Sepsis Detection by Machine Learning Techniques Poster

Elena Wang, Master of Statistical Science

03/03/2023

Github repository: https://github.com/ElenaW0528/Elena_Wang_Portfolio

Sepsis Detection by Machine Learning Techniques

Elena Wang, Master of Statistical Science
Duke University

03/03/2023

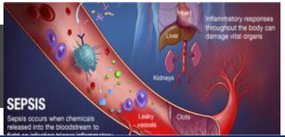
Problem Description and Background

Sepsis: A costly, life-threatening immune disease

Goal: Explore a Machine-Learning Model to precisely classify the Sepsis Condition using a series of vital sign variables and demographic variables to save lives and reduce treatment cost

Databse: We access this EMR data from PhysioNet of 40,336 Patients sent to ICU in 3 hospitals from 2009-2019

PhysioNet
The Research Resource for Complex Physiologic Signals


Heart
Inflammatory responses throughout the body can damage vital organs
Liver
Kidneys
SEPSIS
Sepsis occurs when chemicals released into the bloodstream to
Leaky vessels
Clots

Data Cleaning and Preprocessing

1. Drop the column with every observations missing (ETCO2)
2. Impute missing value in each column by using the previous or next value in dataset order. If both are missing, impute mean value for continuous variable and mode for categorical variables
3. Balance the count of each outcome category by combine 10% of the majority to the minority in each dataset
4. Drop columns that does not convey information of predictions

	Train Set A			Test set B		
	Sepsis Label = 0	Sepsis Label = 1	Total	Sepsis Label = 0	Sepsis Label = 1	Total
Before	18,546(91%)	1,790(9%)	20,336	18,858(94%)	1,142(6%)	2,000
After	1,670(46%)	1,975(54%)	3,645	1,779(59%)	1,249(41%)	3,028

Exploratory Data Analysis (EDA)

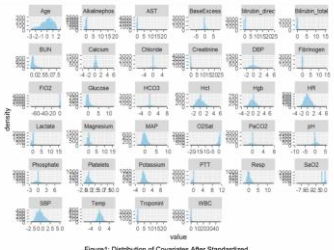


Figure1: Distribution of Covariates After Standardized

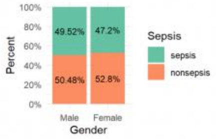


Figure2: Gender vs. Sepsis

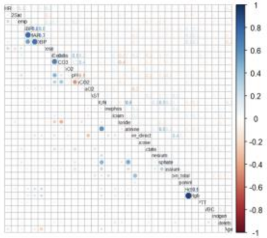


Figure3: Correlation Plot

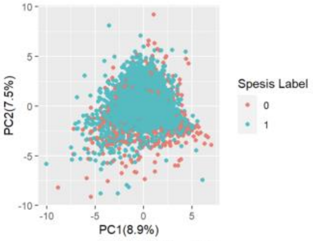


Figure4: PC1 vs PC2

Cluster plot

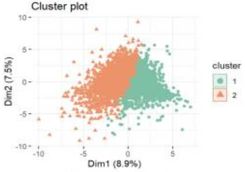


Figure5: K-means clustering

```
      1    2
0  1675 1759
1  1125 2114
```

EDA Summary:

1. Standardization of Covariates
2. Sepsis and non-sepsis patients are evenly distributed in gender
3. Most predictors have low correlations
4. Both Principal Component Analysis and K-means clustering indicate no clear separation among patients

Method Introduction

CV Logistic LASSO Regression to Select 15 Features (lambda: 0.0038)

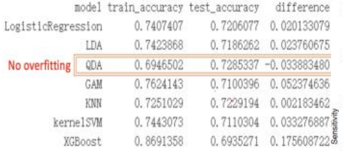
Machine Learning Classification Models:

1. Logistic Regression
2. Linear Discriminant Analysis (LDA)
3. Quadratic Discriminant Analysis (QDA)
4. Logistic Generalized Additive Model (GAM)
5. CV K-nearest-neighbor (KNN) (k=28)
6. CV Polynomial Kernel Support Vector Machine (SVM)
   (gamma: 0.1, cost: 0.1, degree: 3)
7. CV XGboost
   (gamma: 0.15, max depth: 6, eta: 0.1647, minimal child weight: 19)

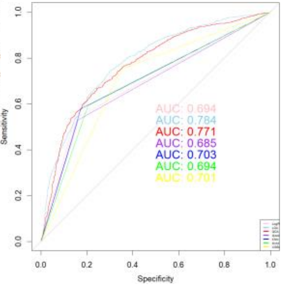
Evaluations of the methods:

- Training versus testing error and accuracy
- Area under the curve and Receiver operating characteristics
- Models' efficiency and variability

Models' Analysis



```
                     model train_accuracy test_accuracy   difference
        LogisticRegression      0.7407407     0.7206077  0.020133079
                       LDA      0.7423868     0.7186262  0.023760675
No overfitting         QDA      0.6946502     0.7285337 -0.033883480
                       GAM      0.7624143     0.7100396  0.052374636
                       KNN      0.7251029     0.7229194  0.002183462
                 kernelSVM      0.7443073     0.7110304  0.033276887
                   XGBoost      0.8691358     0.6935271  0.175608722
```



QDA
✓ no overfitting
✓ highest testing accuracy
✓ higher AUC score
✓ less of computational power and memory

Model Consistency & Variability – GLMM

a. QDA vs. GLMM – potential correlation in patient-specific differences
fixed effect: same predictors
random effect: patients' ID
GLMM: 0.7855 AUC, 0.72 Accuracy
QDA: 0.7711 AUC, 0.7285 Accuracy -> similar
-> consistent fixed effects & small random intercept term

b. Model vs. Model by GLMM – potential correlation in models' differences
response variable: 0-1 indicator for each models' correct prediction
fixed effect: model identities
random effect: patients' ID
-> log odds/ odds ratio of each model making a correct prediction OR difference in the log odds between pairs of models
-> Hypothesis test – ANOVA model -> likelihood ratio test -> p-value

Current Conclusions and Future Work

Real Predictors: Doctors would consider statistically important predictors

System Satisfactory: Good fitness of assumptions and extensive algorithms to avoid problems such as overfitting as well as comprehensive comparisons for evaluation (73% accuracy)

Future Improvements:

1. Device limitations for complicated models such as SVM, XGBoost
2. Simplify XGBoost to balance Bias-Variance and Overfitting
3. Paradox between algorithms and real analysis: feature engineering methods such as lasso and stepwise do not consider clinical significance