# INFSCI 2160: Data Mining

# R Lab

Haomin Xie & Yunshu Liang

## 1. Introduction & Problem definition

For the Data Mining R Lab, our team includes Haomin Xie and Yunshu Liang.

The goal of the lab is using dataset provided by UPMC to predict patients that are at risk to progress to ESRD within 2 years, based on data mining techniques. We use R to implement XGboost algorithm to make a prediction on this problem.

## 2. Data Preview

The dataset has 143 features that can be used in the model.

There are three types of variables: categorical, continuous, and numeric.

## 3. Feature analysis & Selection

Firstly, we deal with the data which has more than 30% null data value. We have tried two methods to deal with them. First one we set all the data of the columns into 0. The other one we eliminate them. After analyzing the data, there are 29 columns that have more than 30% null values. And details can be seen in the RMarkdown file.

## 4. Model training

We applied xgboost to fit the model.

Firstly, we use 7:3 proportion to separate the dataset into train and test. And further separate the train dataset into train and validation sets for the cross-validation.

Secondly, set the parameters for the xgboost with the binary: logistic objective and gbtree booster. Furthermore, in order to enhance the model operability and deal with the unbalanced target values, we set the scale_pos_weight around 38 based on the scale of the positive values and the negative values.

Finally, we modify the model performances using two indicators: the Youden's J statistic and the observation of the distribution of the predicted probabilities. Youden's J modifies the model with relatively high sensitivity and specificity. While by observing the different thresholds performances, we can modify the model as required.

**5.Risk Categories**

We used the probabilities given by the model, combined with the methods mentioned in the paper, to identify the different risk categories of patients.

Low Risk: 0% to 9.9%
Intermediate Risk: 10.0% to 19.9%
High Risk: 20% or more

**6. Model evaluation & Deployment strategy**

We use confusion matrix and AUC to evaluate our model.

**1) For overall performance and general diagnosis**

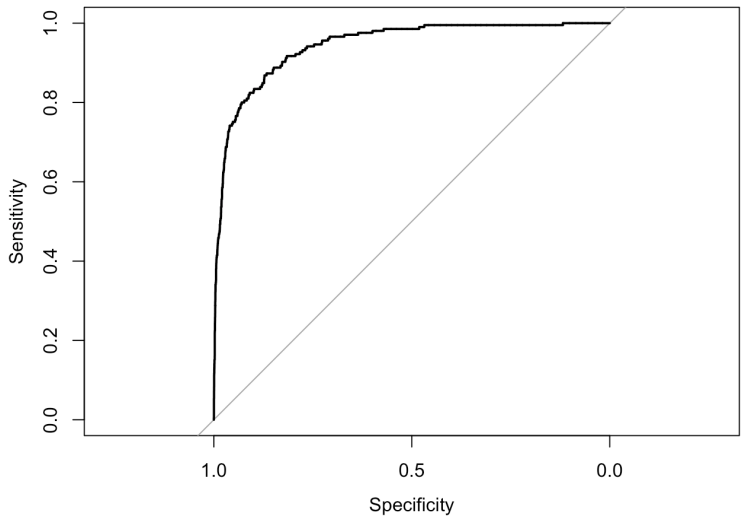
A computed Youden's J value is the best parameter for the model general performance.

And we use it as the threshold. The matrix is shown below. The positive and negative predicted value are relatively high. Therefore, this can be used for a general diagnose. By this means, if a patient is diagnosed as progress and not progress to ESRD within 2 years, there is a high probability that the patient does in the situation.

```
Confusion Matrix and Statistics


          0    1
0 7462  781
1   27  178
```

Below graph is our ROC curve. The corresponding AUC is `0.9439`.



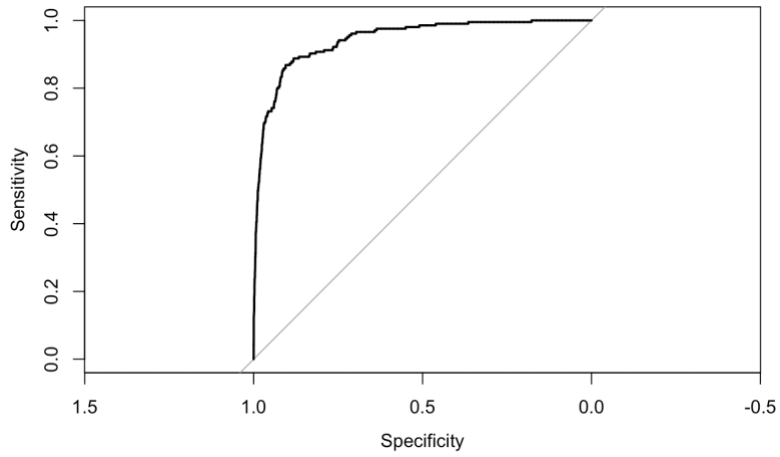
**2) For the purpose for the correct probability of patients as progress to ESRD within 2 years**

The distribution of the predicted probabilities is separated into 5 bins, combined with the histogram, we can derive the thresholds with some bias when applied into the model. And if the facility wants to maximize the predicted value and does not want leave out any chance that a patient has a high risk. We can apply the threshold that maximize it. The confusion matrix is below.

In this case, we are not allowed to predict the patients with a predicted result 0. On the other hand, we are more confident to predict the patient with a predicted result 1. Therefore, this can be used in a more sensitive and reliable way to predict the patient progress to ESRD within 2 years.

Confusion Matrix and Statistics

| | 0 | 1 |
|---|---|---|
| 0 | 6738 | 1505 |
| 1 | 20 | 185 |




The details can be seen in the RMarkdown file. We provided two thresholds mentioned above in the file.

## 7.Conclusion

```
               Accuracy : 0.8195
                 95% CI : (0.8111, 0.8276)
    No Information Rate : 0.8
    P-Value [Acc > NIR] : 2.974e-06

                  Kappa : 0.1588
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9970
            Specificity : 0.1095
         Pos Pred Value : 0.8174
         Neg Pred Value : 0.9024
             Prevalence : 0.8000
         Detection Rate : 0.7976
   Detection Prevalence : 0.9757
      Balanced Accuracy : 0.5533

       'Positive' Class : 0
```

```
               Accuracy : 0.9044
                 95% CI : (0.8979, 0.9105)
    No Information Rate : 0.8865
    P-Value [Acc > NIR] : 6.331e-08

                  Kappa : 0.2769
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9964
            Specificity : 0.1856
         Pos Pred Value : 0.9053
         Neg Pred Value : 0.8683
             Prevalence : 0.8865
         Detection Rate : 0.8833
   Detection Prevalence : 0.9757
      Balanced Accuracy : 0.5910

       'Positive' Class : 0
```

Two summaries of confusion matrixes are shown respectively.

Based on the different requirements, the model can be deployed.