Machine Learning Engineer Nanodegree

Survivability of ICU Patients with Severe Sepsis Shock

Sangeet Saurabh

July 14, 2017

Domain Background

Sepsis and severe sepsis (sepsis accompanied by acute organ dysfunction) are leading causes of death in the United States and the most common cause of death among critically ill patients in non-coronary intensive care units (ICU). Recent data suggest the annual cost of hospital care for patients with septicemia is $14 billion in United States. Also in the United States, the incidence of severe sepsis is estimated to be 300 cases per 100000 population. Approximately half of these cases occur outside the ICU. A fourth of patients who develop severe sepsis will die during their hospitalization. Septic shock is associated with the highest mortality, approaching 50%. The cumulative burden of organ failure is the strongest predictor of death, both in terms of the number of organs failing and the degree of organ dysfunction.Therefore, sepsis and severe sepsis are important public health problems.

Sepsis is a life-threatening condition that arises when the body's response to infection causes injury to its own tissues and organs. Most commonly, the infection is bacterial, but it may also be from fungi, viruses, or parasites. Common locations for the primary infection include lungs, brain, urinary tract, skin, and abdominal organs. Risk factors include young or old age, a weakened immune system from conditions such as cancer or diabetes, major trauma, or burns. Common signs and symptoms include fever, increased heart rate, increased breathing rate, and confusion. There also may be symptoms related to a specific infection, such as a cough with pneumonia, or painful urination with a kidney infection. In the very young, old, and people with a weakened immune system, there may be no symptoms of a specific infection and the body temperature may be low or normal, rather than high.Severe sepsis is sepsis causing poor organ function or insufficient blood flow. Insufficient blood flow may be evident by low blood pressure, high blood lactate, or low urine output. Septic shock is low blood pressure due to sepsis that does not improve after reasonable amounts of intravenous fluids are given.

Sepsis usually is treated with intravenous fluids and antibiotics. Typically, antibiotics are given as soon as possible. Often, ongoing care is performed in an intensive care unit. If fluid replacement is not enough to maintain blood pressure, medications that raise blood pressure may be used. Mechanical ventilation and dialysis may be needed to support the function of the lungs and kidneys, respectively. To guide treatment, a central venous catheter and an arterial catheter may be placed for access to the bloodstream. Other measurements such as cardiac output and superior vena cava oxygen saturation may be used. People with sepsis need preventive measures for deep vein thrombosis, stress ulcers and pressure ulcers, unless other conditions prevent such interventions. Some might benefit from tight control of blood sugar levels with insulin.

Problem Statement

Although U.S. hospitals pay out over $22.2 billion annually for sepsis treatment, the disease still causes 20-45% of death in hospitals. By the time the physical manifestations of sepsis start to appear, it's sometimes too late to help the patient. Somewhere, buried in the data – in the blood values, the blood-pressure values, the heart rate, the temperature – is a prediction that this patient is heading toward an overwhelming infection and is likely to die. If we could predict that very early, it will make it less likely the patient would die.

A predictive model that determines which patients are most vulnerable can really help ICUs focus their Physicians and staff on the most vulnerable patients. Potentially, such selective treatment can reduce mortality without increasing spending. This machine learning model is being designed to predict death within 30 days of ICU admission for the patients who have been diagnosed with Sepsis/septic shock.

Datasets and Inputs

For this study, the MIMIC-III dataset(https://mimic.physionet.org/) will be used. This dataset is accessible after taking small "Data or Specimens Only Research" course offered by MIT and suggested for use by Udacity. More than 40,000 individual patient entries will be utilized. 1184 patients were admitted with sepsis and thousands more diagnosed by the end of their stay. Most of the focus will be put on the patients admitted with sepsis since there is an associated time of admittance that will aid in finding more effective treatments. Diagnosis data is tied to ICD-9 billing codes that do not have a time associated.

The data was downloaded locally as CSV files. The provided conversion scripts were modified to load the entries into the remote Amazon RDS Postgres database. This database and server is secured and is HIPAA compliant. An Amazon compute EC2 instance will be used to do most of machine learning related heavy lifting. Many of the tables from the database will be used but the most critical are the following:

1. **Admissions:** The ADMISSIONS table gives information regarding a patient's admission to the hospital. Information available includes timing information for admission and

discharge, demographic information, the source of the admission, and so on. More detailed information available here - https://mimic.physionet.org/mimictables/admissions/ .

2. **Patients:** Contains all charted data for all patients. This provides basic information about the patient including date of birth, date of death (if applicable), gender, etc. More detailed information available at - https://mimic.physionet.org/mimictables/patients/ .

3. **Labevents:** The LABEVENTS data contains information regarding laboratory based measurements. The process for acquiring a lab measurement is as follows: first, a member of the clinical staff acquires a blood from a site in the patient's body. Next, the blood is bar coded to associate it with the patient and timestamped to record the time of the fluid acquisition. The lab analyses the data and returns a result within 4-12 hours.

4. **D_labitems:** D_LABITEMS contains definitions for all ITEMID associated with lab measurements in the MIMIC database. All data in LABEVENTS link to the D_LABITEMS table. Each unique LABEL in the hospital database was assigned an ITEMID in this table, and the use of this ITEMID facilitates efficient storage and querying of the data.

