Predicting Mortality Within 30 days for Sepsis-III Patients Using a Voting Classifier

Prepared by: Eric Rodriguez, Isidro Romille Pride, and Rachel Daniel
Team 3
Target Audience: Hospital Executive

Executive summary

Predicting mortality after sepsis diagnosis is an important and developing area of research that presents unique challenges, which include but are not limited to: a lack of available data that reveal early insights into patient mortality, an absence of predictive accuracy in the current state of the art models that aim to predict sepsis mortality, and sepsis being improperly diagnosed by healthcare professionals. Our project is designed to make use of available data that is collected within 24 hours of a patient being admitted to a hospital in order to predict mortality from sepsis occurring within 30 days of diagnosis. Our goal is to create a robust model that can be deployed in hospital settings for sepsis-III patients. Robustness in our case can be defined as having high specificity, high precision, and having a harmonic balance between the two for both the positive and negative class (high F1 score). Our voting classifier model, which is an ensemble of Logistic Regression, Random Forest, and XGBoost methods, was able to achieve an accuracy of 76%, and an F1 score of 0.54 overall for predicting mortality.

A key challenge that we faced in our analysis was the strong imbalance within the data between the positive and negative classes. Due to this class imbalance, the model ultimately wasn’t able to train on enough data points associated with the positive class to achieve a genuinely strong performance. However, our project still reveals valuable insights for healthcare data scientists dealing with sepsis mortality. We go into areas of suggested research for improvements in ML performance for similar models below, which we believe can help provide the data needed to make these models more reliable.

Introduction

Sepsis is one of the leading causes of death within hospitals as it its treatment is challenging, and it can quickly lead to organ dysfunction and failure (AAMC News, 2023). Sepsis occurs when the body's immune system responds aggressively to infection (commonly hospital-acquired) and attacks vital organs (Mayo Clinic, 2023) It is important to research and explore various machine learning methods to aid in early intervention and treatment of patients diagnosed with this deadly condition. This report presents an ensemble of various machine learning methods represented in a voting classifier model that has the potential to predict if death will occur within 30 days of sepsis-III diagnosis based on metrics collected within a patient's first 24 hours within the ICU. The model has the potential to be used by medical professionals to identify patients with specific pre-existing factors and health measurements that are related to 30-day death and perform additional intervention when necessary.

The data utilized within this report comes from the MIMIC-III database, which is a freely available database containing deidentified medical data. The dataset, which was sourced by Hou et al. (2020) who previously explored a similar idea, includes first-admission data for adult patients diagnosed with "sepsis", "severe sepsis", and "septic shock." The features we chose to include from the provided dataset are pre-existing conditions and factors, such as diabetes and age, and in-hospital metrics that may indicate severity of illness, such as SOFA score and WBC count. A major challenge that we faced was an imbalance in classes within the dataset, where the majority of the data was the surviving class, represented by a 0 or negative classification. This was an issue for us since we are targeting the non-surviving class or the positive class represented by a 1. We attempted to tackle this class imbalance problem with a multitude of informed data-driven approaches, including under sampling and oversampling, while leaving the validation set in its original form to evaluate performance in an unbiased manner. We also recognized that the data is particularly biased in regard to age and race as the patient population is primary older (over 60 years old) and white. These characteristics of the data should be accounted for and addressed in future model and sepsis research.

Data Analysis and Approach

We utilized a multifaceted approach to feature selection and feature engineering that included statistics (e.g., ANOVA, chi-square test), visualizations, such density plots and bar charts, and trial and error. We recognized the importance of strong features within this model, since certain features that have an incredibly strong correlation with our outcome variable unfortunately had to be excluded due to potential for data leakage. One of these variables is length of stay (LOS). The reason this had to be excluded was because by the time this variable is known, the patient would have already finished their stay at the hospital, thus it may have already been too late to intervene.

We first began by performing EDA in Python, where we created density plots for numerical features that we believed would be strong predictors of sepsis 30-day mortality, given our domain knowledge on the subject. These features included heartrate (min and max), urine output, creatinine, and SOFA scores. Through our density plots, we were able to confirm that these features did indeed have noticeable differences in the distributions between the positive and negative class. They were the first predictors that we included in our binary classification model.

