Early Sepsis Diagnosis using Machine Learning

Matjaž Muc
Faculty of Computer and Information Science
Večna pot 113, 1000 Ljubljana
email: mm1706@student.uni-lj.si
code: https://github.com/Matjaz12/Early-Sepis-Diagnosis

***Abstract*—Sepsis is a potentially life-threatening condition that occurs as a response to body's response to infection. It might cause tissue damage, organ failure or can even cause death. Internationally, an estimated 30 million people have developed Sepsis and more than 6 million people die from it each year. Hence coming up with a good model that might be able to solve this issue becomes imminent and important.**
In this paper we aim to analyze and evaluate several algorithms in hopes to answer some questions we face.

Keywords:
Overview of sepsis symptoms and reason on why sepsis is hard to diagnose.
Solutions to common problems in medical data.
Comparison of several models and strategies.

I. Introduction

Early detection of sepsis is extremely important for improving sepsis outcomes. Each hour of delayed treatment can increase mortality rate by 4-8%, mortality rate typically ranges from 27-42%. The condition is known to be one of the hardest medical conditions to diagnose.

Immediate challenges we encounter during the analysis come from the nature of the data-set which consist of real medical data and is therefor far from perfect. The data-set is heavily skewed, containing only 2% of sepsis positive patients. Data-set contains about 50% of features that have more than 90% of missing values. After exploring the feature relations we realize that the boarded between the two classes is not well defined, which is to be expected since the condition is hard to diagnose, even more in most patients it is not very different from non-septic infection. We should also realise we are classifying between patients in Intensive care units (ICUs), therefor their health condition is severe regardless of them having sepsis or not. We attempt to solve the imbalance problem by using strategies presented in paper [kononenko pape]. In order to solve the in-completeness of data, we pick features that differ between classes and impute the missing values for them. In this paper we will explore and evaluate performance of a logistic regression classifier, random forest model and neural networks. We also try to perform One-class classification, by learning features of non-septic patients in order to distinguish between the two.

In this paper we start by taking a look at the general data information such as its origin, reliability, and structure. We continue speaking on the analysis of the data, we present our discoveries and challenges we faced. Afterward we introduce solutions to missing values problem and data-set imbalance problem. We review results of a few distinct models. TODO We review results of the One Class Classification strategy. Finally we present paper discoveries in the Results section.

January 5, 2022

II. Materials and Methods

A. Related work

Md. Mohaimenul Islam, Tahmina Nasrin, Bruno Andreas Walther, Chieh-Chen Wu, Hsuan-Chia Yang, Yu-Chuan Li,Prediction of sepsis patients using machine learning approach: A meta-analysis,Computer Methods and Programs in Biomedicine,Volume 170, 2019, Pages 1-9,ISSN 0169-2607.

X. Wang, Z. Wang, J. Weng, C. Wen, H. Chen and X. Wang, A New Effective Machine Learning Framework for Sepsis Diagnosis, in IEEE Access, vol. 6, pp. 48300-48310, 2018, doi: 10.1109/ACCESS.2018.2867728.

Silvia Cateni, Valentina Colla, Marco Vannucci,A method for resampling imbalanced datasets in binary classification tasks for real-world problems,Neurocomputing,Volume 135,2014,Pages 32-41,ISSN 0925-2312.

B. Data Overview

The data is obtained from Beth Israel Deaconess Medical Center. It was collected for 40,336 patients and posted on the Physio net Challenge 2019 and is available for the public.
The data consists of a of hourly vital sign summaries, laboratory values, and patients demographics. Data-set includes over 790215 data points. The data is extracted from the Electronic Medical Record (ERM) and underwent series of pre-processing steps prior to being published.
As mentioned before the data-set has an hourly time sequence record for each patient. Each patient has 40 measurements spaced between one hour period, the current state of patient may change at any of 40 time instances.

We can approach the classification problem in two ways. In this paper we focus on the second approach.

1) Taking in the time component.
   We can take in the time component as a part of the data-set, this provides us with additional information of

how attributes of a patient change over time. Following this approach would require us to perform time series classification, using Long Short Term Memory Neural Network. This presents a need for a lot of computing resources, since the number of data-points is significant.

2) Ignoring time component.
We can completely ignore the time component, resulting in a loss of any information about attribute dynamics. In this case we would treat each hourly measurement as an individual patient.

