

The background of the slide is a dark, grainy photograph of a hospital room. In the center, there is a patient bed with a person lying on it, covered with a white sheet. To the left of the bed, a medical cart with various equipment is visible. In the background, there are white vertical blinds covering a window. The overall atmosphere is somber and clinical.

Jason Kim  
Metis

May 14<sup>th</sup>, 2021

# Sepsis Classification



## Problem

---

- CDC:
  - 1.7 million adults in the US develop sepsis
  - ~270,000 Americans die as a result of sepsis
  - 1/3<sup>rd</sup> of patients who die in hospitals have sepsis

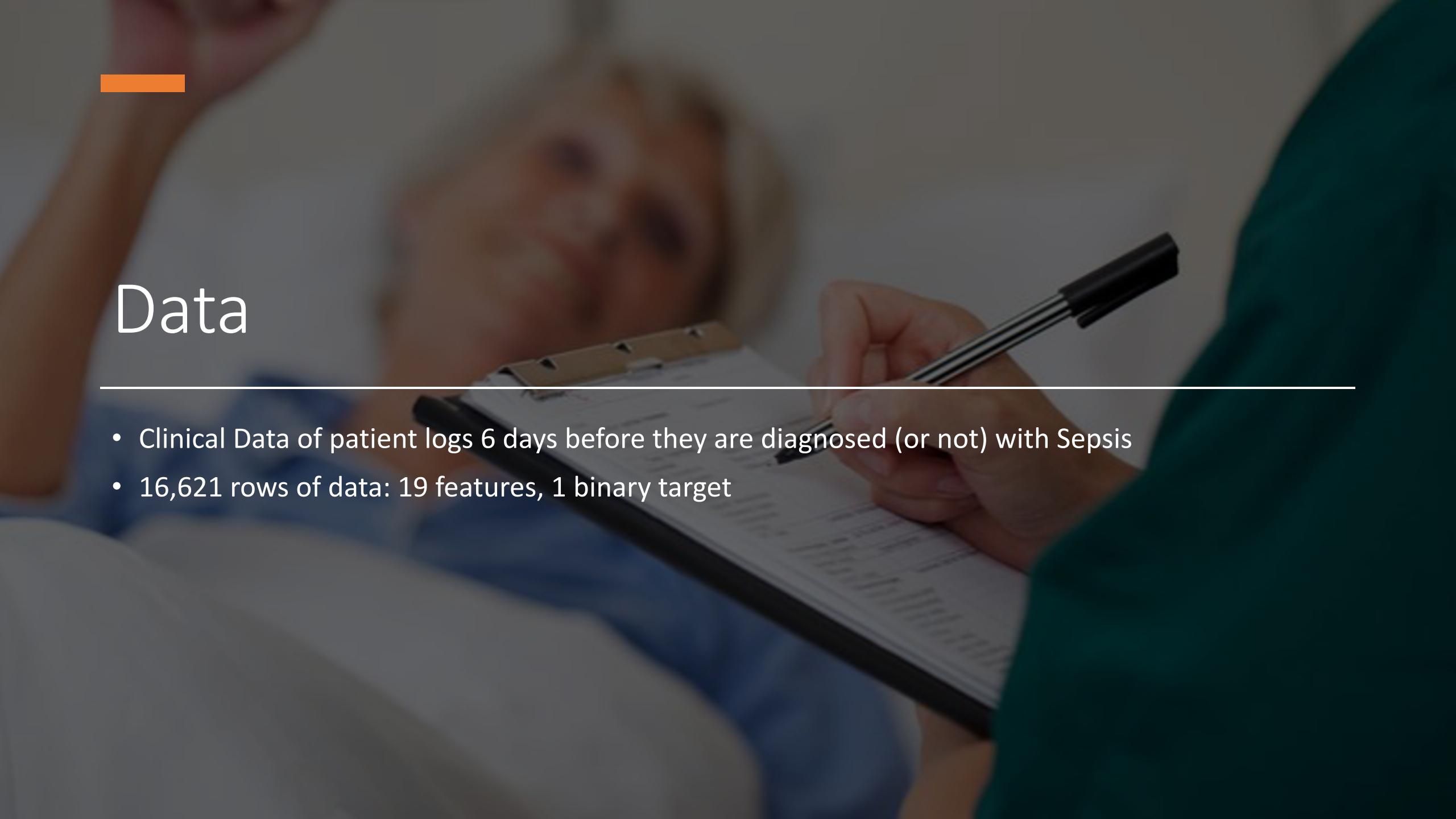
# Goal

Develop a classification mode to predict Sepsis in a patient



## Tools

- Data Obtained from Kaggle dataset
- Pandas, Numpy: Data Manipulation
- Visualizations: matplotlib, seaborn
- Models: sklearn

