By Aisling Casey
Capstone Presentation
Springboard School of Data
May 2021

Predicting Sepsis in ICU Patients

Earlier intervention for better health outcomes

Problem Identification

Context
Sepsis is a leading cause of death in US hospital patients

Solution Space
Classification of pre and non sepsis patients

Success Criteria
Accurate classification of pre and non sepsis patients in test set

Data Source
Hourly data from 40,336 ICU patients in 2 hospitals

Problem Statement: Early intervention in sepsis patients can lead to better health outcomes. Is it possible to predict sepsis in ICU patients hours before clinical diagnosis?

Data Structure & Source

Time (Hours)	Vital Signs 1-8	Laboratory Values 9-34	Demographics 35-40	Sepsis Label 41
t_0	…	…	…	0
t_1	…	…	…	1
….	…	…	…	0
t_n	…	…	…	0

Data made available by Physionet Computing in Cardiology Challenge 2019

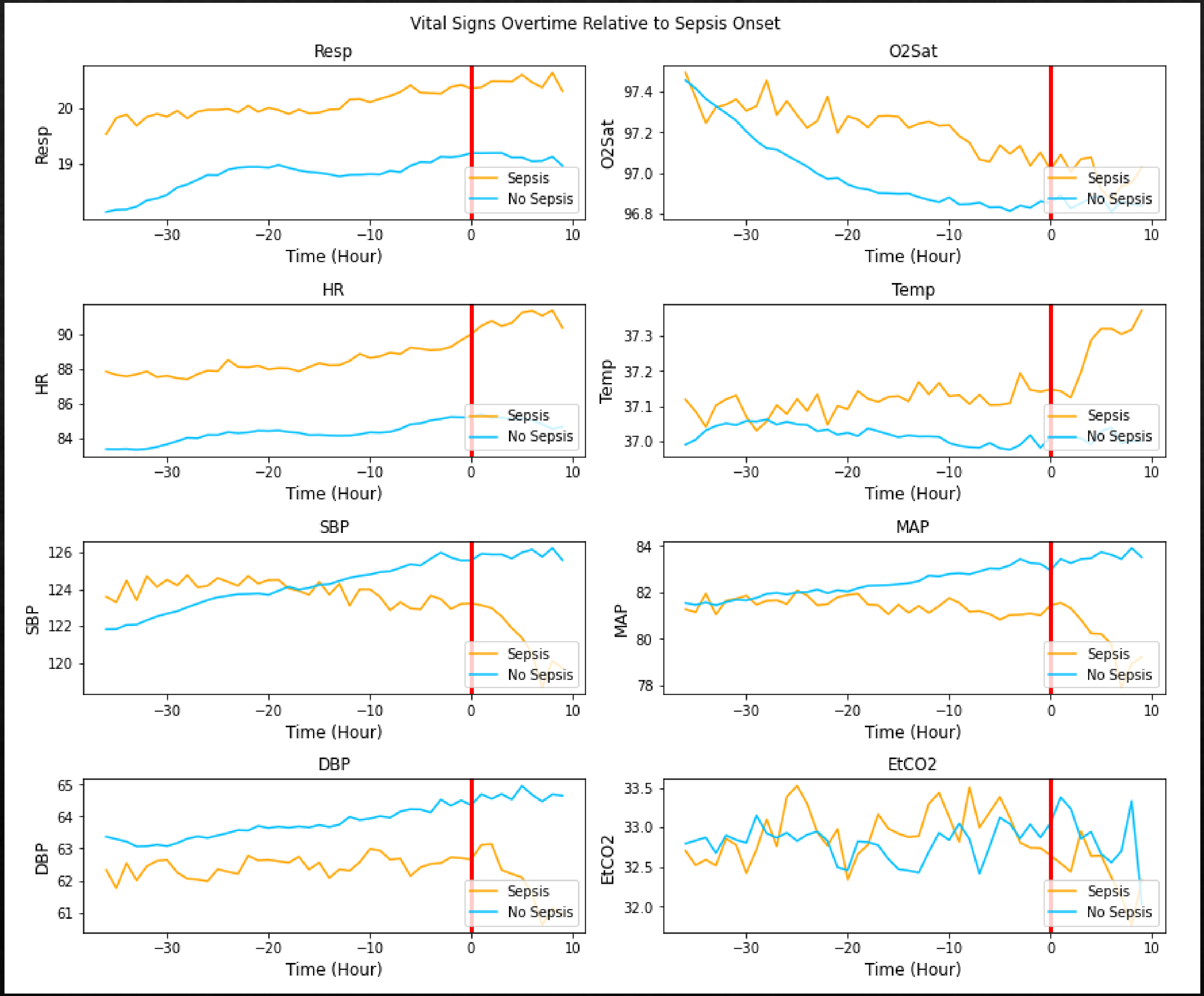
Second classifier column, pre-sepsis, added

