Stroke Predictions

May 2023

Heng-Tser Tsai

Agenda

- Data Dictionary
- Exploratory Data Analysis (EDA)
- Machine Learning Models
- Next Steps

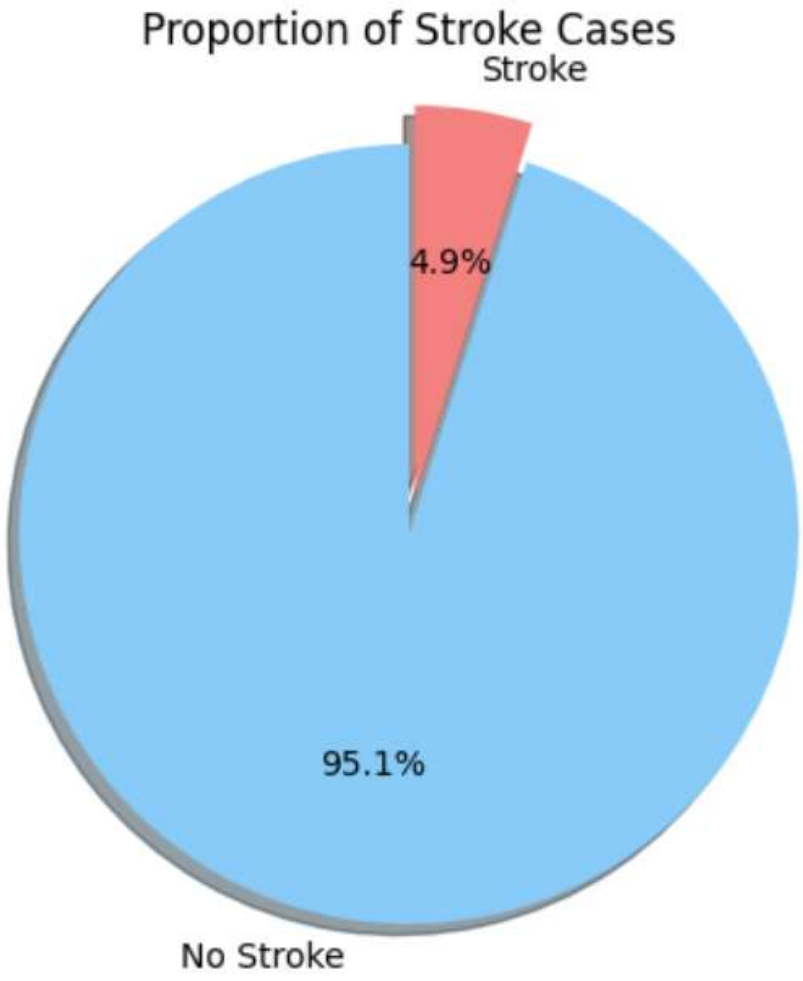
Data Dictionary

Column Name	Data Type	Description
id	Integer	Unique identifier
gender	Object	"Male", "Female", "Other"
age	Float	Age of patient
hypertension	Integer	0 if the patient doesn't have hypertension, 1 if the patient has hypertension
heart Disease	Integer	0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
ever_married	Object	"No" or "Yes"
work_type	Object	"children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
Residence_type	Object	"Rural" or "Urban"
avg_glucose_level	Float	average glucose level in blood
bmi	Float	body mass index
smoking_status	Object	"formerly smoked", "never smoked", "smokes" or "Unknown"*
stroke	Integer	1 if the patient had a stroke or 0 if not, target

