Sepsis Prediction in ICU

Vyshak Sugunan

Agenda

1. Project Goal
2. Exploratory Analysis
3. Predictive Models
4. Model Selection
5. Findings and Conclusions
6. Next Step

Project Goal

To predict sepsis development during a patient's ICU stay

- Sepsis is a life-threatening condition that arises when the body's response to infection causes injury to its own tissues and organs
- #1 cause of death in hospitals, worldwide
- Chances of survival reduces considerably after going into sepsis. So early detection, monitoring and treatment is crucial

Exploratory Analysis

Data Overview

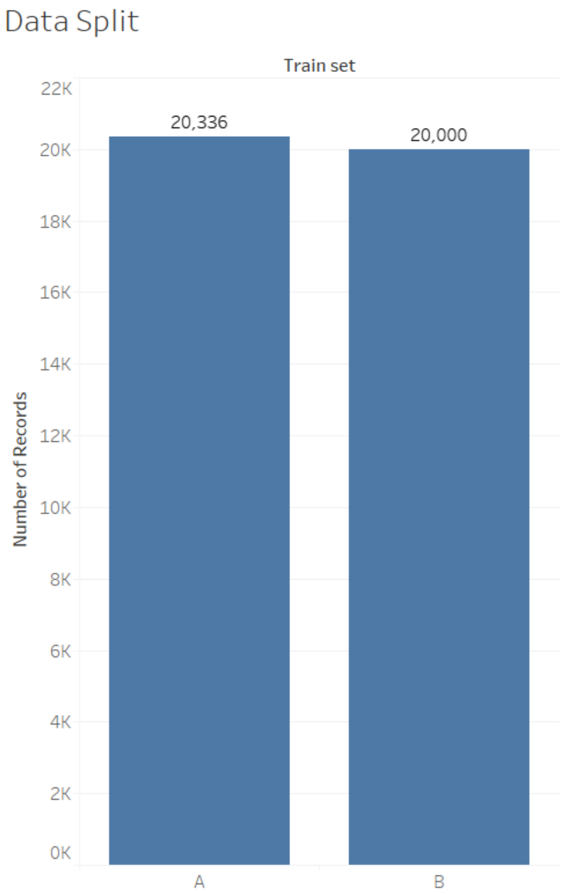
Data

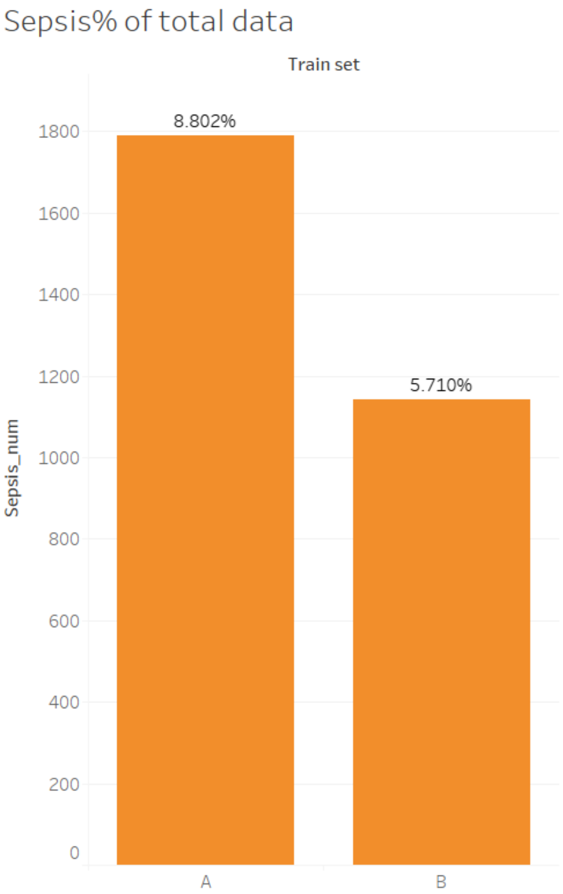
40,336 patients

41 features

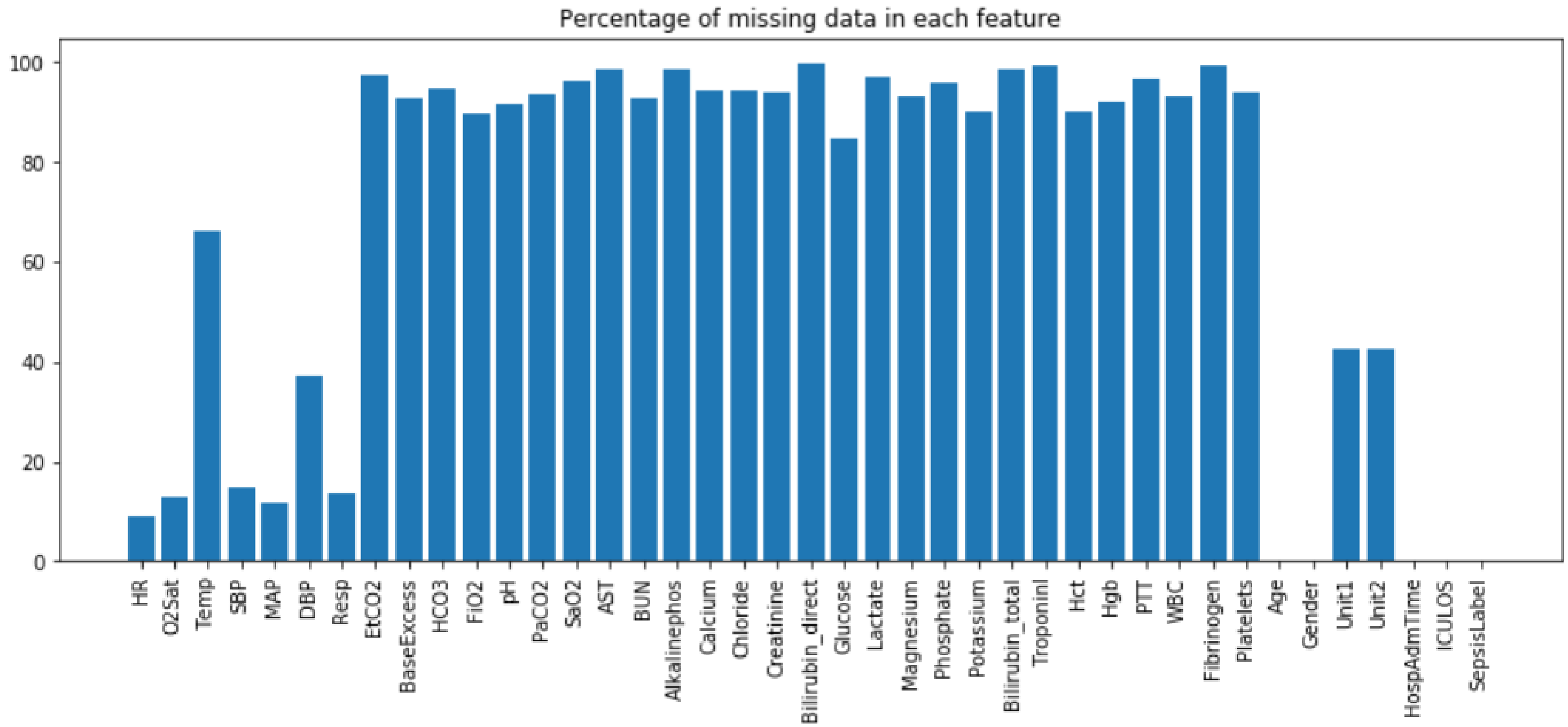
34 Vitals and medical info, 2 demos,
4 hospital info and Sepsis flag