Sepsis Prevalence

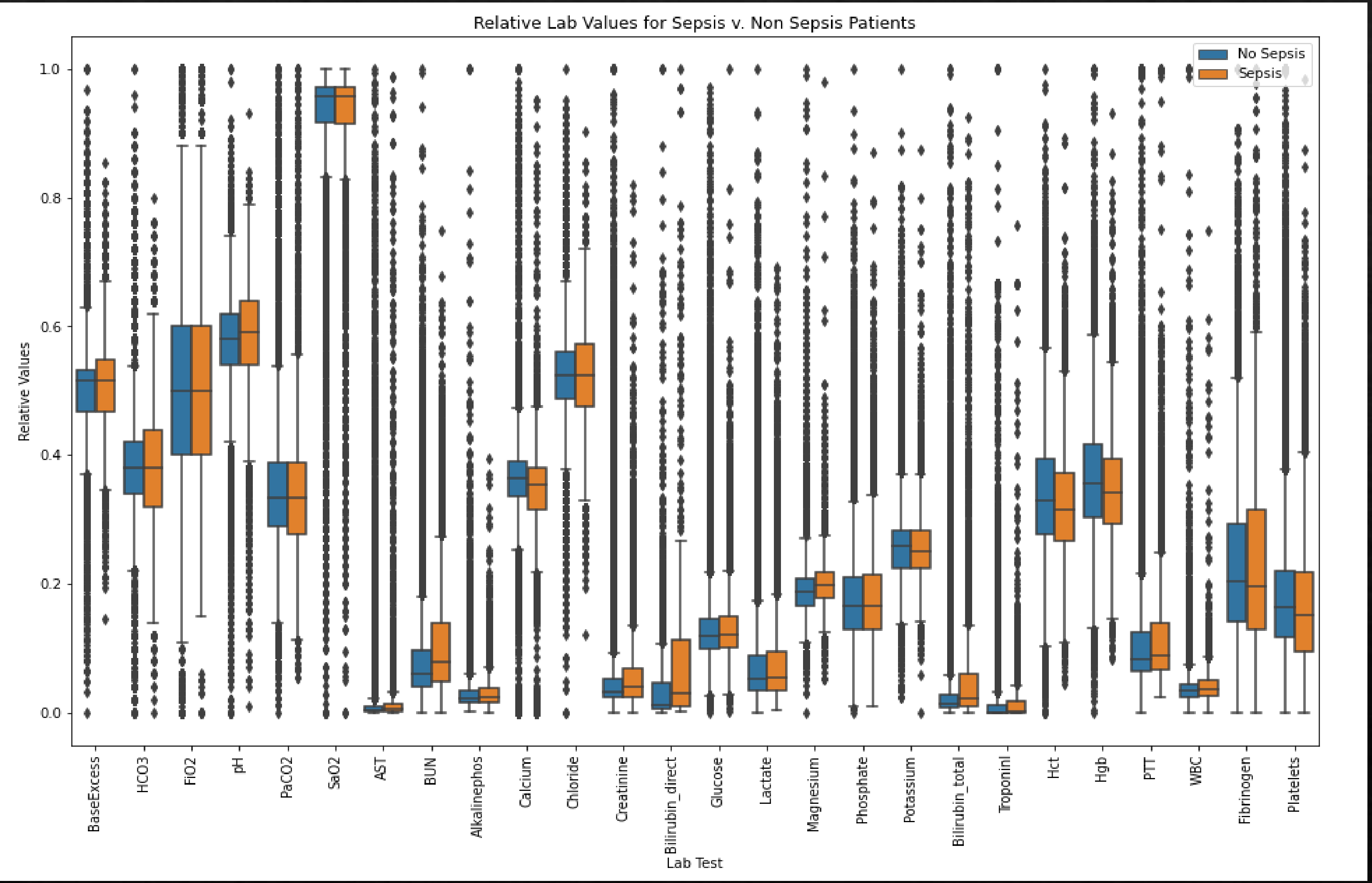
Of **40,336** patients available in the data set, **7.27%** develop sepsis at some point during their hospital stay.

