Final Report
Predicting Sepsis in ICU Patients

By Aisling Casey – 05/19/2021

Table of Contents

Overview

Problem Statement

Sepsis is a leading cause of death in US hospital patients. Sepsis occurs when.. "when the body's response to infection causes tissue damage, organ failure, or death"[1][3]. Prompt intervention in sepsis patients can improve the likelihood of their condition improving significantly, while unnecessary treatment in non-sepsis patients drains limited hospital resources.

The purpose of this project was to create a model that could predict if an ICU patient would develop sepsis in the next six hours using hourly vital sign, lab and demographic data. Using the data science method, I created a model for predicting if sepsis would occur within the next 3 hours, as well as a model that would predict if the patient had sepsis or not.

Results

Unfortunately, I was not able to create a useful model from this problem. The logistic regression, decision tree (random forest & gradient boost) and SVM models had very little skill in properly classifying unseen sepsis patients, both in the sepsis and pre-sepsis stages. Suspected reasons for this and next steps are discussed in the Modeling & Future Directions section of this report.

Methods

Data Wrangling

The data came from a data science competition[1], which provided the hourly data for over 40,000 different ICU patients that included vital sign (e.g. heart rate), lab and demographic data, along with a sepsis label indicating if the patient had sepsis or not at that time.

Variable Type Column #	Vital Signs 1-8	Laboratory Values 9-34	Demographics 35-40	Sepsis Label 41
t_0	…	…	…	…
t_1	…	…	…	…
….	…	…	…	…
t_n	…	…	…	…

Figure 1: Data structure of one patient's dataset; there are as many rows as hours the patient spent in the ICU. There are 40,366 patient datasets in total.

Exploratory Data Analysis

Of 40336 patients available in the data set, 7.27% develop sepsis at some point during their hospital stay. Of the 1552210 data points in the data set (each one representing an hour) 1.8% occur while a patient has sepsis. So, this is a very imbalanced data set.

