Sepsis Detection by Machine Learning Techniques

Elena Wang, Master of Statistical Science

03/03/2023

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

Abstract

Sepsis is a syndromic response to infection and is always life-threatening and costly to treat. Because of the complexity of the early syndrome of sepsis, it is hard to find the early identification of this disease, so it attaches great importance to this early recognition. This report found that the QDA model has the best performance on the prediction among all the machine learning models we tried, with accuracy equal to 72.85%, which would contribute to the application of Sepsis prognosis and treatment.

1. Problem Description

Sepsis is a life-threatening condition that occurs when the body's response to infection causes tissue damage, organ failure, or death *(Singer et al., 2016)*. According to WHO and CDC's data, nearly 270k+ people die from Sepsis, and 24 billion would be spent on the treatment of Sepsis in U.S hospitals each year. It could be life-saving and cost-saving if the Sepsis condition of a patient could be diagnosed in advance by using several vital signs. Our goal is to use Electronic Medical Record(EMR) data which is sourced from ICU patients from 3 separate hospitals, to develop the best machine-learning algorithm in order to predict each patient's sepsis status by using selected vital signs. Our data is provided by the Physionet database, which is a repository of freely-available medical research data. Since this widely-used database is managed by MIT lab and sponsored by NIH, we could consider that it is an authentic and practical real-world medical record and could be used to represent the information of potential Sepsis patients. It is reported that K-nearest-Neighbour (KNN) has satisfied Sepsis predictions*(Perng et al., 2019)* while GB-boost and Random Forest may have better performance*(Ke li et al.)*. Therefore, there are many mature machine learning algorithms and a range of programming software in use for machine learning algorithms in the diagnosis of Sepsis status. After conducting each algorithm, we will compare the test error rate and train error rate as well as compare the accuracy and efficiency for different models. Therefore, We believe our actionable proposed model could be used in the future and real world to help reduce the odds of having Sepsis and save costs.

2. Data Description and Preprocessing

This Sepsis dataset, which is labeled by Sepsis-3 clinical criteria, contains Sepsis conditions, 8 vital physical signs, 26 laboratory values, and 6 demographic features with Gender as categorical and others as numeric for each hour of 40,336 patients' ICU stay in Emory University Hospital and Beth Israel Deaconess Medical Center from 2009 to 2019. The original dataset was separated by Physionet into training A (20,336 subjects, 41 variables) and testing B (20,000 subjects, 41 variables), in which each subject records time-series information to detect Sepsis binary outcome. Some of the variables are described in **Table1**.

SepsisLabel 1 if t>=t_sepsis - 6 and 0 if t<t_sepsis-6, while for non-sepsis patients is 0.	Heart rate (beats/min)	Gender (0 for Female)	Temperature (°C)	systolic BP (mm Hg)	Mean arterial pressure (mm Hg)	Diastolic BP (mm Hg)	Respiration rate (breaths/min)
End tidal carbon dioxide (mm Hg)	ICU length of stay	Bicarbonate (mmol/L)	Fraction of inspired oxygen (%)	Unit1/Unit2 (Administrative identifier for ICU)	carbon dioxide pressure from blood (mm Hg)	Oxygen saturation (blood,%)	Time between hospital and ICU admission

Table 1. Variable Table

2.1 Missing Values and Outliers (details in Python Notebook)

Firstly, Unit1, Unit 2, ICU length of stay, and Time between hospital and ICU admission variables are dropped since these variables did not convey information that could be used to prognose the Sepsis situation. Second, the End-tidal carbon dioxide variable has every observation missing (MCAR) and was dropped from both the train and test set. There are more than 50,000 missing values in each column except for Age, Gender, and Sepsis Label, and each value is based on neighbors for each subject, in which the probability that X is missing depends on the unobserved value of X itself (MNAR) in the biomedical time series view; therefore, we decided to impute the missing values in each column by using the last or next records of the missing values for each patient. If both the previous and next values are also missing, we will replace the mean for continuous variables and the mode for categorical variables.
For the outliers, several observations have values larger or smaller than the mean plus or minus 1.5 interquartile range, respectively. However, we considered that it is normal for patients in ICU to have extreme values for vital signs, so we did not remove outliers to ensure all observations are valid. We also checked for extreme values, like 999, and did not find this value.

