diego

COGS109

Instructor: Eran A Mukamel

6/4/2021

Introduction

According to the World Health Organization, as the second largest “killer”, stroke was responsible for the deaths of 814000 people (WTO, 2020). Additionally, its causing factors are very complicated; a research by John Hopkins Medicine has shown that smoking status, overweight, high blood pressure, and heart disease are the dominant risk factors of stroke. Therefore, our group thinks that clearly identifying the potential red flags that cause strokes, thereby accurately predicting the probability of getting a stroke in advance so that medical treatments and precautionary care can be prepared in time is crucial.

We will use the Stroke Prediction Dataset from Kaggle (https://www.kaggle.com/fedesoriano/stroke-prediction-dataset) to decompose and reasonably predict patients’ likelyhood to get strokes based on various physical and social factors by implementing logistic regression and cross-validation. This dataset contains 5110 observations and 10 predictors, which is sufficient for us to construct a simple logistic regression model.

With regard to the dataset, we recognized several assumptions about it, which are in fact, basically common sense. For instance, we have found strong correlations between smoking status or heart diseases and chance of getting a stroke. Meanwhile, we have discovered that social factors such as work types--more precisely--employment status as it is portrayed in this dataset, and marriage status do not contribute significantly to our prediction, which matches our common sense as well.

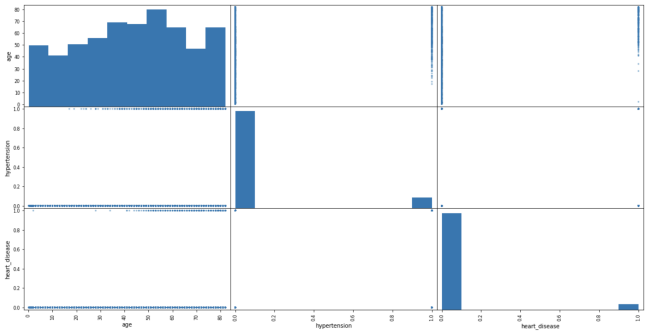
Ultimately, we hope our findings can be clinically useful and informative for people who have rare medical knowledge to understand their risk of getting strokes.

Hypothesis

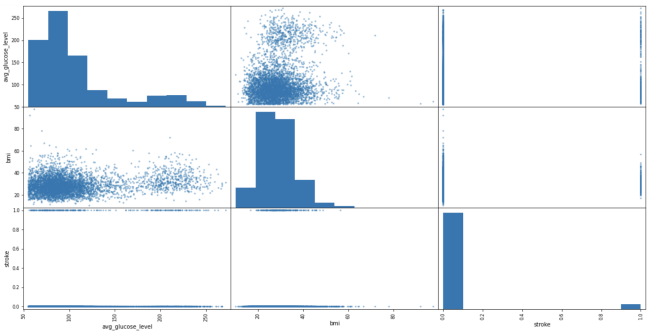
We hypothesize that the most accurate model that predicts the onset of stroke should incorporate the following predictors: age, heart disease, BMI, hypertension, and smoking.

Methods

Since the center focus of this project is to predict one's chance of getting a stroke on top of various predictor variables, the one method that we implemented throughout to build the classifier is logistic regression, which is one of the most common and suitable classification methods when handling multiple predictors and outputting relatively accurate binary prediction results. Before proceeding on to training our logistic classifier, we first checked the validity of



deploying logistic regression by screening our dataset and showing that the predictors do not exhibit multicollinearity, as it is depicted in the two diagrams. Absence of multicollinearity is a key assumption to performing logistic regression.



Since this is the only classification model we relied on, our group put much effort into model selection by implementing train-test cross validation (reducing bias) and hyperparameter tuning (promoting sensitivity and specificity) in order to obtain a classifier that has the highest f-1 score while maintaining a significant interpretability and predictive accuracy. As the dataset contains multiple binary or ternary variables, we first convert them as dummy variables, and discard useless, insignificant, and null predictors or data points before processing it. In addition, in order to acquire the most practical classifier, we also performed oversampling and downsampling to this extremely imbalanced data set; and this balancing indeed makes our model specificity and sensitivity higher, but with a prediction accuracy reduction trade-off.

