Predicting MRSA Infections in Hospital Environments

Student: Ahmed Hussain
Mentors: Dr. Parantapa Bhattacharya, Prof. Anil Vullikanti

Network Systems Science and Advanced Computing Division (NSSAC)

Background

Healthcare Associated Infections (HAIs) are infections which can occur while a patient is receiving healthcare for another condition

- Daily, about 1 in 31 hospital patients contracts an HAI[1]

MRSA (Methicillin-resistant *Staphylococcus aureus*) is a staph bacteria "superbug"

- Most frequently transmitted by direct skin-to-skin contact or contact with shared items[1]
- Usually starts as a skin infection that can appear anywhere on a patient's body. Early symptoms of MRSA in a person can include a bump that is red, swollen, and hot

Proprietary data which outlines a patient's information when they for any healthcare-related visit at a hospital (Ivy)

- Length of stay, inpatient/outpatient, number of visits
- Network features (e.g. provider, # of MRSA contacts)
- Demographics of patient
- Antibiotics used during stay
- Devices used during stay
- Surgeries administered
- Dialysis used
- ICU visits

Classifier Performance and Model Selection

Feature Extraction

- Large dataset (`73463 x 102`) with all features

Data Pre-Processing

- Dropped columns with near-zero variance
  - Some surgery types and department locations
- Train/test split, test size for dataset is 20%
- Yeo-Johnson power transform
  - Versus Box-Cox power transform (Y-J supports negatives)
  - Conforms data to a Gaussian distribution in order to stabilize variance

Hyperparameter Tuning

- GridSearchCV: 5-fold cross-validation

	Precision	Recall	F1	ROC AUC	Accuracy	Bayes Factor
Logistic Regression	0.46	0.61	0.53	0.748	0.85	0.858
SVM	0.48	0.60	0.53	0.747	0.86	0.917
AdaBoost	0.50	0.66	0.57	0.779	0.86	1.017
GBDT	0.67	0.45	0.54	0.709	0.89	2.036
MLP	0.59	0.44	0.50	0.694	0.88	1.411

Figure 1: Scoring each of the models with their best hyperparameters.

- Precision, recall, F1 score, ROC-AUC, accuracy, Bayes' factor (ratio of true positives to false positives)
  - Accuracy used when true positives and true negatives are more important, F1 score (precision & recall) used when false negatives and false positives are more important
  - ROC AUC is measure of the ability of a classifier to distinguish between positive and negative classes

