Stroke and Health Status

Sagar Bansal, Haochen Jiang, Liuqing Wang, Snow Wang

Agenda

- Introduction
- Data Characteristics
- Supervised Learning
  - Logistics Regression
- Unsupervised Learning
  - Cluster Analysis
- Discoveries & Limitations

Introduction

Introduction

Motivation of Analysis

- Stroke
  - Blood vessel blockage causes insufficiency of blood supply to the brain
  - **5th** cause of death and leading cause of disability in the U.S.
  - **140,000** deaths per year, **40** patients per second

Project Goals

- Examine relationships
- Perform inferential statements
- Interpretation
- Find relationships between observations

Data Characteristics

Data Description

Original dataset

- Observational & cross-sectional
- 12 variables with 43400 observations

Dataset after cleaning

- **7** variables were chosen based on our primary interest
- **29072** observations

Basic statistics

Qualitative Variables

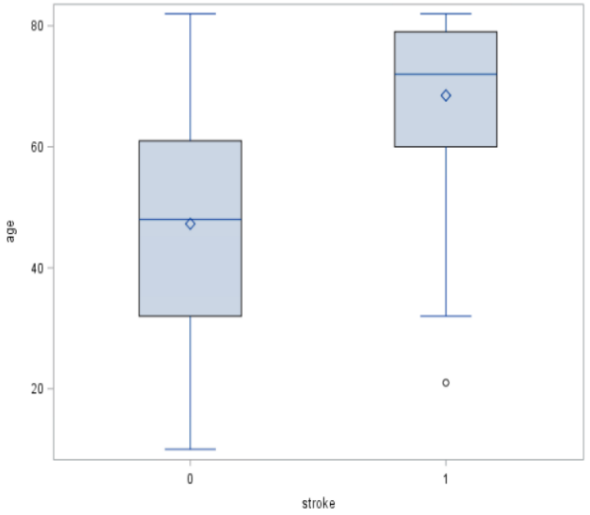
Variable	Description	Percent of Sample Data
STROKE	1 if an observation has had a stroke experience	1.88%
	0 if an observation has never had an stroke experience	98.11%
HBP	YES if an observation has hypertension	11.15%
	NO if an observation does not have hypertension	88.85%
HD	YES if an observation has heart disease	5.21%
	NO if an observation does not have heart disease	94.78%
SMOKING_STATUS	NEVER SMOKED if an observation has never smoked	54.16%
	FORMERLY SMOKED if an observation smoked before but has quited now	24.42%
	SMOKES if an observation is currently smoking	21.42%

Quantatitive Variables

Variable	Description	Mean of Sample Data	Minimum of Sample Data	Maximum of Sample Data
AGE	Age of an observation	47.67	10	82
AVG_GLUCOSE_LEVEL	Average glucose level of an observation	106.40	55.01	291.05
BMI	Body Mass Index of an observation	30.05	10.1	92

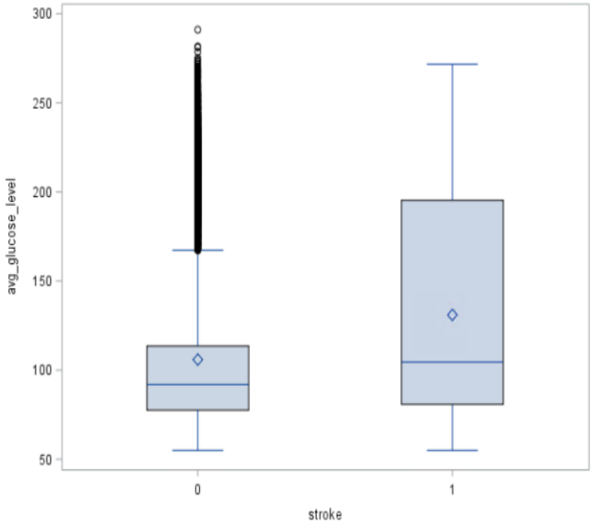
- Observed an imbalance issue but it is an accurate description of reality