Results

Data Overview and Processing

The table on the top right is the raw dataset that we directly downloaded and imported from Kaggle. This dataset contains 5110 observations with 10 predictors including 201 missing observations from the "bmi" column. The first process was to handle the unavailable data points. We chose to directly drop these observations to simplify the model building progress. In addition, as it is noted by the publisher on Kaggle, the "unknown" entries in "smoking status" column suggest that the information was unavailable, i.e. NaN, from the patient, therefore, our group decided to discard these unknown data from our data set as well. After that, we assigned all categorical predictors to dummy variables and eliminated one useless predictor "gender_type_other" (for the reason that there is only one "True" in this column, which leads to perfect separation error during model fitting). Afterward, we found that this dataset is extremely unbalanced such that it contains 3246 observations of non-stroke patients and only 180 stroke patients; therefore, we decided to perform downsampling and oversampling to our dataset and were left with 2163 non-stroke patients and 1298 stroke patients in our new sample, with much reduced bias in comparison to our imbalanced, original data sample.

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
...	...	...	...	...	...	...	...	...	...	...	...
18234	Female	80.0	1	0	Yes	Private	Urban	83.75	NaN	never smoked	0
44873	Female	81.0	0	0	Yes	Self-employed	Urban	125.20	40.0	never smoked	0
19723	Female	35.0	0	0	Yes	Self-employed	Rural	82.99	30.6	never smoked	0
37544	Male	51.0	0	0	Yes	Private	Rural	166.29	25.6	formerly smoked	0
44679	Female	44.0	0	0	Yes	Govt_job	Urban	85.28	26.2	Unknown	0

5110 rows × 11 columns

Model Construction and Selection

Unlike many other regression models, we have 2 essential parameters that we must tune in order to reach the best possible model: one is the the number of predictors, which associates with model complexity and over\underfitting issue; another one is the Bayesian decision threshold, which will determine the sensitivity as well as the specificity of our model. Therefore, our group

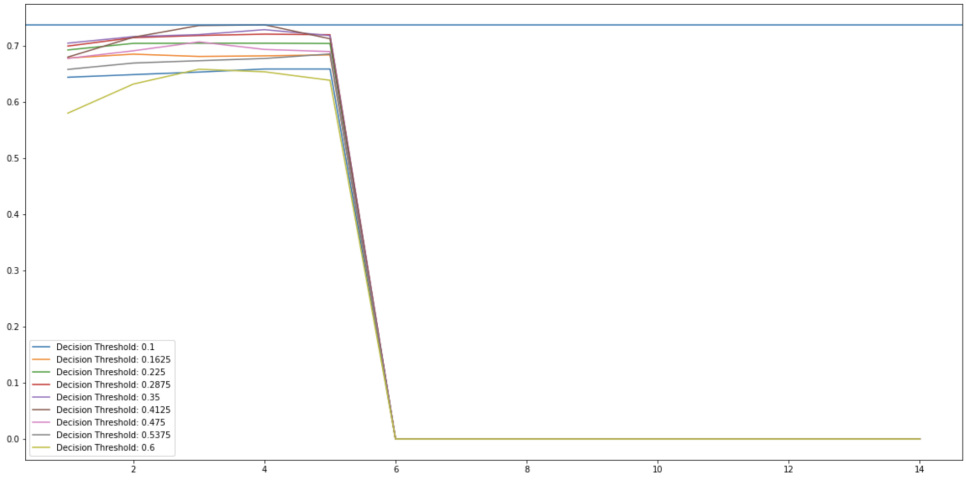
performed stepwise model selections to ensure we would be able to arrive at the closest-to-best model with highest runtime efficiency. During the model fitting process, we first split and reserved ⅔ data for training and ⅓ for testing, which is an essential step of cross validation since we are testing models with multiple complexities at different steps, and cross validation can prevent us from overfitting a model or fitting a model with high bias. Then, in order to maintain high interpretability and statistical significance, we drop the models with any p-value that is greater or equal to 0.05, which we recognized should follow along the conventional alpha threshold that determines the statistical significance. Moreover, we tested the f-1 score performance of each model with 9 different Bayesian decision thresholds that are uniformly distributed over the [0.1, 0.6] interval--whose upper bound, as we figured, should be enough to demonstrate the idea that a prediction is more likely to be a stroke than not at 60%--in order to find out the best decision threshold in addition to the best model we could obtain.

