Machine Learning Nanodegree Capstone Project Report:

Sepsis Risk Classifier

by Wen-Yun Chang

February 8th, 2020

https://github.com/wynchang/MIMIC-sepsis

1 Definition

1.1 Project Overview

This project is to build a binary classification system to classify patients' risk of sepsis based on their age, gender, and relevant chart events (blood pressure, temperature, heart rate and white blood count) at each hospital admission. Selected linear and non-linear classifier algorithms will be used to train the models to predict if the patient of each hospital admission has sepsis risk or no risk. The best performing models will be compared against benchmark.

1.2 Domain background

Sepsis is a person's extreme response to an infection, and can arise from any situation, such as surgery or normal infection. It can turn deadly very quickly if not recognized and treated in time. Crude mortality rate of sepsis is 34.7%. In the hospitals, clinicians have to be always on the lookout for sepsis, as early intervention is known to produce better clinical outcomes.

In 2015, the Centers for Disease Control and Prevention (CMS) added timely treatment of sepsis as a quality measure. In addition to the high mortality rate, treatment of sepsis costs 27 billion dollars per year.

1.3 Problem Statement

If we can access a patient's sepsis risk from admission condition, it can alert clinicians to increase checks for sepsis. The purpose of this project is to build a Sepsis Risk Classifier based on the patient's gender, age and charts taken during a hospital admission.

According to the Systemic inflammatory response syndrome (SIRS) criteria, vitals and lab readings such as heart rate, body temperature, white blood count (WBC) and blood pressure are potential indicators of sepsis. Moreover, in a report by Robert A Balk, the Sepsis criteria has been evolving since 1980s as clinicians try to determine better criteria. Machine learning models could assist clinicians as a complementary surveillance to identify sepsis.

1.4 Evaluation Metrics

The aim is to achieve high sensitivity and low specificity with this model, that is minimizing false negatives, as we cannot afford to let sepsis cases go undetected. In exchange, higher false positives are tolerated. The model built will be optimized for low false omission rate (FOR), where FOR = FN / (FN+TN).

2 Analysis

2.1 Data Exploration

The MIT MIMIC-III Clinical Database contains 12 years of de-identified critical care patient data collected at Beth Israel Deaconess Medical Center. Besides admission records with diagnosis, there are vitals, procedures, laboratory results, prescription and other information.

Since sepsis is a leading cause of death at the intensive care units, this critical care clinical dataset will be relevant for sepsis analysis. The MIMIC-III Clinical Database contains admission records with diagnoses. Those diagnoses including sepsis will be flagged as sepsis outcome. This will support supervised learning, as well as provide the test data for validation.

Out of 58,976 Intensive Care Unit (ICU) admissions, 3% or 1,783 have a sepsis diagnosis. It is noted that this is not a balanced dataset. Since elective admission type and newborn deliveries have very low sepsis rate, emergency and urgent admission types will be the focus of the project.

Count of ROW_ID	**Sepsis_Flag**	
Admission Type	**0**	**1**
ELECTIVE	99.99%	0.01%
EMERGENCY	95.85%	4.15%
NEWBORN	100.00%	0.00%
URGENT	97.23%	2.77%
Grand Total	**96.98%**	**3.02%**

