Future directions. In this project1, I mainly focused on xgboost, as it can utilize gpu, and every epoch of hyperparameter process took much less time, so I can fine tune the model with different set of parameters. But after I modified parameters like tree method, scale pos weight, I found that the f1 score would not be significantly improved to above 0.90, the f1 score was still between 0.86 and 0.88, and different preprocessing data method would not help xgboost to improve. Xgboost is a relatively fast way to build a usable model for prediction in a limited time.

And then I tried on random forest for binary classification, random forest was running on cpu only and every epoch of hyperparameter process took much longer time, 10x time longer than every epoch of xgboost, but I saw the potential ability to improve f1 score to close or even above 0.90 if I have more time to fine tuning. I saw the trends that if I increased the epochs of hyperparameter process, the f1 score was slowly improving, but every epoch took about 2 seconds so running 500 epochs were needed to wait for over 15 minutes. I didn’t have much time and opportunity to fine tune and explore further.

In random forest I also found that if I used the max features as sqrt, the f1 score will be better that the use all features to help the model to make the split, include all bio feature data at every node will not help.