Next, we created charts and graphs with the 30-day expire flag as our response variable and various predictors as our independent variables. The purpose of this exercise is to visualize patient risk in terms of historical rates of patients dying within 30 days with different ranges of clinical variables. Ultimately, this analysis would help inform our feature selection and feature engineering. Using our tableau visualizations, we were able to clearly segment patients into groups based on risk. We initially visualized the type of sepsis, SOFA groupings, age groupings, BUN and WBC against the 30-day expire flag. The differences in the distributions reveal critical trends in kidney function and infection severity among patients who did not survive, as seen from the paired boxplots for BUN and WBC levels displayed below in figure 1. Namely, WBC and BUN scores for patients who died within 30 days have higher medians, and more extreme upper and lower hinges than those who survived. This visualization provides insight and informs feature selection for our model by revealing a clear pattern in our patient population. The ability to filter for specific ranges of BUN and WBC values in the dashboard allows users to focus on identifying the distributions of these markers in the surviving and non-surviving patient population, excluding outliers. This granular view can help users better understand how variations in these markers correlate with outcomes, offering a valuable perspective on the relative risk posed by elevated or decreased levels.

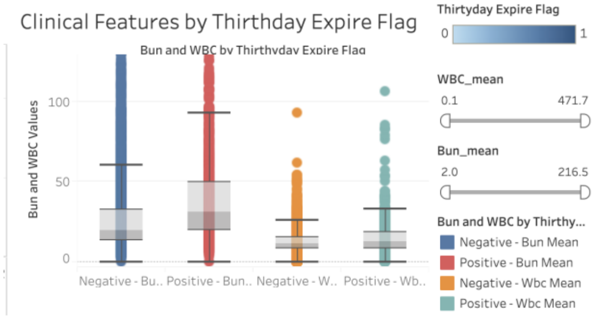


Figure 1. BUN and WBC levels for each class.

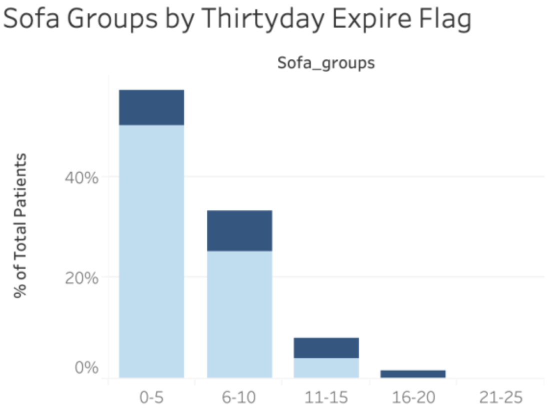


Figure 2. Distribution of each class within SOFA levels.

Similarly, the breakdowns of sepsis type and SOFA score illustrate the relationship between disease severity and mortality. Understanding that higher SOFA scores correlate with higher mortality helps validate the importance of this metric in managing sepsis-III patients and in our ML model. As demonstrated within the chart displayed in figure 2, there are more non-surviving patients within the higher sofa scores, represented by the dark blue color. The dashboard's breakdowns of SOFA scores illustrate a clear difference in risk levels across different categories, reinforcing the role of SOFA as a critical tool for assessing patient severity. Clinicians use the Sequential Organ Failure Assessment (**SOFA**) score to predict ICU mortality based on lab results and clinical data. The visual evidence of its strong association with 30-day mortality in this analysis underscores its value in sepsis care. This validation can further strengthen confidence in using SOFA scores to prioritize monitoring and care strategies, ensuring that the most vulnerable patients receive the necessary attention. Age, one of the present on admission features that can be used in modeling, also shows clear risk stratification - patients over 80 have the highest mortality rates, while the 61-80 age group represents the largest patient cohort. This age-based insight suggests that including age groups rather than age as a continuous variable in a predictive model could be beneficial, as it allows the model to capture the distinct risk profiles associated with each age range. By categorizing age into groups, the model can better differentiate the varying levels of mortality risk among different age cohorts, leading to more accurate predictions and tailored interventions for each group. Overall, the dashboard aids in identifying patterns that offer clinicians a deeper understanding of which patients are most at risk and why. By revealing the underlying factors that drive poorer outcomes, this analysis enables the development of a model that is able to identify patients who might benefit most from early intervention and targeted care strategies.