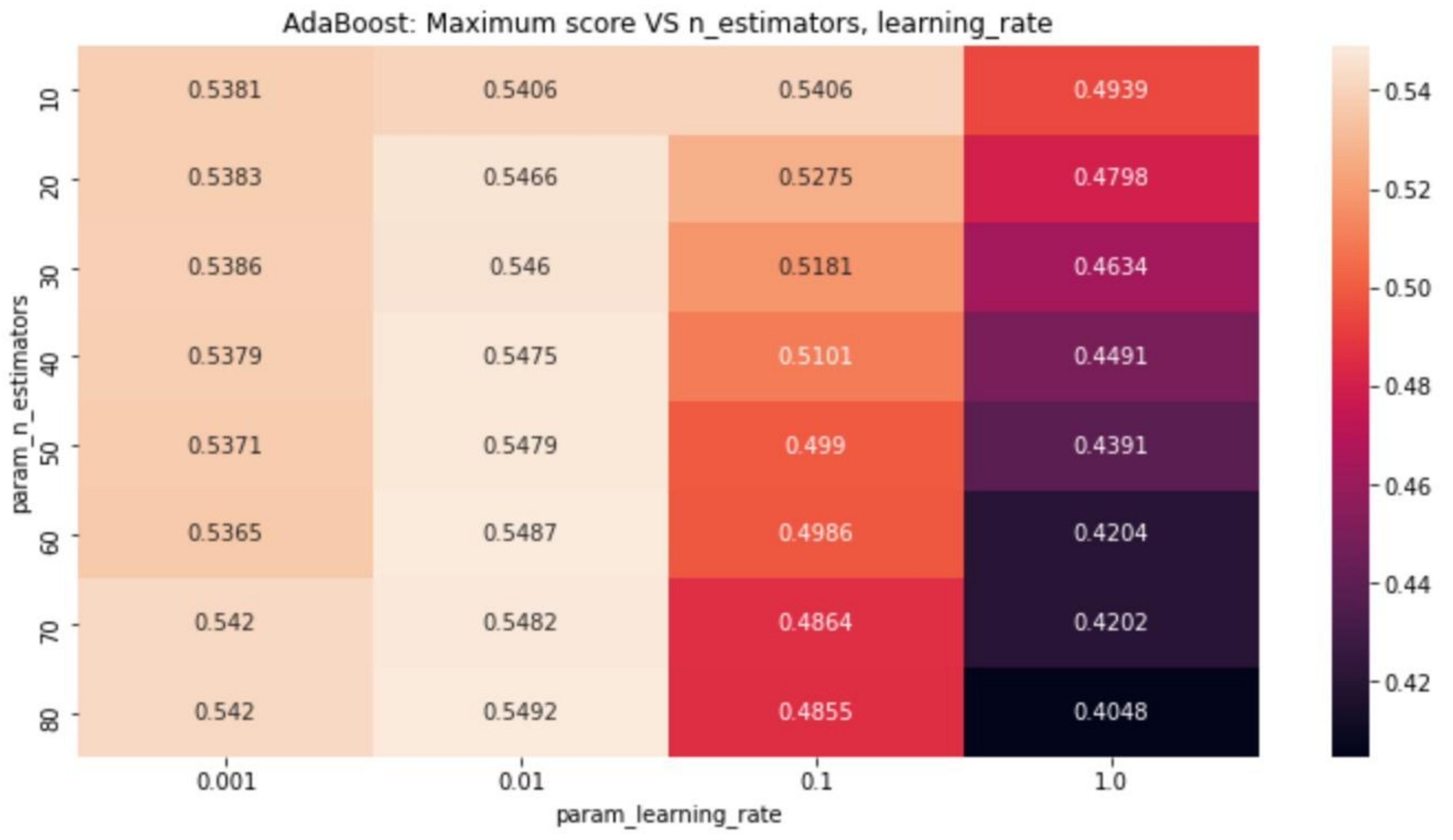

Figure 2: Heatmap of AdaBoost's score using GridSearchCV's *`cv = 5`**. Deviance loss function, subsample is 0.75, max depth is 7.***

Project Goal

Develop machine learning models which can predict with maximum accuracy whether a patient has contracted MRSA

- Determining classifier performance through hyperparameter search
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Adaptive Boosting Decision Tree (AdaBoost)
  - Gradient Boosting Decision Tree (GBDT)
  - Multilayer Perceptron (MLP) Classifier

Predict whether a test result for MRSA is negative or positive

- Score each model on this basis using multiple measures

Determine which datasets and features contribute most to classifier accuracy

- Ablation study
  - Remove each dataset one at a time
  - Determine which removal had the biggest effect on classifier performance

Ablation Study

Practice of removing each set of features and scoring the model

- Used Adaptive Boosting Decision Tree (AdaBoost)
- Same scoring system as hyperparameter search phase

	Precision	Recall	F1	ROC AUC	Accuracy	Bayes Factor
All - Net	0.44	0.48	0.46	0.690	0.84	0.776
All - AB	0.50	0.68	0.58	0.788	0.86	0.996
All - Surg	0.45	0.72	0.55	0.790	0.84	0.813
All - Dev	0.49	0.68	0.57	0.786	0.86	0.972
All - Dial	0.48	0.68	0.56	0.783	0.86	0.925
All - ICU	0.46	0.69	0.55	0.780	0.85	0.867
All - Demo	0.48	0.58	0.53	0.742	0.86	0.929

Figure 3: Retraining AdaBoost without each feature set; Net=Network, AB=Antibiotic, Surg=Surgery, Dev=Device, Dial=Dialysis, ICU=ICU Visits.

Results

- Least important data
  - Antibiotics: What antibiotics a patient has been given does not affect their likelihood of contracting MRSA, considering MRSA is a superbug resistant to antibiotics (specifically methicillin). This is the least useful dataset out of those assessed.
  - Devices: The devices used for the patient (implanted or otherwise) don't contribute much to MRSA infection likelihood.
- Most important data
  - Network: Clearly, which provider (physician, nurse, etc) and how many infected people a patient comes in contact with over the course of 7- or 14-day durations heavily impact the likelihood of a given patient being infected with MRSA. This appears to be the most impactful dataset of the ones assessed.
  - ICU: Removing whether a patient visits the ICU hinders the model's performance. Either the ICU physically causes a lot of infections, or MRSA carriers have a higher likeliness to visit the ICU (unlikely, symptoms have a 10-day incubation period[2]).

References

[1] Centers for Disease Control and Prevention. (2016, March 4). Healthcare-associated infections | HAI | CDC. https://www.cdc.gov/hai/index.html

[2] Mayo Foundation for Medical Education and Research (MFMER). (2020, December 1). MRSA infection - Symptoms and causes. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/mrsa/symptoms-causes/syc-20375336

UNIVERSITY *of* VIRGINIA
BIOCOMPLEXITY INSTITUTE