Of the **1,552,210** data points in the data set, each representing an hour, **1.8%** occur while a patient has sepsis.

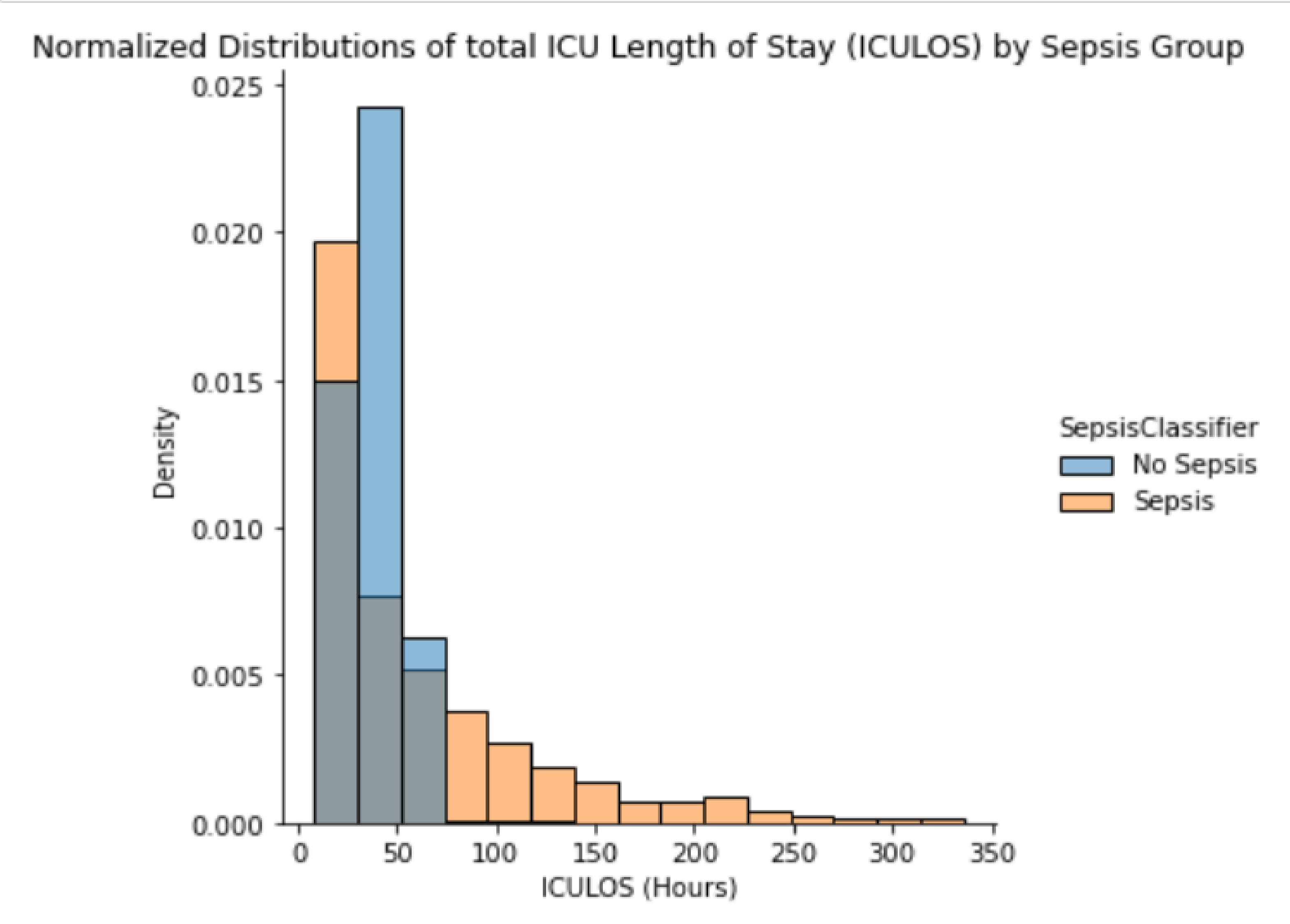
Exploratory Data Analysis: Vital Signs



Exploratory Data Analysis: Normalized Lab Values