Vital Signs

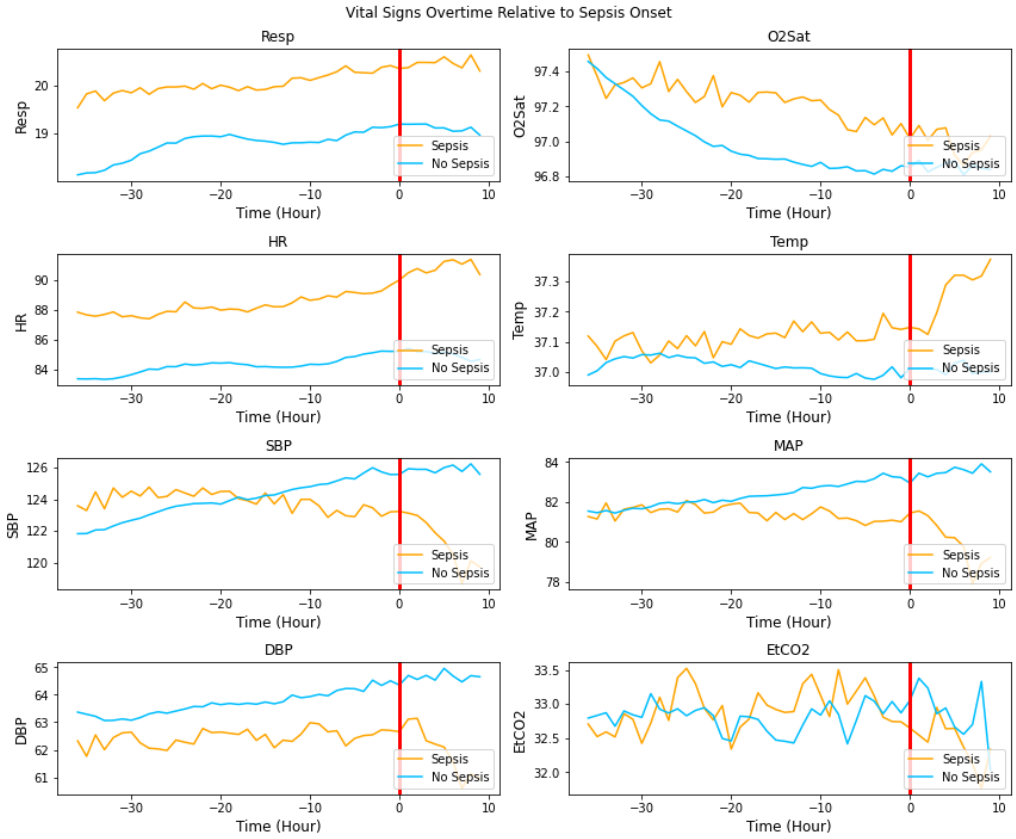

Figure 2: Average vital sign values relative to sepsis onset (red line) for sepsis patients compared to a random selection of values for non sepsis patients.

As seen in figure 2, there are some clear differences in the average time course of vital signs between sepsis and non-sepsis patients. In particular, sepsis patients have consistently higher average respiration & heart rate at any point compared to non-sepsis patients. Decrease in blood pressure is clearly seen in all sepsis patients post sepsis onset; pre-sepsis onset, SBP and MAP shows the strongest downward trend a few hours out, with DBP not as clear. O2Sat is a bit puzzling, as you'd expect it to clearly go down for sepsis patients, more so than non-sepsis patients, but the non-sepsis patients have a starker downward trend.

