Stroke Prediction and Analysis Using Machine Learning

Nur Ahmadi (n.ahmadi16@imperial.ac.uk)

1 Introduction

Stroke is among major causes of death and long-term disability worldwide. It is of great importance to predict the risk of having stroke for better prevention and early treatment. This brief report presents my attempt to develop a machine learning (ML) model to accurately and quickly predict whether or not a person suffered stroke based on the Kaggle stroke dataset [1]. The dataset contains 43400 rows (samples) and 12 columns (variables), ten of which are used as the input variables and described in Table 1. The remaining two variables are 'id' (excluded from our experiments) which corresponds to patient's ID and 'stroke' which has a value of 0 (no stroke) or 1 (stroke) and is used as the output variable (target). The dataset exhibits a highly imbalanced class with 97.88% corresponding to class 0 and only 2.12% corresponding to class 1. There exists missing values of 3.37% and 30.63% in 'bmi' and 'smoking_status' variables, respectively. In addition to providing performance benchmark across various ML models, this report also examines which features are useful for stroke prediction.

Table 1 | Hyperparameter search spaces during the optimisation of deep learning decoders.

No	Variable	Definition
1	gender	Gender of patient ('Male', 'Female', 'Other')
2	age	Age of patient
3	hypertension	0: no hypertension, 1: suffering from hypertension
4	heart_disease	0: no heart disease, 1: suffering from heart disease
5	ever_married	Whether or not ever married ('Yes', 'No')
6	work_type	Type of occupation ('Private', 'Self-employed', 'Govt_job', 'children', 'Never_worked')
7	Residence_type	Area type of residence ('Urban', 'Rural')
8	avg_glucose_level	Average glucose level (measured after meal)
9	bmi	Body mass index
10	smoking_status	Patient's smoking status ('never smoked', 'formerly smoked', 'smokes')

2 Methods

Samples in all variables associated with the indices of missing values in 'smoking_status' were removed from the data. The missing values in 'bmi' were replaced instead of removed by the average of 'bmi' values to avoid loosing significant number of samples (30.63%). To deal with the highly imbalanced class, Synthetic Minority Oversampling Technique (SMOTE) was applied by synthesising new samples from the minority class. Undersampling technique was not used because it would lead to very small number of samples. The balanced data were then split into training set (80%), validation set (10%), and testing set (10%). Training set was used to train the model; the validation set was used to optimise the model's hyperparameters (if required); the testing set was used for final performance evaluation. All features were standardised to zero mean and unit variance before being fed to the model. A total of 7 ML models were implemented and benchmarked which include: (1) Singular Vector Machine (SVM), (2) Gaussian Naive Bayes (GNB), (3) Logistic Regression (LR), (4) Decision Tree (DT), (5) Random Forest (RF), (6) LightGBM (LGBM) and (7) XGBoost (XGB). To examine which features are useful for prediction, tree- and permutation-based feature importance[2] were used. For performance evaluation, I used two different groups of metrics: (1) sensitivity, specificity, and area under the curve (AUC), and (2) precision, recall, and F1 score. More detailed description of the method is provided in supplementary material "Stroke Prediction and Analysis - Notebook.(pdf/ipynb)".

[1] https://www.kaggle.com/asaumya/healthcare-dataset-stroke-data

[2] Breiman. Random forests. *Mach. learning*. 45, (2001)

3 Results and Discussion

Empirical results showed that the XGB model yielded the highest performance (AUC = 1.00, F1 = 0.98), followed by the RF model (AUC = 0.99, F1 = 0.97) as illustrated in figure 1. Feature importance scores associated with these two highest performing models are shown in figure 2. Across different models and feature importance methods, 'age' was consistently observed as the most important feature for prediction. There was difference in the order of feature importance after 'age'. Tree-based feature importance which was calculated from the training data produced higher degree of difference (figures 2(a) and 2(c)) than permutation-based feature importance which was calculated from the test data (figures 2(b) and 2(d)). The higher degree of difference in tree-based feature importance might be due to the different underlying principle (behaviour) of XGB and RF algorithms. These results may indicate that permutation-based feature importance provide more robust method for evaluating the most useful/informative features.

