STROKE PREDICTION ANALYSIS

Volha Puzikava
June, 2022

Overview

Stroke Statistics:

- the 2nd leading cause of death globally
- responsible for 11% of total death
- the leading cause of serious long-term disability

Every:

- 40 seconds someone gets a stroke in the US
- 3.5 minutes someone dies of stroke in the US
- year 795,000 people in the US get a stroke

THIS PROJECT:

1. predicts if patients will develop stroke in their lifetime
2. identifies key factors leading to stroke

Outline

- Business Problem
- Data
- Data Preparation and Exploration
- Modeling
- Evaluation
- Recommendations
- Conclusions

Business Problem

The World Health Organization wants to more frequently monitor people prone to stroke in order to prevent the illness incidences.

Goals:

1. analyze the stroke dataset;
2. identify the key factors that likely increase the occurrence of stroke;
3. provide predictive recommendations and suggestions.

Data

- was taken from [kaggle website]
- provided 11 clinical features for predicting stroke effect:

    - ❑ gender,
    - ❑ age,
    - ❑ marital status,
    - ❑ work type,
    - ❑ residence type,
    - ❑ smoking status,
    - ❑ hypertension,
    - ❑ heart disease,
    - ❑ average glucose level
    - ❑ body mass index

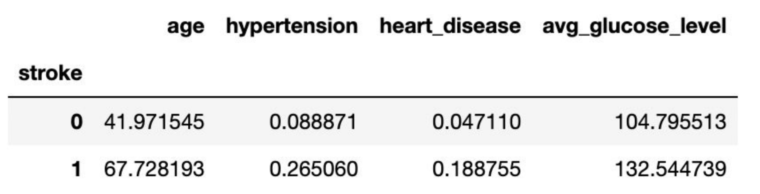
- contained information about 5,110 patients

Data Preparation and Exploration

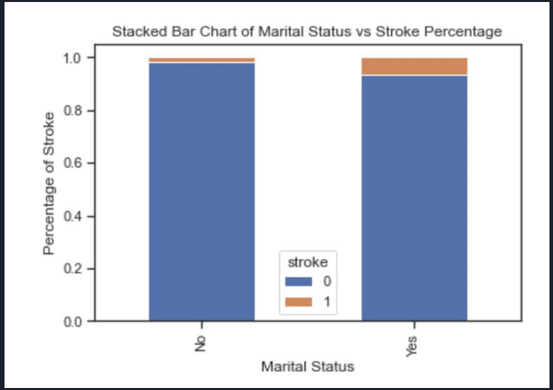
1. The stroke incidences were compared among clinical features.

   Factors that influence on the stroke occurrence:

   - age (average is 67)
   - hypertension and/or heart disease
   - high glucose level
   - marital status (married)
   - work type (self-employed, private or government jobs)
   - smoking status (smoke or smoked in the past)



stroke	age	hypertension	heart_disease	avg_glucose_level
0	41.971545	0.088871	0.047110	104.795513
1	67.728193	0.265060	0.188755	132.544739



Data Preparation and Exploration Cont’d

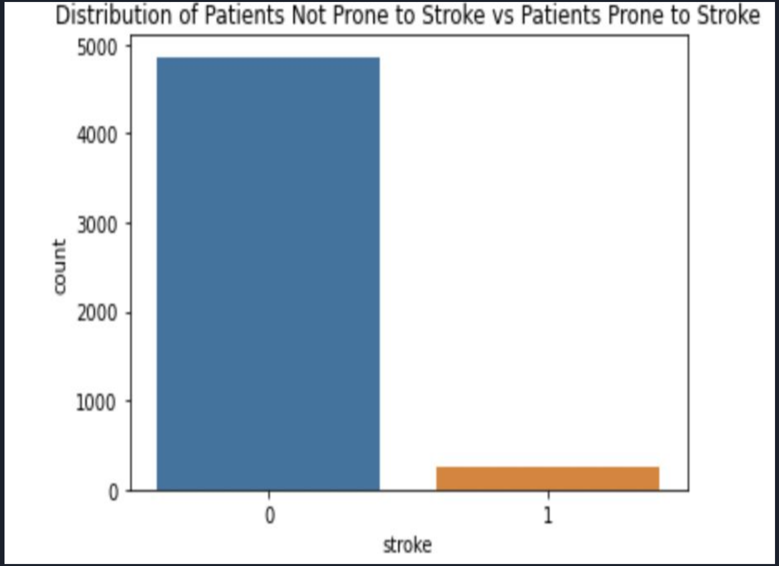
2. Missing values were replaced with the column mean value