Exploratory Data Analysis: ICU Length of Stay



Feature Engineering

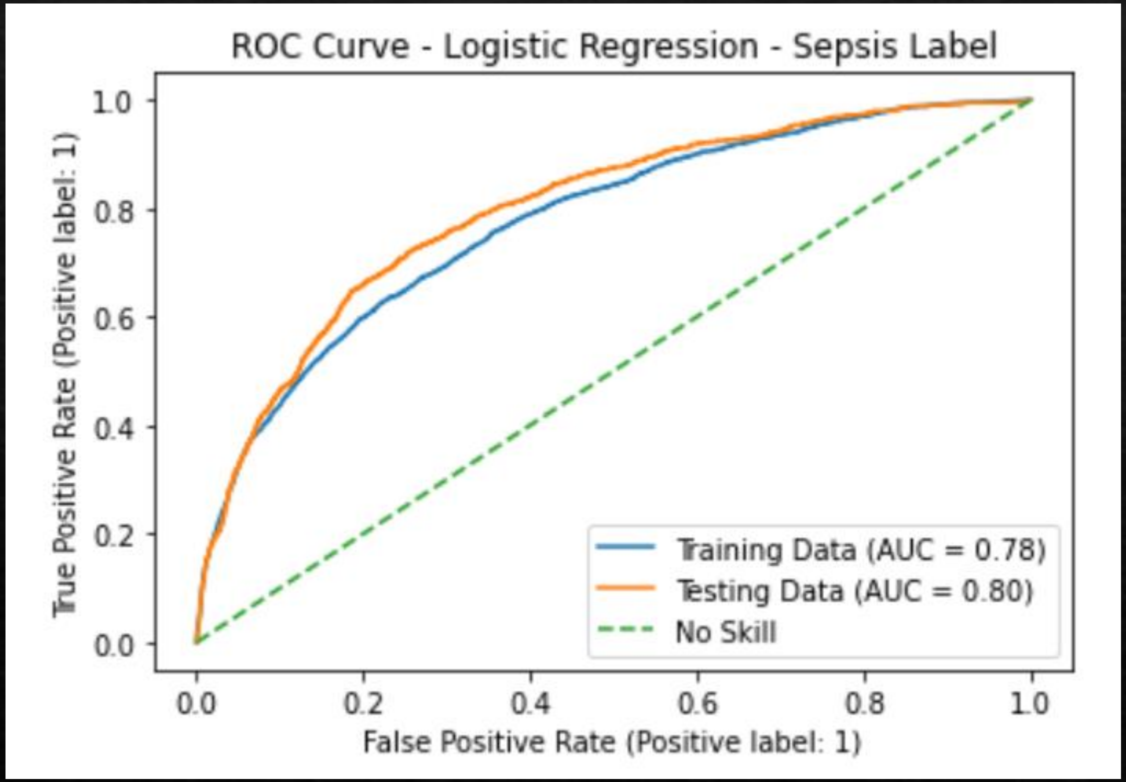
- Interpolation & forward filling
  - Lab & Vital Sign Data
- Changes in vital sign column
  - Past 1, 2, 3 Hours
- Lab value indicator, forward filled

