Early Detection of Sepsis Using Physiological Data

karan sindwani Jul 5, 2019 · 7 min read ★

What is Sepsis ?

infection. Sepsis occurs when the body's response to these chemicals is out of balance, triggering changes that can damage multiple organ systems.

Sepsis is caused by infection and can happen to anyone. Sepsis is most common and most dangerous in:

- Older adults
- Pregnant women
- Children younger than 1
- People who have chronic conditions, such as diabetes, kidney or lung disease, or cancer
- People who have weakened immune systems

Statistics

- In USA, 270,000 people die from sepsis each year
- Internationally , 6 Million people die from sepsis each year
- US hospitals spend 24 Billion each year on sepsis (13 % of Health Budget)
- Each hour of delay in treatment can roughly increase mortality by 4–8 %

> *Source : https://www.mayoclinic.org/diseases-conditions/sepsis/symptoms-causes/syc-20351214*

Objective

The goal of this blog is the early detection of sepsis using physiological data. The early prediction of sepsis is potentially life-saving, and we aim to predict sepsis 6 hours before the clinical prediction of sepsis. Conversely, the late prediction of sepsis is potentially

Challenge Data

The Challenge data repository contains one file per patient (e.g., training/p00101.psv).

Each training data file provides a table with measurements over time. Each column of the table provides a sequence of measurements over time (e.g., heart rate over several hours), where the header of the column describes the measurement. Each row of the table provides a collection of measurements at the same time (e.g., heart rate and oxygen level at the same time).

HR	O2Sat	Temp	...	HospAdmTime	ICULOS	SepsisLabel
NaN	NaN	NaN	...	-50	1	0
86	98	NaN	...	-50	2	0
75	NaN	NaN	...	-50	3	1
99	100	35.5	...	-50	4	1

Features:

- **Vital Signs** : Heart Rate, Temperature , Blood Pressure, Respiratory rate,
- **Laboratory Values :** Platelet Count, Glucose , Calcium etc
- **Demographics :** Age, Gender, Time in ICU , Hospital Admit time

Label :

0 (Non-sepsis) and 1 (Sepsis)

> *Note :Feature description and data can be downloaded from https://physionet.org/challenge/2019/*

records do not have a time-label associated with them, so that opens the scope of interpreting it as a non-temporal problem (ignoring the time component)

There are two ways in which one can approach this problem:

1. **Temporal Approach** : Take into the account the time component for the data. Sepsis is diagnosed for each patient at each hour using the past data.
2. **Non-temporal Approach** : 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)

For this blog I am going to talk about only the Non-temporal approach .

Non-Temporal Approach

In this approach we ignore the time component associated with each patient hourly record and treat them as independently and identically distributed.

Train-Validation-Test -Split

The data repository has data from two hospitals and a total of 40 thousand patients. The actual number of records would be higher as a patient could have stayed in the hospital for a variable amount of time.

Splitting these records to train , validation and test. While splitting I have made sure that each patient is fully contained in exactly one of the splits.

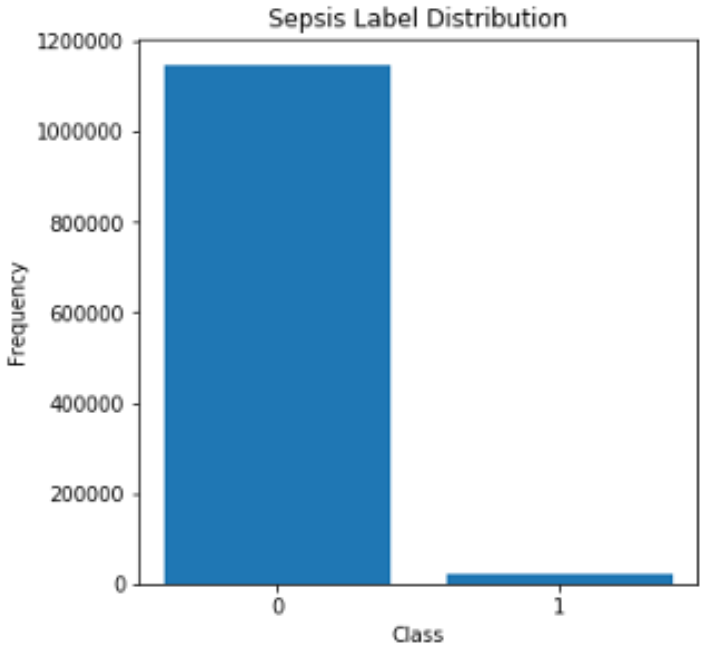
- **Train** : 30K Patients
- **Test** : 5K Patients
- **Validation** : 5K Patients

Exploratory Data Analysis

After performing descriptive data analysis on the train data, these were the concerns that highlighted

