STATS 503 Data Challenge Summary

ChengYu Ko

In this data challenge, we aimed to get the prediction of sepsis, which caused millions of people die each year, to make the antibiotic treatment earlier. About data preprocessing, the main difficulty was to address the problem of missing values and imbalanced labels in this Longitudinal data. First of all, after spliting the data into training, validation, test sets, we filled out the missing values of the demographics variables, `ID`, `Age`, `Gender`, directly (as they are constant over time). Then we performed the ***mean imputation*** within each patient on all of the variabes in the three sets. Also, we employed the ***expanding windows*** to deal with the time-stamp of each person- taking each column's mean within the patient's records as the patient's feature. Next, we dropped the administrative identifier variables, `Unit1` and `Unit2`, and performed the ***multivariate imputation by chained equations*** to estimate the best predictions for the missing values. Finally, to reduce the problem arose by the imbalanced data, we applied the ***synthetic minority oversampling technique*** to the training data by by generating new synthetic examples close to the other points (belonging to the minority class) in feature space. We included *37* variables as the predictors to make the sepsis prediction.

After making the 5-fold cross-validation for ***Logistic***, ***Random Foest***, ***Adaboost with Decision Trees***, and ***Support Vector Machine*** models, we found that the best prediction coming from the ***Random Forest*** model in terms of the lowest ***balanced error rate (BER)*** = *0.1939* and the ***Adaboost with Decision Trees*** model in terms of the highest ***area under the receiver operating characteristic curve (AUC)*** = 0.8528 in validation set. Moreover, the ***Random Forest*** model classified *96* negative cases to be positive, and *145* positive cases to be negative; the ***Adaboost with Decision Trees*** model classified *27* negative cases to be positive, and *188* positive cases to be negative. However, after we fitted the ***Random Forest*** model to the test set, it predicted *2131* positive cases out of *6490* patients, which is not consistent with the distributions of training and validation sets. Therefore, in the test results, we decided to use the predicted labels as our `outcome` column and the predicted probabilities as our `score` column in the ***Adaboost with Decision Trees*** model.