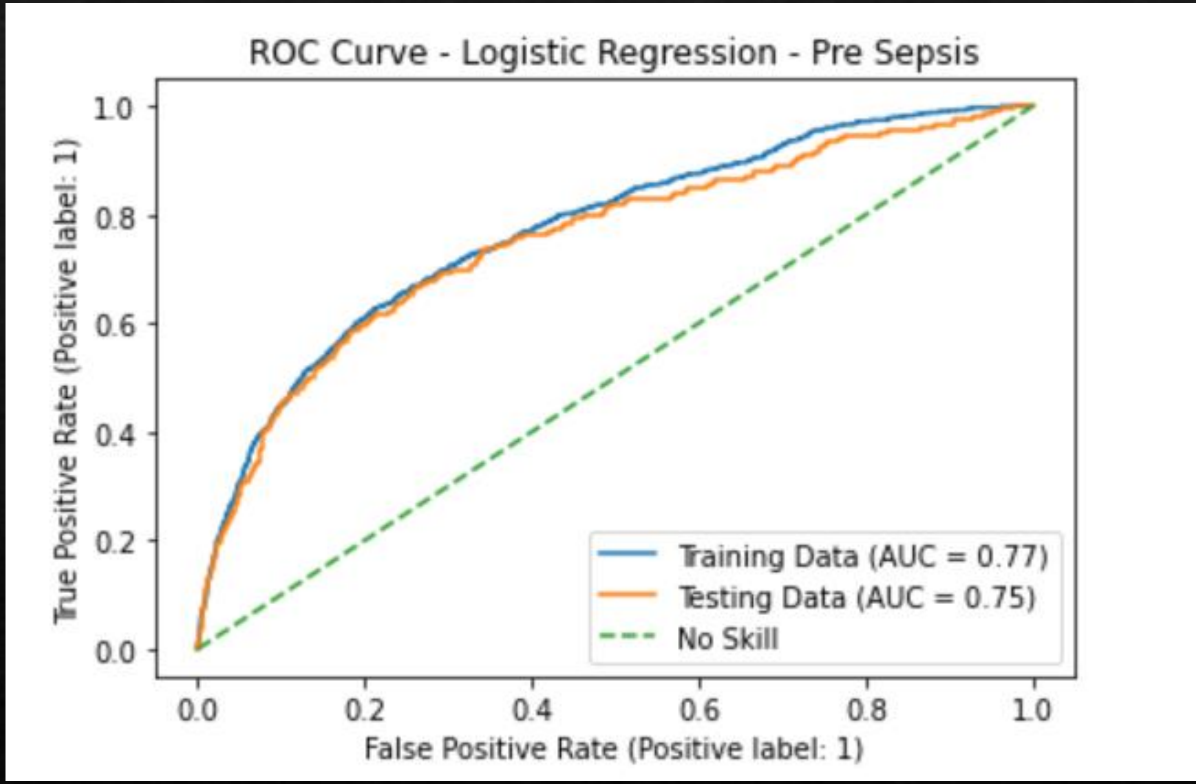
	d3_HR	d2_HR	d1_HR	HR
0	0.0	0.0	0.0	97.0
1	0.0	0.0	0.0	97.0
2	0.0	-8.0	-8.0	89.0
3	-7.0	-7.0	1.0	90.0
4	6.0	14.0	13.0	103.0
5	21.0	20.0	7.0	110.0
6	18.0	5.0	-2.0	108.0
7	3.0	-4.0	-2.0	106.0
8	-6.0	-4.0	-2.0	104.0
9	-6.0	-4.0	-2.0	102.0
10	-2.0	0.0	2.0	104.0

Classification Report: Testing Data, All Models

	Model 1: Logistic Regression		Model 2: Random Forest		Model 3: Gradient Boost		Model 4: SVM	
Performance Metric	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

