DIABETES, HYPERTENSION AND STROKE

KRISTIE KOOKEN

STROKE

STAGGERING LOSS	• Leading cause of death and severe disability
DIABETES and HYPERTENSION	• Combined loss of productivity 600B USD • Established risk factors of Stroke
PREDICTORS?	• Different health behaviors, risk factors and diabetes and hypertension

BACKGROUND

- Stroke
  - 3rd leading cause of death
  - >795,000 strokes annually with 140,000 deaths
- Diabetes
  - 8th leading cause of death
  - Affects 11.6% (38.4M) of the US population
  - 1 in 5 don’t know have diabetes
  - Risk of stroke is 2 times higher
- Hypertension
  - In 2021, hypertension was a contributing factor in about 700k deaths
  - ~50% of adults have hypertension
- Combined
  - About 6 out of 10 of people who have diabetes also have high blood pressure

PURPOSE

- To explore risk factor to predict Stroke
  - Different health behaviors and physical characteristics
  - Diabetes and Hypertension

ANALYSIS

Data Source/Cleaning

- Behavioral Risk Factor Surveillance System (BRFSS) 2015
  - 18 columns
  - Data cleaning & wrangling using python
  - No outlier, no missing imputation

EDA

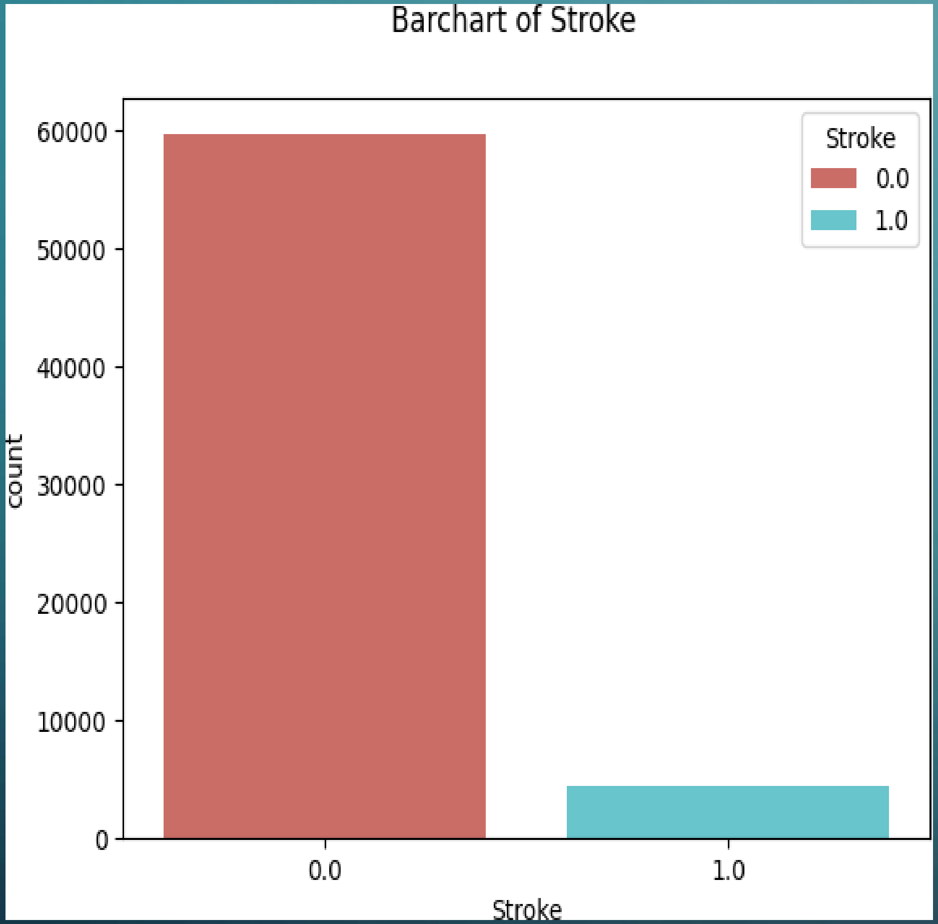
Increase in Stroke:

- Difficulty in walking with BMI 26 vs 32
- High BP with BMI 27 vs 35
- Diabetes with BMI 27 vs 33

ML Logistic Regression

- Dummied
- SMOTE
- RFE (recursive feature elimination)
- Logistic regression
- Confusion matrix
- ROC plot