Features are listed bellow. Note that the sepsis label is set to "positive" 6 hours in advance. This insures all of our predictions are made ahead of time, decreasing the probability of too late diagnosis.

$$t >= t_s - 6 \longrightarrow SepsisLabel = 1 \quad (1)$$

$$t \ < t_s - 6 \longrightarrow SepsisLabel = 0 \quad (2)$$

$$t_s \ ... \ time \ of \ sepsis \quad (3)$$

1) HR Heart rate [beats per minute]
2) O2Sat Pulse oximetry [%]
3) Temp Temperature [deg C]
4) SBP Systolic BP [mm Hg]
5) MAP Mean arterial pressure [mm Hg]
6) DBP Diastolic BP [mm Hg]
7) Resp Respiration rate [breaths per minute]
8) EtCO2 End tidal carbon dioxide [mm Hg]
9) BaseExcess Excess bicarbonate [mmol/L]
10) HCO3 Bicarbonate [mmol/L]
11) FiO2 Fraction of inspired oxygen [%]
12) pH [pH]
13) PaCO2 Partial pressure of carbon dioxide from arterial blood [mm H]
14) SaO2 Oxygen saturation from arterial blood [%]
15) AST Aspartate transaminase [IU/L]
16) BUN Blood urea nitrogen [mg/dL]
17) Alkalinephos Alkaline phosphatase [IU/L]
18) Calcium Calcium [mg/dL]
19) Chloride Chloride [mmol/L]
20) Creatinine Creatinine [mg/dL]
21) Bilirubin direct Direct bilirubin [mg/dL]
22) Glucose Serum glucose [mg/dL]
23) Lactate Lactic acid [mg/dL]
24) Magnesium Magnesium [mmol/dL]
25) Phosphate Phosphate [mg/dL]
26) Potassium Potassiam [mmol/L]
27) Bilirubin total Total bilirubin [mg/dL]
28) TroponinI Troponin I [ng/mL]
29) Hct Hematocrit [%]
30) Hgb Hemoglobin [g/dL]
31) PTT Partial thromboplastin time [seconds]
32) WBC Leukocyte count [count/L]
33) Fibrinogen Fibrinogen concentration [mg/dL]
34) Platelets Platelet count [count/mL]
35) Age [years]
36) Gender Female [0 / 1]
37) Unit1 Identifier for ICU unit [MICU] [0 / 1]
38) Unit2 Identifier for ICU unit [SICU] [0 / 1]
39) HospAdmTime Time between hospital and ICU admission [hours since admission]
40) ICULOS ICU length of stay [hours since ICU admission]
41) SepsisLabel [0 / 1]

C. Data Analysis

First we should address a common problem when dealing with real world medical data, that being the imbalance of data-sets. Often the number of negative patients is much larger than the number of positive patients. This presents challenges when fitting our model and evaluating its performance. As mentioned before our data-set consists of only 2% of sepsis-positive patients.

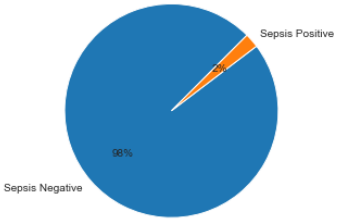

Fig. 1. Ratio between sepsis-positive and sepsis-negative patients

While finding data distribution, we found that the data-set contains lots of missing values, this again is common with medical data. Due to the cost of particular measurements or their in-accessibility at a particular time some measurements are missing. Its also worth noting that the patients are in hospital for different reasons therefore some measurements may have a higher priority than others. Some measurements are also taken only in case patients state worsens.

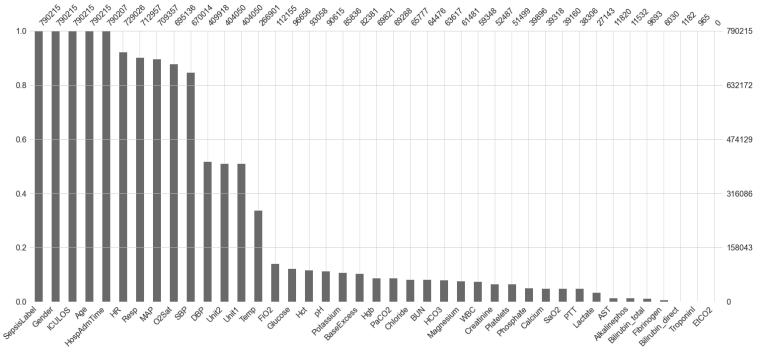

Fig. 2. Percentage of missing values

Next we'll take a look at the distribution of attribute values and how the distribution differs between sepsis-positive and sepsis-negative patients. Using this information we'll try to decide which attributes to keep and which to remove.