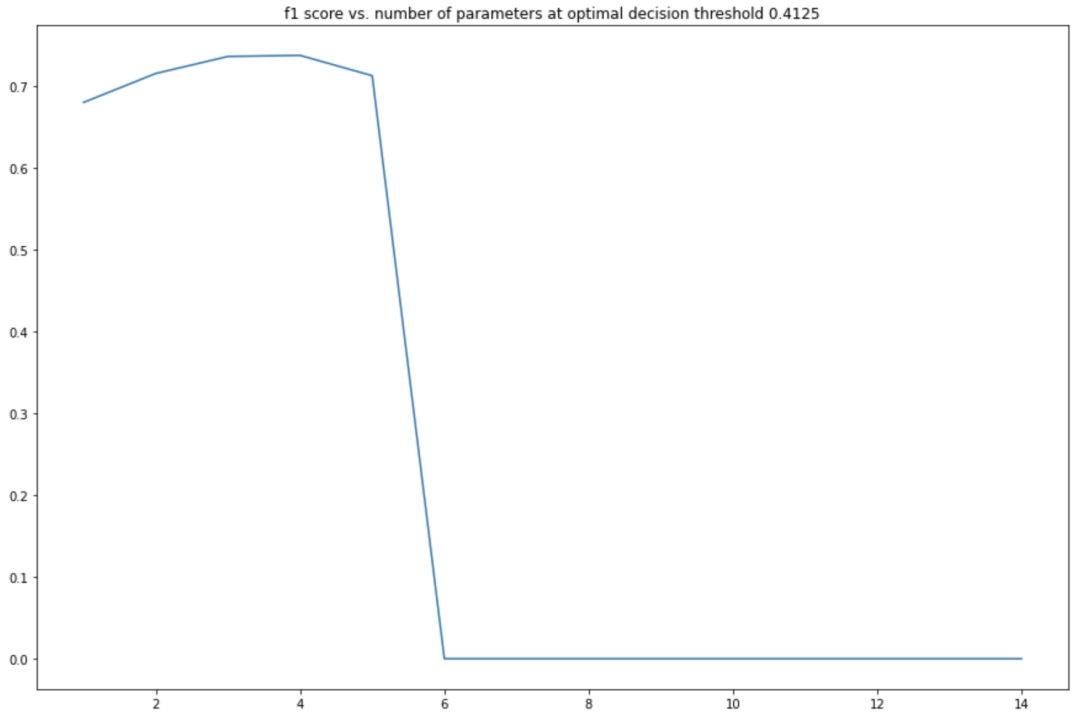
This table is the final model competition table. The first column represents the 9 different Bayesian decision thresholds. The rest of the columns each represent the best f-1 scores of models with the corresponding number of features as noted by the column name, trained under different decision thresholds. Interestingly, any model with more than 5 predictors is invalid due to large p-values, which suggests that many of the predictors are not statistically informative, which is consistent with our assumptions since we expected the power of predictors like work type and marriage status to be not that highly associated with stroke. According to the model competition table, the one model that stands out is the one with 0.4125 decision

	threshold	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	0.1000	0.6438	0.6485	0.6530	0.6585	0.6585	0	0	0	0	0	0	0	0	0
1	0.1625	0.6780	0.6849	0.6808	0.6819	0.6836	0	0	0	0	0	0	0	0	0
2	0.2250	0.6924	0.7041	0.7044	0.7044	0.7039	0	0	0	0	0	0	0	0	0
3	0.2875	0.6993	0.7143	0.7181	0.7205	0.7195	0	0	0	0	0	0	0	0	0
4	0.3500	0.7045	0.7161	0.7197	0.7285	0.7179	0	0	0	0	0	0	0	0	0
5	0.4125	0.6797	0.7150	0.7358	0.7371	0.7123	0	0	0	0	0	0	0	0	0
6	0.4750	0.6773	0.6909	0.7067	0.6933	0.6895	0	0	0	0	0	0	0	0	0
7	0.5375	0.6578	0.6691	0.6732	0.6772	0.6852	0	0	0	0	0	0	0	0	0
8	0.6000	0.5799	0.6317	0.6581	0.6535	0.6384	0	0	0	0	0	0	0	0	0

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-5.7543	0.269	-21.408	0.000	-6.281	-5.227
age	0.0748	0.004	19.299	0.000	0.067	0.082
smoking_status_smokes	0.3908	0.128	3.052	0.002	0.140	0.642
avg_glucose_level	0.0054	0.001	5.900	0.000	0.004	0.007
heart_disease	0.3245	0.157	2.066	0.039	0.017	0.632

threshold and 4 predictors (the two figures below further visualize the selection and comparison process); if we pull out the predictors that we pre-saved during the iteration, the four corresponded predictors are age, smoking status, average glucose level, and heart disease. The figure on the left of the previous page is the model summary.