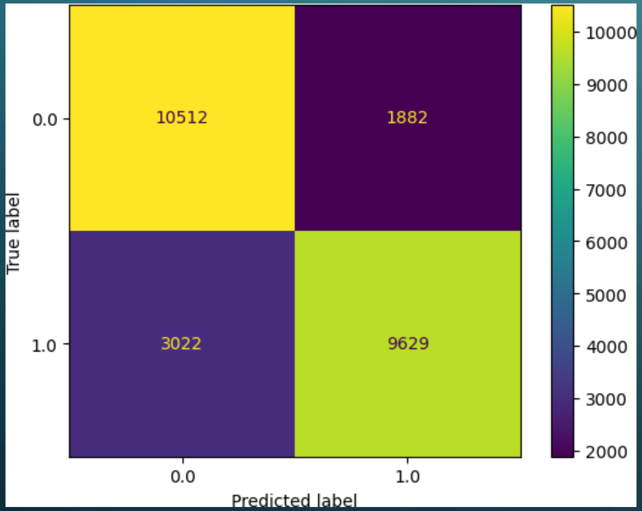
DEEPER DIVE INTO ANALYSIS



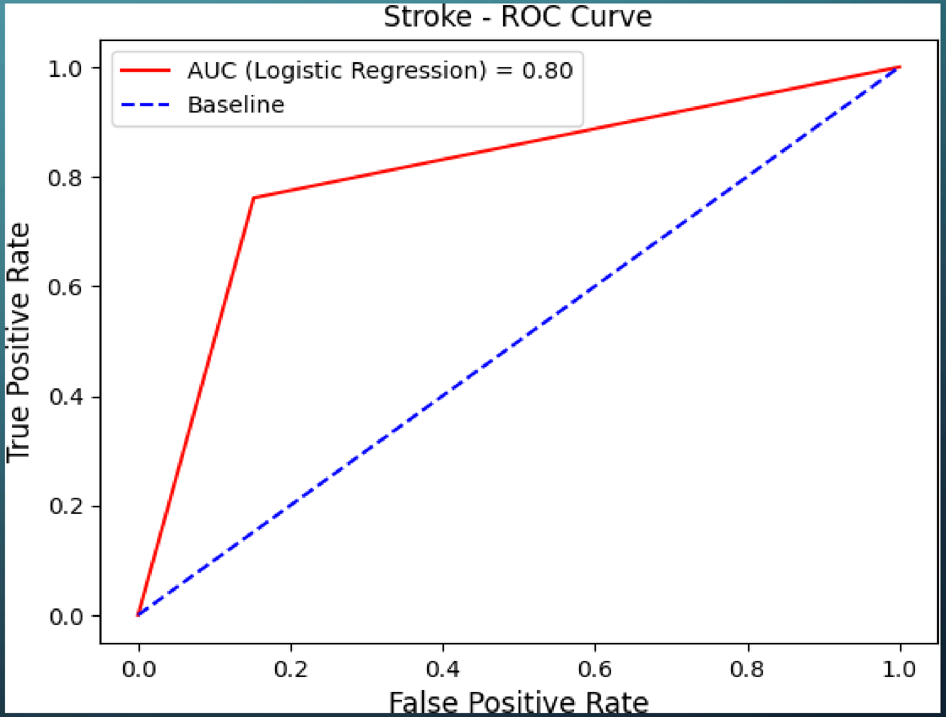
- DUMMIED, any variable that had >2 categories was dummied
- Binary variables had values of 0/1
- SMOTE
  - Balancing technique to up sample the number of strokes in the data
  - After – total rows = 83482
  - # of Y/N each Stroke = 41741

MODEL RESULTS

- 80% accuracy
  - Precision and Recall were also acceptable
    - Ranging 0.76 to 0.85
  - Confusion Matrix



ROC Plot



CONCLUSION

- Overall model results within range for acceptability
- Caution in interpreting results as indicative of predictability of stroke
  - Further analysis and data are needed
    - Additional survey years
    - Time variables to better understand the length of time to having the event of stroke

CONSIDERATIONS

- Data cleaning happened as expected
  - Data was very clean and there is no traceability on how this cleaning happened compared to the original survey
  - Is SMOTE the best way to handle sample imbalance? More robust data is likely the best way.
  - With a larger sample, additional modeling techniques can be explored in order to return the most robust result

BACKUP SLIDES

VARIABLES IN ANALYSIS

Variable	Decode
Age	13-level age category (_AGEG5YR see codebook) 1 = 18-24 9 = 60-64 13 = 80 or older
Sex	Patient's gender (1: male; 0: female).
HighChol	0 = no high cholesterol 1 = high cholesterol
CholCheck	0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years
BMI	Body Mass index
Smoker	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes
HeartDiseaseorAttack	Coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes
PhysActivity	Physical activity in past 30 days - not including job 0 = no 1 = yes
Fruits	Consume Fruit 1 or more times per day 0 = no 1 = yes
Veggies	Consume Vegetables 1 or more times per day 0 = no 1 = yes
HvyAlcoholConsump	Adult men >=14 drinks per week and adult women>=7 drinks per week: 0 = no 1 = yes
GenHlth	Would you say that in general your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor
MentHlth	Days of poor mental health scale 1-30 days
PhysHlth	Physical illness or injury days in past 30 days scale 1-30
DiffWalk	Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes
Stroke	Have you ever had a stroke. 0 = no, 1 = yes
HighBP	0 = no high, BP 1 = high BP
Diabetes	0 = no diabetes, 1 = diabetes