Understanding Patient Survival in the ICU

Team: Tetra Tech Guys

2024-03-07

Outline

- Introduction
- Description about the Dataset
- Results from EDA and Modeling

Introduction

- We want to classify patient survivability in the ICU with variables that are measured in the first 24 hours of admission.
- What variables are most significant in predicting a patient's survivability?
- Can we improve the predictability of the current APACHE scoring system?
- To learn about these variables and their relation to patient survival, we'll use
  - Exploratory Data Analysis (EDA)
  - A Series of Statistical Classification Models

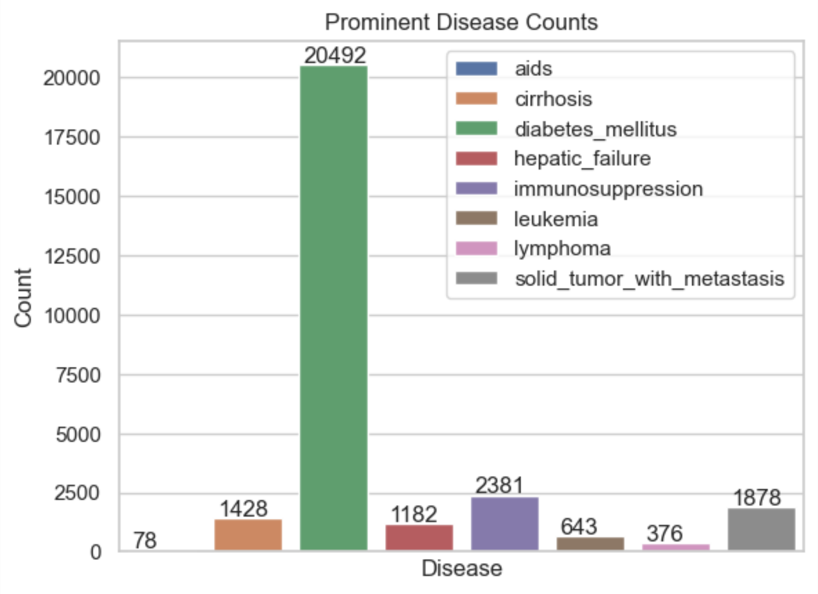
APACHE Scoring System

- Most widely used patient risk assessment is APACHE II Score created in 1981 which gives the patient a score from 0 to 71.
- Acute Physiology Score (Measured in the first 24H)
- Age
- Chronic Health Points
- Although the dataset doesn't include all measures needed to calculate the APACHE score, but it was not so good
- Could we do better?

Description about the Dataset

- Data is from Kaggle, collected from a series of ICU in the US in 2021.
- Our dataset contains 84 columns and 91713 rows.
- All medical measurements are taken and recorded for both the first hour and first 24 hours.

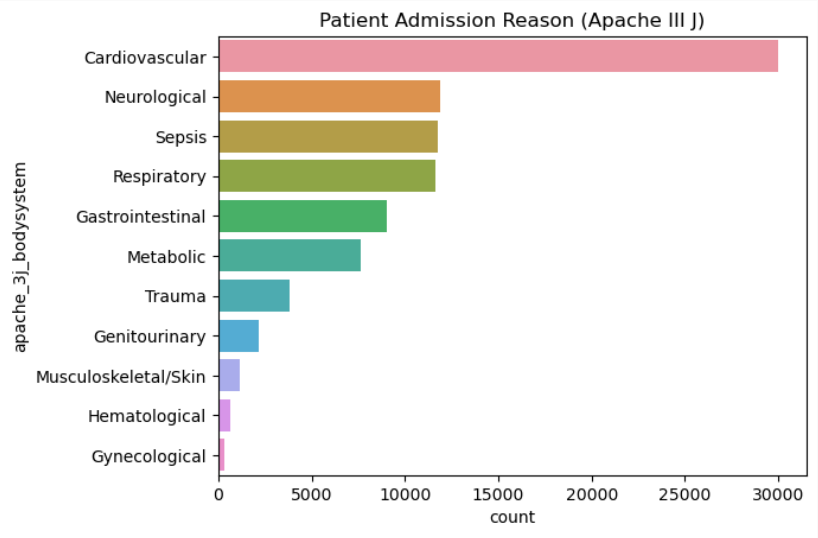
Things to consider about our dataset (Disease skewness)



Things to consider about our dataset (Demographics)

- Our data set has an over representation of white, older, male, and diabetic patients
- 78% of patients are white
- Median age of 65
- 54% are male
- 23% are diabetic
- This can lead to both accuracy issues as we delve into sub-classes and generalization of results issues.