Lab Values

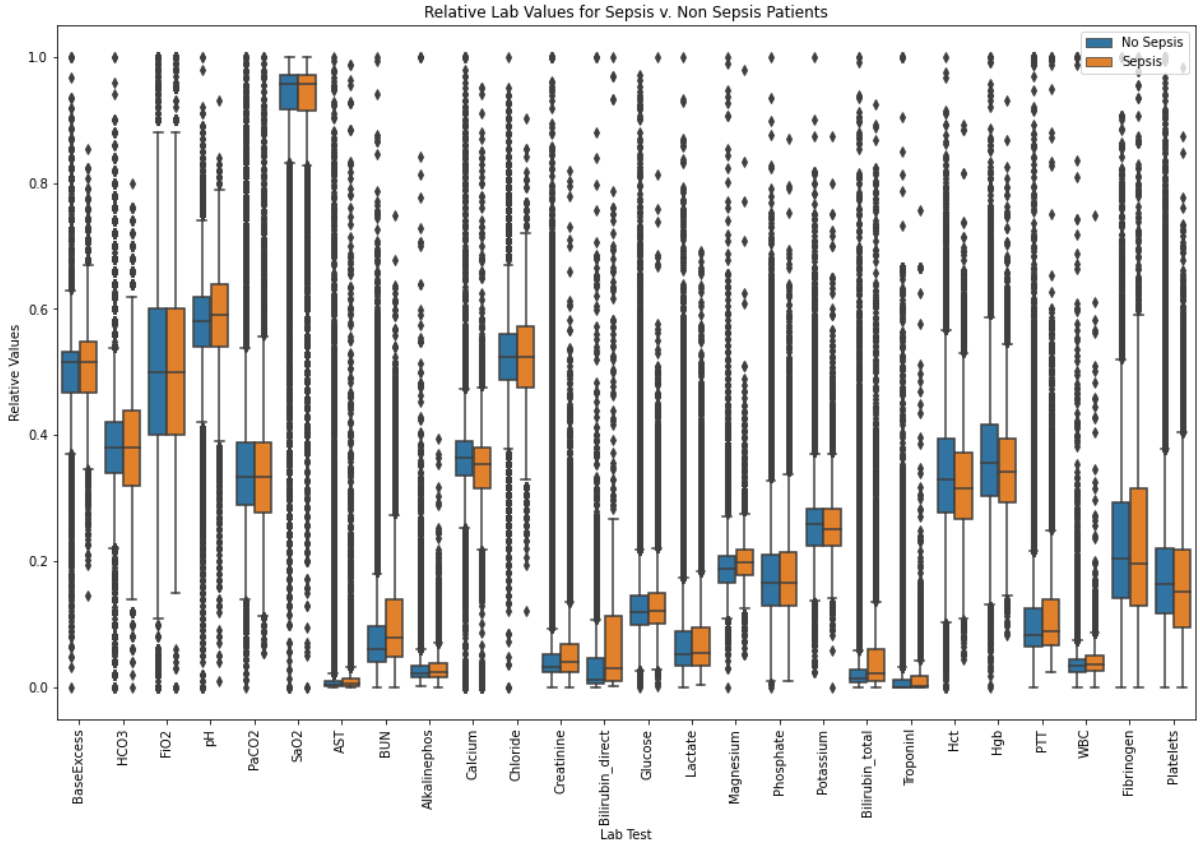

Figure 3: Distribution of relative lab values for sepsis vs. non sepsis patients.

As seen in figure 3, almost no clear pattern exists between the groups of any of lab values. For example, while there is a visual difference in distributions between groups for Bilirubin direct, that could easily be attributed to the small number of lab values available (i.e. noise). Bilirubin_total is the only exception, having lab values for over 30% of patients, and a seemingly different distribution.

Demographic Data

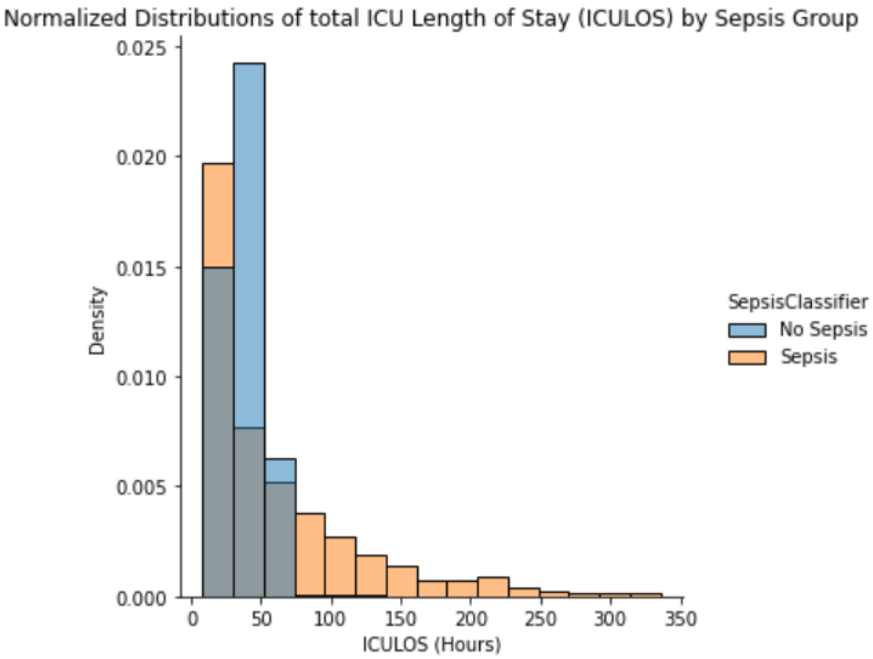

Figure 4: Normalized Distribution of ICU length of stay, sepsis vs. non-sepsis patients

Marginal differences existed between most demographic data, such as 7.71% of males having sepsis compared to 6.71% of females. By far the largest difference was ICU length of stay; a greater share of Sepsis patients stayed longer than 50 hours in the sepsis vs. non-sepsis patients, as seen in figure 4.

Preprocessing

The following changes were made to the data to prepare it for modeling, in this order:

1. Removal of outliers in laboratory & vital sign data
2. Interpolation of vital sign data
3. Forward filling of lab value data
4. Filling in remaining vital sign & lab data with median data
5. Addition of indicator variable column for lab values (if that lab value was present for that patient currently or anytime moving forward)
6. Addition of change in vital sign columns, for previous one, two and three hours.
7. Addition of pre-sepsis label; 1 if in 3 hour period before sepsis onset, 0 otherwise.
8. Log transform of skewed distributions
9. Prevention of data leakage in test & train sets.
10. Min/Max scaling based on test dataset
11. SMOTE - Synthetic Minority Oversampling Technique (for some models)

Modeling

Logistic regression, random forest, gradient boost and SVM models were created for both the sepsis and pre sepsis data. Randomized cross validation was used to determine optimal hyperparameters; sequential folding was used to prevent data leakage within the folds.

