STROKE PREDICTION

UCSF

Classification model to predict stroke in patients

Sameera Bellary
May 2022

INTRODUCTION

- According to CDC, 5% of deaths in the world are caused by strokes
- Stroke is a disease that affects the arteries leading to and within the brain
- A stroke occurs when a blood vessel that carries oxygen and nutrients to the brain is either blocked by a clot or bursts (or ruptures).
- Age and high hypertension are the main factors of a stroke.

Types of stroke

Ischemic stroke

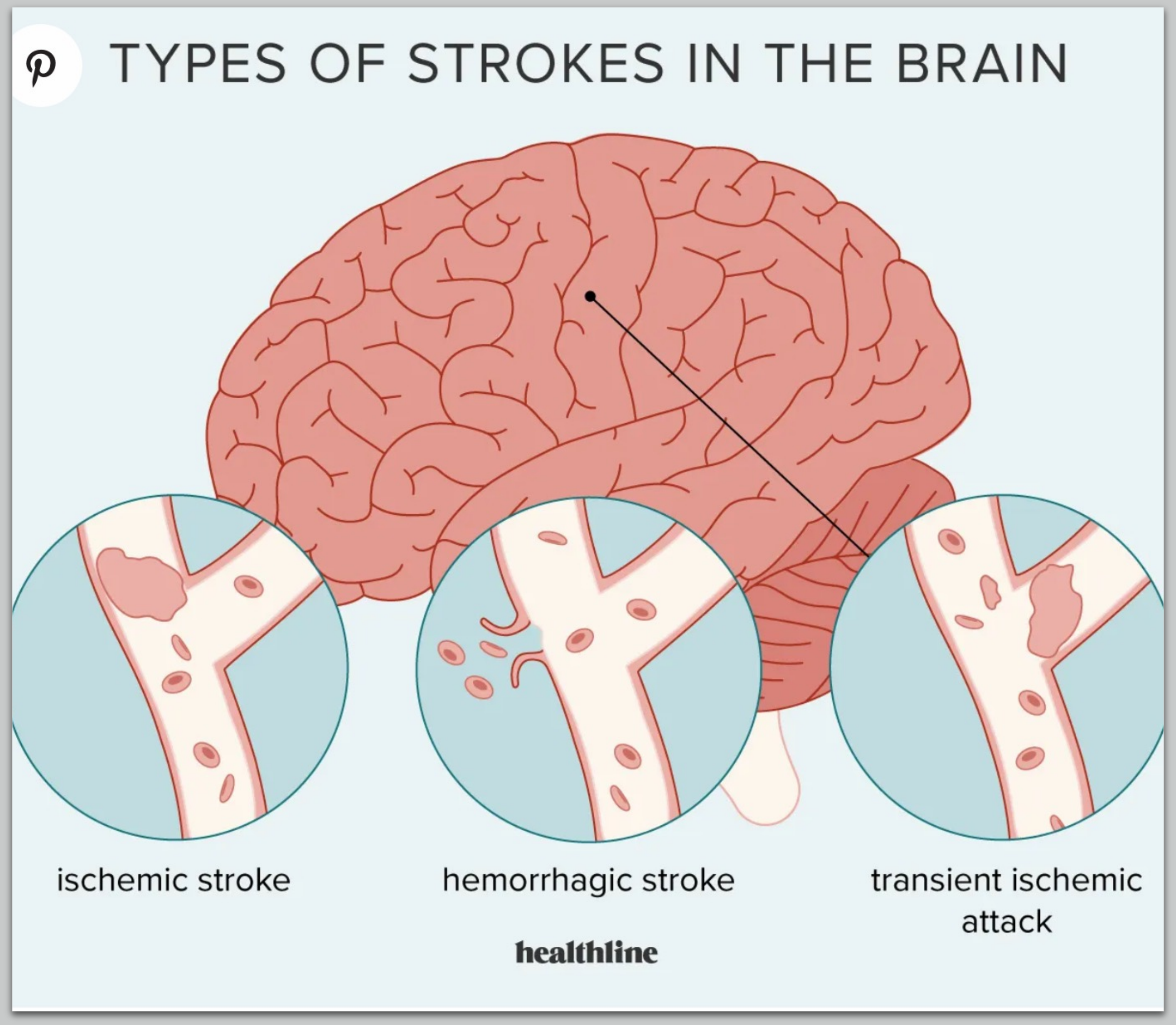
Caused by a blockage or clot in a blood vessel in your brain. The blockage can be caused when a substance called plaque builds up on the inside wall of an artery.

Hemorrhagic stroke

Caused when an artery in the brain breaks open. The interrupted blood flow causes damage to your brain.