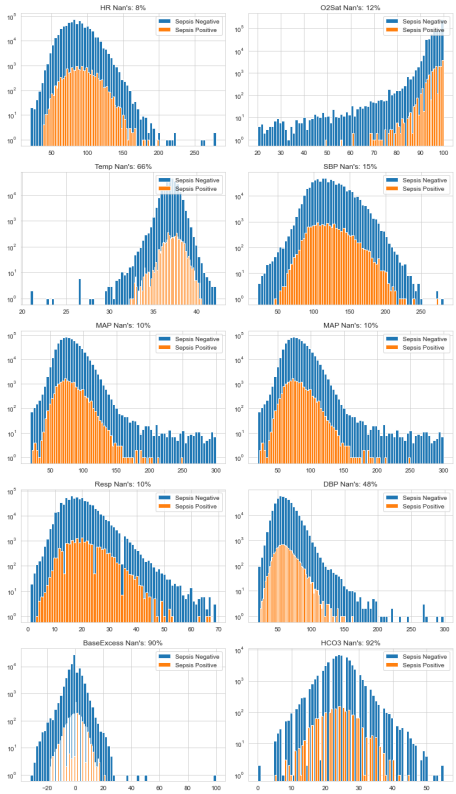

Fig. 3.

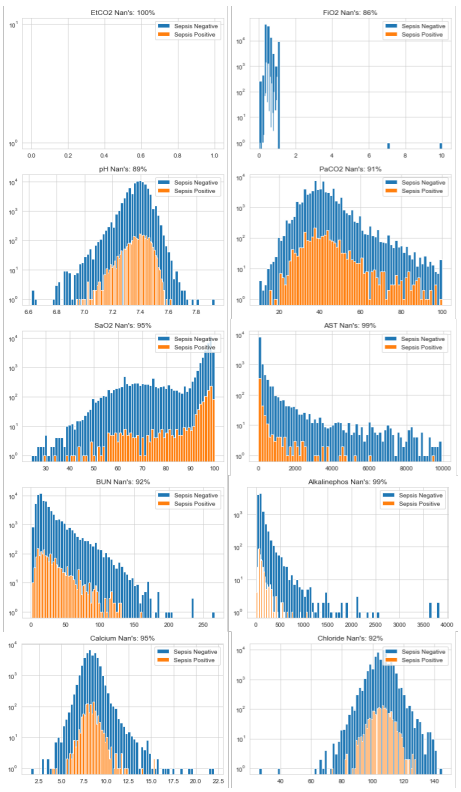

Fig. 4.

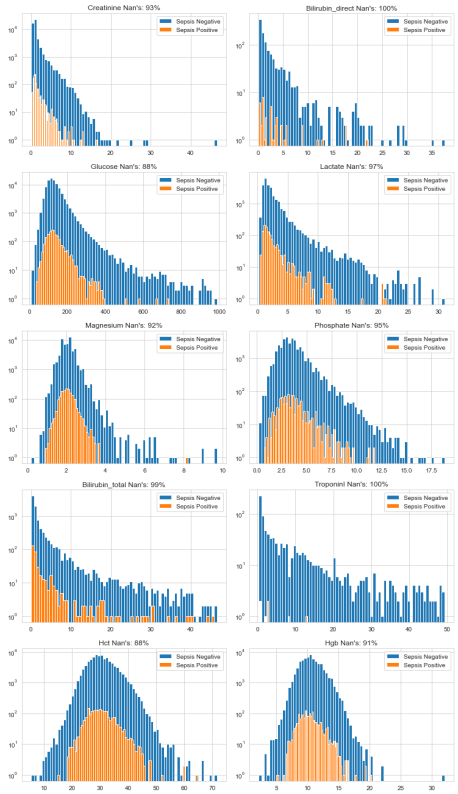

Fig. 5.

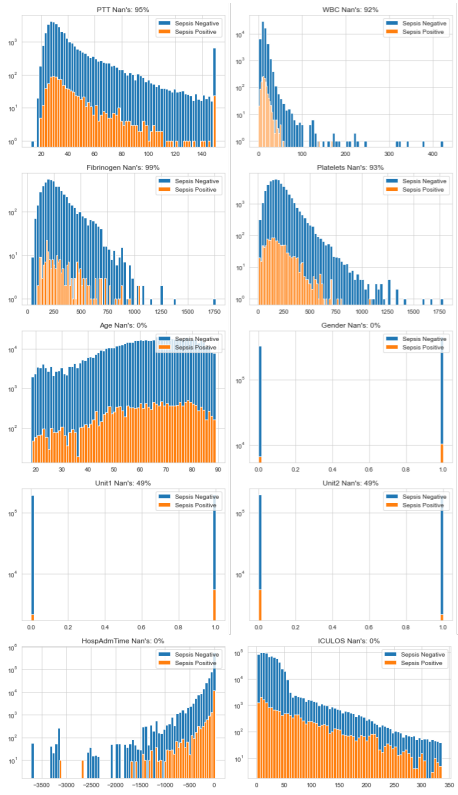

Fig. 6.