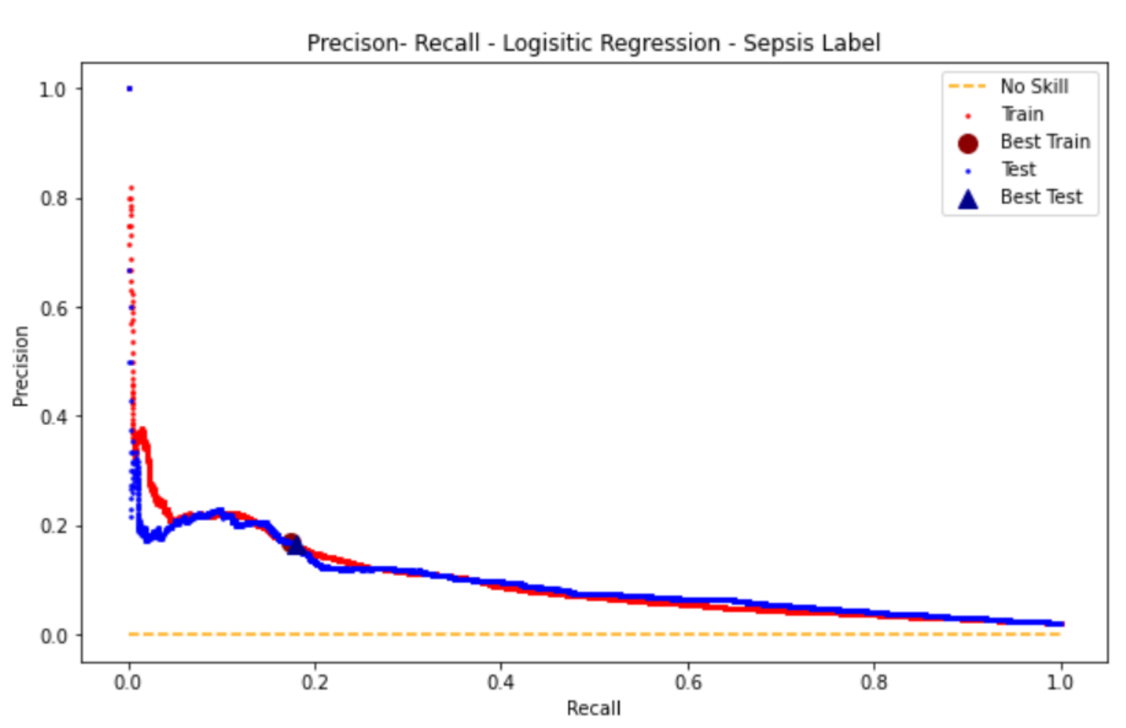
Model Results: ROC Curves

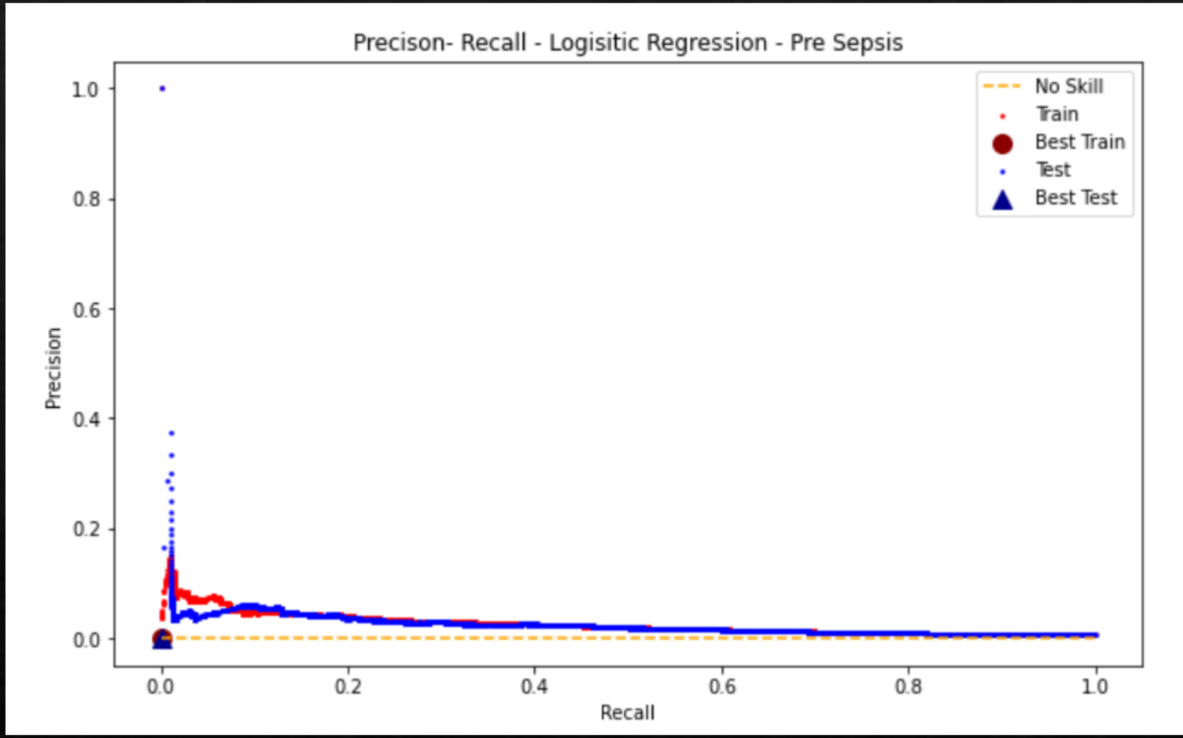




Model Results: Precision–Recall Curves

Better characterization of model performance on imbalanced classes