The f-1 scores of all our models share the same trend that it would first increase as the complexity increases, then plateau when the complexity gets too high. This trend is more clearly presented by the graph above about models at different complexity with respect to a single f1 score. The reason why the curve suddenly drops to 0 at complexity equal to 6 is because in our

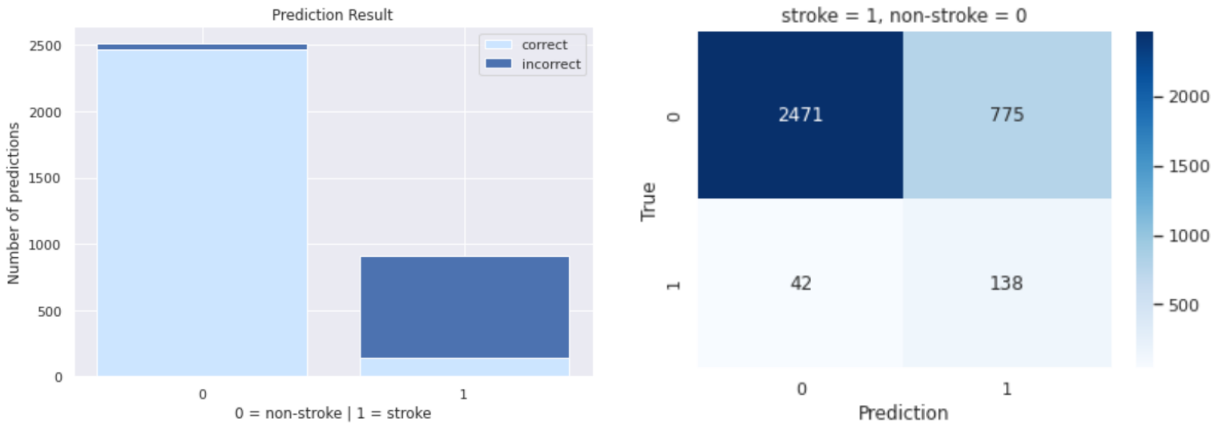
design of stepwise parameter selection, we made it so that when a trained model has one parameter whose p-value exceeds the 0.05 limit, it is automatically discarded without further operation. Then, at complexity of 6, the model has already had one of its parameters being insignificant, and the models of complexity higher than 6 do so as well, their f1 scores are consequently all set to 0. We believe this implementation, along with the observable flattening pattern in f1 score curves, should serve as enough evidence and support of preventing overfitting.

Model Estimation

Gathering all the information we have, we can construct the classifier in the following way--if P is greater than 0.4125, we will say this person is highly likely to get stroke, where P is given by:

$$P = \frac{e^{-5.7543 + 0.0748 * age + 0.3908 * smoke + 0.0054 * glucose\_level + 0.3245 * heart\_disease}}{1 + e^{-5.7543 + 0.0748 * age + 0.3908 * smoke + 0.0054 * glucose\_level + 0.3245 * heart\_disease}}$$

* "smoke" and "heart_disease" will be replaced with 1 or 0: 1 = smoke/has heart_disease



By fitting the model back to the dataset before oversampling and downsampling, we get 76.153% accuracy. Based on the above bar plot and confusion matrix, it is not hard to find that most of the mistakes are from false-positives (false-positive rate: 23.9%), in which we incorrectly marked non-stroke patients as potential stroke patients. This is caused by the imbalanced dataset, the specific reasons will be discussed in following sections.

Discussion

Model/Result Interpretation

As indicated in the summary data of our best model, all predictor parameters have positive coefficients with statistically significant p-values. This suggests that when fixing all other predictors constant, as we increase, say the "age" predictor, by 1 year, the classifier will output 5.3747e-4 increased likelihood in getting a stroke. For "avg glucose level", per unit increase in average glucose level leads to 3.7488e-5 increase in stroke likelihood; for "smoking", if the patient is a non-smoker, the likelihood of stroke decreases by 2.2448e-3; for "heart disease", if the patient hasn't previously had heart conditions, the likelihood decreases by 1.9224e-3. All values are calculated using the P expression outlined above.

In response to our initial hypothesis, predictors "age", "heart disease", and "smoking" indeed have significant effects on predicting stroke. However, "bmi" and "hypertension" turn out not to be significant contributors to the best model, and they are replaced with the predictor "avg glucose level".

Moreover, besides just stroke, the four factors are in general red flag indicators of all types of diseases. For instance, "age" is highly associated with neurodegenerative diseases like Parkinson's, Alzheimer's, etc. "avg glucose level" is in close relation with hyperglycemia, which leads to complications such as cardiovascular diseases, kidney failures, etc. "heart disease" on its own is already daunting. Finally, "smoking", as to our common sense, leads to chronic pulmonary diseases and at worst, lung cancers.

Hence, we would like the viewers and consumers of the information in this report to be especially alerted if he or she has any combinations of the four red flag factors. He or she should prioritize his or her health and well-being at once.

