Prediction of Recovery / Death for COVID-19 Patients

Tianling Wang

1. Problem Definition

The death rate for COVID-19 is pretty high at around 3.4% globally. So I want to explore what factors affect whether a patient could recover from the disease and try to predict the final outcome of a confirmed patient.

2. Data Preprocessing

Based on EDA of the dataset, I found 7 potential features which could be essential for a patient's recovery/death outcome.
(1) case in country: If there are more cases in the country, the medical resources would be more limited.
(2) location: Since the dataset is rather small, I use country rather than city as the feature. After exploring the count in each country group, I choose 5 countries with the highest confirmed numbers and use one-hot encoding to create 5 dummy variables.
(3) gender
(4) age
(5) day to hospital: time between system onset to hospital visit
(6) exposure time: time between exposure start to exposure end
(7) visiting Wuhan

As for the label, I drop all data that has not yet received outcome, and only remain those with recovery / death outcome.

3. Model

I used XGBoost for the classification. One reason is that the dataset is quite small, I can control overfitting by tuning the parameters with lower tree depth and higher L1 and L2 regularization. Another reason is that there're missing values in some columns, XGBoost can automatically handle this problem.
The accuracy of the model is 85% and the AUC is 81%. See the visualization in the notebook for feature importance and the influence of each feature.

4. Next Step

In the next step, I would like to use the model to predict the outcome for patients who are still in hospital. With this prediction, the hospital could better allocate the resources and try to save lifes for vulnerable patients in advance.