2.2 Dealing with Imbalanced Outcome

To avoid the situation that our machine learning model always classified the outcome into the majority group and improve future models' efficiency, it is necessary to make sure that our outcome has a balanced Sepsis situation for this one response variable "SepsisLabel". The original train set and test set had a very imbalanced Sepsis situation. Therefore, we extracted 10% of the majority group for the Sepsis condition in each dataset and combined it with the minority group. The result before and after this process is shown in **Table2**. Afterward, each training and testing outcome is approximately balanced for further modeling. However, it is important to note that balancing the outcome variable in the training and test sets can help prevent certain models from always predicting the majority class, but it may not necessarily improve performance in the real world. It is often a good idea to evaluate the performance of a model on both balanced and imbalanced datasets and to use other metrics such as precision, recall, and F1 score to evaluate performance of the minority class for modeling analysis.

	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%)	20,000
After	1,670(46%)	1,975(54%)	3,645	1,779(59%)	1,249(41%)	3,028

Table 2. Result of Data Pre-processing

After the data pre-processing procedure, we add the patient ID for each patient, and the final dataset used for further analysis is training A(3,645 subjects, 37 variables) and testing B(3,028 subjects, 37 variables).

3. Exploratory Data Analysis

This part will dive into the data performance to provide brief views and assumptions for future modeling. First, standardization is a helpful transformation to be applied as mandatory for some models to work correctly, which can center the data to a mean of 0 and scale to a standard deviation of 1. Below distribution plots after standardization in **Figure1** (Appendix) indicates that all numerical predictors are in the same scale. Some of them, such as Hct, Hgb, HR, etc., look approximately normalized, which means that models requiring strict assumptions could be used later, for example, Logistic Regression with a penalty, KNN, Support Vector Machines (SVM), etc. **Figure2** displays a good separation between a categorical predictor Gender and objective SepsisLabel, which shows that sepsis/non-sepsis patients are evenly distributed in each gender.

Logistic regression requires basic assumptions such as the independence of errors, absence of multicollinearity, etc, which needs to check the correlation between the continuous variables. After correlation analysis, it can be identified that most predictors have a low correlation with each other, which satisfies the requirement (**Figure3,** Appendix).

To observe the performance of response variables, we executed some unsupervised machine learning techniques. Principal Component Analysis (PCA) method will reduce the dimension of the features down to 2 and observe whether there is any evidence of clustering for sepsis and non-sepsis patients. **Figure4** observes that there is no clear separation between the patients, and the first PC only explains 8.91% of the variance, which gives intuitive non-linear algorithms such as kernel SVM. K-means visualization (**Figure5**) also can approve non-linear findings with k=2, and there are 1759 non-sepsis vs. 2114 sepsis patients in cluster1 while there are 1675 non-sepsis vs. 1125 sepsis patients in cluster2, which means that cluster1 could be labeled as sepsis and cluster2 could be labeled as non-sepsis.

