MSBD 6000B Project 1

Name: NG, Yui Lun

Student ID: 20039916

The final model used in this project is Random Forest without any pre/post-processing. Using the tuned parameter (Table 1), the model can achieve around 95% accuracy in 5-fold cross-validation.

max_depth:21	min_samples_leaf: 1
max_features: 11	min_samples_split:2
bootstrap': False	criterion: "entropy"
n_estimators: 91	

Table 1: Tuned Parameter for Random Forest

Best 5-Fold

AUC Score (Train): 0.952831

	precision	recall	f1-score	support
0	0.955	0.967	0.961	369
1	0.956	0.938	0.947	275
Avg/Total	0.955	0.955	0.955	644

Worst 5-Fold

AUC Score (Train): 0.941157

	precision	recall	f1-score	support
0	0.958	0.953	0.955	403
1	0.922	0.929	0.926	241
Avg/Total	0.944	0.944	0.944	644

During the testing phase, SVM on original training data can achieve 85% accuracy while SVM on normalized training data can raise the accuracy to 92%. Random Forest can get a better accuracy (around 93% with the default parameter giving by the RandomForestClassifier) and usually more robust to irrelevant features. Random forest is better than decision tree because decision tree will overfit the training set when the trees are very deep while random forest can average the deep tree to reduce the variance.