One year Post-Stroke Prediction on Cognitive Impairment: A Machine Learning Approach

Muhammad Solihuddin Muhtar, Faizul Hasan, Alex P.A. Nguyen, Jason C. Hsu, Hsiao-Yean Chiu

ABSTRACT

Background

Cognitive impairment following stroke has wide prevalence ranging from 25% to 81%.[1] Further, stroke and the subtypes, including ischemic stroke, transient ischemic attack and intracerebral hemorrhage, significantly increase the long-term risk of dementia after 5 and 10 years. The incidence rate of post-stroke dementia increases yearly, though the relative risk gradually decreases.[2]

Objectives

The study aims to predict the dementia development one year after stroke diagnose.

Methods

The data source were from Taipei Medical University Clinical Research Database (TMUCRD) from January 2004 to September 2017. The inclusion, exclusion and outcome criteria were selected based on ICD9 and ICD10 codes. We included all patient with history of stroke, insomnia, cognitive impairment and other codes related with the diseases. We excluded psychiatric disorder, sleep apnea, traumatic brain injury, cancer, Parkinson's disease, and cognitive impairment from the outcome. The outcomes were mild cognitive impairment, Senile dementia, uncomplicated, senile dementia with delusional or depressive features, Senile dementia with delirium, dementia in conditions classified elsewhere, alzheimer's disease, Frontotemporal dementia, and Senile degeneration of brain. The dataset were trained across multiple algorithm such as SVM, XGBoost, Catboost, LightGBM, random forest, etc., by the help of pycaret library to obtain the best metrics and performance.

Results

Preliminary result performed on the smaller set of data (only choose one year post stroke patient, 453 out of 4935 patients). LightGBM algorithm gives the best AUC metrics on the 10 fold cross validation training. The scores are 0.8201 for accuracy, 0.8054 for AUC, 0.1691 for precision, 0.3200 for recall, and 0.2089 for F1. In the current result, we only use ICD and gender/age for the features. We can see the performance in Figure 1. The true positive still less than the false negative. This is to be expected since we use common features between two labels. The Figure 2 shows the optimal threshold is pretty low, 0.27, much lower than default 0.5 for binary classification. Figure 3 shows the distribution of the data based on label. We can see that some positive labels are overlapping (thus, have the same features) with the negative ones.

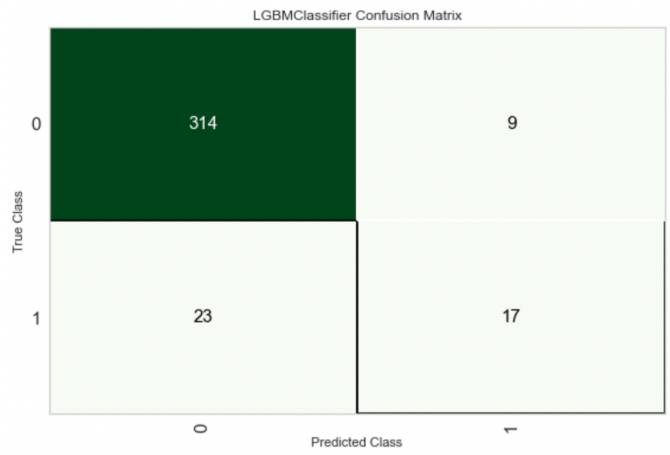

Figure 1. Confusion matrix

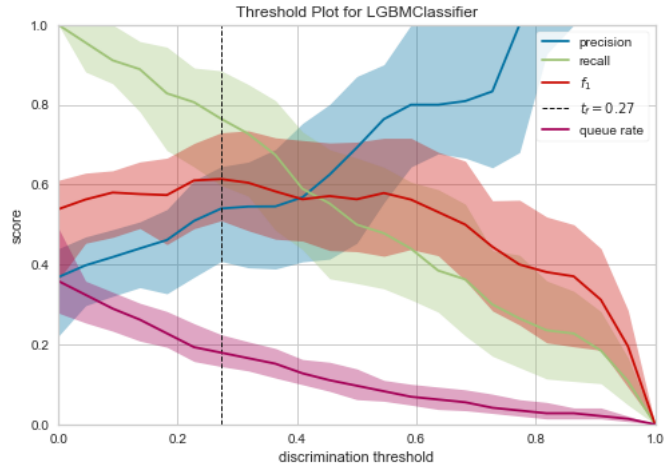

Figure 2. Optimal threshold

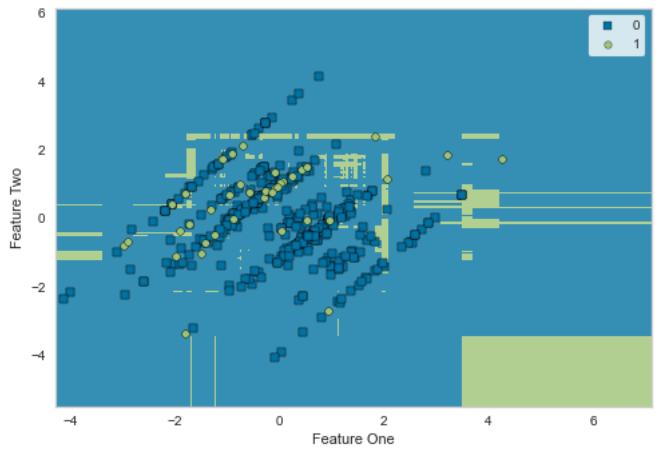

Figure 3. Decision boundary

Conclusion

The current model is able to determine whether a patient will develop cognitive impairment next year or not, though the probability is still very low, using only gender, age and ICD code. Further features engineering will be conducted to improve the performance, especially to increase the true positive and to reduce the false negative numbers.

References

1. del Ser, T., Barba, R., Morin, M. M., Domingo, J., Cemillan, C., Pondal, M., & Vivancos, J. (2005). Evolution of cognitive impairment after stroke and risk factors for delayed progression. *Stroke, 36*(12), 2670-2675
2. Li, C. H., Chang, Y. H., Chou, M. C., Chen, C. H., Ho, B. L., Hsieh, S. W., & Yang, Y. H. (2019). Factors of post-stroke dementia: A nationwide cohort study in Taiwan. *Geriatr Gerontol Int, 19*(8), 815-822. doi:10.1111/ggi.13725