A close-up photograph of a medical professional's hands. The person is wearing a dark green scrub top. They are holding a black pen over a white clipboard with a grid pattern. The clipboard has some handwritten and printed text, though it's mostly illegible. The background is blurred, showing a patient's face and shoulder.

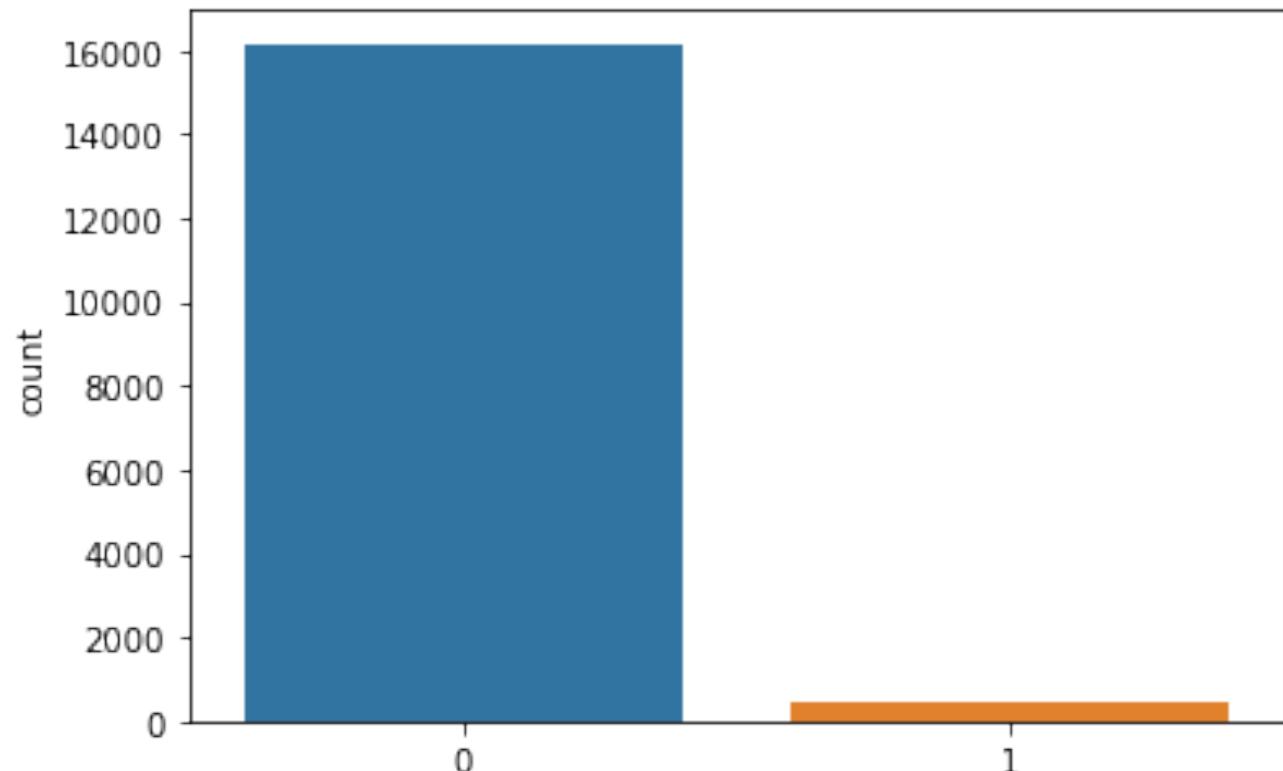
# Data

---

- Clinical Data of patient logs 6 days before they are diagnosed (or not) with Sepsis
- 16,621 rows of data: 19 features, 1 binary target