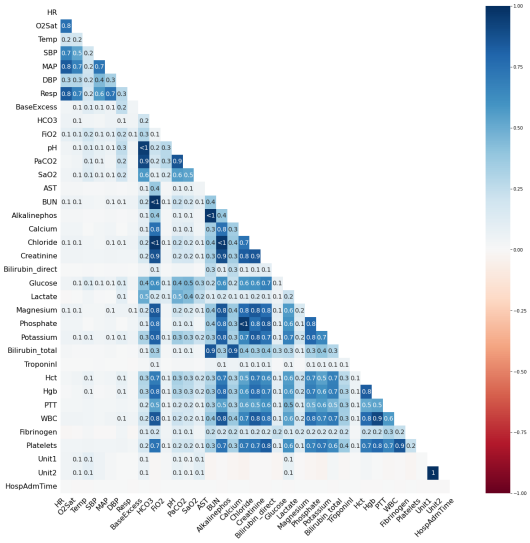

Fig. 7. Feature correlation map

Looking at the distribution plots we notice that non-septic patient attributes tend to cover a larger spectrum of values. Here we point out a few features we may consider removing based on their statistical properties.

FiO2 is represented by somewhat discrete values with bi-modal distribution, this and the fact that only 20% patients have this attribute is enough of a reason to remove this feature. Further investigation shows this feature is strongly correlated with Bilirubin direct, which is missing for about 96% of patients, hence we keep only Bilirubin total.

There is a high correlation between DBP (diastolic BP), MAP (mean arterial pressure), and SBP (systolic BP). Mean arterial pressure is known (source) to be primary indicator of patient state in near septic conditions, thus, both DBP and SBP may be disregarded as MAP is calculated from these two features. Hemoglobin (Hgb) and hematocrit (Hct) values happened to be highly correlated. Research articles about sepsis diagnosis mostly concentrate on hemoglobin out of these two. Let's stick with experts and leave just Hgb. We also remove EtCO2, since it has no values and both Unit1 and Unit2. At this point we realize that an expert in domain of sepsis diagnosis may help decide which features are relevant and which carry the most value for our problem.

Finally lets take a look at a few interesting scatter plots. Note that we ignore most of them, since the total number of possible scatter plots is given by the following formula.

$$m = \frac{n * (n-1)}{2} \tag{4}$$

$$n \ ... \ number \ of \ features \tag{5}$$

In our case m = 780. Here we select pairs of features that tend to separate the classes the best. We notice that among the best scatter plots, their does not exists a pair of features that completely separates the two classes.

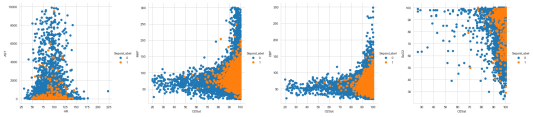

Fig. 8. Scatter plot

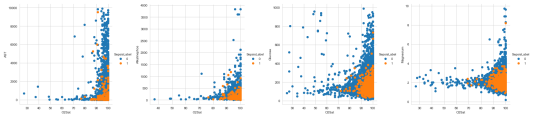

Fig. 9. Scatter plot

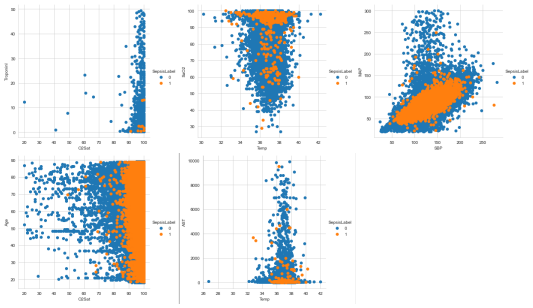

Fig. 10. Scatter plot

Here we should spend a few words on how sepsis is typically diagnosed and why its diagnosis is difficult. The following paragraph is taken from paper[why hard to diagnose]

Infection is typically identified by three sets of information.