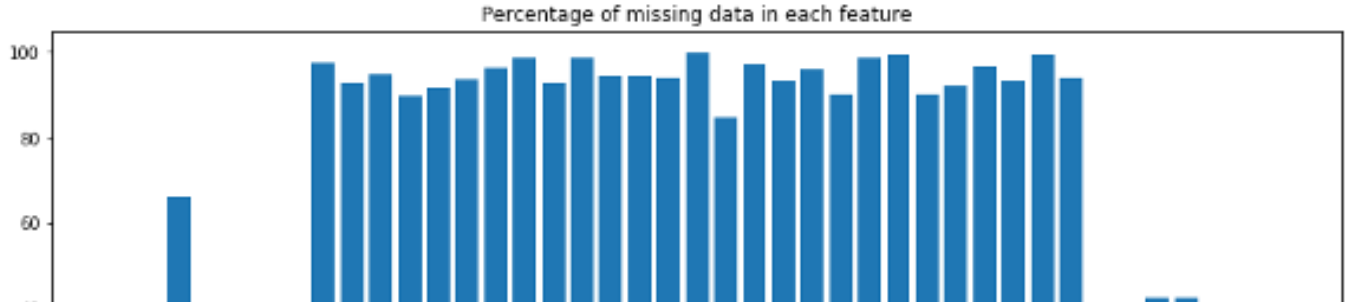
Concerns

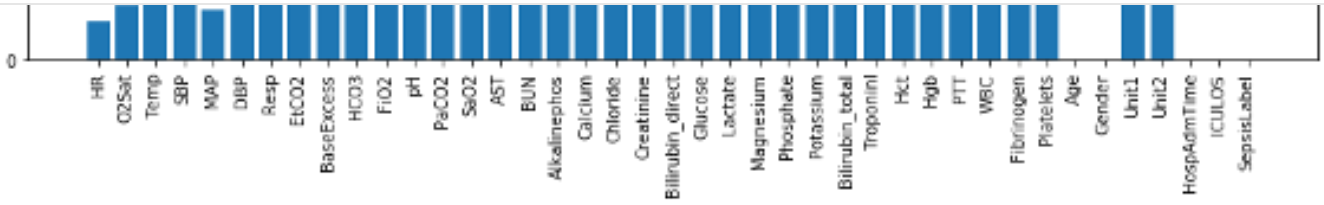
Extremely Imbalance data : As we can see from the bar plot, the records are extremely imbalanced (Less than 1 % vs 99 %+) with the minority class being Sepsis (1).



Label Distribution

Missing Data : High Percentage of missing data in most of the features





> *Note : Detailed EDA and baseline can be found on https://github.com/kskaran94/Sepsis_Identification*

Handling Class Imbalance

There are various pre-defined ways of handling class imbalance in machine learning, which have proven to be successful in many scenarios.

However most of them can not be applied in this problem. As we have an extreme rare imbalance , undersampling the majority class would lead to 99% data loss.

Oversampling or SMOTE(Synthetic Minority Over-sampling Technique) can applied to the minority class, but it would not be a conceptually correct idea as we are dealing with real world health care records and we would want to preserve the original distribution of data .

For Sepsis Identification non-temporal approach, we are going ahead with the original distribution of data and choosing an appropriate evaluation metric for modeling the data.

Handling Missing Data

There are various pre-defined ways of imputing continuous data such as Median , Mean etc, which have proven to be successful in many scenarios.

However continuous data imputation can not be applied in this problem. It would not be a conceptually correct idea as we are dealing with real world health care records and we would want to preserve the original distribution of data .

For Sepsis Identification non-temporal approach, instead of imputing continuous features. We engineer categorical features out of existing continuous features and impute the missing with a new category.

Feature Selection

The data has 40 features which can broadly be classified into

- **Demographics**
- **Vital Signs**
- **Laboratory values**

After doing research on Sepsis from credible sources like www.cdc.gov, Symptoms of Sepsis are High Fever, Abnormal Blood pressure, High respiratory rate. These symptoms give us a direction that features like Heart Rate, Temperature and Blood pressure may be important while predicting sepsis.

Also sepsis is mostly prevalent either in infant or Old patients. This makes age an important feature.

As pointed in the previous section, we cannot handle missing data in the usual way. And the features in the category of Laboratory values have 90% or more missing data. Imputing these features even after engineering them as categorical would lead to features with low variance. Hence they may not add much information to the model

Therefore the features with more than 80% of missing data are ignored.

Feature Engineering

The Feature engineering for the selected features has been described below

- **Heart Rate** : Converted to Categorical (Normal, Abnormal, Missing)
- **O2Stat** : Converted to Categorical (Normal, Abnormal, Missing)
- **Temperature** : Converted to Categorical (Normal, Abnormal, Missing)

- **Age** : Converted to Categorical (Old,Infant, Child/Adult)
- **Gender** : Unchanged
- **HospAdmTime(Hospital Admission Time)** : Unchanged
- **ICULOS(ICU Length of Stay)**: Unchanged

> *Note : Code for Feature Selection and Engineering can be found on https://github.com/kskaran94/Sepsis_Identification*

Model Training and Evaluation

Evaluation Metric

Choosing a suitable/useful evaluation metric is important than we think. In this case choosing a metric like accuracy would not be useful, as a model which predicts majority class would have high accuracy.