Limited occurrence of Sepsis in available data

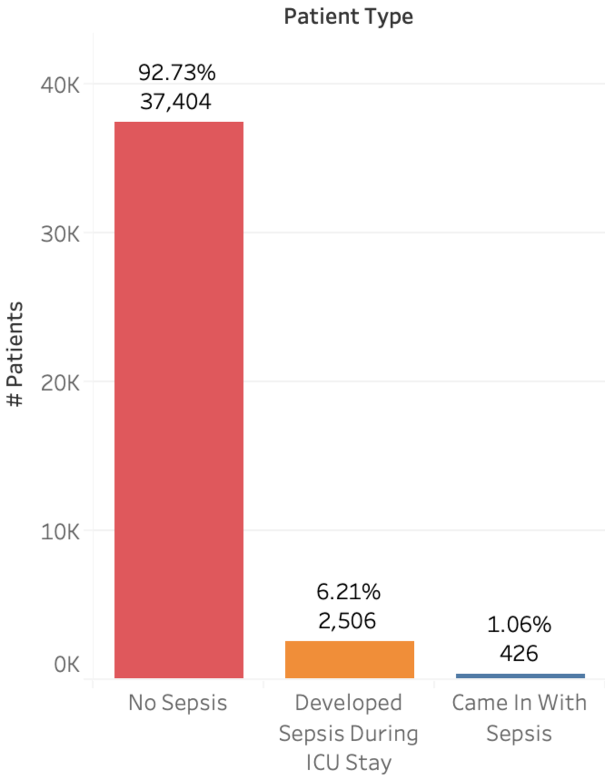




Large number of NA data

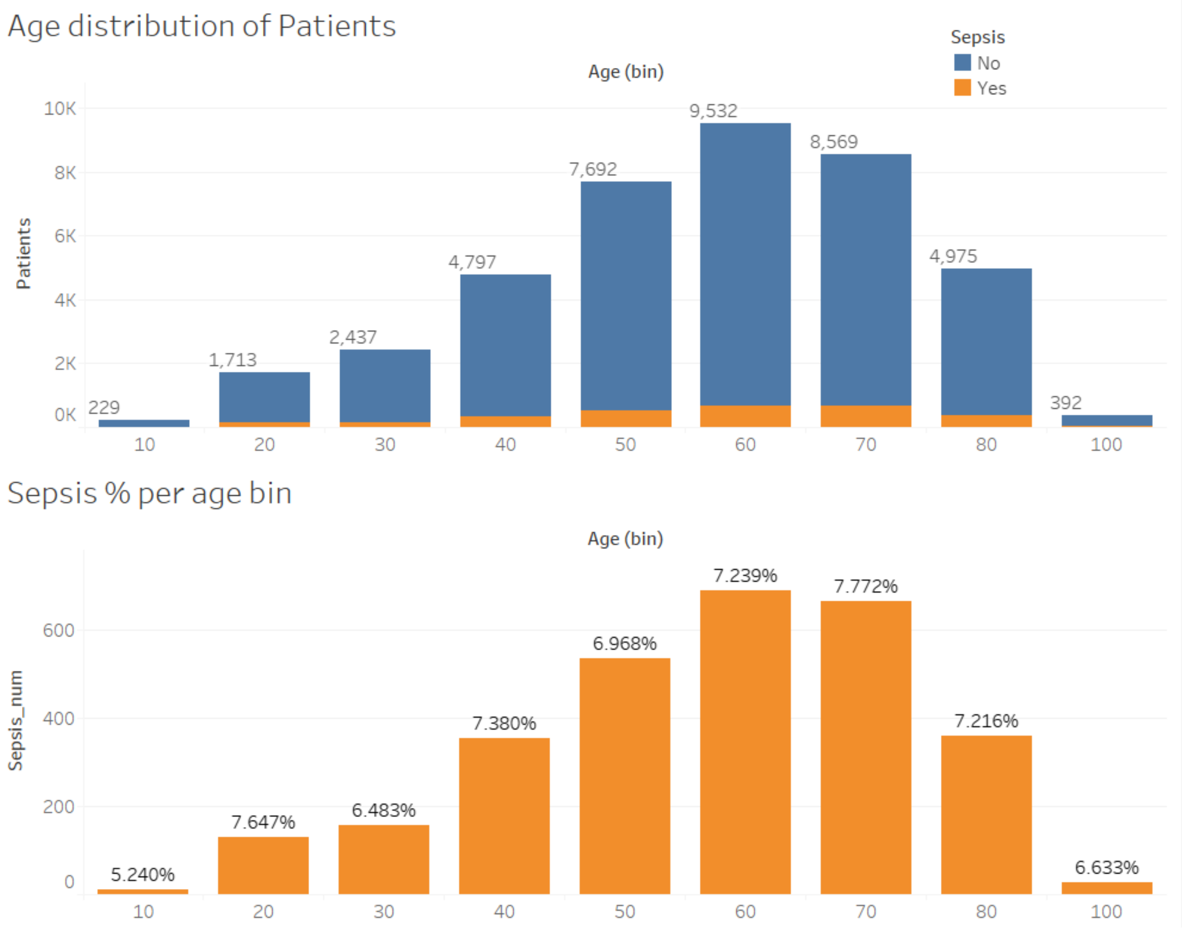


Three Distinct Types of Patients

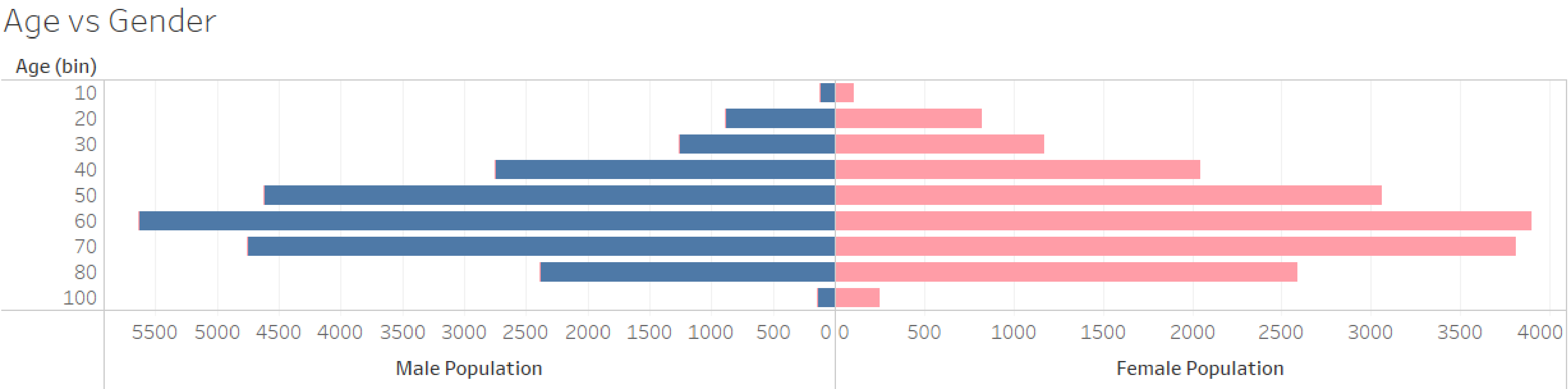


1. Came into ICU with Sepsis
2. Developed Sepsis during ICU stay
3. Never had Sepsis

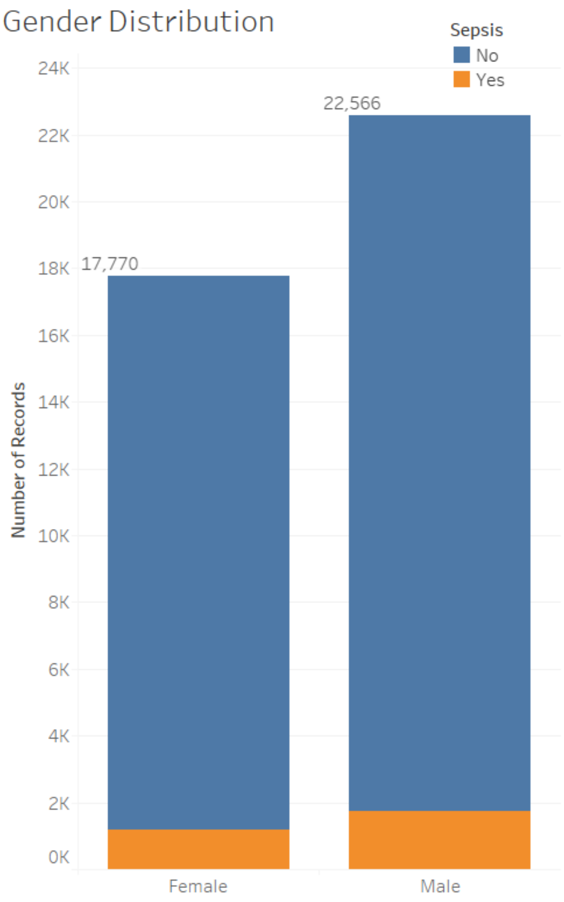
Older adults have more ICU Cases and Sepsis cases

