AKSHAT PORWAL
akshatp

STATS 503: Modern Multivariate Analysis

DATA CHALLENGE REPORT

Objective

The objective of this data challenge is to predict the development of sepsis in patients admitted to the ICU using clinical vital signs, laboratory measurements, and demographic information. The dataset includes records from 21,634 patients, split into a training set of 15,144 patients and a test set of 6,490 patients. The prediction variables include approximately 40 variables, including vital signs, laboratory values, and demographics. The prediction results will be evaluated using the balanced error rate (BER) and the area under the receiver operating characteristic curve (AUC). The ultimate goal of this project is to develop a model that can accurately predict sepsis in ICU patients, which could potentially aid in early intervention and improve patient outcomes.

Data Compilation

For each patient, a statistical summary consisting of mean, standard deviation and range was created. These statistics were used to combine the hourly patient data into a single row, i.e., making one data point for each patient. Furthermore, the process was iterated for every patient and a single data frame was compiled which had the data for all the patients.

Handling Missing Values

The compiled dataframe was checked for null values and the attributes/predictors with more than 90% of missing values were dropped. The files *train_outcome* and *test_nolabel* were then used to segregate the patients in the training and the testing data respectively based on the ID attribute. The training data was then splitted into *sub-training* and *sub-testing* data to evaluate the model performances.

Furthermore, median imputation was performed separately on sub-training and sub-testing data as some indicators were not present for all the patients. Imputation was performed after train-test split to avoid data leakage.

Applying Models

It was observed that the data was highly imbalanced. For imbalanced data, accuracy is not a good evaluation metric since it can be misleading. Precision, recall, F1-score, and AUC-ROC are more

appropriate metrics as they consider the different costs of false positives and false negatives. Therefore, the models that were taken into consideration were:

- **RandomForest:** Random Forest can handle imbalanced data by assigning class weights and using undersampling/oversampling techniques, resulting in improved classification performance.
- **ADABoost**: ADABoost adjusts weights of misclassified instances, assigns more weight to minority class instances, and uses decision stumps as weak classifiers to handle imbalanced data.
- **XGBoost:** XGBoost can handle imbalanced data by adjusting class weights, using early stopping, and employing regularization techniques, leading to improved classification performance.
- **Histogram Gradient Boosting:** Histogram Gradient Boosting can handle imbalanced data by using adaptive binning, regularization, and gradient-based boosting, resulting in improved classification performance.

From the above models, it was observed that Histogram Gradient Boosting gave the highest AUC Score of 0.87, whereas the Random Forest gave an AUC Score of 0.84, ADABoost gave an AUC Score of 0.84 and XGBoost gave an AUC Score of 0.86. Confusion Matrix was also plotted for Histogram Gradient Boosting and it was observed that the model misclassified 365 cases as Type - 2 Error.

Final Model & Prediction

Histogram Gradient Boosting algorithm was chosen for the final prediction on the test data. Similar to the training data, the test dataset was obtained by merging *test_nolabel* with the combined data frame based on '*ID*'. The entire training data was then imputed with median and the test data was also imputed with median separately. Furthermore, the X and Y attributes were segregated from the training and testing data.

After implementing the Histogram Gradient Boosting algorithm, the final predicted result was,

- 516 patients from the test data were predicted to have sepsis.
- 5794 patients from the test data did not have sepsis according to the prediction.

The prediction probabilities were also calculated for each label for every test row. The highest probability decides the Outcome label.