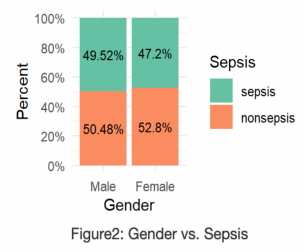

Figure2: Gender vs. Sepsis

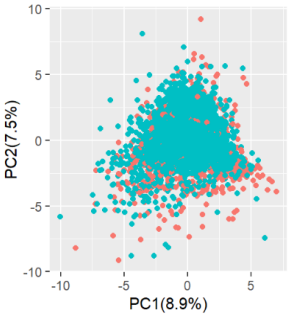

Figure4: PC1 vs PC2

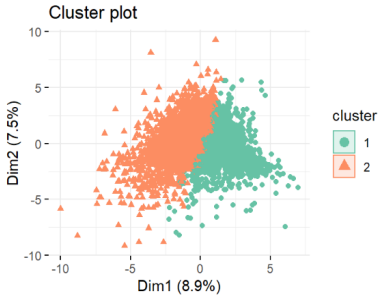

Figure5: K-means clustering

4. Core Methods:

4.1. Cross-Validated(CV) Logistic LASSO Regression

Lasso penalty will estimate the regression coefficients by maximizing the log-likelihood function with the constraint that the sum of the absolute values of the magnitude of coefficients (L1 norm). We first conduct Logistic Lasso Regression to do feature engineering for the classification problem since Lasso penalty is able to shrink relative trivial coefficients shrink to exactly 0 to avoid overfitting in further modeling and logit link (more in method2) could categorize the predicted probabilities for a binary classifier. We used cross-validation in the training dataset to tune the best lambda parameter=0.0038, which minimizes the error rate to avoid overfitting problems and obtain precise predictions. The selected 15 important variables: O2Sat, Temp, MAP, Resp, FiO2, pH, BUN, Lactate, Magnesium, Phosphate, Hgb, PTT, WBC, Fibrinogen, Platelets.

4.2 Logistic Regression

The first model Logistic Lasso Regression is based on Logistic Regression, and simple logistic regression has similar algorithms without penalty. Here we apply logistic regression to the reduced covariates to compare the performance.

4.3 Linear Discriminant Analysis (LDA)

The goal of LDA is to find a linear combination of features that characterizes or separates sepsis/non-sepsis objects in maximization by projecting onto new axes created. From the above EDA, the dataset has good assumptions for LDA, such as homogeneity of variance, no multicollinearity, and multivariate normality. The method first measures the distances between the centroid of each category and the centroid of all data, afterward, maximizing the distance between each category and the control point while minimizing the scatter for each category: $(u_1 - u_2)^2/(s_1^2 + s_2^2)$,

$(u_1 - u_2)^2$ *is distance between means,* $(s_1^2 + s_2^2)$ *is variation within each category*

4.4 Quadratic Discriminant Analysis (QDA)

We used a common covariate matrix to model covariance matrices for all classes in LDA while it is possible to allow each class to have its own covariance matrix, and everything else keep the same, which leads to QDA. This stimulates us to try this model since each class's covariance matrix can differ.

4.5 Logistic Generalized additive model (GAM)

Figure7 (Appendix) randomly shows the scatter plots of predictors and indicates our predictors are nonlinearities, which allows us to try the GAM model to retain the additive structure of linear models. As with Logistic Regression, we could use a logit link to deal with classification problems. Moreover, the dataset still satisfies the model assumptions, such as independence between dependent variables and consistency variance. The mathematical way could be written as follows:

$log(p(X)/(1 - p(X))) = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}) \ldots + f_p(x_{ip})$

4.6 K-Nearest Neighbors (KNN) with CV

KNN algorithm is a simple, easy-to-implement supervised method that can be solved for classification problems. The core characteristic of KNN is that the closer two given points are to each other, and the

more related and similar they are, and it has less strict assumptions than linear models. Here we tune parameter K with a cross-validation method in the training set and select the best one (k=28) to evaluate the testing set.

4.7 Kernel Support Vector Machine (SVM) with CV

Above visualization illustrates that the feature points are non-linear separated, which guides us to implement "polynomial" kernel SVM, especially for the particular case where one class is placed in an inner region surrounded from all sides by points of another class. We could tune gamma, cost, and degree parameters for the polynomial kernel; however, this model efficiency is extremely slow and time-consuming, so we only try several parameters in the cross-validation process here. The result gives us the best parameters are gamma=0.1, cost=0.1, and degree=3.