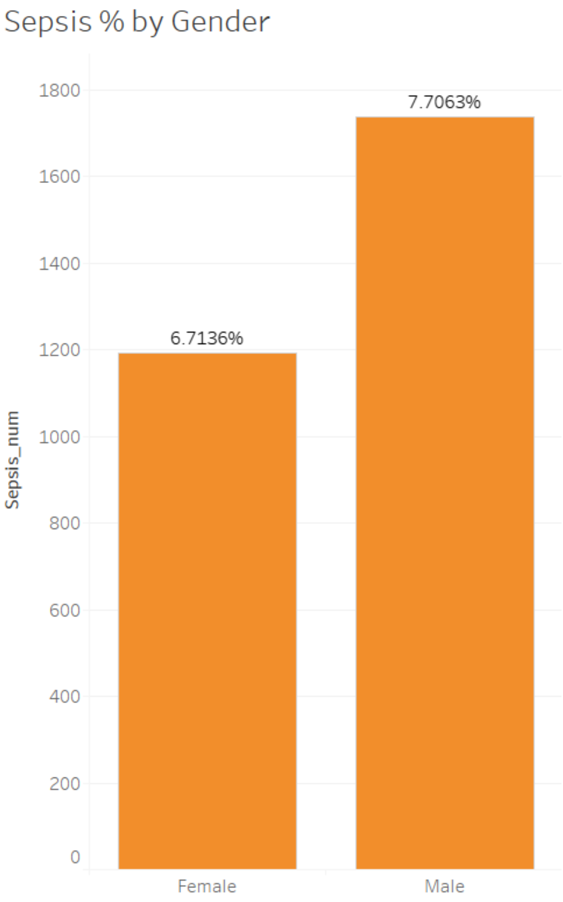


Similar Age distribution for Males and Females

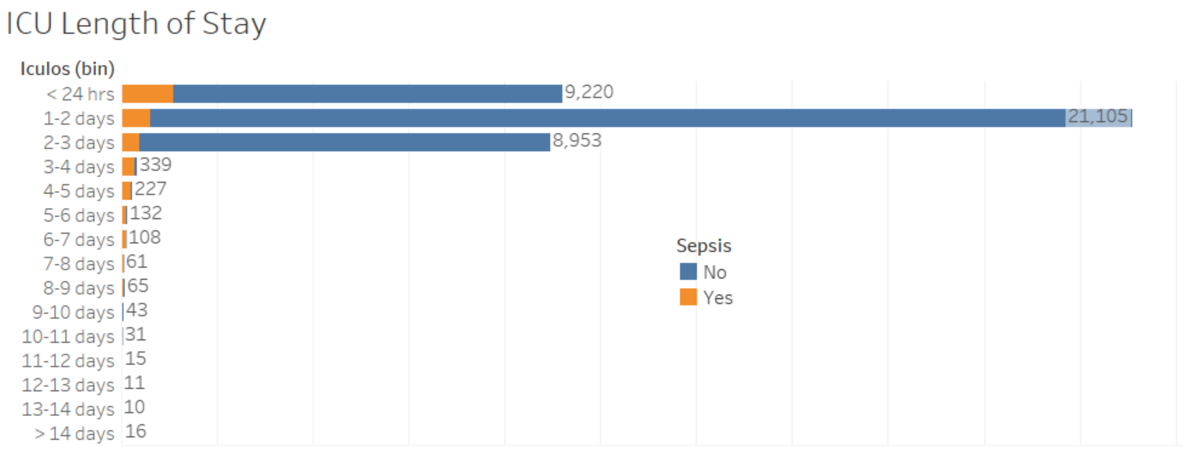


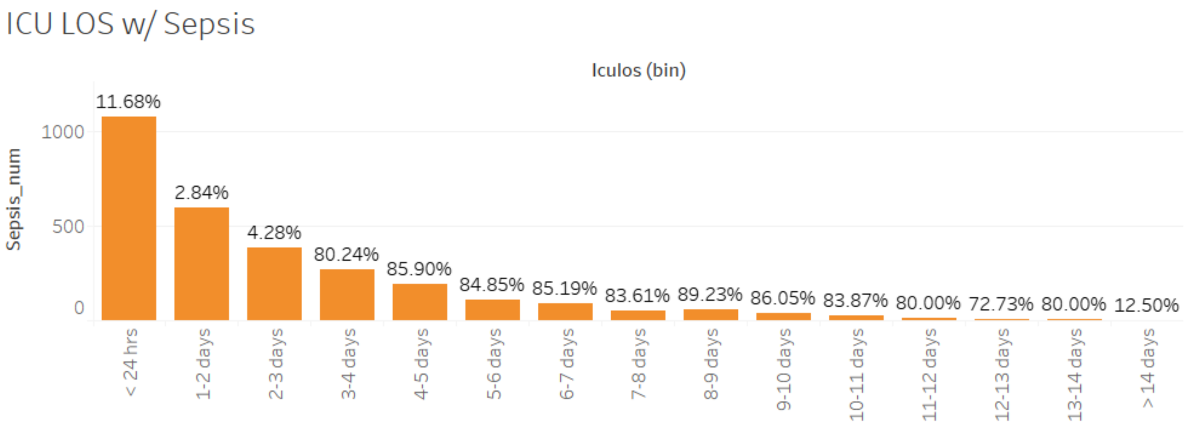
Males represent more ICU patients and higher Sepsis %





