Stroke Prediction

Using machine learning to predict the onset of stroke

Josh Wang
11/21/2022

Image from Unsplash

Background

A stroke occurs when the blood flow to the brain is blocked. It is the **5th cause of death and a leading cause of disability** in the United States.

A stroke is an emergency situation, and it is important to be able to predict the onset of stroke. (Healthcare, insurance, etc.)

Image from Unsplash

The Data (from Kaggle)

features

age

gender

blood glucose

smoking status

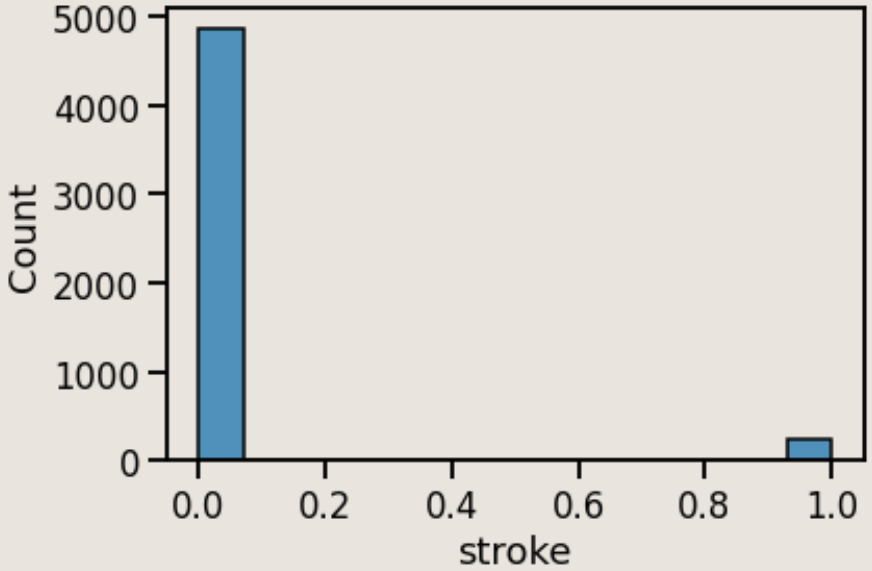
heart disease

etc.

target

stroke (Y/N)

The Data – Class Imbalance



Data is heavily skewed.

Only 5% of the datapoints are positive for stroke.

{0: 95%, 1: 5%}

Methodology

Data

from Kaggle

Cleaning

Preprocessing

One hot encoding
Scaling
Missing data imputation

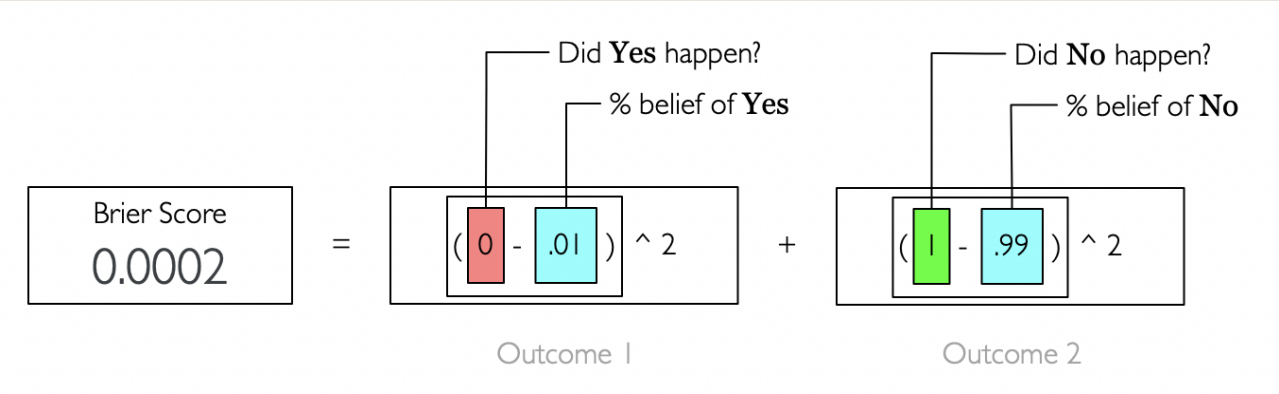
Modeling

Under/oversampling
RandomizedSearchCV
GridSearchCV

Deployment

Evaluation Metric – Brier Skill Score

Brier Score is the sum of the squares of the residuals. (similar to Mean Squared Error)
Measures the accuracy of probabilistic predictions (lower is better):



Brier Skill Score (BSS) = 1 – BS/BS^{ref} (higher is better)

BSS>0, model better than baseline
BSS=0, model has no skill
BSS<0, model worse than baseline

Image from Medium

Results

Cross-validation scores (higher is better):