4.8 XGBoost with CV

Firstly, We continuously split the features to grow a decision tree. In the next turn, the error of the tree in the previous round will be considered when growing a new tree. Each turn will grow one tree, and after k trees are constructed, the scores for each subject will be gained by adding the scores for k trees. This score will be considered as the prediction of this subject. Loads of parameters could be tuned in this model, such as minimum loss reduction (gamma), learning rate (eta), depth of each tree, etc. We tried 5-fold cv and decided to use gamma=0.15, maximal depth=6, eta=0.1647, and minimal child weight=19.

5. 1 Models' Analysis

First, the models could be compared by training and testing accuracy for classification problems calculated by the number of correct predictions divided by the total number of predictions made. Comparison between the accuracy of training and testing set also could explore the overfitting problem and relative bias-variance tradeoff occurrence, where the model with high variance pays much attention to training data and does not generalize on the data which have not seen before; as a result, such models perform very well on training data but has higher error rates on the test data. The below table compares the accuracy across various models we have introduced above, and the result shows that except for QDA, other models have a bit of an overfitting issue, in which the training accuracy is greater than testing accuracy (difference>0). Nevertheless, it is good to see that the difference is not large in most models, which concludes that we do not have significant overfitting problems. XGBoost is a scalable, distributed gradient-boosted decision tree and provides loss update in each iteration; therefore, the model is much more complicated than others, which could cause overfitting problems. Secondly, the ROC curve is a graph showing the performance of a classification model at all classification thresholds, and the relative AUC measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1), which means that higher AUC provides better models. **Figure8** indicates that LDA has the highest AUC (0.784) while GAM has the lowest one (0.685). Finally, Model efficiency is also necessary for memory storage and computational power evaluation. As models' analysis above, kernel SVM and XGBoost are more complicated and time-consuming, although XGBoost is designed for leverages of hardware and software. In conclusion, we would choose QDA as the final model with relatively high test accuracy (72.85%, shown in **Table 3**) and AUC (0.771).

Model	Train accuracy	Test accuracy	Difference	Overfit
Logistic regression	74.07%	72.06%	2.01%	yes
LDA	74.24%	71.86%	2.38%	yes
QDA	69.46%	72.85%	- 3.39%	no
GAM	76.24%	71.00%	5.24%	yes
KNN	72.51%	72.29%	0.22%	yes
Kernel SVM	74.44%	71.10%	3.33%	yes
XGBoost	86.91%	69.34%	17.56%	yes

Table 3. Accuracy for Each Method

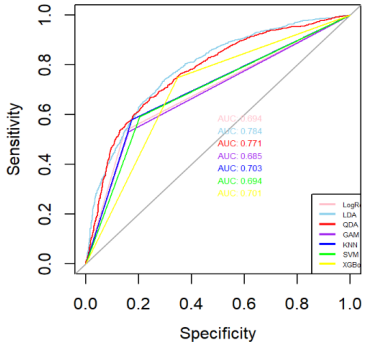

Figure8: ROC-AUC Comparison

5.2 Model Consistency & Variability - Generalized Linear Mixed Model (GLMM)

A GLMM allows for the modeling of the variability within the patient population by including patient-specific random effects in the model. This approach takes into account the potential correlation within patients and allows for the estimation of the variation in patient responses to the predictor variables. To assess the performance of the QDA model in the GLMM framework, GLMM is fitted with the same predictor variables as the QDA model but with a random intercept term for patients (SubjectID). This random intercept term will capture any patient-specific differences that are not accounted for by the fixed predictor variables. The performance of the QDA model in the GLMM framework can be evaluated by examining the estimated coefficients for the fixed effects and the variance of the random intercept term. Here the fitting GLMM model gave 0.7855 AUC compared to QDA's 0.7711, and accuracy is given by 0.72 compared by 0.7285 in QDA, which means that the estimated coefficients for the fixed effects remain consistent, and the variance of the random intercept term is small and suggests the QDA model is still performing well in the presence of patient-specific differences.