For imbalance data problems one can choose either Average_precision or F1_weighted. Since they give a complete representation of the confusion matrix.

We will be going ahead with average precision as the primary metric but still look at other metric like precision and recall.

Machine Learning Models

Converting the Categorical features to OnehotEncoding and scaling the Continuous features.

Performance of different models can be seen in the table below

	Model	Average_Precision
0	Logistic_Regression	0.028

2	Decision_Tree	0.030
3	Random_Forest	0.020
4	LightGBM	0.014

Deep Learning Models

Auto-encoders have proven to be useful for anomaly detection use-cases which involves high class imbalance.

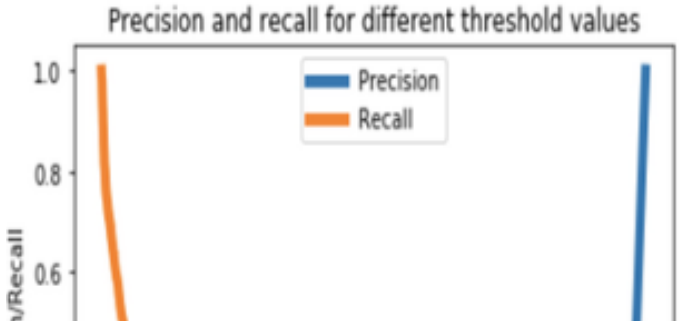
This blog post offers a comprehensive approach for using auto-encoders for extremely rare classification

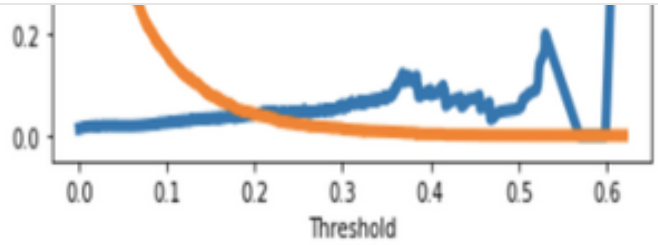
> *Blog link : https://towardsdatascience.com/extreme-rare-event-classification-using-autoencoders-in-keras-a565b386f098*

Auto-encoders were expected to perform better than traditional Machine Learning models as they are modeling the behavior of positive class and treating the negative class as an anomaly.

Auto-encoders increased the average precision to 7 percent, which is a good number for the health care domain. Considering the case we are not compromising neither false positives nor false negatives.

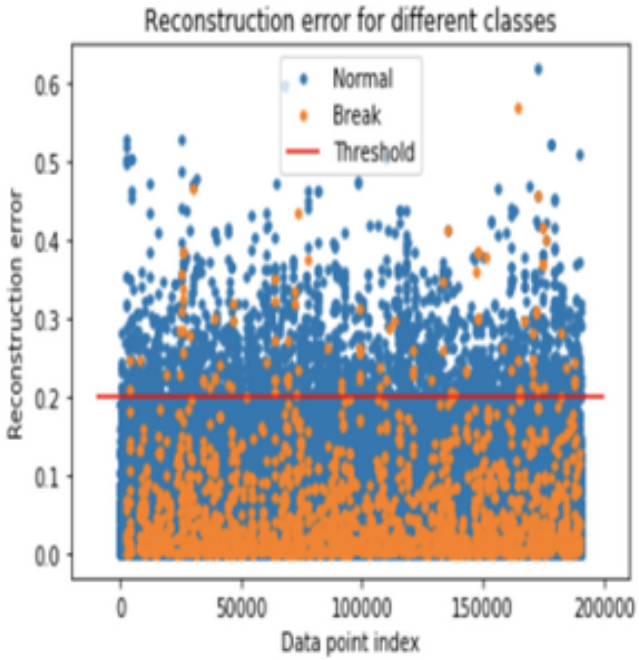
Plotting a precision-recall curve shows that over any threshold , both precision and recall don't take higher values. This explains our relatively low average precision value.





Precision-recall curve

We can confirm our thinking by plotting the reconstruction error for a single threshold. As we can see the red line (threshold) cannot perfectly divide the data.



> *Note : Code for Model Training and Evaluation can be found on https://github.com/kskaran94/Sepsis_Identification*

Conclusion

After looking at both Machine and Deep Learning models we can conclude that we need to add more features or data , for a better model performance. Or even switch to

Connect with me on LinkedIn : https://www.linkedin.com/in/karansindwani/

Machine Learning Data Science Sepsis Health Analytics Deep Learning