5. **microbiologyevents:** this table provides information about whether or not an infection is present, how it was obtained, and when.Note The MICROBIOLOGYEVENTS table does not contain cultures from samples taken outside the ICU.

Since our machine learning model will be supervised learning classification problem, the target must be calculated using the information available in the data. The goal is to predict survivability within 30 days of admittance. That data is not available by default. But the admissions table has

a date of death (dod) as well as date of admission to ICU. Using these 2 fields, survivability within 30 days of admittance is calculated and used as targe.

Solution Statement

The solution to is to categorize patients survivability at the time of admittance. This is calculated by considering: patient's demographic, vitals and infection information. A supervised learning model will be trained off of this information along with the calculated target value, which is the patient survival over 30 days from admittance.

The general strategy to solve this problem is to build a machine learning model using right set of features available within MIMIC-III database. Here are the steps that will be taken to solve the problem -

1. From MIMIC-III database, all the patients who have been diagnosed with Sepsis are filtered out.

2. Sepsis diagnosis customers will be categorized into 2 categories - patients who died within 30 days of admission and patients who survived past 30 days.

3. Following categories of features are analyzed to figure out the right set of features to build machine learning model -
    a. Patient's age, gender, insurance information etc.
    b. Patients vitals collected within 48 hours of admission
    c. Patients infection that's reported within 24 hours of admission

4. Once features are finalized and all the data are available, multiple models using different Machine learning algorithim will be developed, analyzed and a comparative analysis between them will be done. Goal will be to figure out the best performing and most generalizable solution.

Benchmark Model

Quite a few studies have been done to predict survivability of a Sepsis diagnosed patient. I am using the metrics as described in https://www.ncbi.nlm.nih.gov/pubmed/23442987 -

"Survival after sepsis was predicted with an accuracy of 80% by the NN model, which used only information collected at the time of the diagnosis of sepsis. The development of multiple organ failure after the diagnosis of sepsis was predicted accurately (81.5%) with either the MLR or the NN model. Both the MLR and the NN methods depended on the interpretation of a likelihood quantity, requiring the choice of a threshold to make a survival prediction. The accuracy of the MLR models was very sensitive to the

threshold value. The accuracy of the NN models was not sensitive to the choice of threshold, because they generated likelihood predictions that were distributed far from the middle range where the threshold was placed. “

While this study was done in 1996, there have been few follow-up studies. The few follow-up studies had a similar success rate with slightly different attributes and test periods. Due to the detail of this study, most focus will be put on the work of Flanagan et al. for the benchmark.

Evaluation Metrics

Several different machine learning models will be developed to predict survivability of Sepsis patient. Results of machine learning model will be compared against the benchmark model based upon study https://www.ncbi.nlm.nih.gov/pubmed/23442987. Goal will be to produce results better than benchmark study showed in 1996.

Several different solution models will be designed and compared in order to see the difference in accuracy. Using standard classification accuracy is a great metric for quantifying performance. While the feature set is highly complex, the target attribute is boolean and easily assessed. Therefore computing the model’s accuracy at classifying patients by boolean survivability is accurate, simple, and understandable. Also, it can be easily compared against the benchmark study described above.

Project Design

Goal of this project is to predict if a sepsis shock patient will survive after 30 days of ICU admission. It’s a classification problem. In order to solve this classification problem, we will try out multiple machine learning models and figure out which one works the best. We will try out following machine learning algorithms -

1. Logistic Regression
2. Support Vector Machine (SVM)
3. Naive Bayes Classification
4. Random Forest
5. XGBoost
6. Deep Learning
7. Adaboost classifier

Due to the nature of the problem and the relatively large number of features that will likely be involved, the results from the above models will be compared. Here are the steps that will be followed to reach to the final model -

1. **Set up full database:**

(a) Download entire compressed CSV dataset from physionet.org (30GB)

(b) Spin up an Amazon RDS Postgres database and EC2 Compute instance

(c) Modify provided scripts from the MIMIC-III MIT Github repository to load data into the remote RDS Postgres database

(d) Setup docker with all necessary components including SQL alchemy for easy querying and object handling using its ORM

2. Analyze data to understand data and figure out features for ML model:

(a) Gather useful attributes across demographics, vital and infection tables by patient ID and/or hospital stay ID

(b) Analyze demographics, vital and infection data for customers

(b) Figure out the features that will be input for machine learning models

3. **Develop and test ML models**:

(a) Using the feature set above, develop and test machine learning models

(b) Compare results and tweak hyperparameters for better results

4. **Analyze and write up results**

Reference

1. Wikipedia - https://en.wikipedia.org/wiki/Sepsis
2. Epidemiology of severe sepsis - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3916382/
3. Machine learning may help in early identification of severe sepsis - https://www.sciencedaily.com/releases/2017/05/170524100616.htm
4. Sepsis and Machine Learning - https://vision.cloudera.com/sepsis-and-machine-learning/
5. Predicting In-Hospital Mortality due to Sepsis: An Integrative Approach - https://repository.library.brown.edu/studio/item/bdr:697404/PDF/
6. Cloud-based Analytics: Supporting Healthcare's Digital Transformation - https://d1.awsstatic.com/whitepapers/Industries/HCLS/AWS_WhitePaper_21216.pdf
7. Predicting survival of patients with sepsis by use of regression and neural network models - https://www.ncbi.nlm.nih.gov/pubmed/10156949