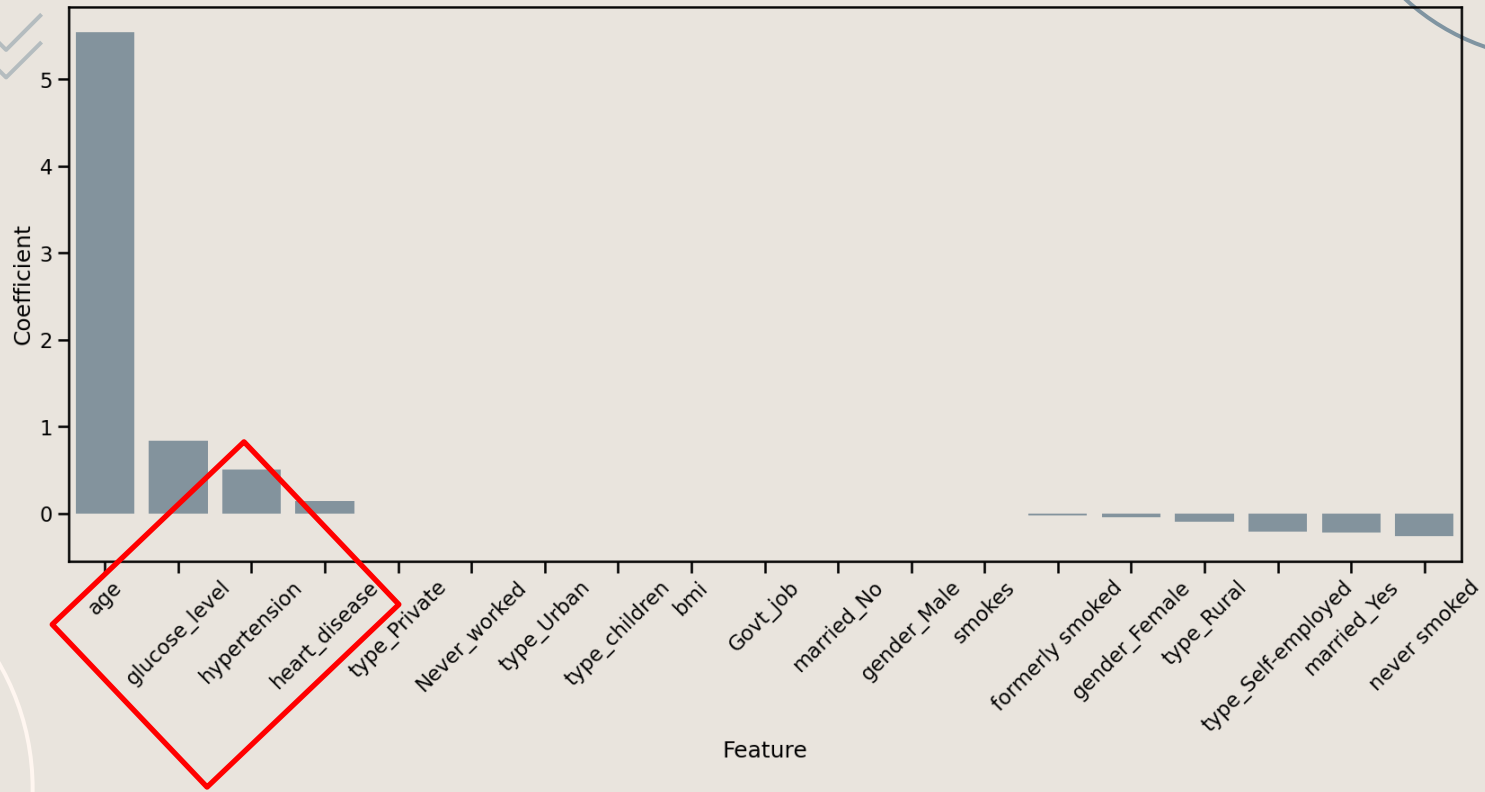
Model	Base Model BSS	Tuned Model BSS
Logistic Regression	0.075	0.076
kNN	0.013	0.049
RandomForest	0.034	0.077
XGBoost	0.021	0.075
SVC	0.007	0.083