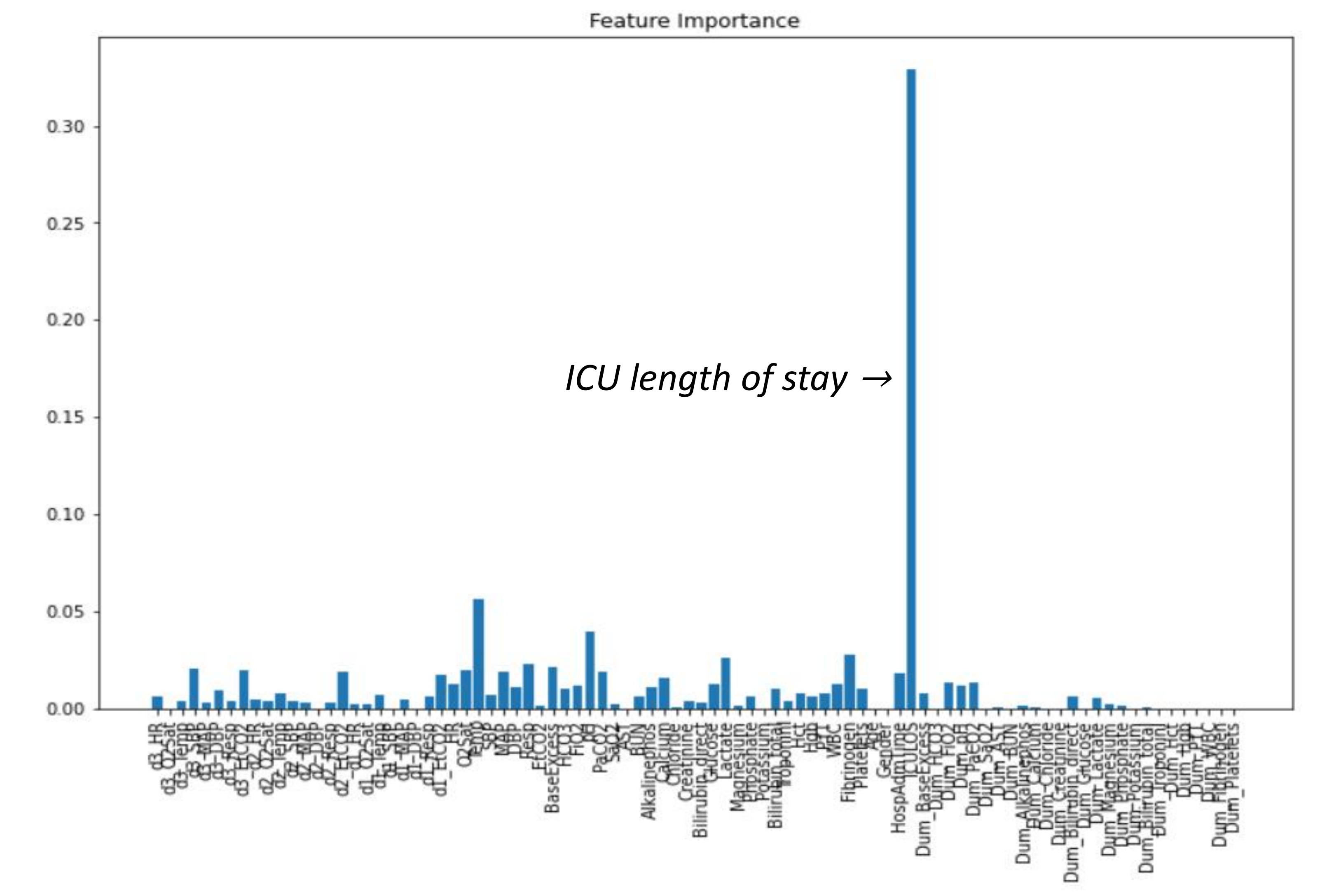


Confusion Matrix:
Sepsis Label Testing Data, Logistic Regression Model

	Actual Non-Sepsis	Actual Sepsis
Predicted Non-Sepsis	56151	1833
Predicted Sepsis	865	251