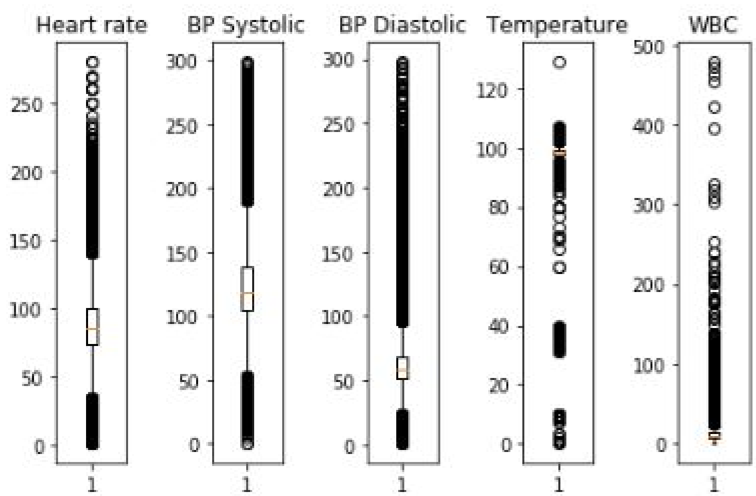
The input features will be focused on the vitals identified by the Systemic inflammatory response syndrome (SIRS) criteria such as heart rate, body temperature, white blood count (WBC) and blood pressure. Also age at admission, gender are extracted and computed from the admission and patient records.

2.2 Exploratory Visualization

This proposal is to build a binary classification system to classify patients' risk of sepsis based on their age, gender, and relevant chart events (blood pressure, at each hospital admission. In the MIMIC-III Patient Database, the files used for this project are:

Data table	Features
ADMISSIONS	Hospital admission identifier (HADM_ID), Patient identifier (SUBJECT_ID), ADMITTIME, ADMISSION_TYPE, DIAGNOSIS
PATIENTS	Patient identifier (SUBJECT_ID), GENDER, date of birth (DOB)
CHARTEVENTS	Chart item (ITEMID), chart item numerical value (VALUENUM), Hospital admission identifier (HADM_ID), Patient identifier (SUBJECT_ID),
D_ITEMS	Chart item (ITEMID),Chart event description (LABEL)

The features we are using from the chart events are heart rate, systolic and diastolic blood pressure, Temperature and white blood count. These are chart events identified by the Systemic inflammatory response syndrome (SIRS) criteria.



2.3 Algorithms and techniques

2.4 Benchmark Model

While there are more sophisticated sepsis alert systems in the market, this project is intended as an academic investigation if chart readings of the SIRS criteria are sufficient to classify if a hospital admission has sepsis risk or no risk. Linear binary classification models will first be built as benchmark. Subsequently, various non-linear classification models are built and measured against the linear model benchmark.

The University of Pennsylvania's research on a Computerized Sepsis Screening and Alert System achieved a predictive value of 74%. It will be a high bar for the best model of this exercise to beat.

3 Methodology

3.1 Data Preprocessing and Feature Engineering

- A subset of the MIMIC-III data – of emergency or urgent admission types -- will be used as input.
- Because the dates are date shifted to de-identify patients, we can only select certain datetime factors for analysis, such as computer the patients' age from their date of birth and date of hospital admission.
- Not all data fields are complete; decisions will need to be made on which rows to exclude, or which will stay in the analysis as "not filled".
- Set Sepsis flag on the cases where diagnosis includes sepsis.
- To further reduce the data set, the input features will be focused on those vitals identified by SIRS criteria.
- Find the maximum and minimum of each numerical chart reading to analyze.
- Convert categorical data to numerical data.
- Normalize the input data.
- The dataset is split into ⅔ train data and ⅓ test data. Because there is only about 4% sepsis cases in the data, the train data is further processed with SMOTE over-sampling to produce a more balanced data set.

3.2 Implementation

1. Define and train model
    - Train the model with supervised learning, first using the linear binary classifiers such as Linear Learner and Linear Support Vector Classifier (SVM).
    - Repeat with non-linear binary classifiers, such as KNeighborsClassifier or Radial Basis Function Support Vector Classifier (RBF SVC), to identify the best model.
    - Optimize hyperparameters.

2. Test model
    - Run the test data set against the trained model.
    - Compare non-linear classifier performance against linear classifier benchmark
    - Measure against published research model.
    - Calculate the false omission rate, accuracy, precision and recall measures.

3.3 Conclusions and recommend model improvements

4 Results

4.1 Model Evaluation and Validation

Our aim is to evaluate models for the lowest false omission rate.

