Sepsis Classification

Jason Kim

Metis

May 14th, 2021

Problem

- CDC:
  - 1.7 million adults in the US develop sepsis
  - ~270,000 Americans die as a result of sepsis
  - 1/3rd 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

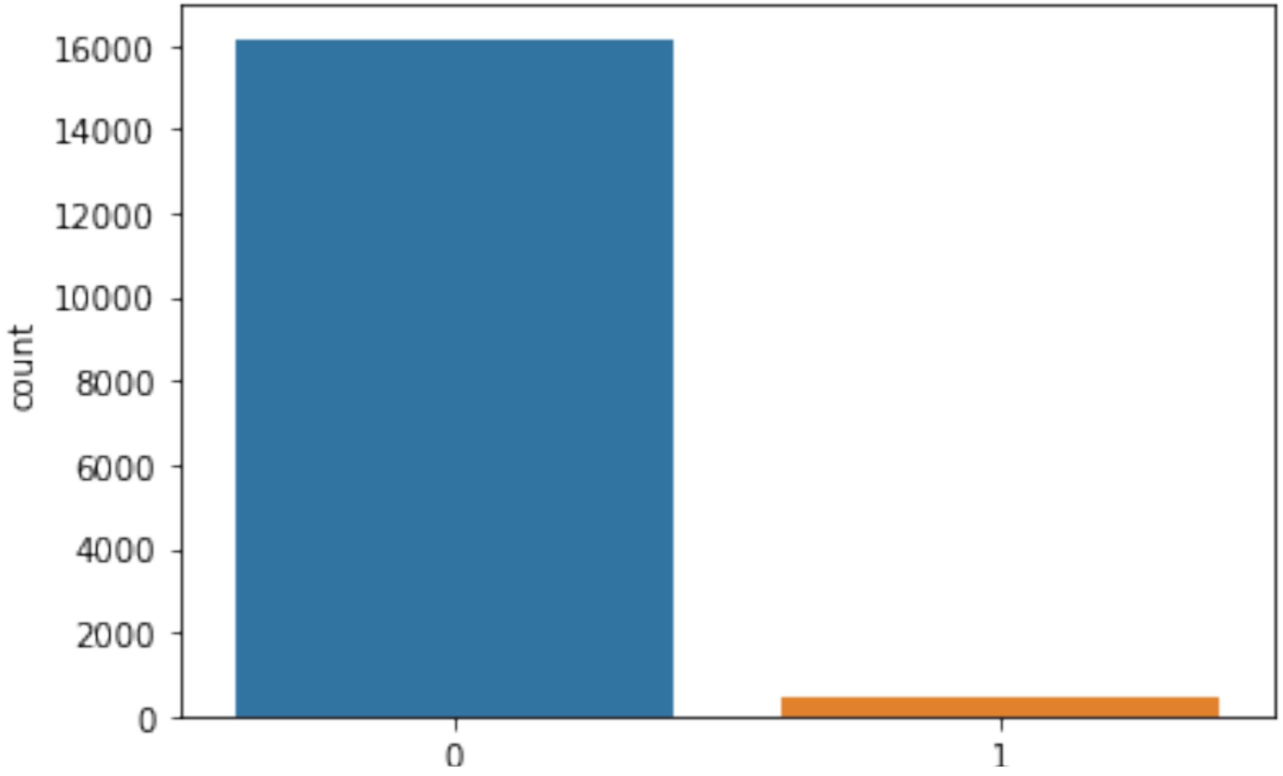
Data

- 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