The ROC curves for all models and testing data had AUC of 0.75 or higher and tended to have similar values for the training data. Sample ROC curves are shown below, for the logistic regression models of sepsis label and pre sepsis.

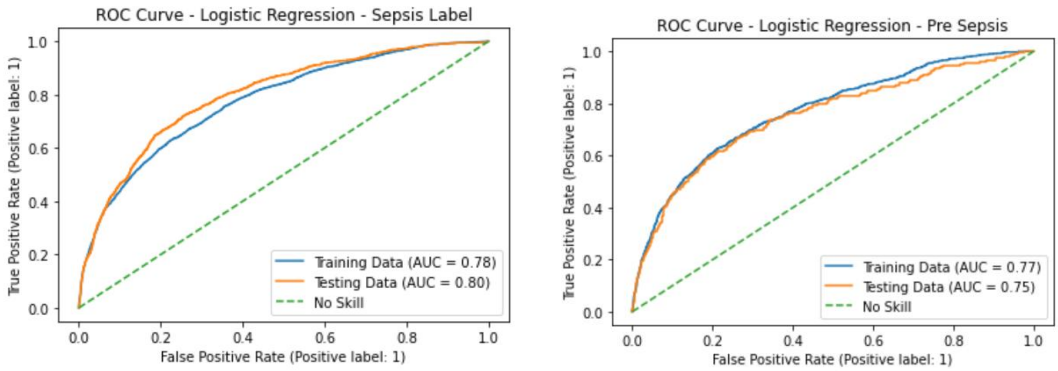

Figure 5: ROC Curves for the logistic regression model of sepsis label (left) and pre-sepsis (right)

But the precision recall curves tell a much different story, because this is such an imbalanced data set. The AUC of precision-recall curves for all testing data is below 0.1. This indicates almost no skill in any of the models. Sample precision-recall curves are shown below, again for the logistic regression models of sepsis label and pre sepsis.

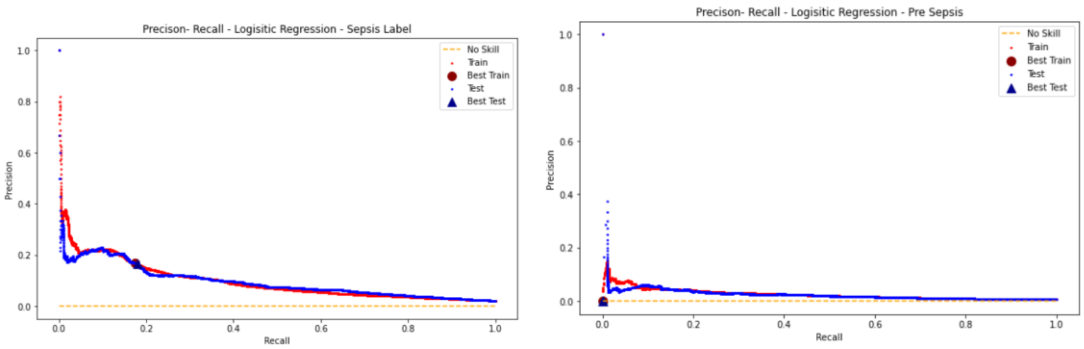

Figure 6: Precision-recall curves for the logistic regression model of sepsis label (left) and pre-sepsis (right)

As shown in the table below, testing data accuracy is quite high for the models, because they largely correctly classify data points as being non-sepsis and non-pre-sepsis. But the precision and recall are extremely low, meaning the models largely fail at properly classifying sepsis or pre sepsis patients.