In terms of our statistical approach, we used ANOVA F-tests for numerical features and chi-squared tests for binary features to generate a level of significance ($p < .05$) for its ability to predict the response variable. These types of tests are often more reliable for feature selection than evaluating p-values for individual predictors after a model has been created, since those p-values within the context of a model might be inflated or deflated based on the presence of other predictors. We didn't select only those features that were statistically significant, however. For example, we included diabetes as a predictor because although it wasn't statistically significant in the chi-squared test, its presence improved our model performance.

We utilized feature engineering techniques as well, using a combination of temporal features, domain knowledge, and feature segmentation. For temporal features, one example we included was the season in which the person was admitted. For a domain knowledge-based feature, we created a feature called "high_risk_combined" that flagged patients who had advanced age, diabetes, low blood pressure, high heart rate, Rapid breathing, and abnormal WBC levels. All these approaches helped improve our final ML model, athough the improvements were not very significant, as most of them were in the top 15 most important features based on SHAP values for the random forest. We implemented feature segmentation for SOFA and AGE, based on our tableau EDA findings.

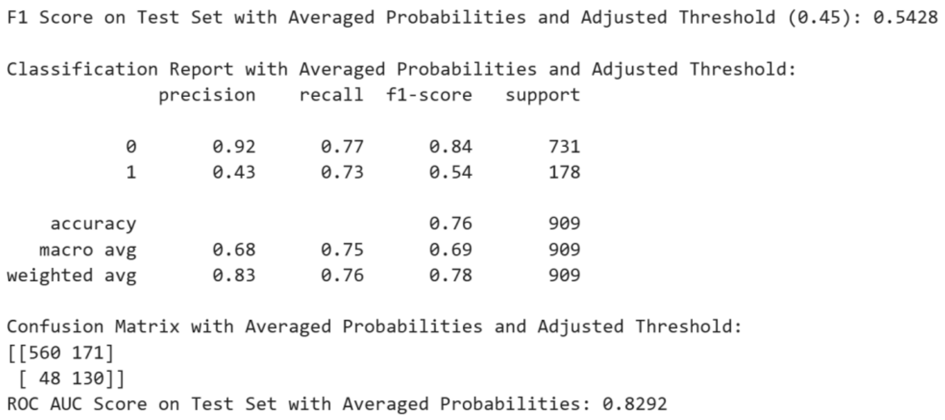
Model Selection

We tested various machine learning methods such as standalone models utilizing logistic regression, KNN, Decision Tree, and Random Forest. We also experimented with more complex model architectures – trying deep learning convolutional neural networks as well as advanced ensemble methods such as stacking/voting classifiers. We ultimately determined that the model that worked best with the data and provided the best results was a voting classifier made up of Logistic Regression, Random Forest, and XGBoost methods. This classifier works well for datasets that have complex interactions that are non-linear, when attempting to improve accuracy from a standalone model, and when working with a dataset that is not super large in terms of number of records.

Our threshold chosen was 0.45. This means that the model favors correctly choosing the positive class over correctly choosing the negative class. While we experimented with using a threshold that optimizes the overall F1 score, we found that our recall for the positive class using this approach was compromised versus using the threshold of 0.45. Our main goal in this task was to improve the recall of the positive class (while maintaining a relatively high precision) as it is more important to accurately predict if a patient will die at the expense of having more false positives.

Results and Analysis

Our Voting Classifier achieved the following performance metrics:



```
F1 Score on Test Set with Averaged Probabilities and Adjusted Threshold (0.45): 0.5428

Classification Report with Averaged Probabilities and Adjusted Threshold:
              precision    recall  f1-score   support

           0       0.92      0.77      0.84       731
           1       0.43      0.73      0.54       178

    accuracy                           0.76       909
   macro avg       0.68      0.75      0.69       909
weighted avg       0.83      0.76      0.78       909

Confusion Matrix with Averaged Probabilities and Adjusted Threshold:
[[560 171]
 [ 48 130]]
ROC AUC Score on Test Set with Averaged Probabilities: 0.8292
```

Figure 3. Voting Classifier model performance metrics

The confusion matrix reveals the instances of false positives, incorrectly predicting death for surviving patients. We identified a model that limited the number of false negatives considering the healthcare context in which our model would be deployed - having more false positives is better than having false negatives and using more resources for false positives is better than losing patient lives due to false negatives.