4.1.1 Linear Models

The Linear Learner algorithm is first used to build the benchmark model. It is then tuned to target precision and target recall respectively. All 3 give comparable false omission rates in the range of 3.7 - 3.9%, but the characteristics differs.

Basic Linear Learner correctly predicts 344 out of 588 true positives. More than half of the negatives are false negatives though. Since precision is only 4.4%, the next model will be tuned to increase precision.

Target precision LinearLearner - while the precision rate barely increased, the number of true negatives is impressive. However, majority of the sepsis cases were not detected.

Target recall LinearLearner has the highest number of true positives. Unfortunately it also has the highest number of false positive, rendering the model ineffective in distinguishing sepsis.

Support Vector Classifier is the linear model with lowest false omission rate of 3.2%. It detects more than half of the sepsis, while also correctly determining many true negatives. This will be the benchmark model.

4.1.2 Non-linear Models

Three non-linear classifiers were tested:

The KNeighborsClassifier has a high 90.3% accuracy, from correctly identifying most of the true negatives. However, it only manages to flag 50 out of the 588 sepsis cases in the test set, The false omission rate of 4.0% is higher than all the linear models.

The XGBoost Classifier and Radial Basic Function Support Vector Classifier both have the lowest false omission rate of 2.9%. XGBoost Classifier has a higher accuracy of 58.4%, and has correctly identified more than half of both the sepsis and non-sepsis cases.

<table>
<tr><th colspan="3">Linear models</th></tr>
<tr>
<td>Metrics for simple,
LinearLearner.

prediction (col) 0.0 1.0
actual (row)
0.0 6294 7439
1.0 244 344

Recall: 0.585
Precision: 0.044
Accuracy: 0.464
False omission rate: 0.037</td>
<td>Metrics for tuned
LinearLearner - target 80%
precision

prediction (col) 0.0 1.0
actual (row)
0.0 11153 2580
1.0 450 138

Recall: 0.235
Precision: 0.051
Accuracy: 0.788
False omission rate: 0.039</td>
<td>Metrics for tuned
LinearLearner - leaders 80%
recall.

prediction (col) 0.0 1.0
actual (row)
0.0 2274 11459
1.0 89 499

Recall: 0.849
Precision: 0.042
Accuracy: 0.194
False omission rate: 0.038</td>
</tr>
<tr>
<td>Metrics for Support Vector
Classifier:

prediction (col) 0.0 1.0
actual (row)
0.0 6343 7390
1.0 213 375

Recall: 0.638
Precision: 0.048
Accuracy: 0.469
False omission rate: 0.032</td>
<td></td>
<td>Metrics for tuned
LinearLearner - target 70%
recall.

prediction (col) 0.0 1.0
actual (row)
0.0 1044 12689
1.0 37 551

Recall: 0.937
Precision: 0.042
Accuracy: 0.111
False omission rate: 0.034</td>
</tr>
<tr><td colspan="3"></td></tr>
<tr><th colspan="3">Non-linear models</th></tr>
<tr><th>KNeighborsClassifier</th><th>XGBoost Classifier</th><th>Radial Basic Function
Support Vector Classifier</th></tr>
<tr>
<td>Metrics for
KNeighborsClassifier:

prediction (col) 0.0 1.0
actual (row)
0.0 12888 845
1.0 538 50

Recall: 0.085
Precision: 0.056
Accuracy: 0.903
False omission rate: 0.040</td>
<td>Metrics for XGBClassifier:

prediction (col) 0.0 1.0
actual (row)
0.0 8011 5722
1.0 236 352

Recall: 0.599
Precision: 0.058
Accuracy: 0.584
False omission rate: 0.029</td>
<td>Metrics for RBF SVC:

prediction (col) 0.0 1.0
actual (row)
0.0 6763 6970
1.0 201 387

Recall: 0.658
Precision: 0.053
Accuracy: 0.499
False omission rate: 0.029</td>
</tr>
</table>

4.2 Justification

Both the XGBoost Classifier and the Radial Basic Function Support Vector Classifier perform better than the linear model benchmark, Support Vector Classifier.