Poor Model Specificity

As mentioned in the "Data Overview and Processing" portion of "Results", our original dataset contains 3246 observations of non-stroke patients and only 180 observations of stroke patients.

This means that, regardless of the predictor variables, if our classifier were just to make predictions by rule of the majority--predicting all patients to be non-stroke patients--it would still achieve a prediction accuracy of 94.75%. However, such a classifier's ability to detect true-positive stroke patients isn't ideal at all (it is at 0% since none of the stroke patients is detected) since it just assumes all patients to be the opposite for simplicity (in analogy, the decision threshold is too high for a prediction to surmount). This contradicts the purpose that we would like our classifier to fulfill. Therefore, we had to "lower" the decision threshold so that some prediction values are now high enough to be cast into the true-stroke class. As the decision threshold is continuously reduced, especially when it is shrunk down to the lower range of the [0.1, 0.6] threshold interval, it becomes inevitable that while true-positive stroke predictions are finally be identified with more ease, lowering the bars simultaneously let the non-stroke predictions pass through, merely because their low likelihoods are no longer "low enough" to be considered non-stroke, thus marking them falsely positive. Given that the initial dataset is so imbalanced, and it still is even after downsampling and oversampling, it is nearly impossible to level off the trade-off between sensitivity, the ability to detect true-positive, and specificity, the ability to detect true-negative results. Nonetheless, since detecting true-positive stroke patients is a matter of greater urgency, and that the issues with specificity and dataset imbalance can be settled with the addition of more "stroke" data, this compromise is necessary.

Data Processing and Model Selection

Another reason that leads to the model being imperfect is the preprocessing that our group has performed on the raw dataset. First we dropped over 1000 thousands observations which reduces the informativeness of our train and test set, especially when the raw dataset was already quite unbalanced. Additionally, the oversampling and downsampling method that we performed may result in overfitting in the cross-validation process since we performed oversampling and downsampling before splitting the data into train and test set. Moreover, the weight of minority class, in this case the stroke cases, are amplified while the effectiveness of non-stroke cases decreased may also lead to higher bias in our result. However, in disease-detecting cases, I would argue that people are usually much more tolerant to false-positives than false-negative since no one wants to be diagnosed as healthy while being actually unhealthy.

Although we have plotted the diagrams between each variable, without performing actual linearity tests, we cannot confidently conclude that a collinear relationship does not exist at all. For example, smoking is undoubtedly a conjoint factor to heart disease and hypertension, but since they are binary variables, the linearity may not be present through plots, as well as age and heart disease and hypertension. Therefore, logistic regression may not be able to reach its best performance. We could have applied more models such as KNN and random forest, but due to time issues and the memory limitations of our environment, logistic regression became the only possible modeling approach.

Furthermore, cross-validation causes higher variance, and we did not choose to perform K-Fold due to the slow calculation time. Also, stepwise selection cannot guarantee to find the best subset. We have tried to perform the best subset selection, it took more than 30 minutes to get a result even though we have reduced the sample size and attempted to train ⅔ data. Consequently, our classifier may consist of high variance and small bias, which may serve as essential factors in terms of the high false-positive rates.

Future Outlook

If we have the opportunity to elaborate on this project more in depth, our group has these two additions in mind:

1. We would like to, as mentioned by the end of our discussion, introduce more “stroke” data and merge them with the existing dataset so that it becomes truly-balanced in nature. As for the source of our “stroke” data, we plan on contacting and cooperating with Jacobs Medical Center, if possible.
2. We would like to perform a more thorough model selection on the complete dataset using the best-subset selection method. This method in theory should lead us to the true best model of all models, and since it is then trained and validated on the complete, balanced dataset, its predictive accuracy will be more promising.
3. We will attempt more models with approved assumptions in order to find the best classifier among more models. Theoretically, logistic regression could outperform other models, but no matter what, we need more support to prove the validity of our predictor.

Potential implication for researchers interested in this topic is that they can focus more on the four features we include in our best model if they would like to elaborate on this project, and they may consider increasing the specificity while maintaining a high sensitivity if they are trying to build a more accurate prediction model.

Reference

Hopkins Medicine. (2021). *Risk Factors for Stroke*. Johns Hopkins Medicine. https://www.hopkinsmedicine.org/health/conditions-and-diseases/stroke/risk-factors-for-stroke

Khan, N., & Ali, A. (2010). Improving the speaking ability in English: The students' perspective. *Procedia Social and Behavioral Science*, *2*(2), 3575–3579. https://doi.org/10.1016/j.sbspro.2010.03.554