# EDA

- Highly imbalanced dataset:
- 2% positives vs 98% negatives
- Remedies:
  - Resampling: Random OverSampling on Train set



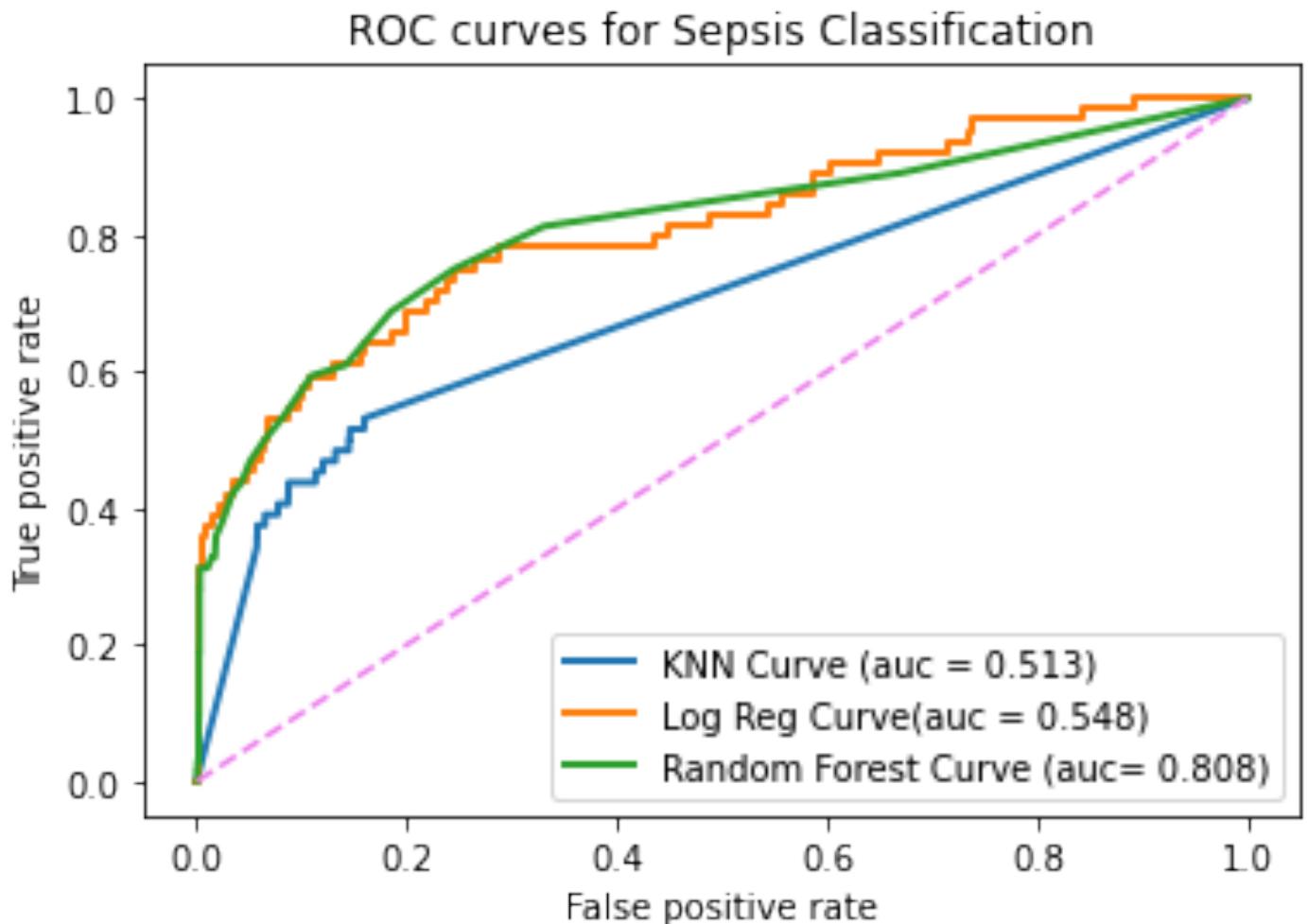


## Methodology

- F beta with beta=2 (emphasizing recall)
  - Recall: Maximize True Positives – not missing too many positive sepsis cases
  - Precision: Minimize False Positives – minimizing patient ICU Length of Stay
- ROC-AUC to compare models