1) Clinical signs and symptoms of a host response: increase in white blood cell count, increase in concentrations of inflammatory markers.
2) The presence of signs of infection:
   respiratory symptoms with abnormal chest auscultation and typical radiographic chest infiltrates, present signs of meningitis.
3) Proven microbiological invasion of a sterile environment.

However, not all mentioned symptoms are always present. E.g an immunosuppressed patient may not develop fever, and a source of infection is sometimes impossible to identify. Early on symptoms of sepsis can also be mistaken for influenza or other infections.

D. Data Imputation

Since features with large percentage of missing values may still hold valuable information for classification we decide to impute missing data by choosing the closest value. This may aggravate the boarder between classes, but completely ignoring these features would very likely make it impossible to differentiate between patients.

E. Data-set imbalance

To balance the data-set we take advantage of two well known strategies: random-oversampling of minority class and Synthetic Minority Oversampling Technique (SMOTE) which produces new synthetic records. These synthetic training records are generated by randomly selecting one or more of the k-nearest neighbors for each example in the minority class.

F. Ignoring the time Component

For the following analysis we treat each hourly record of a patient as a separate patient. We also remove features HospAdmTime and ICULOS indicating time.

1) Logistic Regression: We decide to start with Logistic regression, since it is easy to interpret and very efficient to train. Outputs have a probabilistic interpretation, and the algorithm can be regularized to avoid over-fitting.

Data is first standardised (mean of features is brought to 0 and standard deviation is brought to 1). Results using random oversampling and SMOTE to balance the data-set are displayed on figures bellow.

In order to produces best models, a search for optimal hyper-parameters is performed. Hyper-parameters used include a list of regularization strength inverse values, and two penalty conditions (L1, L2).

Each combination of parameters is evaluated using 10 cross-validation folds. Model that produces best f1-score is selected.

```
Selected Hyper parameters
Parameters: {'penalty': 'l2', 'C': 774.2636826811278}
Score: 0.6624070029981123
```

Fig. 11. Logistic Regression Random Oversampling Hyper-parameters

```
Selected Hyper parameters
Parameters: {'penalty': 'l1', 'C': 0.1}
Score: 0.6625957730317948
```

Fig. 12. Logistic Regression SMOTE Hyper-parameters

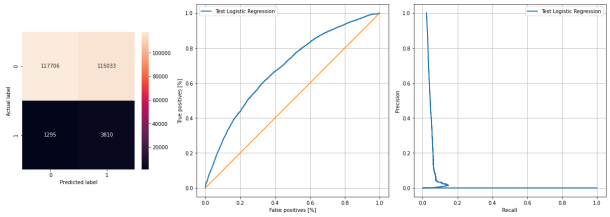

Fig. 13. Logistic Regression Random Oversampling

```
Accuarcy  = 0.5109063083365567
Recall    = 0.7463271302644466
Precision = 0.03205910318655705
F1Score   = 0.06147739374576435
F2Score   = 0.1367915383124017
```

Fig. 14. Logistic Regression Random Oversampling Report

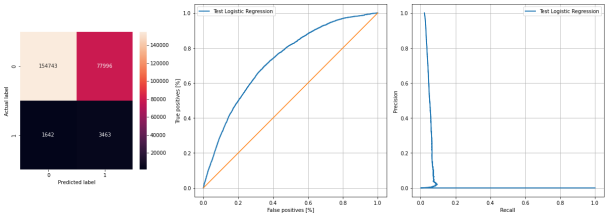

Fig. 15. Logistic Regression SMOTE

```
Accuarcy  = 0.6651670843073612
Recall    = 0.6783545543584721
Precision = 0.04251218404350655
F1Score   = 0.08001016588882214
F2Score   = 0.16995651704472953
```

Fig. 16. Logistic Regression SMOTE Report

TODO:We notice that SMOTE results in much better results compared to over-samling. doesn't outperform the oversampling strategy may be the fact that our data-set is already synthetic to some degree.

In general performance of both is bad, the ROC curve seems promising, but the area under the precision-recall curve is barely above no-skill level.

2) Random Forest: We decide to continue with a random forest, model can avoid the need for adding additional synthetic septic patients, by taking advantage of class weighting parameter of algorithm. We give a larger weight to the minority class. Feature selection is performed to shrink the dimension of feature space. We keep features which provide the most information, this decreases the time needed to train the model and also yields better results. Bellow selected features are listed.