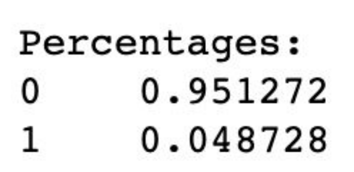
3. All categorical variables were transformed into numbers

4. The strength of the relationship between the variables was visualized (AKA correlation)

Modeling

- Dependent variable was assigned to “stroke” column; all other features served as predictors
- The data was split into training set and testing set to estimate how well the learned model will generalize to new data
- The dataset was imbalanced, so synthetic data was generated in the training set to oversample a minority target class (SMOTE-NC)


Distribution of Patients Not Prone to Stroke vs Patients Prone to Stroke

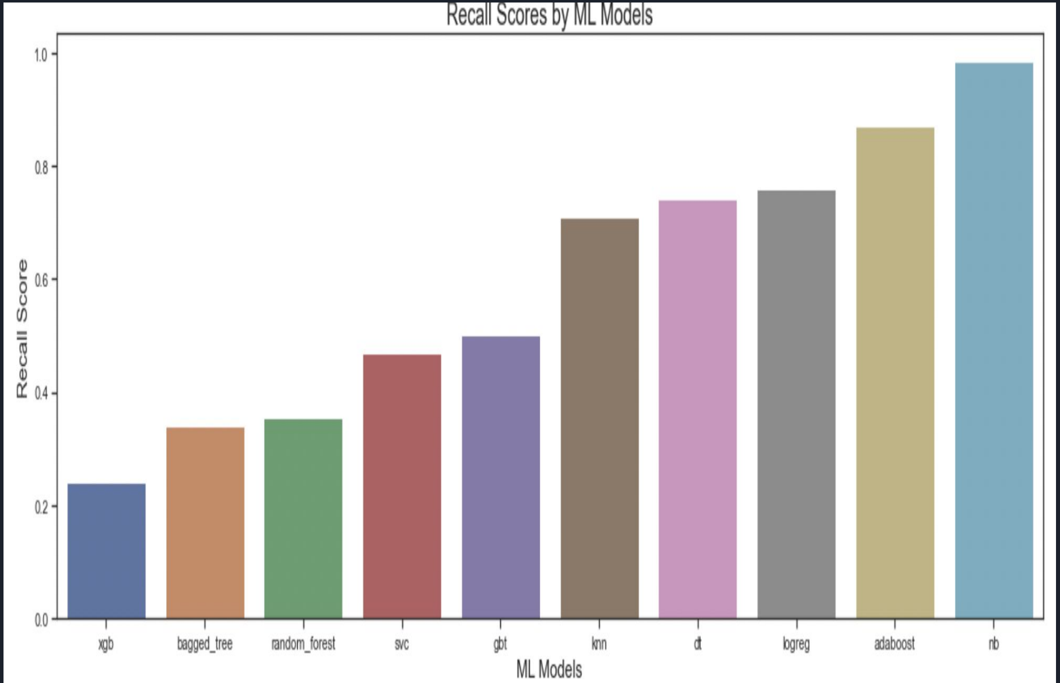


```
Percentages:
0     0.951272
1     0.048728
```

Modeling Cont’d

- The data was represented at the same scale to avoid “leaking” of information from one set to another
- Ten different machine learning models were built, tuned and run
- Models ability to correctly predict the positives out of actual positives (recall score) were checked
- The model with the highest recall score was selected

Evaluation



- the best model is naive bayes model
- it has the highest recall score of 98%

BUT:

- it classifies 68% of patients as prone to stroke
- is 37% accurate

Recommendations

1. pay more attention to people who:

    - over 45 years old,
    - hypertension,
    - heart disease,
    - high glucose level,
    - married,
    - self-employed,
    - smoke,
    - smoked in the past.

2. closely monitor 68% of the patients in order to successfully treat 98% of the ones who will develop stroke

Conclusions

- The model should be utilized only if there is a special kind of treatment or particular monitoring practices developed for the patients prone to get stroke.
- If the stakeholder changes the direction of the research and sets different goals the model should be changed

Next Step:

- gather more information: family history, diet, presence of diabetes, alcohol consumption, etc.

Thank you!

Email: helga.mikel@gmail.com

GitHub: @VolhaP87

LinkedIn: linkedin.com/in/volha-puzikava-2319294a