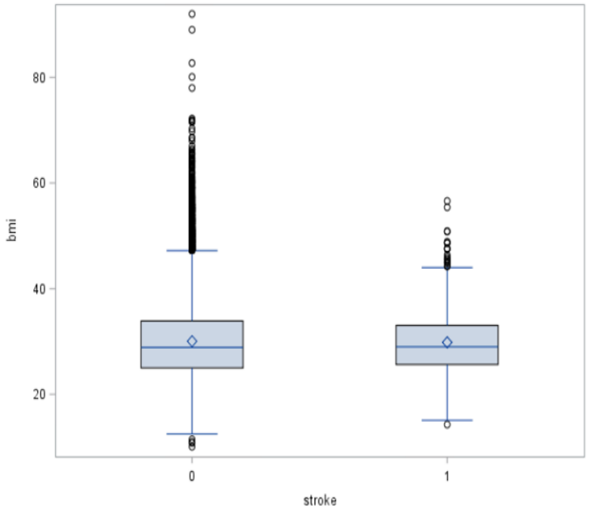
Initial Observations



Age

Average Glucose Level





bmi

Initial Investigation

The FREQ Procedure

Table of stroke by HBP

Frequency Percent Row Pct Col Pct	HBP		
stroke	no	yes	Total
0	25442 87.51 89.20 98.49	3082 10.60 10.80 95.09	28524 98.12
1	389 1.34 70.99 1.51	159 0.55 29.01 4.91	548 1.88
Total	25831 88.85	3241 11.15	29072 100.00

Frequency Missing = 15

The FREQ Procedure

Table of stroke by HD

Frequency Percent Row Pct Col Pct	HD		
stroke	no	yes	Total
0	27129 93.32 95.11 98.45	1395 4.80 4.89 92.02	28524 98.12
1	427 1.47 77.92 1.55	121 0.42 22.08 7.98	548 1.88
Total	27556 94.79	1516 5.21	29072 100.00

Frequency Missing = 15

The FREQ Procedure

Table of smoking_status by stroke

Frequency Percent Row Pct Col Pct	stroke		
smoking_status	0	1	Total
formerly smoked	6919 23.80 97.46 24.26	180 0.62 2.54 32.85	7099 24.42
never smoked	15491 53.28 98.37 54.31	256 0.88 1.63 46.72	15747 54.17
smokes	6114 21.03 98.20 21.43	112 0.39 1.80 20.44	6226 21.42
Total	28524 98.12	548 1.88	29072 100.00

Frequency Missing = 15

Supervised Learning

- Logistic Regression

Model Selection

- Logistic Regression Model
- 6 Predictors
  - 3 qualitative variables
    - Hypertension, heart disease, smoking status
  - 3 quantitative variables
    - Age, average glucose level, bmi (body mass index)
- Response
  - Whether people had strokes or not

Logistic Regression Model and Assumptions

$$Y_i \sim \text{Binomial}\,(n = 1, p = f(E_i))$$

$$\begin{aligned} E_i = \beta_0 &+ \beta_{age}\text{Age}_i + \beta_{HBP}\text{HBP}_{i\,no} + \beta_{HD}\text{HD}_{i\,no} \\ &+ \beta_{average\_glucose\_level}\text{Average\_Glucose\_Level}_i + \beta_{bmi}\text{bmi}_i \\ &+ \beta_{smoking\_status}\text{Smoking\_Status}_{i\,formerly\,smoked} \\ &+ \beta_{smoking\_status}\text{Smoking\_Status}_{i\,never\,smoked} \end{aligned}$$