Important finding from EDA (Bodysystem)



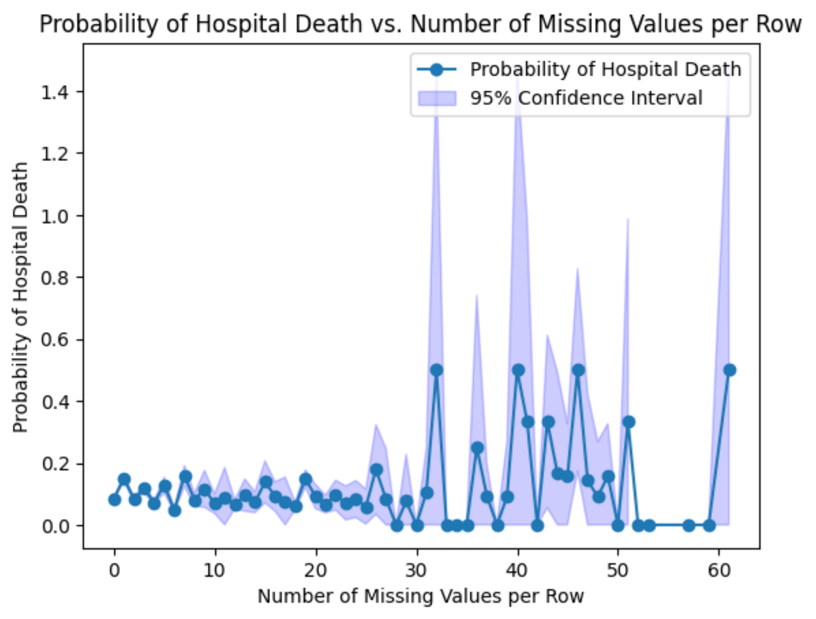
Important finding from EDA

- Cardiovascular issues (APACHE Bodystem) were the most common with 30,000 individual cases.
- People of Hispanic ethnicity die at higher rates than other ethnicity
- Heart rates of patients who ended up dying tended to be nearly 20 BPM more.
- At first glance, our data seems to have really good indicating features of death, with numerous vitals such as blood content(glucose/potassium), weight, height, heart rate, breath rate, blood pressure, and more.

Handling Missing Values

- $\sim$35,000 observations have at least one missing value
- Imputed based on number of missing values in a row:
    - Categorical variables gain a new category "Missing"
    - if $<$ 30 missing values in a row impute with median value
    - if $>=$ 30 missing values, impute with multiple imputation

Handling Missing Values



Method of Evaluating our models

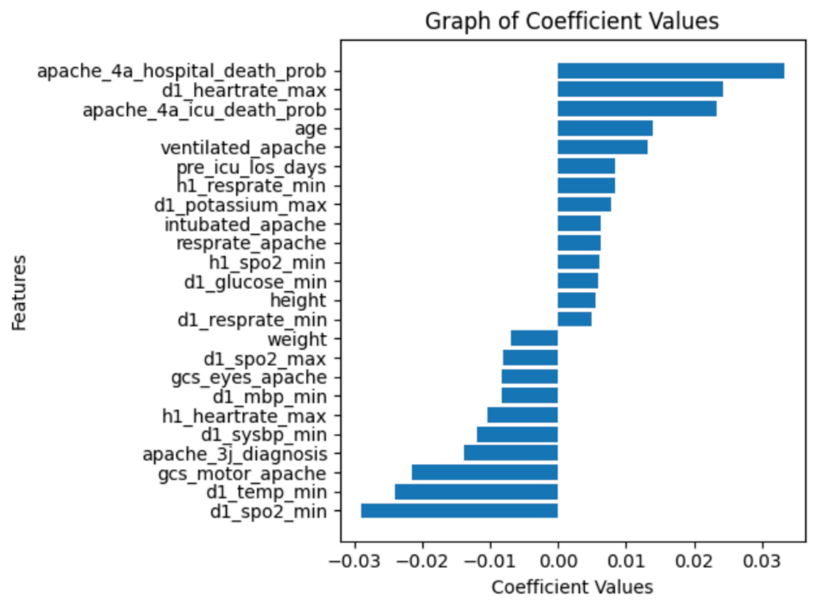
- Dealing with a highly unbalanced dataset (~8% of patients passed).
- For our results, we will be using Precision and Recall, combined into the F1-score, to assess the performance of our model.
- Precision measures the proportion of individuals who actually passed among those we predicted would.
- Recall measures the proportion of people who passed that we correctly predicted.

Results

Feature Selection

- Subjectively combined multiple statistical feature selection methods through analysis of common features to get the most important variables in predicting patient survival.
- Feature selection methods including PCA, ridge regression, LASSO, forward selection, and univariate testing.

Feature Selection



Our own Models and Results

Table 1: Model Results

Model	Precision	Recall	F1 Score
Baseline Regression	0.6337	0.1753	0.2748
Logistic Regression	0.722	0.230	0.349
Decision Tree	0.526	0.278	0.364
KNN	0.332	0.495	0.397
RandomForest Classifier*	0.471-0.800	0.144-0.500	0.232-0.531
XGBoost*	0.563-0.739	0.217-0.500	0.313-0.536

* XGBoost and RandomForest Classifier was run with subgrouping different models based on body system

Takeaways

- We are not able to predict patient survival very well
- Our model accuracies are slightly better than the baseline APACHE model logistic regression
- Each unique model had similarly poor scoring metrics, indicating "capped out" score based on our available features
- We do know that none of the data's features in their current state have a highly explanatory relationship with patient survival, but each has a statistically significant, trivial effect on survival

Next Steps

- Explore if model accuracy issues are due to a skewed dataset with a dominant class and underrepresented classes.
- Missing values or things to consider?
- New models within subgroup (gender, age, race, bodysystem) while avoiding ethical concerns
- Becoming a doctor and find new features to measure

Takeaways Cont.

- We did a fantastic job exploring our data thoroughly. We got great insights into the data and really pulled the most we could from it. I think the bake off was great.

Preguntas? Fragens?