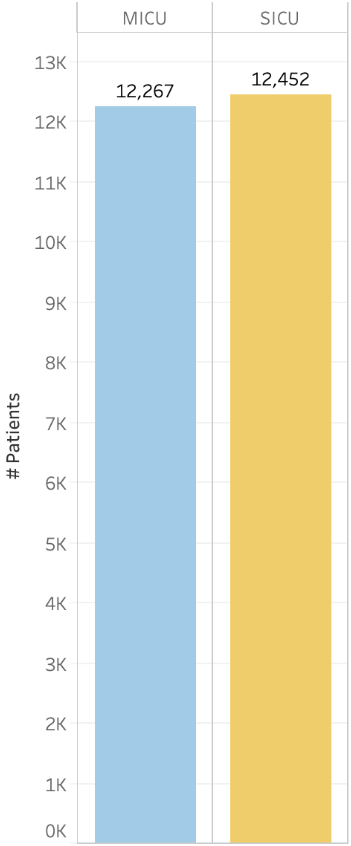
ICU Length of Stay distribution

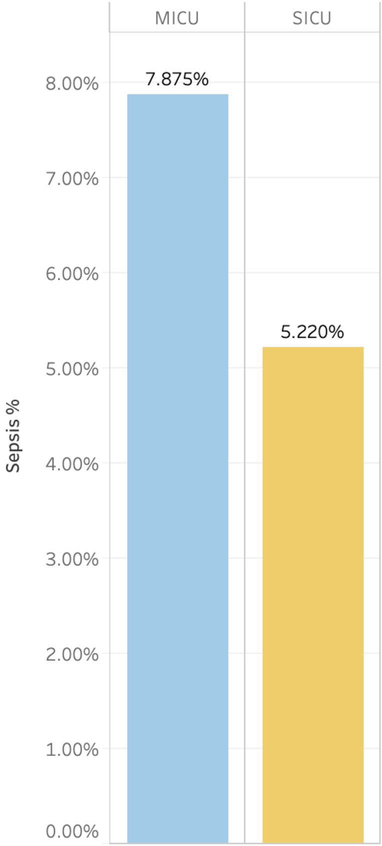




MICU vs SICU patient split is almost equal, but MICU had higher Sepsis %

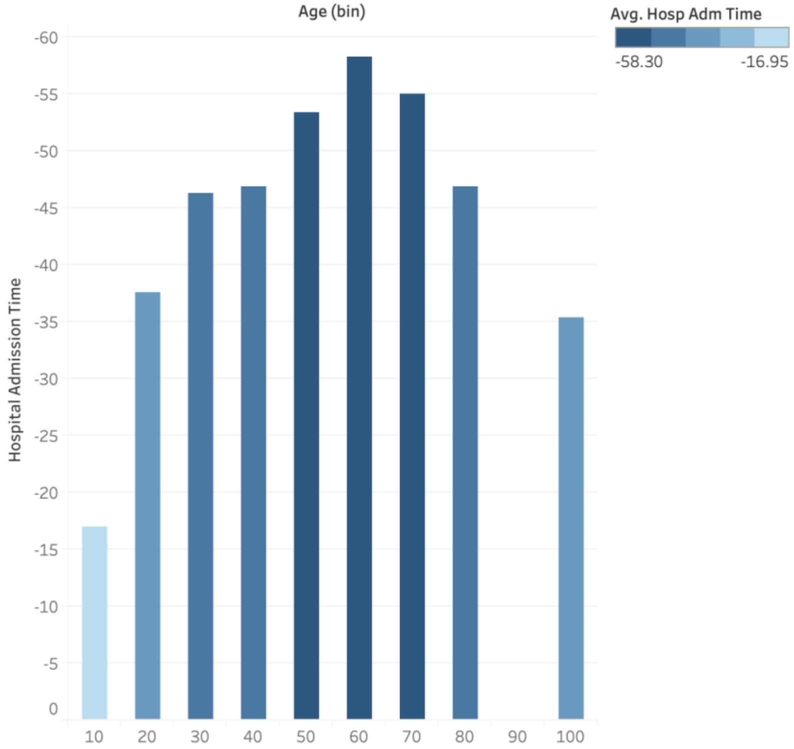
Patients

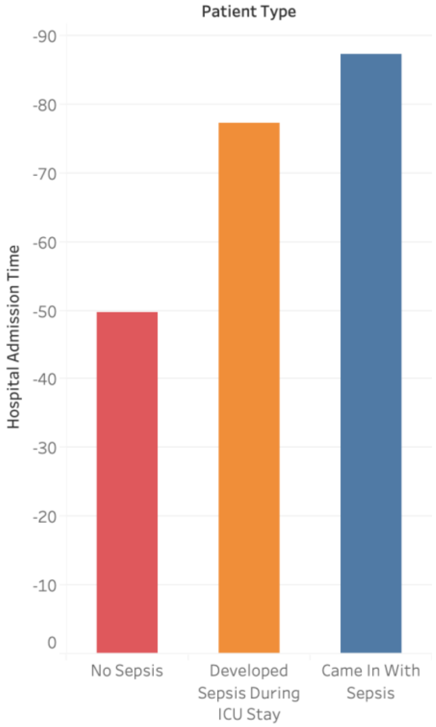




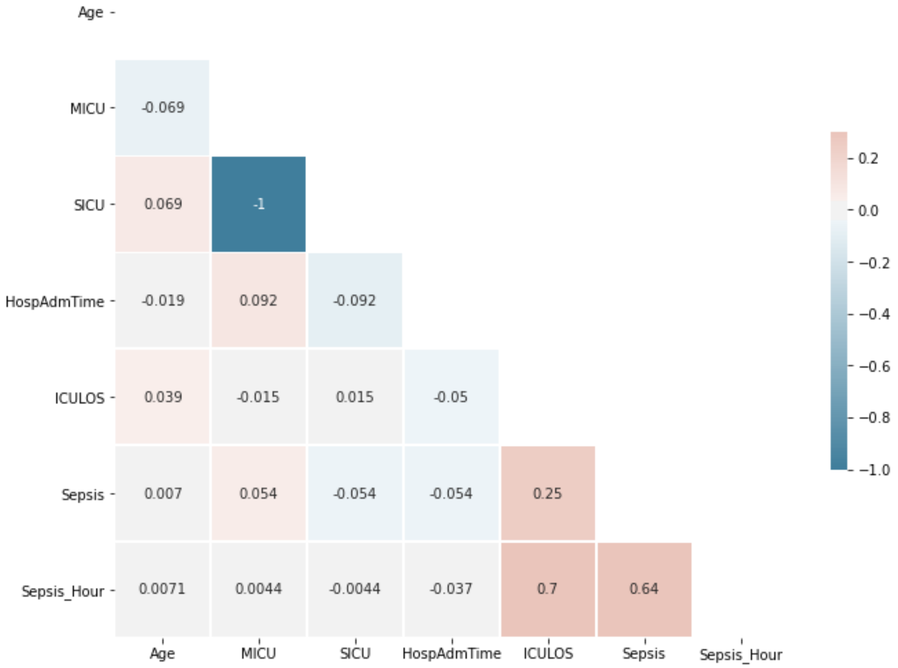
% Sepsis

Older patients and Sepsis patients have longer Hospital Admission Time

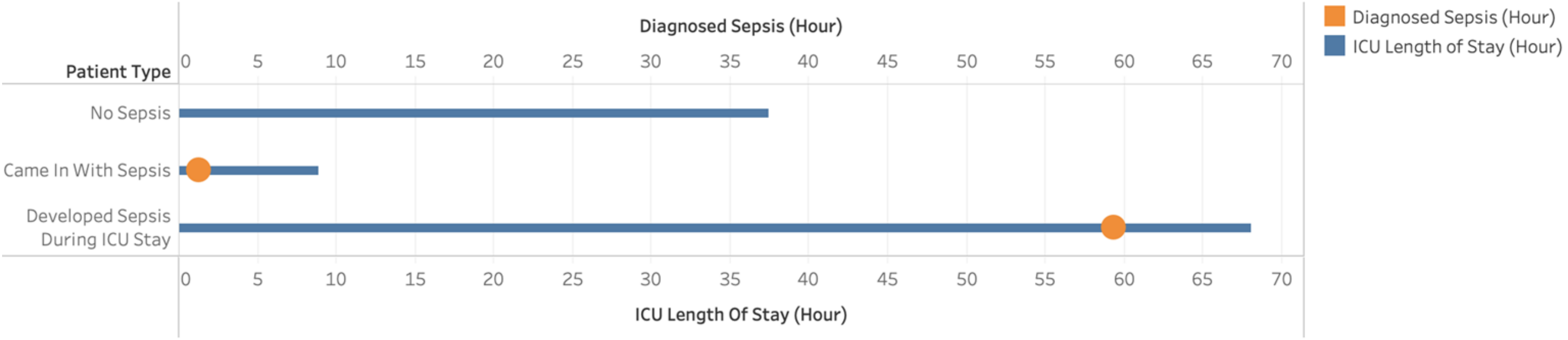




ICULOS and Sepsis Hour strongly correlated



