# Information extraction using XGBoost

In this project, I use both R and Python, so you will find two folders there, one for R, the other for Python notebook.

Basically, I use R to draw the majority of plots and use Python to fit all models. The plots are saved under the ‘images’ folder in R folder.

I have created a PowerPoint file for project illustration.

The analysis pipeline is in the 2nd slide of PowerPoint, I highly recommend read the PowerPoint first then go to my codes.

In my code scripts, I try to train the model using three different samples (original, combined, SMOTE), details are discussed in PowerPoint. With each sample, the code structure is almost the same. So the codes may look like a little lengthy, If I get enough time, I should actually make a wrapper for them to make the codes look tidy.

I highly recommend do not run all the script especially for the XGBoost model training part, it is really time-consuming in my computer (like 1 hour).