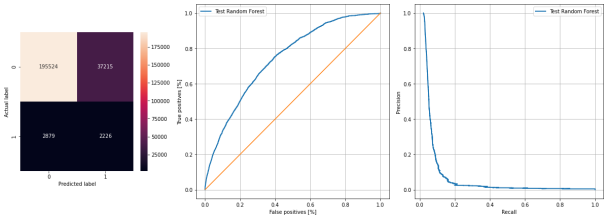

Fig. 17. Random Forest Baseline

```
Accuarcy  = 0.8314273221102908
Recall    = 0.4360430950048972
Precision = 0.05643873126949114
F1Score   = 0.09994163336775469
F2Score   = 0.18593073954661635
```

Fig. 18. Random Forest Baseline Report

```
Selected Features
['HR', 'Temp', 'MAP', 'Resp', 'PaCO2', 'BUN', 'Chloride', 'Creatinine', 'Glucose', 'Potassium', 'Hgb', 'PTT', 'WBC', 'Platelets', 'Age']
```

Fig. 19. Random Forest Selected Features

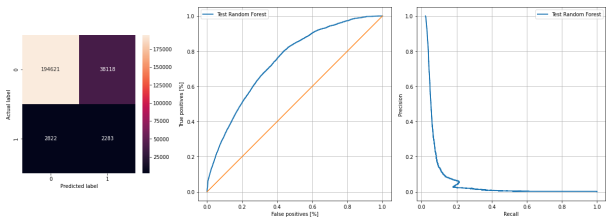

Fig. 20. Random Forest Feature Selection

```
Accuarcy  = 0.8278703688131717
Recall    = 0.4472086190009794
Precision = 0.056508502264795424
F1Score   = 0.10033841691205556
F2Score   = 0.18768188619062495
```

Fig. 21. Random Forest Feature Selection Report

We find that Random Forest produces the largest AUC so far. Since the test data-set is imbalanced we should also pay attention to precision-recall curve. Area under precision-recall curve has increased quite a bit. From the confusion matrix we notice the algorithm has little trouble classifying sepsis negative patients, but still struggles to classify sepsis positive patients.

3) One Class Classification: One-class classification algorithms were originally designed for anomaly detection. In theory this strategy can be effective for imbalanced classification data-sets where there are very few examples of the minority class and or data-sets where there is no coherent structure to separate the classes that could be learned by a supervised algorithm. One-class models ignore the task of discrimination and instead focus on deviations from what is normal or what is expected.

Assuming this may be the case in our problem we pick Isolation Forest algorithm. It is based on modeling the normal data in such a way to isolate anomalies that are both few in number and different in the feature space. The main benefit of the algorithm is its speed, when compared to one class classifiers based on support vector machines.

Tree structures are created to isolate anomalies. The result is that, isolated examples have a relatively short depth in the trees, whereas normal data is less isolated and has a greater depth in the trees.

A search for optimal hyper-parameters is performed using 10 cross validation folds. Bellow the selected hyper parameters and best f1-score achieved are listed.

```
Selected Hyper parameters
Parameters: {'n_estimators': 100, 'max_samples': 300, 'max_features': 5, 'contamination': 0.1, 'bootstrap': True}
F1 Score: 0.016664523354793938
```

Fig. 22. Baseline Isolation Forest

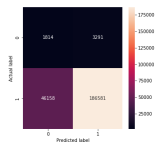

Fig. 23. Baseline Isolation Forest

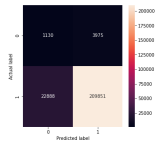

Fig. 24. Best Isolation Forest

```
F1 Score report on sepis positive patients
Isolation Forest F1 Score: 0.068
Best Isolation Forest F1 Score: 0.078
```

Fig. 25. One class classification Report

Comparing the best septic f1-score to best f1-score of random forest we conclude that on this data-set, classic classification methods outperform the one class classifiers.

4) Neural Network: Based on current results it appears that our problem is highly non-linear. In this section we try multiple methods using Neural Network with two hidden layers and a dropout technique to reduce over-fitting on the training data-set. All methods below share the same neural network architecture. Data is first standardized, model is trained using training data-set and validation data-set. Metrics used to optimise during training include: TP, FP, TN, FN, Accuracy, Precision, Recall, ROC and PRC.

```
Layer (type)                 Output Shape              Param #
=================================================================
dense_51 (Dense)             (None, 500)               15500

dropout_34 (Dropout)         (None, 500)               0

dense_52 (Dense)             (None, 250)               125250

dropout_35 (Dropout)         (None, 250)               0

dense_53 (Dense)             (None, 1)                 251

=================================================================
Total params: 141,001
Trainable params: 141,001
Non-trainable params: 0
_________________________________________________________________
```

Fig. 26. Neural Network Architecture

5) Baseline Model: The baseline model is fitted using experimentally picked batch size of 60 and 100 epochs.