Patients who developed Sepsis likely left ICU (no more records) within 8 hours after Sepsis diagnosis



Predictive Models

Methodology

01 Files Aggregation

40k individual hourly patient medical records.

02 Data Cleaning

NAs: > 40% feature columns were dropped.
Fill NA with median values.
Dropped non-meaningful columns.

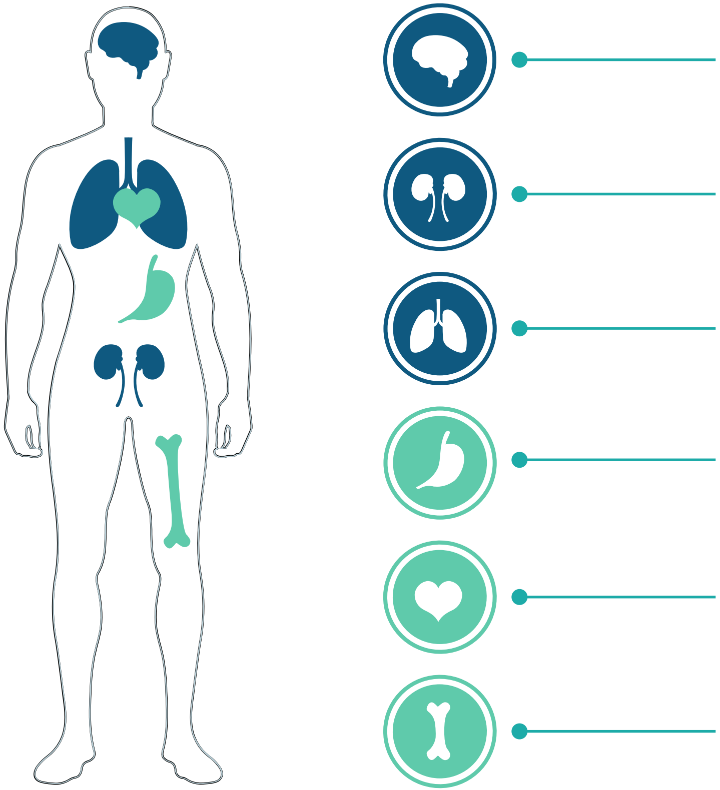
03 Modeling with un-/aggregated data

Unaggregated: per patient per hour.
Aggregated: per patient.