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

EDA
- TARGET

- Highly imbalance
- Classification data

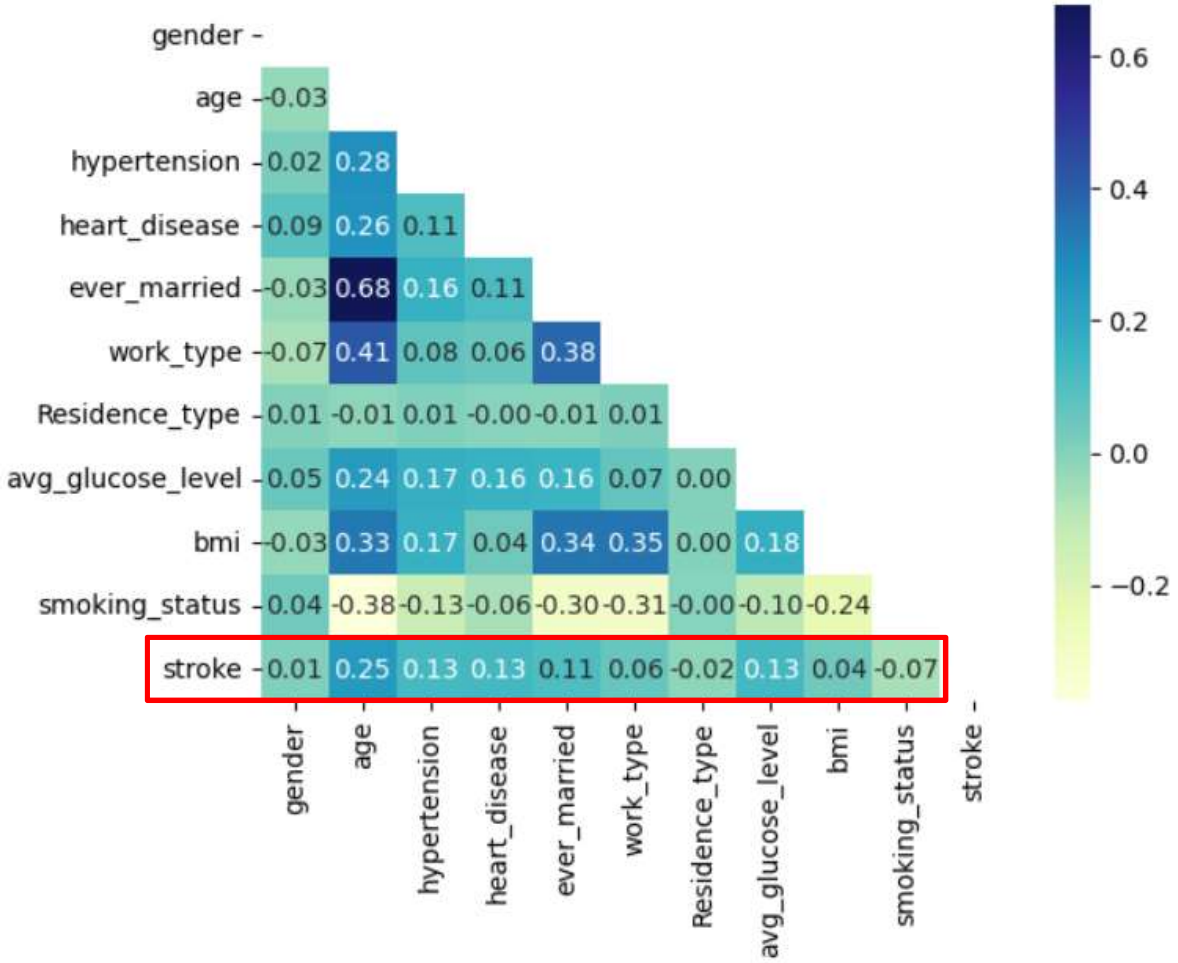


EDA

- Heatmap

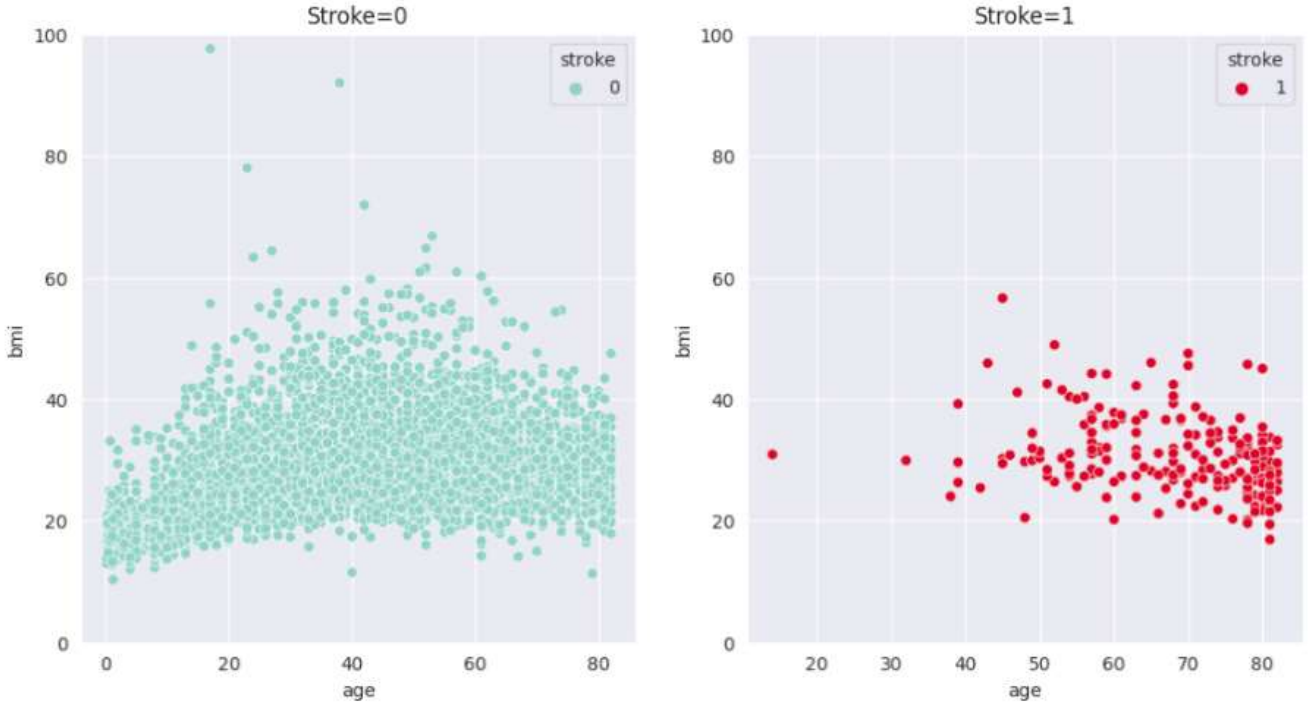
Stroke **vs**

- **age**
- hypertension
- heart_disease
- ever_married
- avg_gluclose_level



EDA

- Scatter Plot

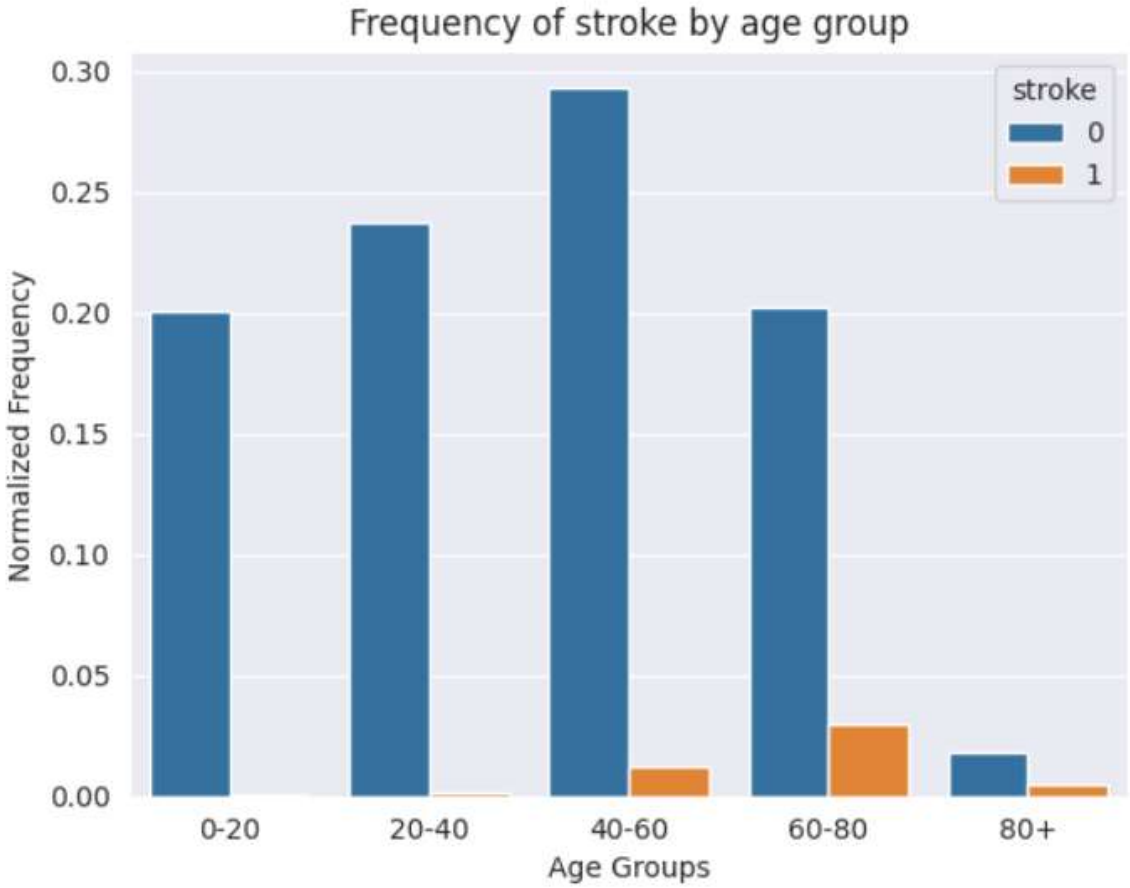


- Strokes occurs on older population especially after 50 years old.
- High BMI is not highly correlated to stroke.

EDA

- Bar Plot

- Chances of stroke increase as you age.
- According to this data, people generally do not have strokes.



MACHINE LEARNING

- Models and Results

	Test F1	Test AUPRC	Test TP
SMOTE Logistic Regresion	0.268	0.459	0.738
Tuned SMOTE Logistic Regresion	0.265	0.452	0.725
Oversampling Tuned Logistic Regresion	0.254	0.443	0.712
Oversampling SVC	0.23	0.367	0.562
SMOTE SVC	0.24	0.371	0.562

- Goal = successfully predict the stroke
- The value of true positive (TP) on testing data is emphasized
- Top 5 models are all oversampling models out of 20 models.

SUMMARY

- For 50 + year old, it is suggested to actively check the symptom of stroke to raise the awareness of it.

- Other character could lead to stroke:
  - female
  - married
  - living at urban
  - working in private sector
  - never smoked

Your **Life Style** Matters!

- Oversampling Logistic Regression model has the best opportunity on predicting TP.