Transient ischemic attack (TIA)

Caused by a small clot that briefly blocks an artery. It is sometimes called a mini-stroke or warning stroke. The TIA symptoms usually last less than an hour.



GOAL

- UCSF stroke clinic wants to predict which patient is at a higher risk for a stroke.

- The risk can be classified as
  - No risk
  - High risk

Methodology

- Data Set
  - Around 5,000 entries of patient data
  - 16 predictors – categorical and numerical
  - Categorical features were binarized
  - Target Variable – Stroke
    - 1 == Yes (had a stroke)
    - 0 == No (no stroke)
    - Highly imbalanced
    - 95.75% Negatives (0) to 4.25% Positives (1)

Tools

Pandas – Clean, Explore and Feature Engineering

Scikit-Learn – Build different Classification models and perform cross validation, variable selection and regularization

Matplotlib/ Seaborn – Visualizing data exploration, modeling and results

Python 3.8.5 – to run all of the above

Results

Seven individual models and two ensembles built :

1. **Logistic Regression** (regularization optimized for AUC metric with class weight adjustment)
2. **Logistic Regression** (regularization optimized for log loss metric with class weight adjustment)
3. **Logistic Regression** (regularization optimized for AUC metric)
4. **Random Forest Classifier** (hyperparameters optimized for AUC metric)
5. **Random Forest Classifier** (hyperparameters optimized for log loss metric)
6. **XG Boost Classifier** (hyperparameters optimized for AUC metric)
7. **XG Boost Classifier** (hyperparameters optimized for log loss metric)
8. **Ensemble Voting Classifier** (combining the last five models)
9. **Ensemble Voting Classifier** (combining the first two models)

Goal : Maximize Recall (True positive rate, or TPR)

Recall = TP/(TP + **FN**)

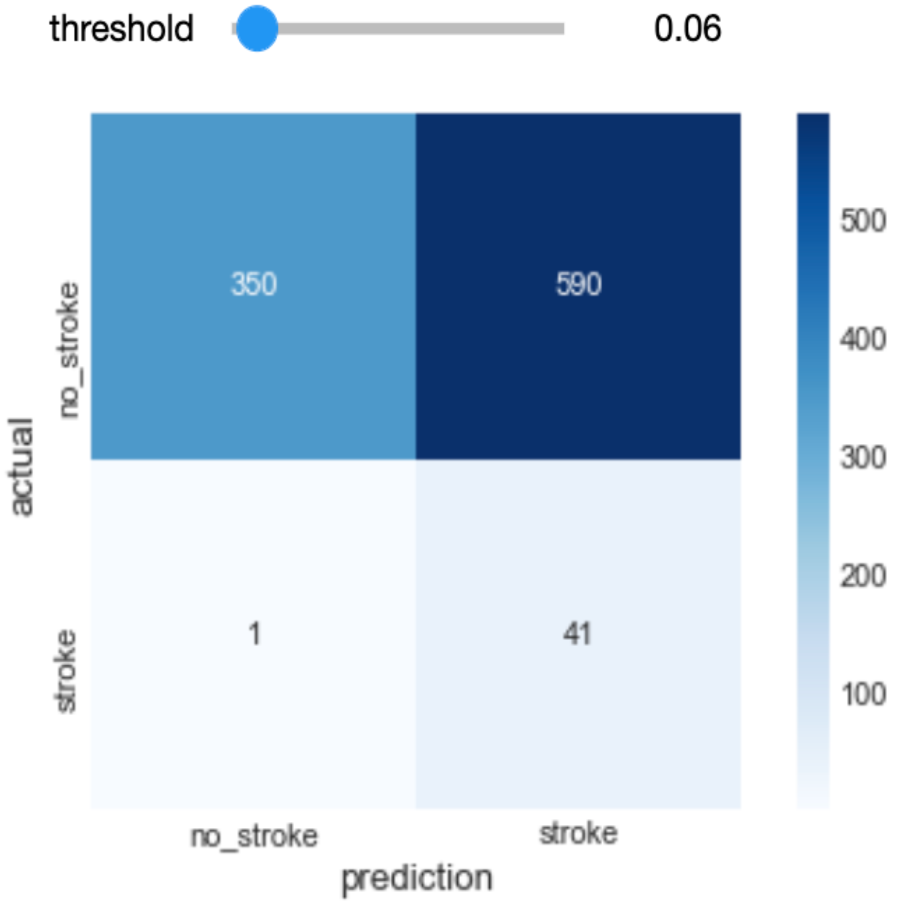
Minimize false negatives, so the misses are very few high risk patients