04 Model Selection

Accuracy, Recall, Precision and ROC/AUC scores are used. Higher recall is preferred.

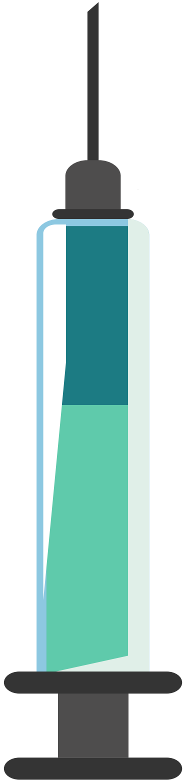
Selected features list



1. HR
2. O2 Saturation
3. Spontaneous bacterial peritonitis (SBP)
4. Mean arterial pressure (MAP)
5. Diastolic Blood Pressure (DBP)
6. Respirations
7. Age
8. Gender
9. Hospital Admission Time
10. ICULOS - Intensive Care Unit Length of Stay

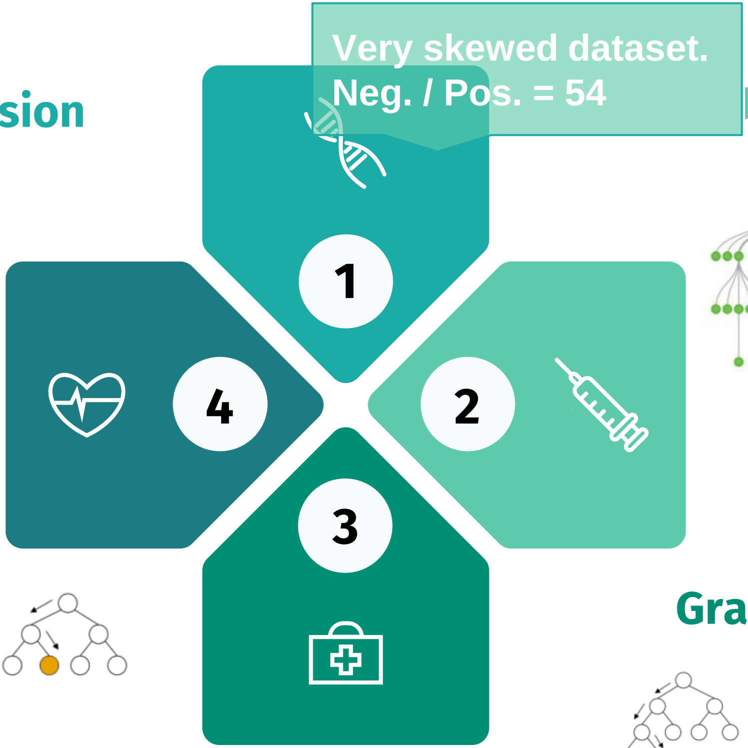
#1 Non-temporal Approach Unaggregated Data

Ignore the time component and treat record as independently and identically distributed.

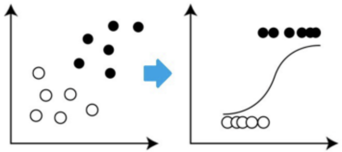


This approach would help in predicting Sepsis at each hour for any patient (with or without patient past data).