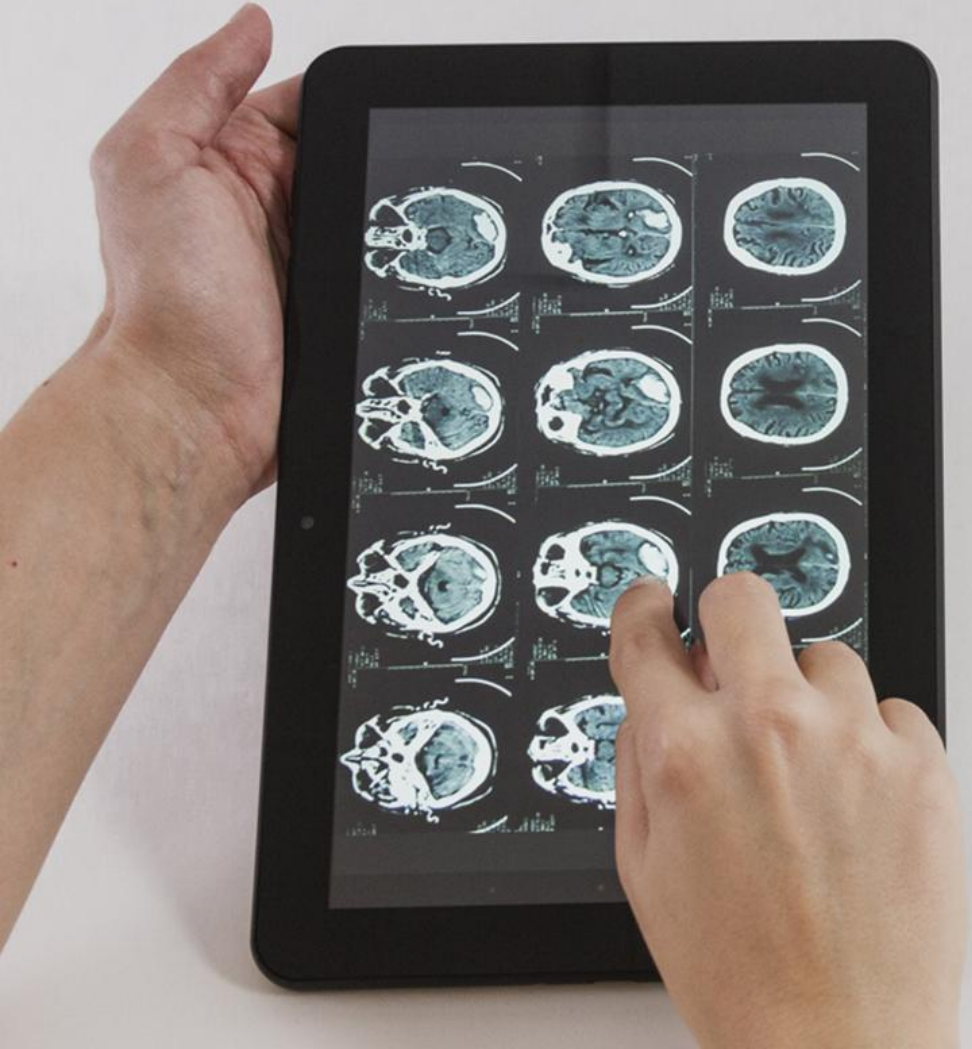
Probabilistic outputs were calibrated with Platt Scaling when necessary.

SVC wins. The BSS of SVC on the holdout data is 0.113

Results

Feature coefficients from Logistic Regression:





Product Deployment: A Web App to Predict Stroke Onset

(powered by Streamlit)

Image from Unsplash

Example CSV input file

Upload your input CSV file

Drag and drop file here
Limit 200MB per file • CSV

Browse files

Gender

Male

Age

30

0 130

Hypertension

0

Heart Disease

0

Ever married

Yes

Work Type

Private

Residence Type

Rural

Average Glucose Level

100

40 300

Stroke Prediction App

This app predicts the probability of the onset of stroke!

Data obtained from the Stroke Prediction Dataset on Kaggle by fedesoriano.

Prediction powered by SupportVectorClassifier on sklearn.

User Input features

Awaiting CSV file to be uploaded. Currently using example input parameters (shown below).

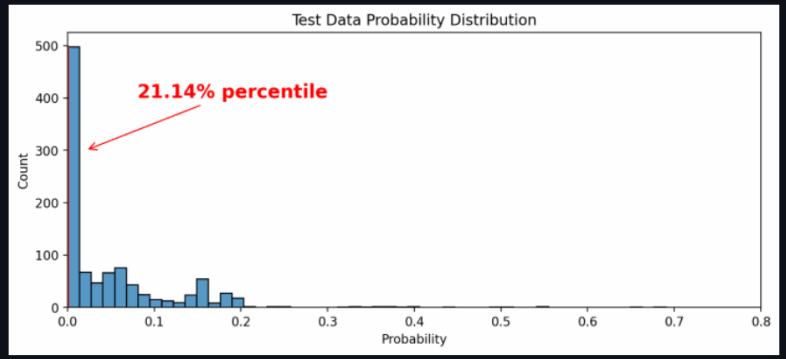
	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_t
0	Male	30	0	0	Yes	Private	Rural

Prediction Probability

The probability of stroke onset is

0.001

See where your probability lands among the test data:



Thank you

The code for this project can be found here:

https://github.com/JoshJingtianWang/Stroke_Prediction

Image from Unsplash