Final Script_Lucy Yuan

Given the progress of the project, the main Machine Learning Analysis is based on 3.5 million rows of data – a sample of the full database.

[Click to show Hypothesis]

Our hypothesis is that, patients who have medical underlying conditions are more likely to run into serious situations, even death from COVID infection, and are supposed to be classified as high risk group.

[Click to show Data Characteristics]

After eliminating the negative cases, we got more than 1.4 million rows of data and 41 columns.

[Click to show Preprocessing Focus]

Furthermore, we eliminated the columns with irrelevant info, and encoded all the medical conditions, and target column ["Date of Deceased"], removed the outliers and scale down the ['age'] column into new age groups.

[Click to show Preliminary Features]

Before fitting the model, we used R to perform a preliminary feature analysis. The result shows almost all the features/medical conditions are significantly relevant to the target ['Death']. More details will be disclosed by Safaa.

[Click to show Models Selection]

As we found only 11% of the data is in 'High risk' class, we are focused on the following models and different resampling method to handle the imbalanced dataset.

Given the use case of the project, we would rather be more aggressive in classifying patients than miss any potential high risk patients. So Recall is the metrics we are focusing on.

<Flip slide>

Based on the listed models, we started with the K fold(n_splits=10) cross validation on 30,000 rows of data, and compared Recall of different models with resampling methods.

The result shows both SMOTE and Undersampling performed well across all models while SMOTEENN was showing large amount of variance in Recall. Comparing with the validation without any resampling, we can confirm that Resampling significantly improves the model performance.

We also found that the Easy Ensemble and Balanced Random Forest stay consistent across different resampling methods due to natural integration of Undersampling.

<Flip slide>

Furthermore, we increased the sample data to 50,000 rows, and used StratifiedKFold(n_splits = 5) validation and test datasets to evaluate (25% of the sample data) all 7 models with both SMOTE and

Undersampling. Based on the best recall score, we decided to use the Balanced Random Forest as our prediction model.

[Click to show Final Model Selection]

Finally, we ran the 1.4 million rows of data into the Balanced Random Forest, and got the 88% of recall and 86.57% of the balanced accuracy score in almost 1 min.

Before we reach the conclusion, we also did Feature engineering and other optimization attempt. I will hand it over to Safaa for further explanation.