Stroke Prediction --Deep Learning Project Report

Shao Yan Chen
2021/12

Analytic Approach

Purpose
Predict whether a patient will get stroke through their medical record and ML/DL models.

Description
Dataset has 12 features and 5110 observations.
For ML models, we tuned their own parameters to approach an optimal performance.
For DNN models, we referred to some research to build a model with great capability and interpretability.

Model training
Logistic Regression/ KNN/ SVM/ Decision Tree/ Random Forest/ DNN

Final Model Description:

Process workflow
MICE/ EDA/ one-hot encoding/ Normalize data/ train-test split.

Selected Model
DNN

Setting
Hidden layers: 6 Denses(each with 21 neurons),
activation function: tanh,
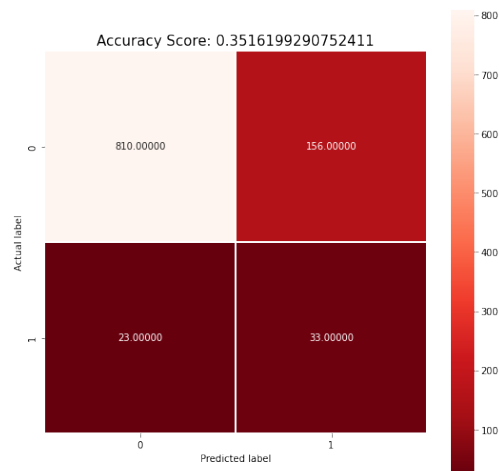output layer: sigmoid,
optimizer: SGD,
learning rate: 0.0005,
loss: binary crossentropy,
batch normalization.

Results

1. Score: 0.35/ Accuracy: 0.82
2. Confusion Matrix:

Accuracy Score: 0.3516199290752411

3. ROC: 0.71, Precision: 0.17, Recall: 0.58, F1: 0.26

Model Understanding

1. Imbalanced data can disastrously affect the performance of model
2. Deep learning can find some pattern which some ML model cannot
3. Since this is the dataset about disease, we cannot solely focus on the performance on accuracy

Conclusion and Discussion for Next Steps

1. The disease prediction performance on structure data would be limited, image data might support the model to predict more accurately
2. Collect more features, try different settings of model
3. Develop process method of imbalanced data is crucial in medical data since the positive and negative cases are always huge imbalanced
4. Long-term follow-up research might be a better way to prevent the stroke