- As FN decreases, TPR or recall increases
- As Recall increases, precision decreases (False positive rate)

Best Model : Voting classifier (Soft)

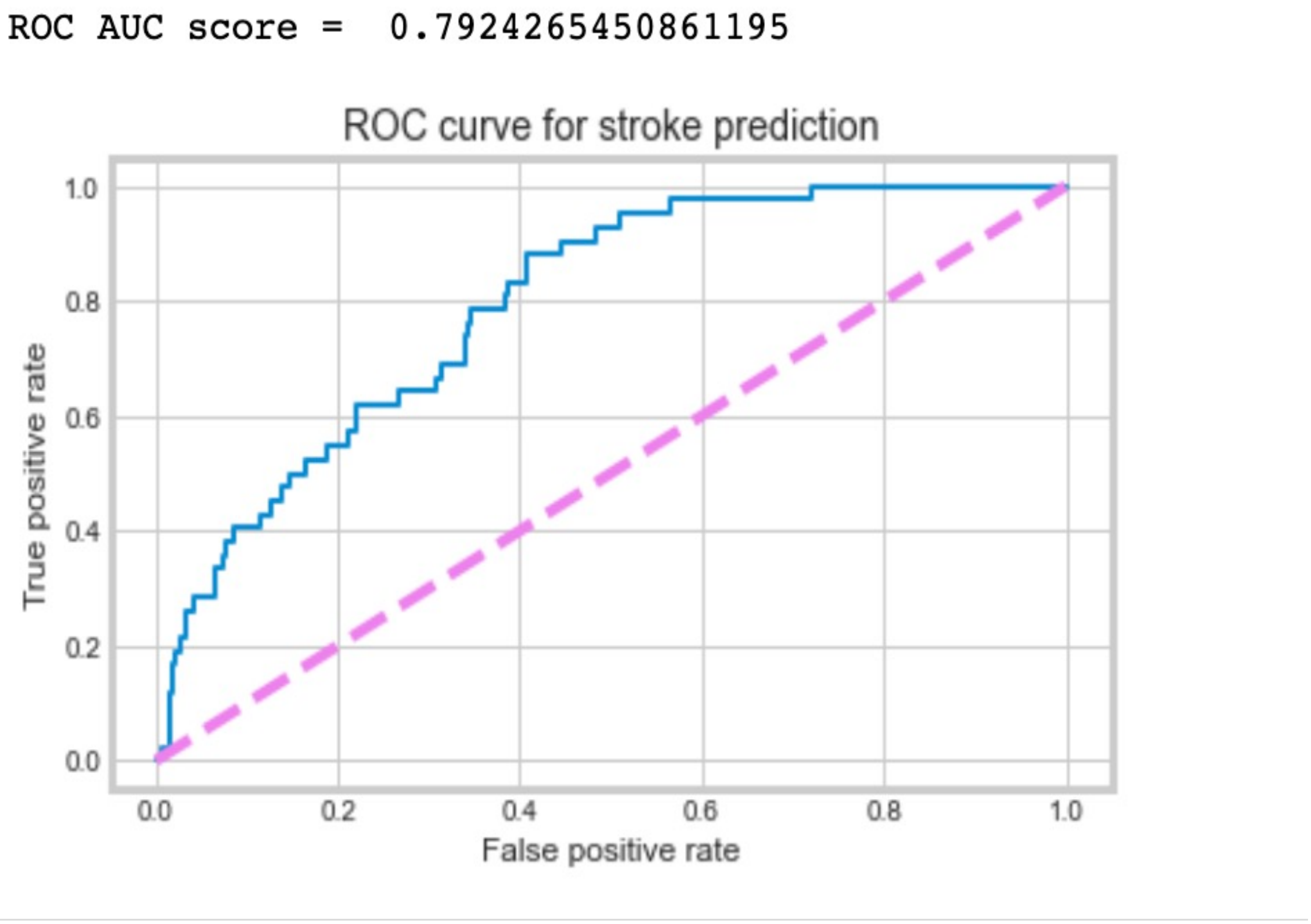
Logistic Regression with weights adjusted

- Accuracy Score – 0.92
- Recall – 0.97
  - False negatives minimized
- At threshold 0.06
  - FN = 1 and FP = 590



AUC Score - 79%

- Precision – 0.1
- Precision decreases
- True positive rate or recall increases



```
ROC AUC score =  0.7924265450861195
```

CONCLUSIONS

Recommendations

- Deploy the model
- Develop a multi class model predicting high, medium and low risk patients

Future Work

- Collect more data
- Collect more features which increase the risk of stroke
- Retrain the model periodically and increase the precision, without decreasing the recall