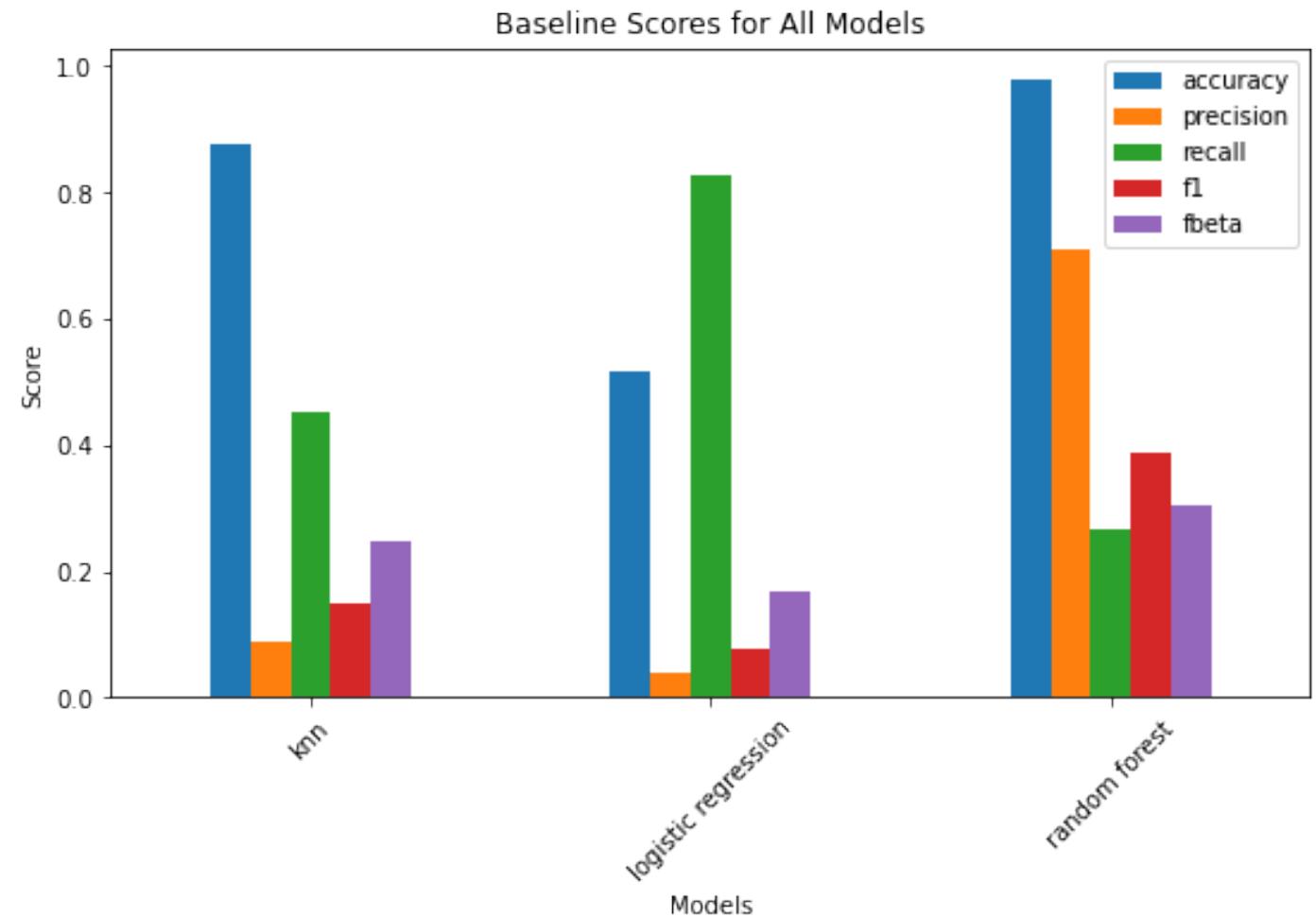
## Choosing a Model

- ROC-AUC curve:
- Random Forest: AUC = 0.808
- Logistic Regression: AUC = 0.548
- KNN: AUC = 0.513