Model Results Testing Data	Logistic Regression		Random Forest		Gradient Boost		SVM	
	Accuracy	Sepsis F1	Accuracy	Sepsis F1	Accuracy	Sepsis F1	Accuracy	Sepsis F1
Sepsis Label	.95	.16	0.96	0.17	0.96	0.18	0.84	0.11
Pre Sepsis	.99	.02	0.99	0.04	0.99	0.01	x	x
Pre Sepsis SMOTE	0.64	0.06	x	x	0.81	0.03	x	x

So why do the models perform so poorly? Insight can be found in plotting the feature importance of the tree models. An example is shown in figure 7, the gradient boost model of pre-sepsis. ICU length of stay (ICULOS) is by far the most important feature in predicting sepsis, according to the model. Nothing else comes close. So there does not appear to be a latent pattern in lab or vital sign variables that the models can use to distinguish sepsis from non sepsis patients in the data's current form.

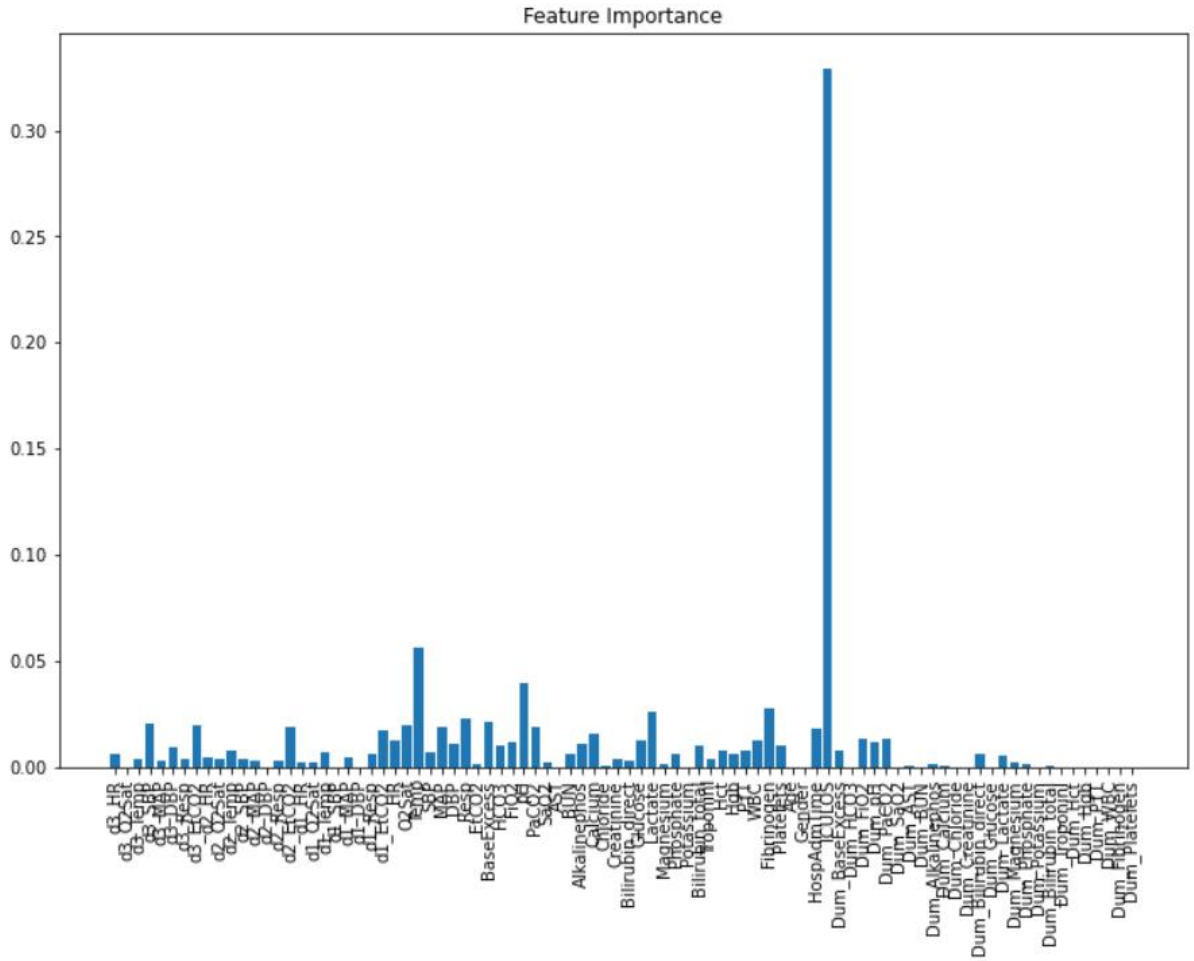

Figure 7: Feature importance in the gradient boost model for pre-sepsis

SMOTE was employed to balance the classes; the results of the gradient boost model of SMOTE is shown in figure 8. The model clearly overfits to the training data, so SMOTE is not useful in this case.

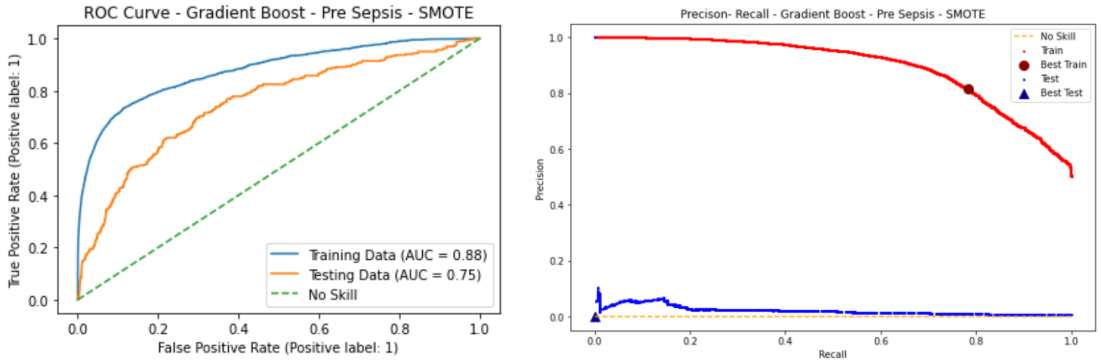

Figure 8: ROC and precision recall curves of the gradient boost model trained on SMOTE data.

Future Directions

During exploratory data analysis, there was no obvious pattern in the difference in lab values between sepsis and non-sepsis patients. There were some differences in the vital signs; but given the imbalanced nature of the classes, this would not necessarily be enough difference. I hoped that by modeling the data with machine learning algorithms such as XGBoost and SVM, latent relationships in the variables would be uncovered that could make classification possible. Unfortunately these models performed poorly at correctly classifying sepsis patients.

