Unique screen name: Bossk
My name: Yuchen (Bobby) Zhang
PUID: 00280-96080
Prof. Bruno Ribeiro
CS 37300, Fall 2017
Purdue University
Nov 30, 2017

Kaggle Competition

Data Preprocessing

1. Transform all continuous variables into numbers
2. Extract numerical information from discrete variables, such as Employment
3. Bijectively map plain string features into relatively small numerical values
4. Compare the range of features in train data with those in test data; then filter out that piece of data entry if outlier exists

Test different models

1. Randomly split the "ToTrain" data into train and test data
2. Use different classifiers to train the data
3. Adjust parameters for each classifier to get the best score of accuracy

Classifier used

1. Decision Tree: recursively split the data using the fields that minimize the information/gini gain until reach max deep or the split is pure.
2. Bagging on DecisionTreeClassifier: since decision tree is unstable, and bagging is to help stabilize unstable classifier (reduce variance).
3. Gradient Boosting: convert weak classifiers to stronger ones, using gradient approach to acquire the score.
4. Voting Classifier: democracy, minority obey the majority.

Voting

1. Use Voting Classifier on the classifiers listed above with the highest accuracy score as the weight for each classifier.
2. Perform the similar concept as voting classifier, but using the score as weight given by the predicted result on "ToPredict" data on public board.