However, the accuracy and AUC scores of all models are not significantly different, in which GLMM also can be considered to evaluate the performance of different models. Since the values for the models are repeated measurements made on each test set subject, "fixed effects" can refer to variables that are assumed to have a fixed relationship with the response variable, and here the response variable is the 0-1 indicator variable for each model used to predict the test set, and the fixed effects are the model identities (i.e., the different models being compared) with the individual ID as random effects. The model would output estimates for the fixed effect coefficients, which represent the log odds of each model making a correct prediction, and exponentiate these coefficients to obtain the odds ratios. Additionally, the difference in the log odds estimates between pairs of models can also be exponentiated. Moreover, the hypothesis test between models' accuracy can be compared to their coefficients by the likelihood ratio test in the ANOVA model given by the p-value of the test indicates the significance of the difference between the two models.

6. Current Conclusions and Future Work

Among all the vital signs recorded, O2Sat, Temp, MAP, Resp, FiO2, pH, BUN, Lactate, Magnesium, Phosphate, Hgb, PTT, WBC, Fibrinogen, Platelets are considered statistically important for the prediction of Sepsis status for ICU patients. This indicates that these signals could be the first to consider if doctors would like to identify whether an ICU patient has a high probability of developing Sepsis or not. In terms of system satisfaction, we evaluated the models by accuracy, AUC-ROC curve, model efficiency and models' variability and concluded that QDA is the best model for this project with high consistency analyzed by GLMM model. The accuracy of sepsis detection could achieve around 73%, which is a good reference in real clinical analysis. Moreover, some limitations could be considered and improved in the future: 1. Due to device limitations, we are not able to tune many parameters in complex models such as kernel SVM and XGBoost, and larger CPU and GPU on the Cluster should be considered later. 2. We have an overfitting problem in XGBoost, but we could simplify the model by a random forest or decision tree. In this case, we should be careful about the model's efficiency. 3. Feature engineering methods would possibly focus less on the predictors that we are interested in the real clinic analysis, such as lasso may shrink those coefficients to 0 or not significant in stepwise, which leads to a paradox in algorithms vs. reality, and should consider the clinical significance in the real analysis.