Our top 15 features from the random forest component of our model, with respect to SHAP values, included the following: urineoutput, lods, elixhauser_hospital, resprate_mean, lactate_mean, bun_mean, aniongap_max, spo2_mea, sofa_0-6, shock_index, age_80-99, sysbp_mean, age_creatinine_interaction, bicarbonate_min, and sodium_max. These features were in line with previous research and what is expected clinically. Urine output is highly predictive of sepsis mortality, as is elixhauser_hospital, which is a metric that measures the overall health risk associated with pre-existing conditions. Our model was able to achieve a similar ROC-AUC as the previous researchers' best model, which achieved an ROC-AUC of 0.85.

Future Research

We suggest selecting a patient population that is diverse in terms of age, race and ethnicity, and if possible, socioeconomic status. It would be helpful to have additional clinical markers that improve sepsis mortality, such as COPD, chronic kidney disease (CKD), and comorbidity indices such as the Charlston Index. The presence of COPD and CKD have been shown to have high predictive power in sepsis mortality. In addition, further hospital data such as whether the patient was admitted in an emergency or voluntarily and the number of previous ICU or hospital visits may be strong predictors. Also, lifestyle factors such as smoking or alcohol usage, different types of cancer (not only metastatic cancer), and more context overall about the patient's condition may be useful. These additional factors would improve the model's predictive accuracy as more specific patient data is included. We postulate that non-sepsis diagnosis codes may provide this context, which can be used as one-hot encoded categorical features.

Conclusion

Even though our sample was predominantly white males, it can be reasonably inferred that many sepsis-III patients might exhibit similar differences in the distribution of BUN and WBC levels, reflecting the general physiological responses to severe infection. It is important to note that while min, max and mean lab result measurements are likely to be measured over the course of a patient's stay, we included min and max values in our feature list because we assumed that at least one measurement for each feature would be known within 24 hours of the patient's stay. Our results outline the importance of strong feature selection over a strong ML algorithm for predicting sepsis mortality. Strong features are more important than the choice of ML algorithm, as illustrated by the fact that despite employing complex model architectures and feature engineering, we still weren't able to achieve a high F1 score overall. While not being able to deploy this model in a hospital setting, due to the low precision score of the positive class, (high number of false positives), our findings are valuable to clinical settings for several reasons. We have shown a clear process for selecting and engineering strong features, using visualizations and statistics, for this binary classification task for future, improved models. We also built upon previous research for this dataset, in which past researchers chose features that were available after 24 hours of the patients' stay, whereas we focused only on features available in the first 24 hours as a way of improving the real-world applicability of our model. We also provided some suggestions for future research, given the limitations of the features and patient population of this dataset. By clearly outlining our goal and methodology, which is to predict sepsis mortality as early as possible through careful feature and model selection, we provide the framework for future research to achieve higher performance in the future.

References

Balch, B. (2023, October 10). *Sepsis is the third leading cause of death in U.S. hospitals. but quick action can save lives.* AAMC News. https://www.aamc.org/news/sepsis-third-leading-cause-death-us-hospitals-quick-action-can-save-lives

Mayo Foundation for Medical Education and Research. (2023, February 10). *Sepsis*. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/sepsis/symptoms-causes/syc-20351214

Hou, N., Li, M., He, L., Xie, B., Wang, L., Zhang, R., Yu, Y., Sun, X., Pan, Z., & Wang, K. (2020). Predicting 30-days mortality for mimic-III patients with sepsis-3: A machine learning approach using xgboost. *Journal of Translational Medicine*, *18*(1). https://doi.org/10.1186/s12967-020-02620-5

Zhang, Y., Xu, W., Yang, P., & Zhang, A. (2023). Machine learning for the prediction of sepsis-related death: A systematic review and meta-analysis. *BMC Medical Informatics and Decision Making*, *23*(1). https://doi.org/10.1186/s12911-023-02383-1

Interactive Tableau Dashboard: https://ericrod12.github.io/