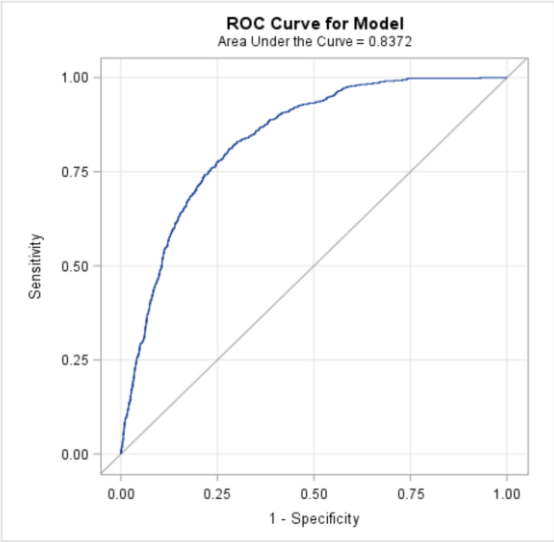
- f is the logistic link function, i.e. $f(E) = \frac{e^E}{1+e^E}$ß
  - Y_i is whether or not patient have stroke or not
  - Age is the age for *i*th patient
  - $HBP_{i\,no}$ is dummy variable for the *i*th hypertension was listed in, where "yes" is the reference level, where HBP stands for hypertension
  - $HD_{i\,no}$ is dummy variable for the *i*th heart disease was listed in, where "yes" is the reference level, where HD stands for heart disease
  - Average_glucose_level is the average glucose level measured after meal for *i*th patient
  - bmi(box mass index) is the body mass index for *i*th patient
  - $Smoking\_Status_{i\,never\,smoked}$ and $Smoking\_Status_{i\,formerly\,smoked}$ are two dummy variables for the *i*th smoking status was listed in, where "smokes" is the reference level.

Check Model Utility

- Misclassification Rate
  - 25.8%
  - Optimal % correct:
    - **74.2%** when cutoff is 0.02

Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.000	548	0	28524	0	1.9	100.0	0.0	98.1	.
0.020	427	21149	7375	121	74.2	77.9	74.1	94.5	0.6
0.040	329	24529	3995	219	85.5	60.0	86.0	92.4	0.9
0.060	222	26294	2230	326	91.2	40.5	92.2	90.9	1.2

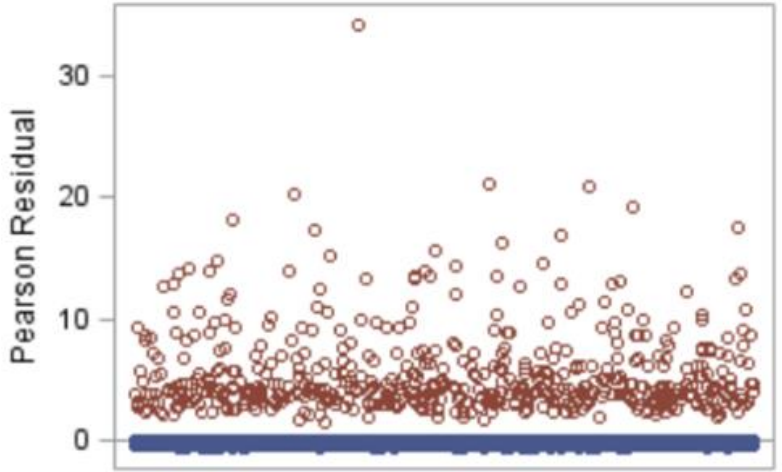
- ROC always above the diagonal line
- AUC=**0.8372** > 0.5
- The model generally performs well



Check Model Validity

- “Goodness” of Fit
  - Hosmer and Lemeshow Test (HL Test)
  - **0.2079** > 0.05
  - No evidence that the model doesn’t fit well.
- Independence Issue
  - No time series structures
  - Residual vs. Index plot
    - No clear patterns

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
10.8921	8	0.2079



Model Interpretation

- Only **bmi** is not statistically significant

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-7.3284	0.3936	346.7463	<.0001
age		1	0.0719	0.00371	374.7961	<.0001
HBP	no	1	-0.4384	0.1005	19.0161	<.0001
HBP	yes	0	0	.	.	.
HD	no	1	-0.6169	0.1125	30.0453	<.0001
HD	yes	0	0	.	.	.
avg_glucose_level		1	0.00384	0.000774	24.5504	<.0001
bmi		1	-0.00783	0.00713	1.2055	0.2722
smoking_status	formerly smoked	1	-0.2695	0.1268	4.5180	0.0335
smoking_status	never smoked	1	-0.2484	0.1183	4.4066	0.0358
smoking_status	smokes	0	0	.	.	.