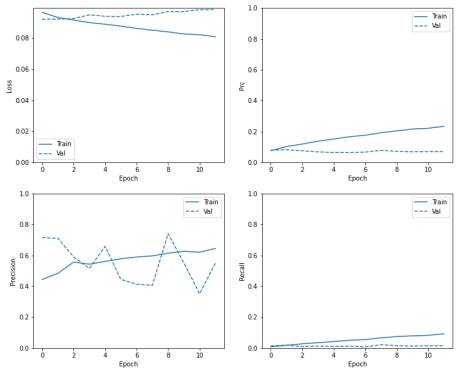

Fig. 27. Baseline model Metrics

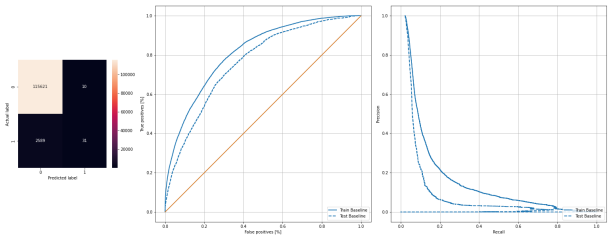

Fig. 28. Baseline model

```
accuracy :  0.9780213236808777
precision :  0.7560975551605225
recall :  0.011832061223685741
auc :  0.7609980702400208
prc :  0.0925351083278656
```

Fig. 29. Baseline model Report

6) Weighted model: To deal with imbalance we compute weights for both classes.

$$W_0 = \frac{1}{neg} * \frac{total}{2.0} = 0.51 \tag{6}$$

$$W_1 = \frac{1}{pos} * \frac{total}{2.0} = 22.96 \tag{7}$$

$$neg \ ... \ number \ of \ sepsis \ negative \ patients \tag{8}$$

$$pos \ ... \ number \ of \ sepsis \ positive \ patients \tag{9}$$

$$total \ ... \ number \ of \ all \ patients \tag{10}$$

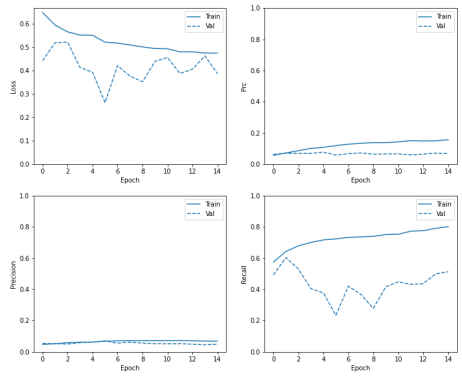

Fig. 30. Weighted model Metrics

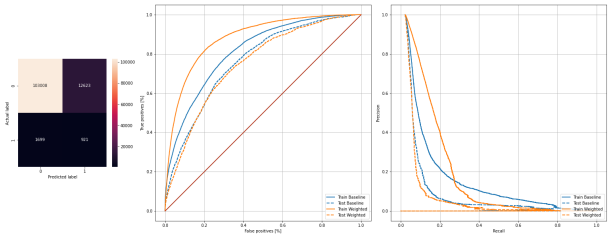

Fig. 31. Weighted model

```
accuracy :  0.8788847327232361
precision :  0.06800059229135513
recall :  0.3515267074108124
auc :  0.7507770657539368
prc :  0.07772089540958405
```

Fig. 32. Weighted model Report

We can that the accuracy and precision of weighted model are lower because there are more false positives, but conversely the recall and AUC are higher because the model also found more true positives. Despite having lower accuracy, this model has higher recall (identifies more septic patients). Medical expert should consider the trade-offs between these different types of errors(recall and accuracy).

7) Class Weights with Feature Selection: In this section we try a simplified version of a method proposed in paper [second paper].We first scale our features, we perform feature selection using Random Forest and Relief algorithm. We use a batch-size of 60 and 100 epochs to train the model.