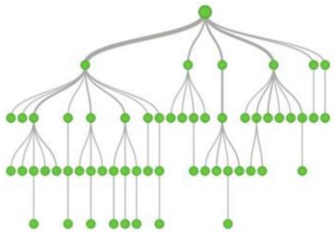
Unaggregated Data - Models used



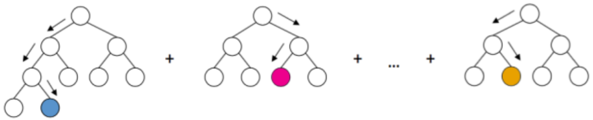
Logistic Regression



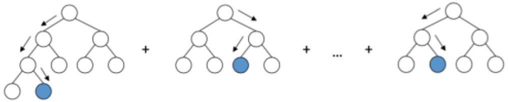
Decision Tree



XGBoost



Gradient Boost



Unaggregated Data - Model Performance

ROC: 70%

Logistic Regression

- Accuracy: 0.982
- Precision: 0.362
- Recall: 0.002

ROC: 70%

Decision Tree

- Accuracy: 0.976
- Precision: 0.362
- Recall: 0.4114

ROC: 80%

Gradient Boost

- Accuracy: 0.912
- Precision: 0.093
- Recall: 0.445

ROC: 85%

XGBoost

- Accuracy: 0.982
- Precision: 0.812
- Recall: 0.026

Unaggregated Data - Model Selection

ROC: 70% | ROC: 70% | ROC: 80% | ROC: 85%

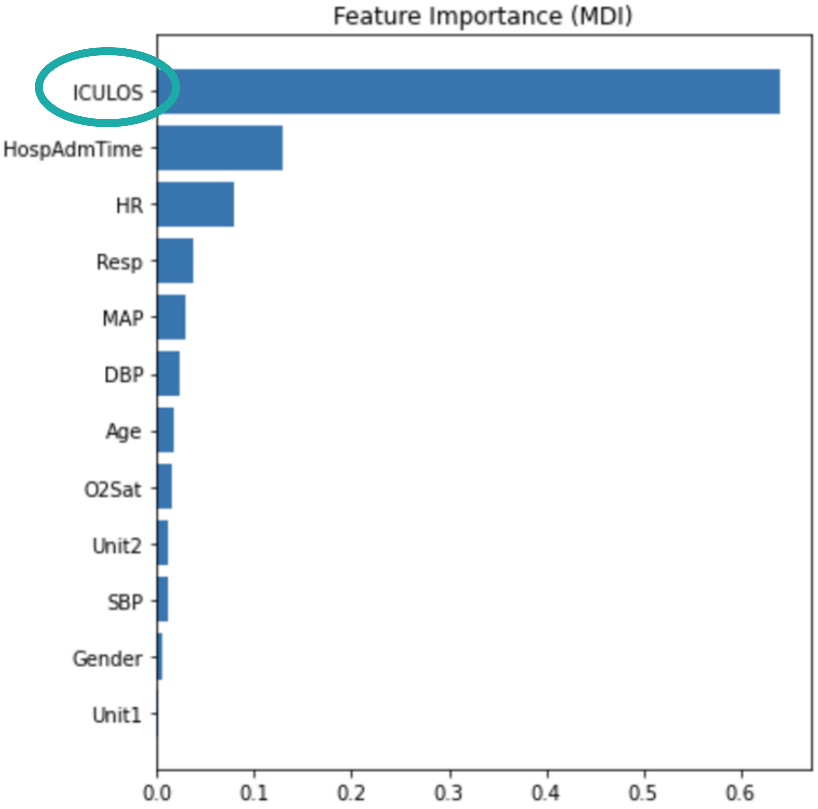
Logistic Regression
- Accuracy: 0.982
- Precision: 0.362
- Recall: 0.002

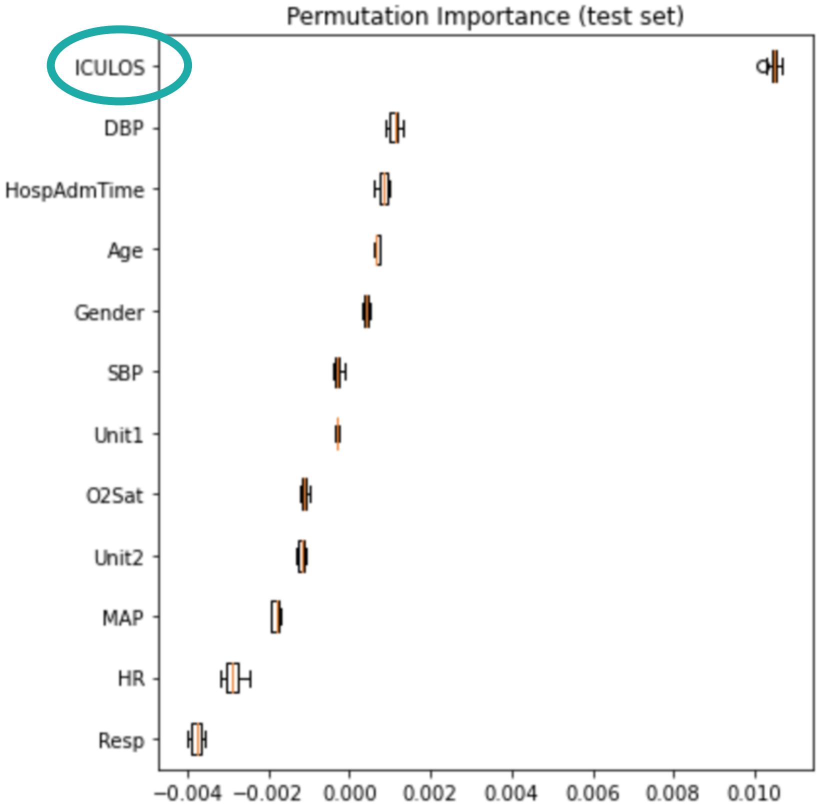
Decision Tree
- Accuracy: 0.976
- Precision: 0.362
- Recall: 0.4114

Gradient Boost
- Accuracy: 0.912
- Precision: 0.093
- Recall: 0.445

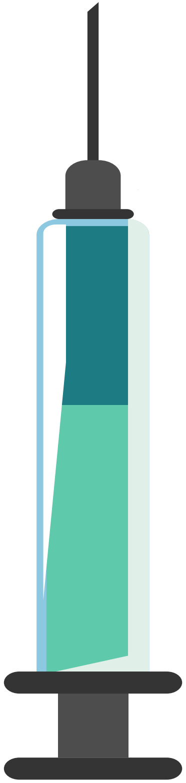
XGBoost
- Accuracy: 0.982
- Precision: 0.812
- Recall: 0.026

Gradient Boost Feature Importance





#2 Non-temporal Approach Aggregated Data

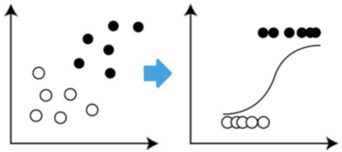


Aggregate data by patient and feature engineer maximum, minimum and variance of the vitals variables

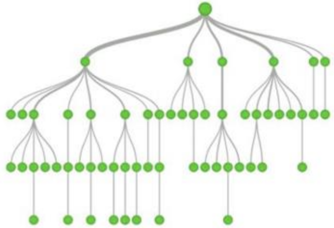
This approach would help capture the fluctuation in the patients' condition for predicting Sepsis.