Model Interpretation

- ***Age*** (quantitative)
  - An increase in age of 1 year is associated with **7.5%** increase in the odds of having a stroke when other variables stay fixed.
- ***Average glucose level*** (quantitative)
  - 1 unit increase in average glucose level is associated with **0.4%** increase in the odds of having a stroke when other variables stay fixed.
- ***Heart disease*** (qualitative)
  - People with no heart disease have **46%** lower odds of having a stroke than people with heart disease when other variables stay fixed.
- ***Hypertension*** (qualitative)
  - People with no hypertension have **35.5%** lower odds of having a stroke than people with hypertension, when other variables stay fixed.
- ***Smoking status*** (qualitative)
  - People who ***formerly smoked*** or ***never smoked*** have at least **20%** lower odds of having a stroke than people who ***smoke*** when other variables stay fixed.

Model Interpretation Cont.

- People who are **older**, **have hypertension**, **heart disease**, **higher average glucose level** tend to have higher odds of having a stroke for the same other conditions
- Impact of **formerly smoking** on the response is almost the same as **never smoked** when other variables remain fixed

Unsupervised Learning

- Cluster Analysis

Model Selection

- Cluster Analysis technique
- Variables used to make clusters
  - Age, average glucose level, bmi (body mass index)
- Linkage Criterion
  - Average Linkage
- Method
  - Agglomerative Hierarchical Clustering
- Validation Variable
  - Whether people had strokes or not

Identifying Jumps

Cluster History

Number of Clusters	Clusters Joined		Freq	Norm RMS Distance	Tie	Calculated Difference Between each cluster
10	CL14	CL21	23527	0.6044		0.0024
9	CL27	CL16	812	0.6144		0.01
8	CL31	CL26	45	0.6522		0.0378
7	CL12	CL9	3875	0.7278		0.0756
6	CL8	OB15044	46	0.8234		0.0956 jump
5	CL10	CL11	25148	0.9104		0.087
4	CL6	CL7	3921	0.9248		0.0144
3	CL43	CL5	25150	1.0142		0.0894
2	CL3	OB2	25151	1.2328		0.2186 jump
1	CL2	CL4	29072	1.7765		0.5437 jump

Finding the best Cluster Configuration

- The 3-cluster and 2-cluster configurations are almost the same
- Not satisfied with using these as our final configuration
  - Clusters 1 and 2 in both cluster configurations have a wide range of values for several variables
- **7 cluster configuration** seems to be most reasonable
  - More specific clusters with narrower value ranges

The MEANS Procedure

CLUSTER	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
1	25150	age	25150	45.7165010	18.5547590	10.0000000	82.0000000
		avg_glucose_level	25150	90.6913726	21.2228056	55.0100000	175.1000000
		bmi	25150	29.4982624	6.9119126	10.1000000	92.0000000
2	3921	age	3921	60.2053048	14.5693252	10.0000000	82.0000000
		avg_glucose_level	3921	207.1743943	23.7953404	148.6400000	291.0500000
		bmi	3921	33.6047947	7.8625129	15.0000000	82.7000000
3	1	age	1	78.0000000	.	78.0000000	78.0000000
		avg_glucose_level	1	135.7300000	.	135.7300000	135.7300000
		bmi	1	89.0000000	.	89.0000000	89.0000000

The MEANS Procedure

CLUSTER	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
1	25151	age	25151	45.7177846	18.5555068	10.0000000	82.0000000
		avg_glucose_level	25151	90.6931633	21.2242837	55.0100000	175.1000000
		bmi	25151	29.5006282	6.9219509	10.1000000	92.0000000
2	3921	age	3921	60.2053048	14.5693252	10.0000000	82.0000000
		avg_glucose_level	3921	207.1743943	23.7953404	148.6400000	291.0500000
		bmi	3921	33.6047947	7.8625129	15.0000000	82.7000000

Initial Observations

- 2 age groups
  - younger people (10-50 years old)
  - older people (51-82 years old)
- **Cluster 1:** younger people with **slightly low** average glucose level and **overweight** bmi
- **Cluster 2**: younger people with **medium** average glucose level and **overweight** bmi
- **Cluster 3**: older people with **high** average glucose level and **obese** bmi.
- **Cluster 4**: younger people with **very high** average glucose level and **obese** bmi
- **Cluster 5**, **6** and **7**: Outliers