TODO: INCASE REUSLT DIFFER INLCUDE BOTH ELSE INCLUDE RELIEF VERISON

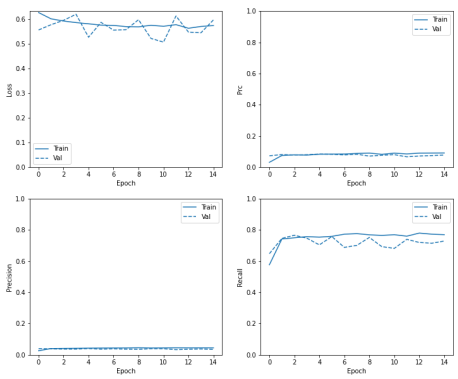

Fig. 33. Weighted model Metrics

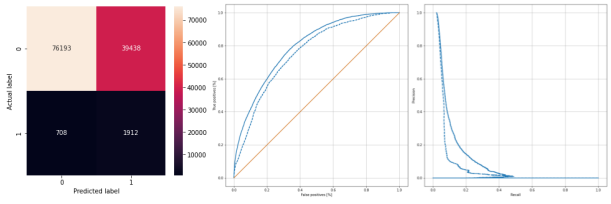

Fig. 34. Weighted model

```
accuracy :  0.66050183773040077
precision :  0.046239420771598816
recall :  0.7297710180282593
auc :  0.7593860626220703
prc :  0.07395583391189575
```

Fig. 35. Weighted model Report

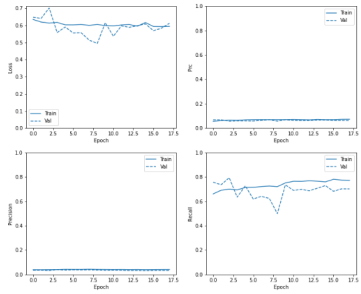

Fig. 36. Weighted model Metrics

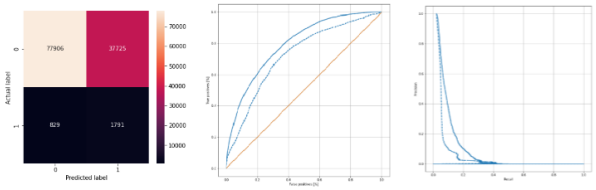

Fig. 37. Weighted model

```
accuracy :  0.6739646792411804
precision :  0.045323412865400314
recall :  0.6835877895355225
auc :  0.721988320350647
prc :  0.0628448948264122
```

Fig. 38. Weighted model Report

Judging by the confusion matrix this are the best results so far, the area under the precision-recall curve has also increased quit a bit.

G. Results

Doseženi rezultati so predstavljeni v poglavju Results. Pri predstavitvi rezultatov je priporočljiva uporaba tabel in slik. Rezultate je potrebno pred staviti objektivno in brez interpretacij. Tudi negativni rezultati so lahko pomembni in kot take jih je zaželeno predstaviti.

TODO: ADD TABLE OF ALL RESULTING METRICS OF DIFFERENT MODELS

III. Conclusion

We conclude that classification on imbalanced data-sets still present a challenge, especially if the data-set also has a large percentage of missing values. Methods such as over-sampling and SMOTE do increase the recall, but the price we pay results in precision and accuracy drop. Effectiveness of SMOTE appears to be insignificant in case data-set is already synthetic too some degree. In our case the effect of random over-sampling was the same as SMOTE. Algorithms that provide the ability to weight minority classes perform competitively, they provide a benefit of fast training, which is an important role when dealing with large data-sets. One class learning approach presents an alternative method, this approach may be more effective if the feature space of the classes where more distinct. The best results where achieved using a neural network in combination with feature selection.

References

[1] Association, Information Košir, Domen Bosnic, Zoran Kononenko, Igor. (2013). The Use of Prediction Reliability Estimates on Imbalanced Datasets. 10.4018/978-1-4666-2455-9.ch035.

[2] Vincent JL. The Clinical Challenge of Sepsis Identification and Monitoring. PLoS Med. 2016;13(5):e1002022. Published 2016 May 17. doi:10.1371/journal.pmed.1002022

[3] F. T. Liu, K. M. Ting and Z. Zhou, "Isolation Forest," 2008 Eighth IEEE International Conference on Data Mining, 2008, pp. 413-422, doi: 10.1109/ICDM.2008.17.

[4] PLoS One. 2015; 10(3): e0118432. Published online 2015 Mar 4. doi: 10.1371/journal.pone.0118432

[5] https://www.physionet.org/content/challenge-2019/1.0.0/

[6] https://machinelearningmastery.com/one-class-classification-algorithms

[7] https://www.tensorflow.org/tutorials/structured$_data/imbalanced_data$

[8] https://machinelearningmastery.com/how-to-control-the-speed-and-stability-of-training-neural-networks-with-gradient-descent-batch-size/

[9] https://machinelearningmastery.com/cost-sensitive-neural-network-for-imbalanced-classification/

[10] https://compsci682.github.io/notes/neural-networks-1/