Measure	Support Vector Classifier (Benchmark)	XGBoost Classifier	Radial Basic Function Support Vector Classifier
False omission rate	3.2%	2.9%	2.9%
Accuracy	46.9%	58.4%	49.9%
Precision	4.8%	5.8%	5.3%
Recall	63.8%	59.9%	65.8%

These models will need further tuning to meet the University of Pennsylvania's research system with a predictive value of 74%.

5 Conclusion

This project has been an attempt to assist clinicians in the detection of sepsis, but focusing on features that are on the SIRS criteria. The MIT MIMIC-III Patient Database has rich information collected over 12 years at the critical care unit. Only a subset of the database is utilized.

Because of the large data files, significant time has been spent on preprocessing the data and extracting only what is needed for our analysis. The date shifts to de-identify the patients need extra preprocessing, and the rule that patients 90 years or older at the time of hospital admission had their date of birth replaced. This results in some estimations made to set these patients as 90 years old, in order to keep them in the analysis.

There are a wide range of binary classification algorithms available. To truly understand how the hyperparameters should be tuned to improve performance will require much more time and practice.

The results from the models trained in this project are what was expected, given that sepsis has been difficult to detect even to experienced clinicians. Non-linear models such as XGBoost Classifier and the Radial Basic Function Support Vector Classifier have been shown to perform better than the benchmark of linear classifiers. Both XGBoost Classifier and the Radial Basic Function Support Vector Classifier have the lowest false omission rate of 2.9%, and are able to detect more than half of the sepsis cases, achieving accuracies of 58.4% and 49.9% respectively.

For future iterations to this project, these are areas to improve on:

- Further adjust hyperparameters to improve model performance. The XGBoost Classifier and the Radial Basic Function Support Vector Classifier models have the potential to improve.
- Revisit the dataset and find more features to improve the model training. This includes laboratory reports and textual notes from clinicians.

6 References

1. Sepsis, Centers for Disease Control and Prevention

   https://www.cdc.gov/sepsis/index.html

2. Balk RA. Systemic inflammatory response syndrome (SIRS): where did it come from and is it still relevant today?. *Virulence*. 2014;5(1):20–26. doi:10.4161/viru.27135
   https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3916374/

3. Hatfield, Kelly M. MSPH1; Dantes, Raymund B. MD, MPH1,2; Baggs, James PhD1; Sapiano, Mathew R. P. PhD1; Fiore, Anthony E. MD, MPH1; Jernigan, John A. MD, MS1; Epstein, Lauren MD, MSc. Assessing Variability in Hospital-Level Mortality Among U.S. Medicare Beneficiaries With Hospitalizations for Severe Sepsis and Septic Shock. Critical Care Medicine: November 2018 - Volume 46 - Issue 11 - p 1753-1760. doi: 10.1097/CCM.0000000000003324
   https://journals.lww.com/ccmjournal/Citation/2018/11000/Assessing_Variability_in_Hospital_Level_Mortality.6.aspx

4. Johnson, A., Pollard, T., Mark, R. (2016). MIMIC-III Clinical Database. PhysioNet. doi:10.13026/C2XW26

5. Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035.

6. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals (2003). Circulation. 101(23):e215-e220.

7. Zuick S, Asch DA, Graustein A, et al. Can a Computerized Sepsis Screening and Alert System Accurately Diagnose Sepsis in Hospitalized Floor Patients and Potentially Provide Opportunities for Early Intervention? A Pilot Study. J Intensive & Crit Care 2016, 2:3.

8. Wikipedia. Sensitivity and specificity. https://en.wikipedia.org/wiki/Sensitivity_and_specificity

9. Sci-kit Learn. Classifier comparison. https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html

10. Society of Critical Care Medicine. Critical Care Statistics. https://www.sccm.org/Communications/Critical-Care-Statistics