The SAS System

The MEANS Procedure

CLUSTER	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
1	23527	age	23527	45.7979343	18.4257246	10.0000000	82.0000000
		avg_glucose_level	23527	87.2814082	17.0930001	55.0100000	135.9400000
		bmi	23527	29.4947592	6.9122666	10.1000000	72.2000000
2	1621	age	1621	44.5533621	20.3078260	10.0000000	82.0000000
		avg_glucose_level	1621	140.2167798	11.1836505	119.6500000	175.1000000
		bmi	1621	29.4806292	6.6278170	12.5000000	61.8000000
3	3875	age	3875	60.4356129	14.4589452	10.0000000	82.0000000
		avg_glucose_level	3875	206.5863381	23.2653195	148.6400000	272.8600000
		bmi	3875	33.5713806	7.7569170	15.0000000	80.1000000
4	45	age	45	39.9333333	8.3758310	19.0000000	52.0000000
		avg_glucose_level	45	255.9486667	12.8352948	228.2400000	281.5900000
		bmi	45	36.5911111	14.1449930	17.3000000	82.7000000
5	2	age	2	30.5000000	10.6066017	23.0000000	38.0000000
		avg_glucose_level	2	63.4650000	9.2843120	56.9000000	70.0300000
		bmi	2	85.0000000	9.8994949	78.0000000	92.0000000
6	1	age	1	80.0000000	.	80.0000000	80.0000000
		avg_glucose_level	1	291.0500000	.	291.0500000	291.0500000
		bmi	1	28.7000000	.	28.7000000	28.7000000
7	1	age	1	78.0000000	.	78.0000000	78.0000000
		avg_glucose_level	1	135.7300000	.	135.7300000	135.7300000
		bmi	1	89.0000000	.	89.0000000	89.0000000

Cluster labeling

Cluster 1 - younger people in comparatively healthy status

Cluster 2 - younger people who struggle with their glucose level

Cluster 3 - older people who struggle with their health status

Cluster 4 - younger people who are in extremely unhealthy status

Checking Model Validity using Stroke variable

Finding: younger people in very unhealthy status have very low risk of having a stroke

- May need more investigation

Frequency Percent Row Pct Col Pct	Table of stroke by CLUSTER							
	CLUSTER							
stroke	1	2	3	4	5	6	7	Total
0	23199	1585	3691	45	2	1	1	28524
	79.80	5.45	12.70	0.15	0.01	0.00	0.00	98.12
	81.33	5.56	12.94	0.16	0.01	0.00	0.00	
	98.61	97.78	95.25	100.00	100.00	100.00	100.00	
1	328	36	184	0	0	0	0	548
	1.13	0.12	0.63	0.00	0.00	0.00	0.00	1.88
	59.85	6.57	33.58	0.00	0.00	0.00	0.00	
	1.39	2.22	4.75	0.00	0.00	0.00	0.00	

Check Model Validity using Chi-Squared Test

- Chi-Squared Test of Association
  - P-value(**<0.0001**) < 0.05
  - Reject the null hypothesis
- Confirms that the proportions of Stroke values are different in each clusters at 5% significance level
- The clusters are reasonable

Statistics for Table of stroke by CLUSTER

Statistic	DF	Value	Prob
Chi-Square	1	193.1872	<.0001
Likelihood Ratio Chi-Square	1	146.9508	<.0001
Continuity Adj. Chi-Square	1	191.4364	<.0001
Mantel-Haenszel Chi-Square	1	193.1806	<.0001
Phi Coefficient		0.0815	
Contingency Coefficient		0.0812	
Cramer's V		0.0815	

Interpretations

- **Age** and **Average Glucose Level** have a great impact on our validation variable
- Older people who struggle with their health status should be prioritized to
  - Create awareness
  - Educate on the severity of stroke
  - Do further medical research
  - Provide customized treatments

Discoveries & Limitations

Discoveries

Logistic Regression

- Age, smoking status, average glucose level, hypertension, and heart disease are significant
- People with heart disease have a higher risk of having a stroke

Cluster Analysis

- Age and Average Glucose Level play important roles
- Older people with high average glucose level are more likely to get a stroke

Limitations

- No specific time of measurement of the average glucose level
  - Lead to standard inconsistency
  - One hour after meal is recommended
- No specific amount of years for whom quitted smoking and started smoking
  - Never smoked and formerly smoked had similar impact on stroke
  - More information needed
- Cluster analysis is subjective
  - No correct answer

Thank you!

Stay Healthy and Safe Everybody!