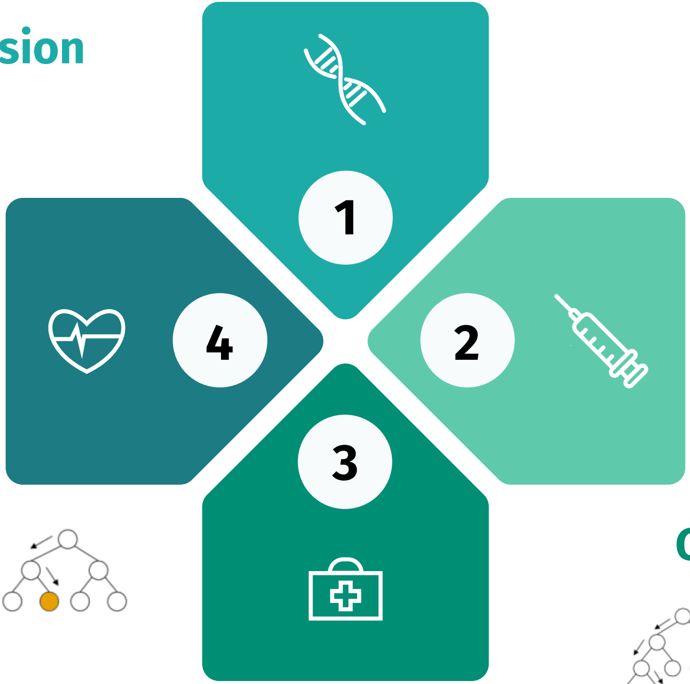
Aggregated Data - Models used

Logistic Regression

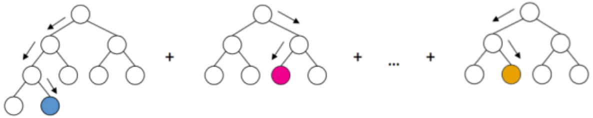


Decision Tree

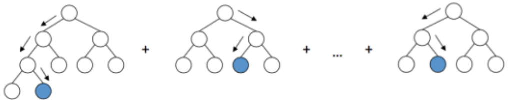




XGBoost



Gradient Boost



Aggregated Data - Model Performance

Logistic Regression

ROC: 69%

- Accuracy: 0.931
- Precision: 0.764
- Recall: 0.095

Decision Tree

ROC: 73%

- Accuracy: 0.921
- Precision: 0.466
- Recall: 0.504

Gradient Boost

ROC: 90%

- Accuracy: 0.943
- Precision: 0.605
- Recall: 0.657

XGBoost

ROC: 89%

- Accuracy: 0.956
- Precision: 0.830
- Recall: 0.501

Aggregated Data - Model Selection

ROC: 69% | ROC: 73% | ROC: 90% | ROC: 89%

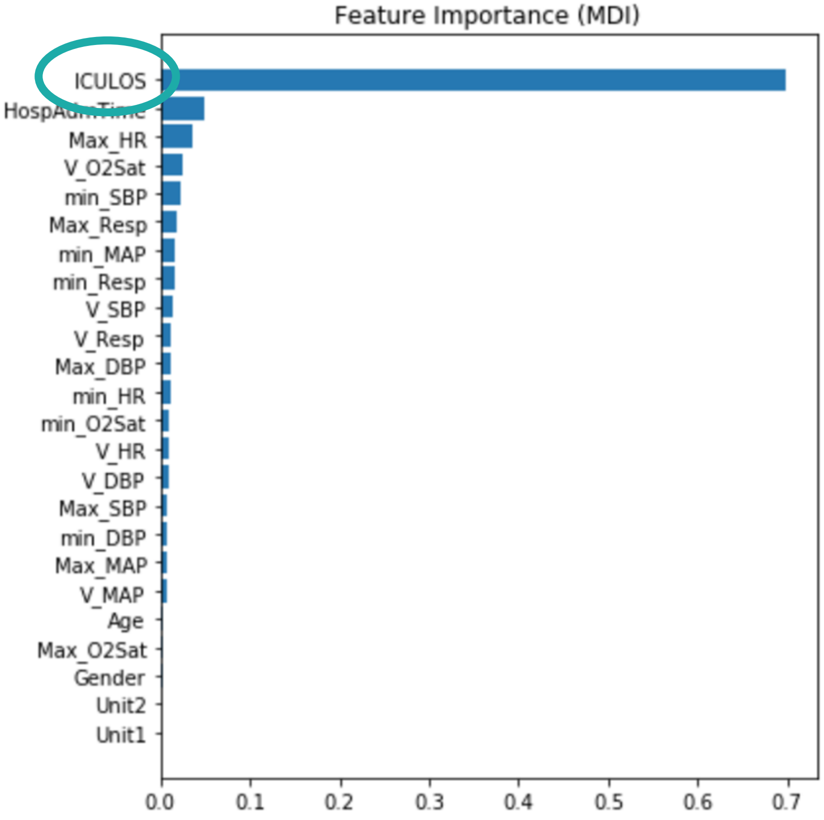
Logistic Regression
- Accuracy: 0.931
- Precision: 0.764
- Recall: 0.095

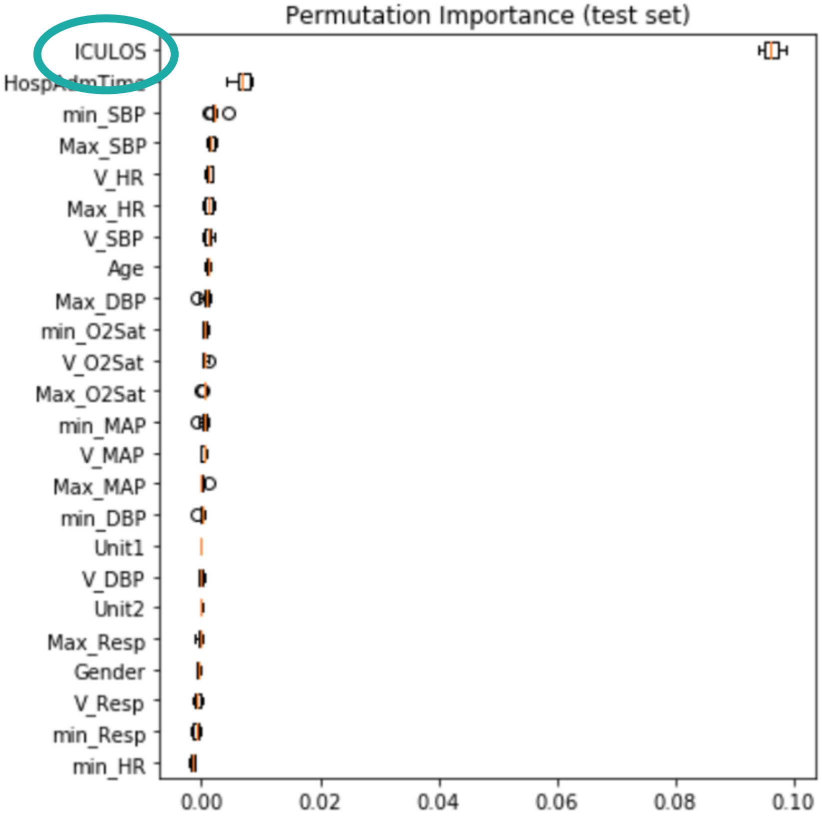
Decision Tree
- Accuracy: 0.921
- Precision: 0.466
- Recall: 0.504

Gradient Boost
- Accuracy: 0.943
- Precision: 0.605
- Recall: 0.657

XGBoost
- Accuracy: 0.956
- Precision: 0.830
- Recall: 0.501

Gradient Boost Feature Importance





Conclusion

- When data is limited, we recommend unaggregated data model with gradient boosting / decision tree to predict sepsis
- When data is more complete, we recommend aggregating the data and then use gradient boosting to predict sepsis
- ICU length of stay and hospital admission time is the most important feature predicting a patient’s sepsis development

Next Steps

- Get a more complete recording of the vital data to assess its importance in sepsis prediction

- Consult with healthcare professionals to consider other factors that can be easily retrieved and included into the model

- Suspect two-way causal effects between ICU length of stay/hospital admission time and Sepsis development

Thank You!