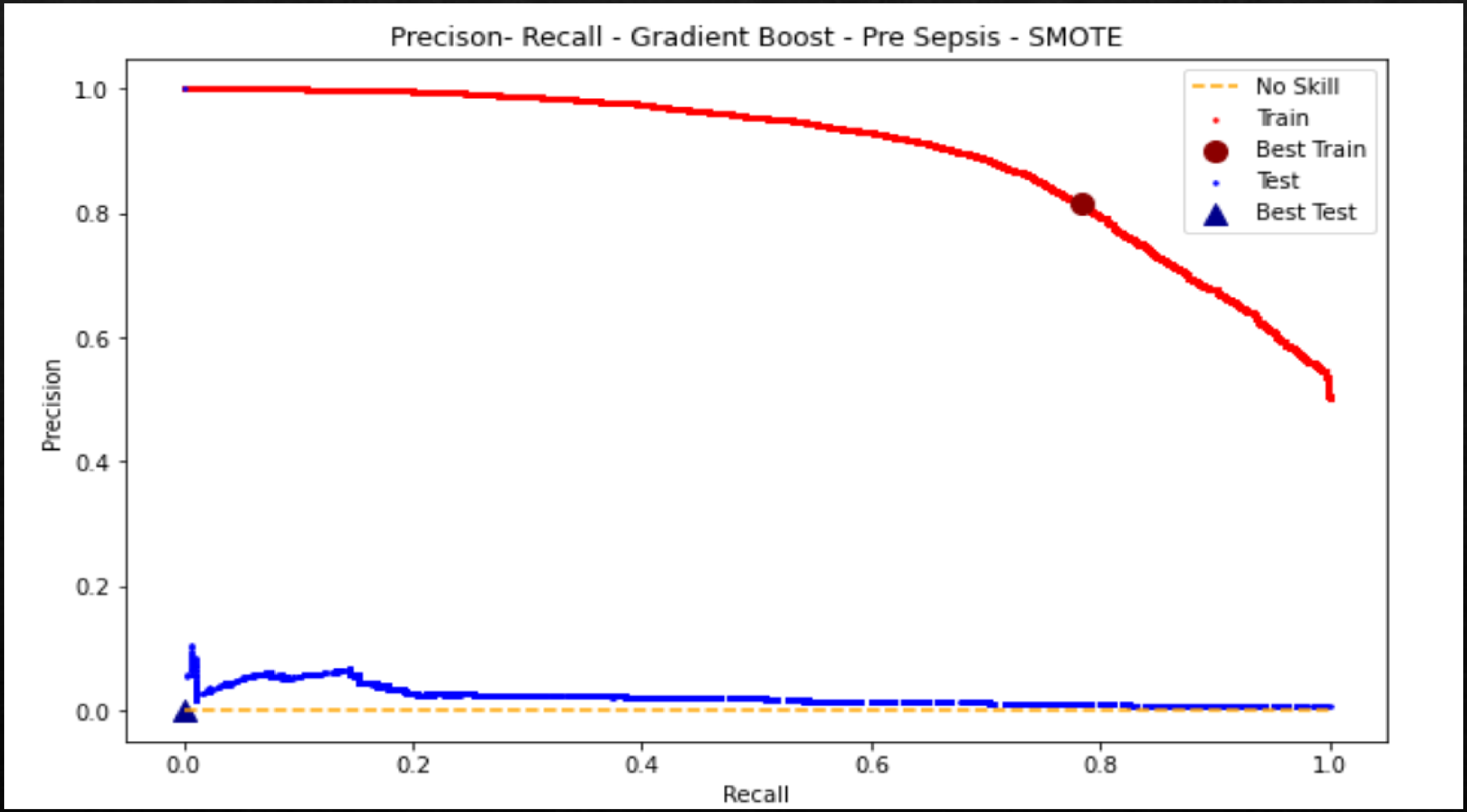
Note: Classification threshold set to 10.1%,
which yields the best F1 test score on the training data set

Feature Importance: Gradient Boost Model



What about SMOTE?

SMOTE: A technique that generates artificial data based on the minority class



Conclusion & Future Steps

- Sepsis is known to manifest very differently in different people; this was evidenced by the lack of distinction in the distributions of laboratory values between groups
- No model proved useful in accurately classifying sepsis patients, neither during sepsis or in the hours before it
- Other groups in the competition were able to build useful models with advanced feature engineering and machine learning algorithms, such as neural nets[1]
- Better classification is possible, but methods more advanced than I am currently able to employ are needed

[1] 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.