Early Prediction of Sepsis using Random Forest from Clinical Data

Eve Johns
Kaiser Permanente | San Diego, CA

OBJECTIVE

Predict sepsis in intensive care units at least six hours before the onset time of sepsis from Sepsis Challenge Competition hosted by the KP Data Science Team.

DATA

The data use for the challenge is from PhysioNet Computing in Cardiology Challenge 2019:

- 40 clinical variables: 8 vital sign variables, 26 laboratory variables, and 6 demographic variables.
- Total 436,513 electronic health records with 9,400 records exhibits sepsis within 6 hours before the onset time of sepsis.
- Each row contains hourly entries of the 40 clinical variables.

DATA PREPROCESSING

Missing Value Imputation

It is expected that large amount of laboratory results are missing based on the nature of clinical data. Therefore, in this work, a two-step missing value imputation approach were used:

1) Sort the ICU length of stay (ICULOS) by ascending order for each patient (ptn), then impute the missing values with the nearest previous valid value.
2) If all the values in the feature were missing, then replace them by zeros.

Feature Engineering

Instead of using the original 40 features from the raw dataset, I added 10 more features derived from some of the laboratory variables according to Sepsis-3 guidelines listed below:

Feature Name	Description
SOFA_FiO2	SOFA score* for Respiratory System
SOFA_MAP	SOFA score for Mean arterial pressure
SOFA_Bilirubin_total	SOFA score for Liver
SOFA_Platelets	SOFA score for Coagulation
SOFA_Creatinine	SOFA score for Kidneys
SOFA	Total SOFA partial score for the patient row record
SOFA_min	Minimum SOFA for the patient
SOFA_max	Maximum SOFA for the patient
SOFA_diff	SOFA partial score total change for the patient
Organ_Failure	Binary indicator for organ failure based on SOFA_diff increased by at least 2 points

Table1. Augmented features

* Sequential Organ Failure Assessment Score

MODEL TRAINING AND OPTIMIZATION

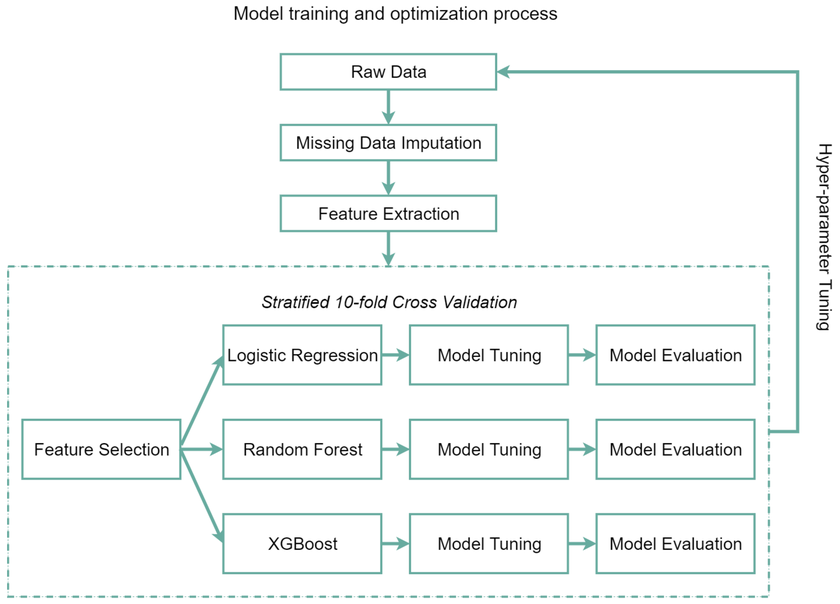

Figure 1. Model training and hyper-parameter tuning process

FEATURE SELECTION

Stratified 10 folds Cross Validation applied on the whole training set prior to the feature selection step. Approximately 90% of the data used for feature selection and classification steps, and the other 10% data was used for validation.

Step Backward Elimination was used for feature selection process using OLS (Ordinary Least Squares) model. After eliminating the features with p-value greater than 0.05, 38 features were kept from the dataset for training.

CLASSIFICATION

The dependent variable of the dataset is categorical so I conducted 3 classification models: Logistic Regression, Random Forest, and XGBoost.

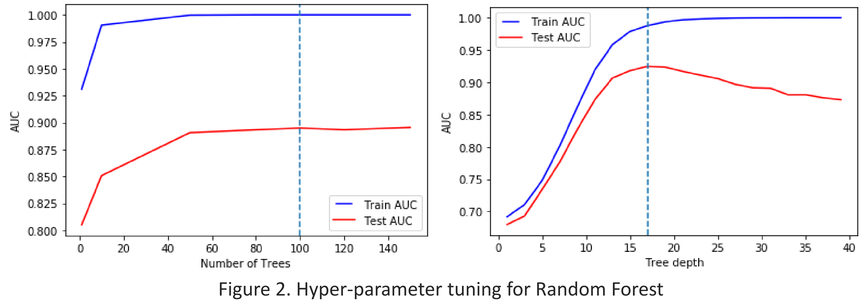

Figure 2. Hyper-parameter tuning for Random Forest

FEATURE IMPORTANCE

Top 10 features from Random Forest model: HospAdmTime (Hours between hospital and ICU Admission), Platelets (Platelet count), WBC (Leukocyte count), Hgb (Hemoglobin), O2Sat (Pulse oximetry %), TroponinI (Troponin I), Unit2 (Surgical ICU), Bilirubin_direct (Direct bilirubin), Calcium, and Magnesium.

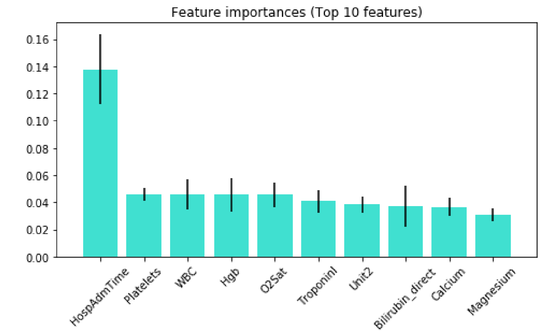

Figure 3. Top 10 important features from Random Forest classification

RESULTS

The Random Forest classifier offers the best performance with respect to accuracy, sensitivity, specificity, and AUC values.

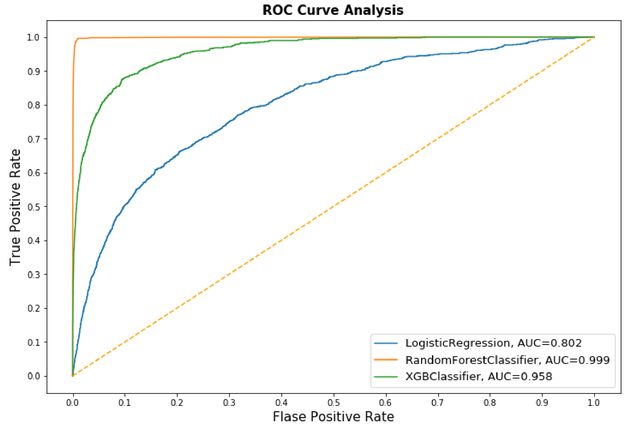

Figure 4. ROC curves comparison among 3 models

Method	Sensitivity	Specificity	Accuracy	AUC
Logistic Regression	0.971	0.757	0.757	0.802
Random Forest	0.995	0.995	0.995	0.999
XGBoost	0.981	0.982	0.982	0.958

Table 2. Models performance comparison