Because this data was apart of a competition, I sought the methods of other groups that were more successful in their model creation. As discussed in[4], one group was able to build a model after heterogenous events embedding and attentional multihead aggregation. So better classification is possible, but methods more advanced than I am currently able to employ are needed.

Citations

[1] Reyna, M., Josef, C., Jeter, R., Shashikumar, S., Moody, B., Westover, M. B., Sharma, A., Nemati, S., & Clifford, G. (2019). Early Prediction of Sepsis from Clinical Data -- the PhysioNet Computing in Cardiology Challenge 2019 (version 1.0.0). *PhysioNet*. https://doi.org/10.13026/v64v-d857.

[2] Reyna MA, Josef CS, Jeter R, Shashikumar SP, Westover MB, Nemati S, Clifford GD, Sharma A. Early Prediction of Sepsis From Clinical Data: The PhysioNet/Computing in Cardiology Challenge. Critical Care Medicine 48 2: 210-217 (2019). https://doi.org/10.1097/CCM.0000000000004145

[3] Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche JD, Coopersmith CM, Hotchkiss RS, Levy MM, Marshall JC, Martin GS, Opal SM, Rubenfeld GD, van der Poll T, Vincent JL, Angus DC. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016 Feb 23;315(8):801-10. doi: 10.1001/jama.2016.0287. PMID: 26903338; PMCID: PMC4968574.

[4] Liu, L., Wu, H., Wang, Z., Lieu, Z, Zhang, M. Early Prediction of Sepsis From Clinical Data via Heterogeneous Event Aggregation. arXiv.org. 2019 Oct 12; 1910.06792v1.