Early Prediction of Sepsis from Clinical Data

Darren Liu

Introduction

Motivation: Sepsis is a life-threatening condition that occurs when the body's response to infection causes tissue damage, organ failure, or death. In the US, nearly 1.7 million people develop sepsis and 270,000 people die from sepsis each year; over one third of people who die in US hospitals have sepsis.

Objectives and goals: The goal of the project is early detection of sepsis using physiological data.

Methodology

The data I used:

The data is sourced from ICU patients in two separate hospital. The data contains 40,336 files and each file contains a table providing a sequence of measurements over time. Each row represents a single hour's worth of data. There are 41 columns including vital signs, laboratory values, Demographics, and the target outcome.

Vital signs (columns 1-8)	
HR	Heart rate (beats per minute)
O2Sat	Pulse oximetry (%)
Temp	Temperature (Deg C)
SBP	Systolic BP (mm Hg)
MAP	Mean arterial pressure (mm Hg)
DBP	Diastolic BP (mm Hg)
Resp	Respiration rate (breaths per minute)
EtCO2	End tidal carbon dioxide (mm Hg)

Methodology

The data I used:

The data is sourced from ICU patients in two separate hospital. The data contains 40,336 files and each file contains a table providing a sequence of measurements over time. Each row represents a single hour's worth of data. There are 41 columns including vital signs, laboratory values, Demographics, and the target outcome.

Laboratory values (columns 9-34)

BaseExcess	Measure of excess bicarbonate (mmol/L)
HCO3	Bicarbonate (mmol/L)
FiO2	Fraction of inspired oxygen (%)
pH	N/A
PaCO2	Partial pressure of carbon dioxide from arterial blood (mm Hg)
SaO2	Oxygen saturation from arterial blood (%)
AST	Aspartate transaminase (IU/L)
BUN	Blood urea nitrogen (mg/dL)
Alkalinephos	Alkaline phosphatase (IU/L)
Calcium	(mg/dL)
Chloride	(mmol/L)
Creatinine	(mg/dL)
Bilirubin_direct	Bilirubin direct (mg/dL)
Glucose	Serum glucose (mg/dL)
Lactate	Lactic acid (mg/dL)
Magnesium	(mmol/dL)
Phosphate	(mg/dL)
Potassium	(mmol/L)
Bilirubin_total	Total bilirubin (mg/dL)
TroponinI	Troponin I (ng/mL)
Hct	Hematocrit (%)
Hgb	Hemoglobin (g/dL)
PTT	partial thromboplastin time (seconds)
WBC	Leukocyte count (count*10^3/µL)
Fibrinogen	(mg/dL)
Platelets	(count*10^3/µL)

Methodology

The data I used:

The data is sourced from ICU patients in two separate hospital. The data contains 40,336 files and each file contains a table providing a sequence of measurements over time. Each row represents a single hour's worth of data. There are 41 columns including vital signs, laboratory values, Demographics, and the target outcome.

Demographics (columns 35-40)	
Age	Years (100 for patients 90 or above)
Gender	Female (0) or Male (1)
Unit1	Administrative identifier for ICU unit (MICU)
Unit2	Administrative identifier for ICU unit (SICU)
HospAdmTime	Hours between hospital admit and ICU admit
ICULOS	ICU length-of-stay (hours since ICU admit)
Outcome (column 41)	
SepsisLabel	For sepsis patients, `SepsisLabel` is 1 if $t \geq t_{\text{sepsis}} - 6$ and 0 if $t < t_{\text{sepsis}} - 6$. For non-sepsis patients, `SepsisLabel` is 0.

Methodology

Tools I used:

Pandas for exploratory data analysis

Matplotlib and Seaborn for plotting

Scikit Learn for modeling and scoring

Methodology

How I used the data:

1. Removing columns with a missing rate higher than 93%.
2. Filling NaN HR, Temp, SBP, Resp, O2Sat and MAP values with the last non-NaN value.
3. Marking the rest NaN values as "missing".
4. Getting the symptoms of sepsis.
5. Creating labels for HR, Age, Temp, Resp, SBP, MAP and DBP columns.
6. Building a base model with a single decision tree
7. Tuning balanced bagging classifier

Methodology

Model Evaluation:

The result will count as true positive if the classifier predicts sepsis **between 12 hours before and 3 hour after sepsis time**, which in this case is between 6 hours before and 9 hours after SepsisLabel turns into 1.

Results

Final scores: (18 features)

Accuracy: 0.610

F1 score: 0.087

ROC AUC: 0.916

Confusion Matrix

Results

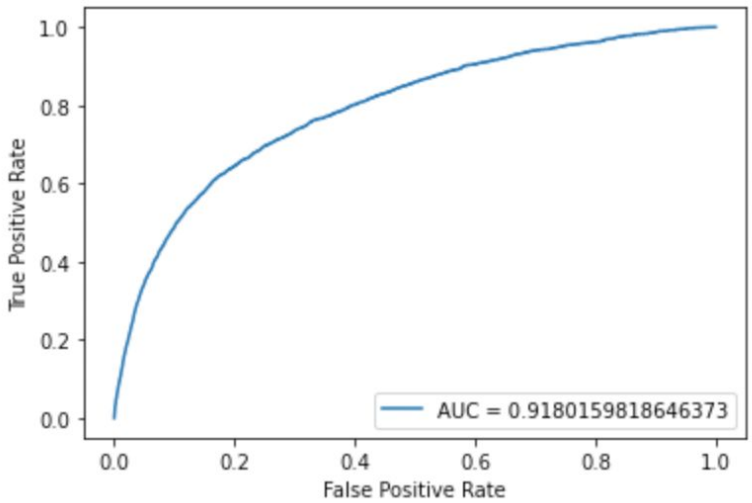
Final scores: (18 features)

Accuracy: 0.610

F1 score: 0.087

ROC AUC: 0.916

ROC Curve



Conclusions

The balanced bagging classifier now is better on predicting true positives than predicting true negatives. After tuning the model, the accuracy and ROC AUC score raised but F1 score is still pretty low.

Future Work

Now I'm taking the data row by row. However, historical data is also important on deciding the patient's condition at the moment. **Including historical data** on training and predicting will be the first future work to do.

Tuning hyperparameters of the classifier and create more accurate labels for more columns may also improve the performance.