Appendix

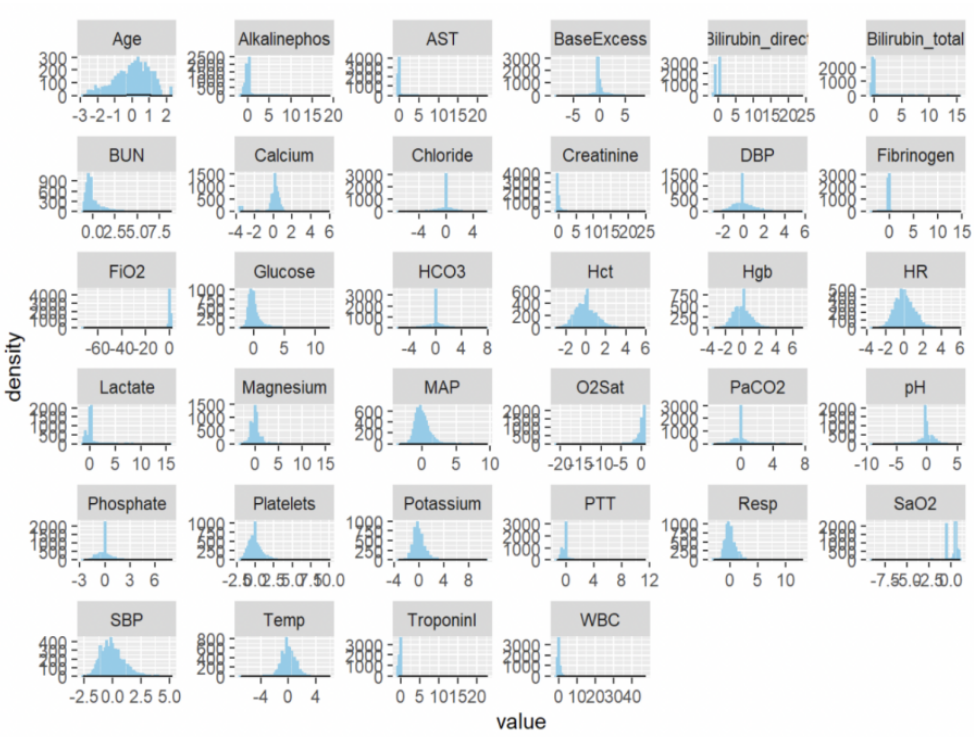

Figure1: Distribution of Covariates After Standardized

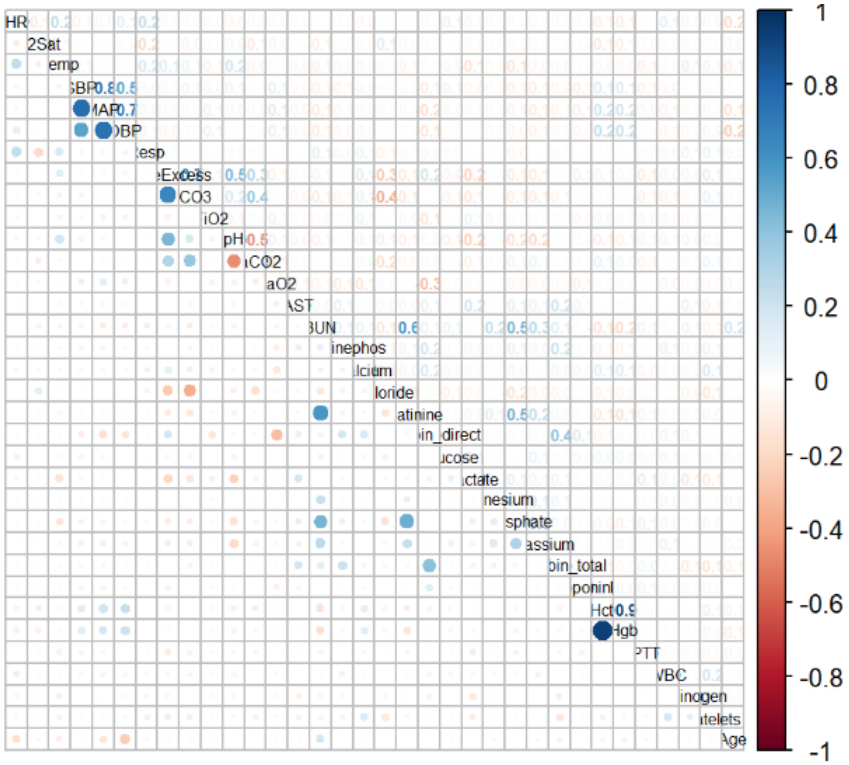

Figure3: Correlation Plot

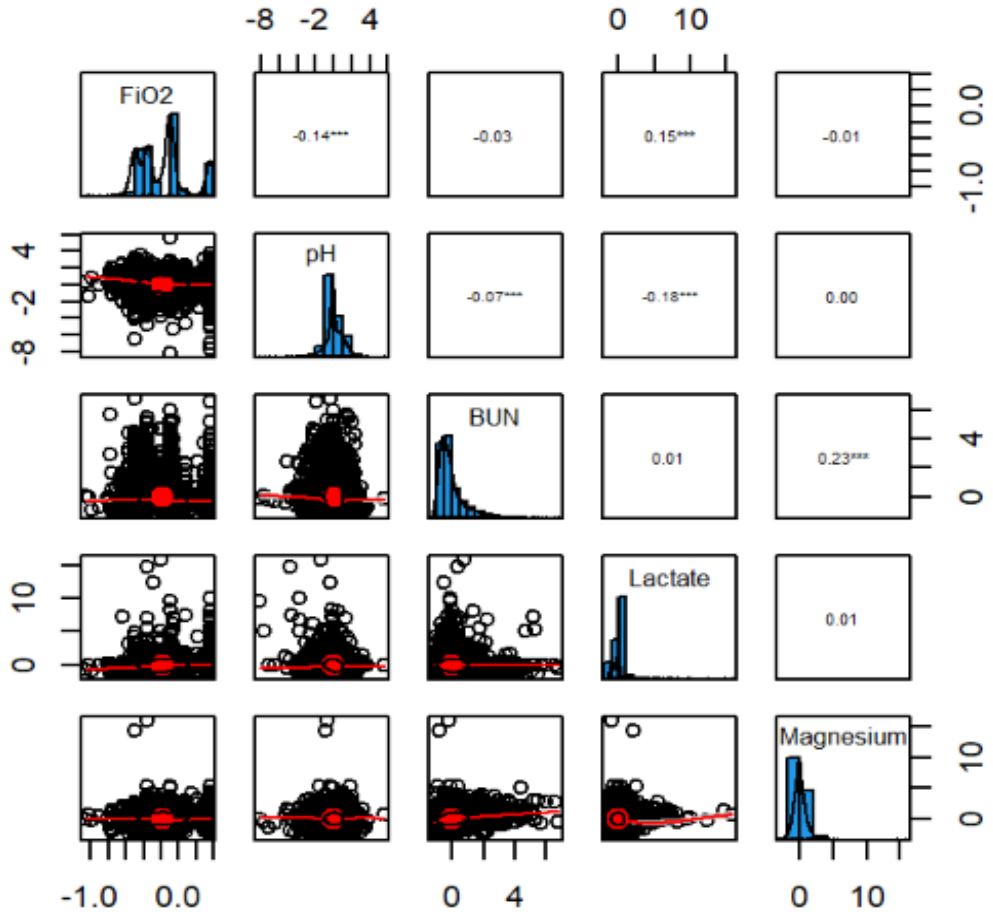

Figure7: GAM Pre-View

Reference

[1] Singer, M., Deutschman, et.al,.(2016). The third international consensus definitions for sepsis and septic shock (sepsis-3). JAMA, 315(8), 801. https://doi.org/10.1001/jama.2016.0287

[2] Deng, H.-F., et.al., (2022). Evaluating machine learning models for sepsis prediction: A systematic review of methodologies. IScience, 25(1), 103651. https://doi.org/10.1016/j.isci.2021.103651

[3] Centers for Disease Control and Prevention. (2022, August 9). What is sepsis? Centers for Disease Control and Prevention. Retrieved November 28, 2022, from https://www.cdc.gov/sepsis/what-is-sepsis.html

[4] World Health Organization. (n.d.). Sepsis. World Health Organization. Retrieved November 28, 2022, from https://www.who.int/news-room/fact-sheets/detail/sepsis

[5] Zhao, X., Shen, W., & Wang, G. (2021). Early prediction of sepsis based on machine learning algorithm. Computational Intelligence and Neuroscience, 2021, 1–13. https://doi.org/10.1155/2021/6522633

[6] Goh, K. H., Wang, L., Yeow, A. Y., Poh, H., Li, K., Yeow, J. J., & Tan, G. Y. (2021). Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nature Communications, 12(1). https://doi.org/10.1038/s41467-021-20910-4

[7] Kim, H. I., & Park, S. (2019). Sepsis: Early recognition and optimized treatment. Tuberculosis and Respiratory Diseases, 82(1), 6. https://doi.org/10.4046/trd.2018.0041

[8] Nemati, S., Holder, A., Razmi, F., Stanley, M. D., Clifford, G. D., & Buchman, T. G. (2018). An interpretable machine learning model for accurate prediction of sepsis in the ICU. Critical Care Medicine, 46(4), 547–553. https://doi.org/10.1097/ccm.0000000000002936