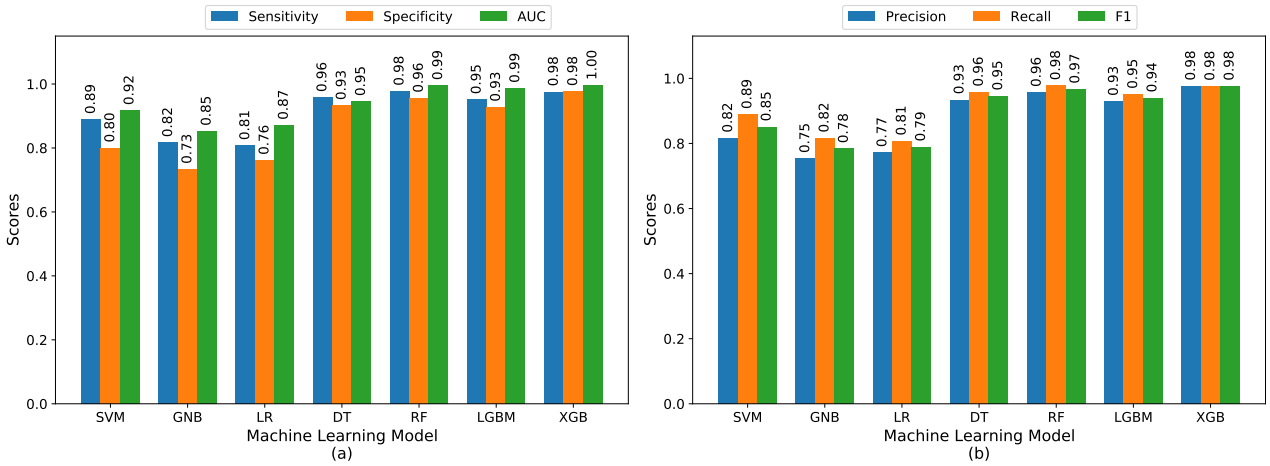

Fig. 1 | Performance benchmark across models using different metrics: (a) sensitivity, specificity, AUC, and (b) precision, recall, F1.

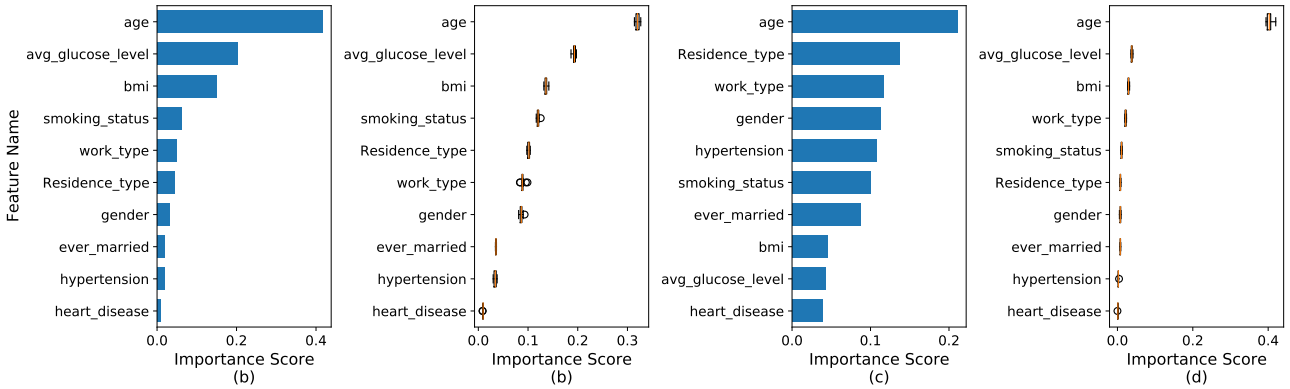

Fig. 2 | Feature importance score. (a)-(b) Tree- and permutation-based feature importance for random forest, respectively. (c)-(d) Tree- and permutation-based feature importance for XGBoost, respectively.

Future work may investigate the effectiveness of other alternative techniques for dealing with highly imbalanced class such Adaptive Synthetic Sampling (ADASYN) oversampling or combination of oversampling and undersampling (e.g. Tomek links) techniques [3]. Additionally, k-fold cross-validation and hyperparameter optimisation could be employed to further improve the accuracy and robustness of the model. In summary, two highest stroke prediction performance were achieved by XGBoost and random forest; three most important features (in descending order) for stroke prediction were 'age', 'avg_glucose_level', and 'bmi'.

[3] https://heartbeat.fritz.ai/resampling-to-properly-handle-imbalanced-datasets-in-machine-learning-64d82c16ceaa