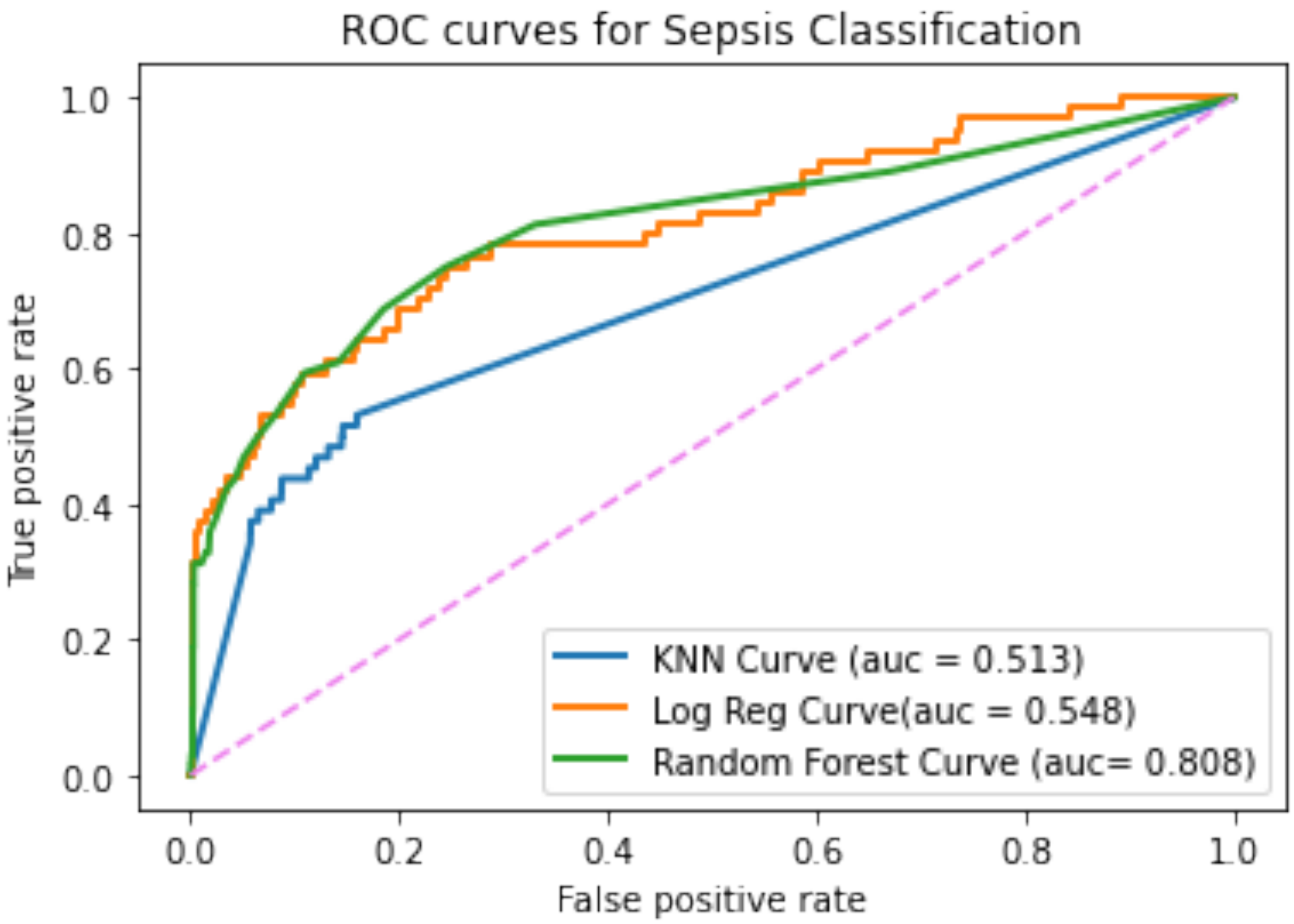
Methodology

- 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

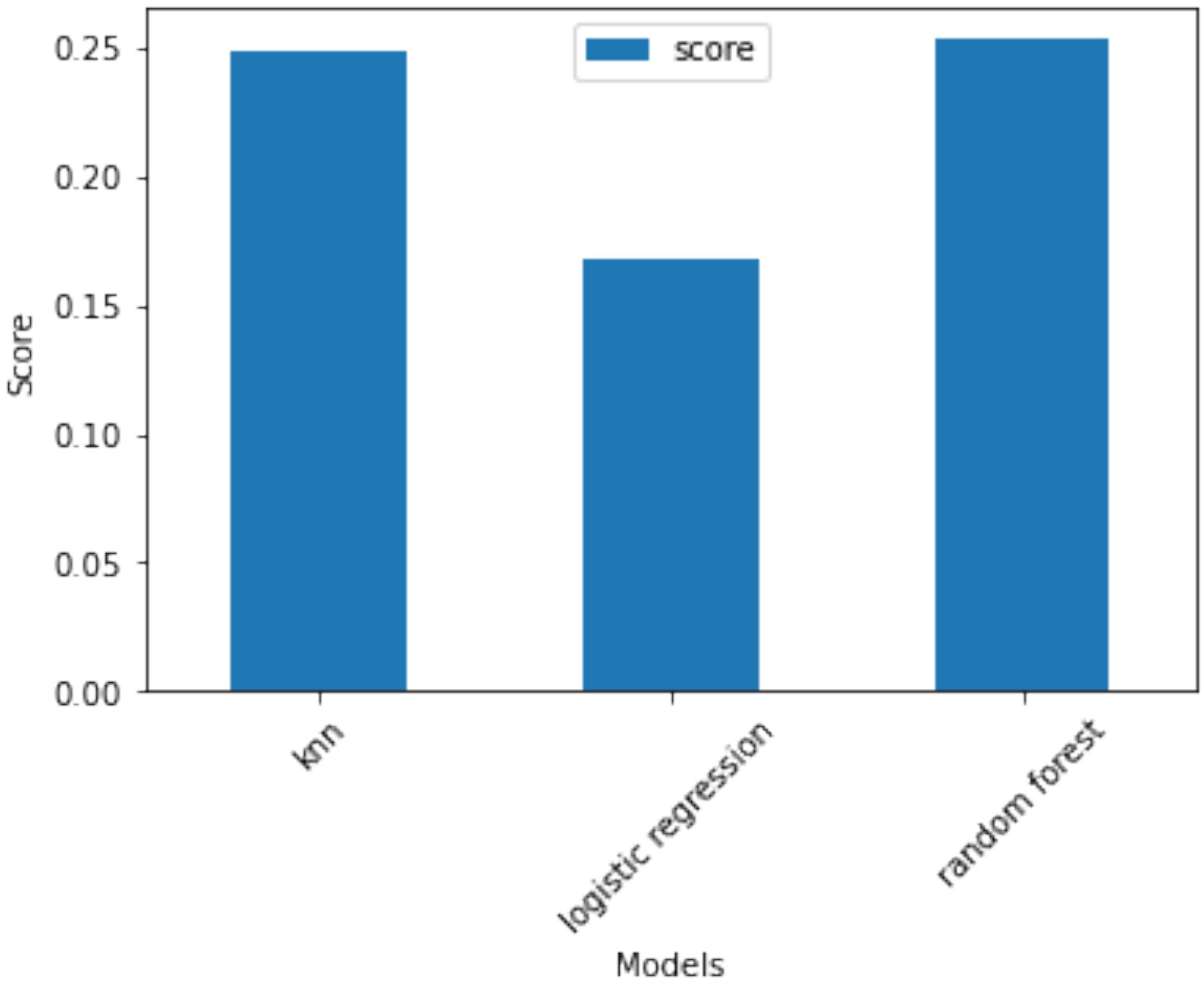
- ROC-AUC curve:

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



Choosing a Model

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

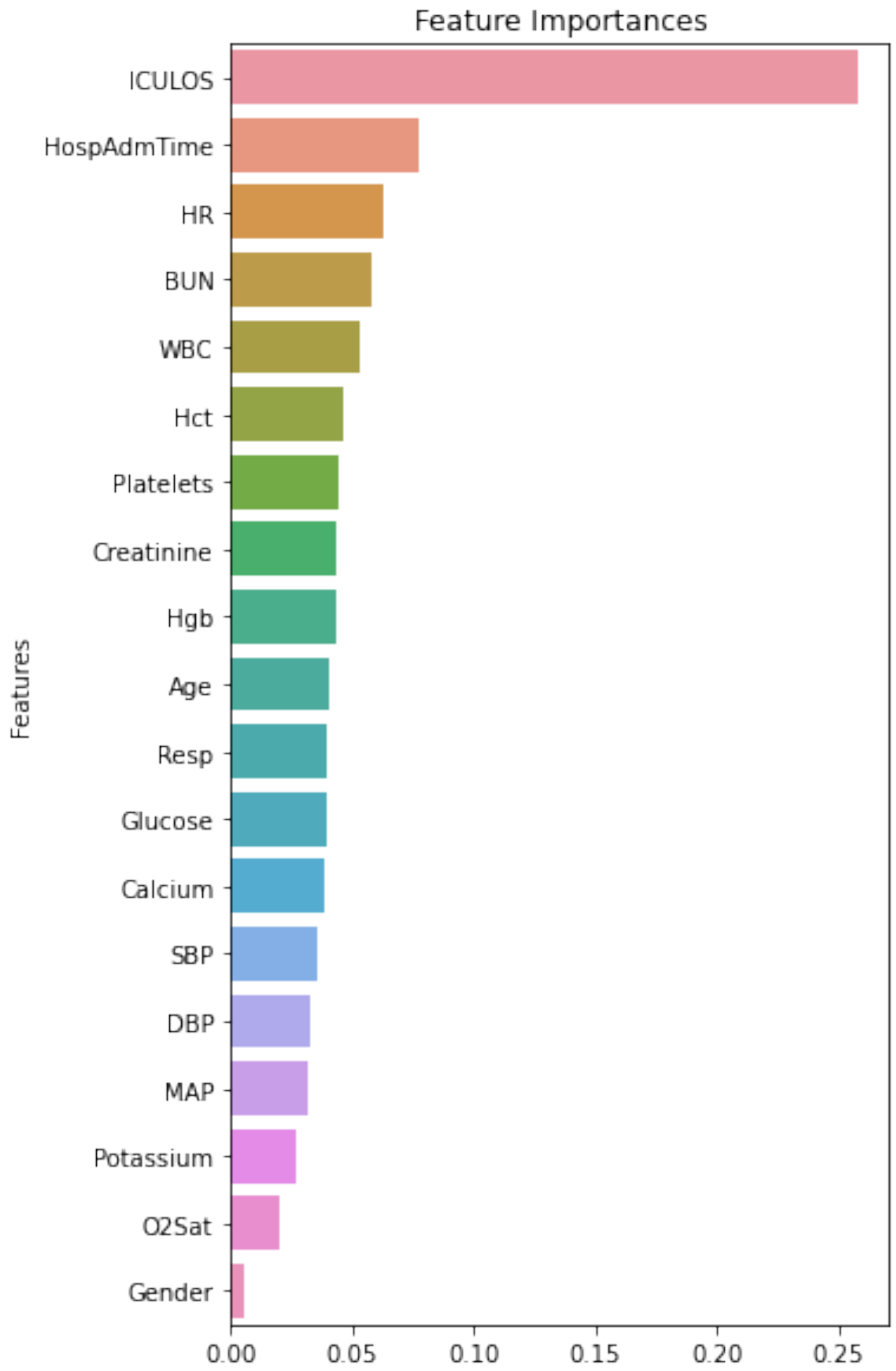


Results

Random Forest F-Beta Score (Baseline):	0.253
Random Forest F-Beta after some hyperparameter tuning with RandomizedSearchCV:	0.287 (increase of 0.034)

Feature Importance

- ICULOS – 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)