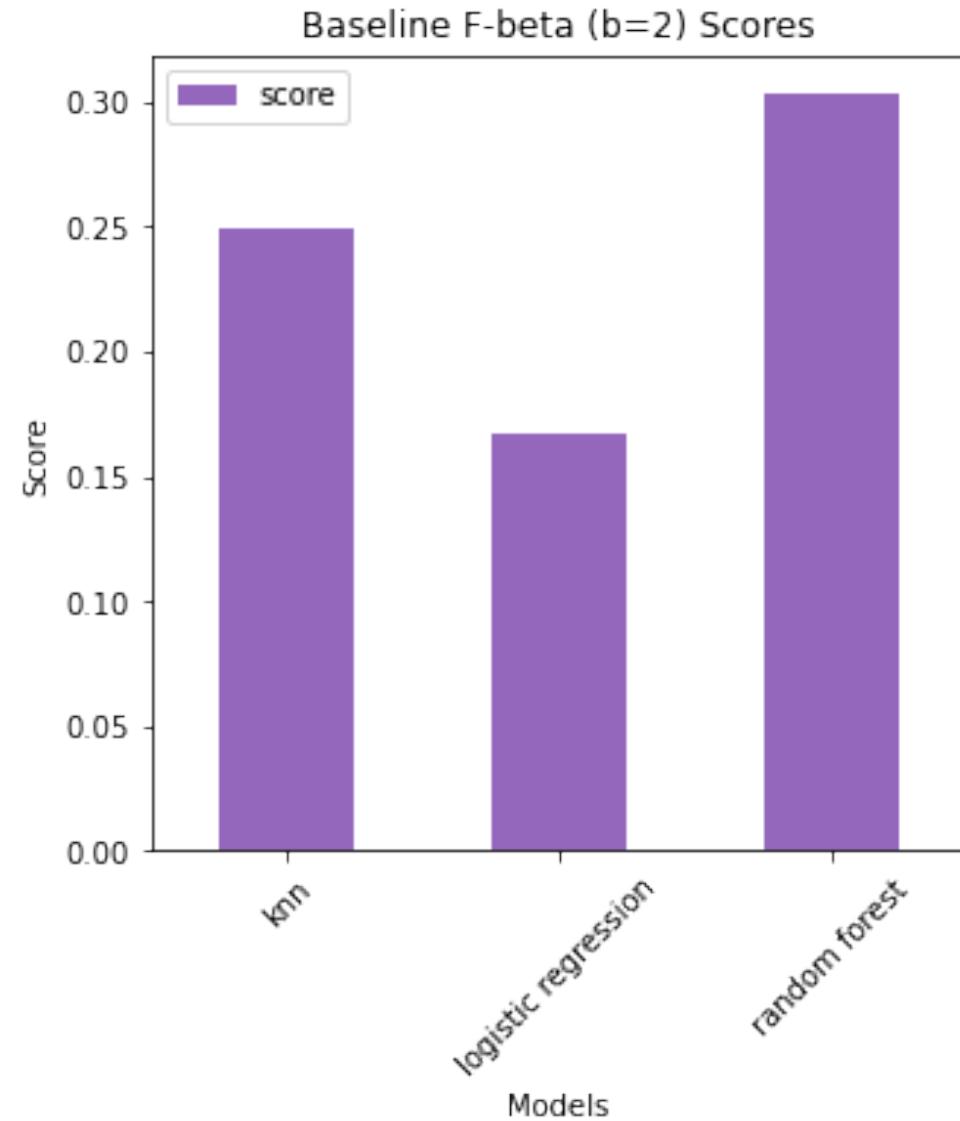
## Baseline Scores

- Highest Scores:
  - Accuracy: Random Forest
  - Precision: Random Forest
  - Recall: Logistic Regression
  - F1: Random Forest
  - F-Beta( $b=2$ ): Random Forest



# Choosing a Model

- Maximizing AUC and F-Beta (b=2)
- F-Beta Scores:
  - Random Forest: 0.253
  - KNN: 0.249
  - Logistic Regression: 0.168



# Results



Random Forest F-Beta  
Score (Baseline):

0.253



Random Forest F-Beta  
after some  
hyperparameter tuning  
with  
RandomizedSearchCV:

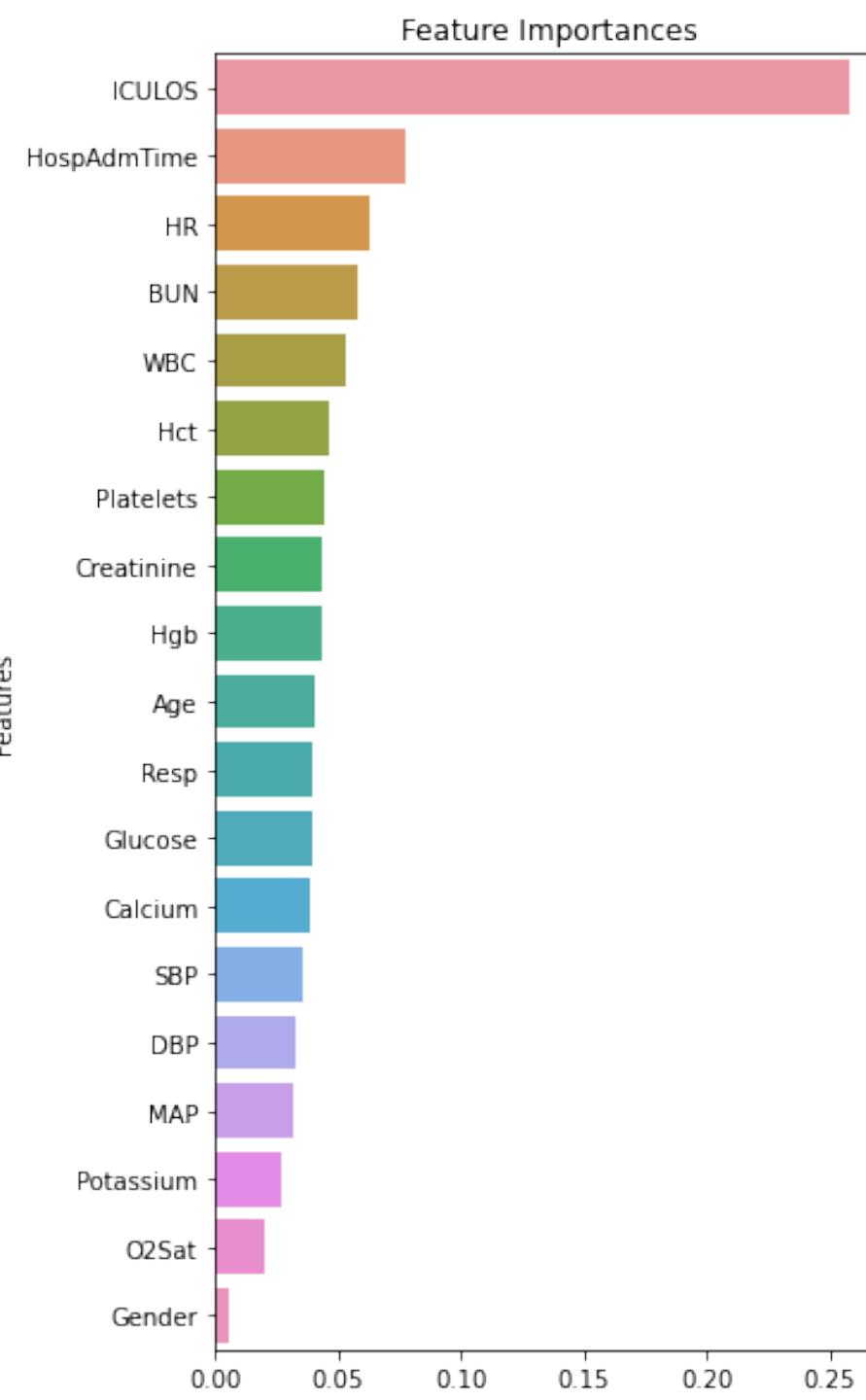
0.287  
(increase of  
0.034)

ICU LOS

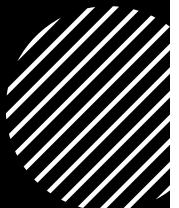
# Feature Importance

---

- ICU LOS – Longer stays in the ICU has seems to have the most significant impact on Sepsis classification



# Future Work



Further hyperparameter tuning



Exploring other models



Better data manipulation  
(imputing methods)