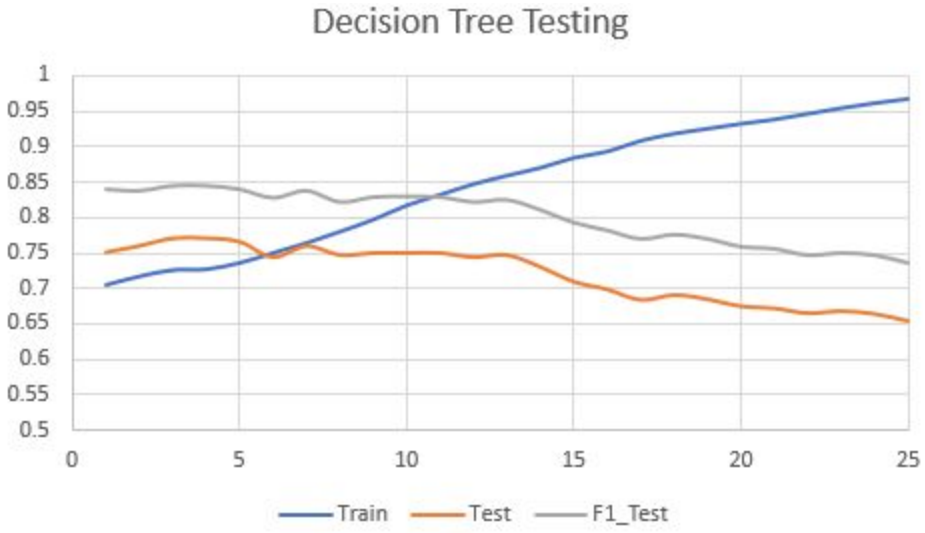
Yunqiao Cai
Ghaith Shan
Sarun Pitaksuteephong
May 11, 2020

CS434 Implementation Assignment #3 Report

1.

a. Plot accuracy vs. # of trees, plus F1_score vs. # of trees.



b. Report the depth that gives the best validation accuracy.

The best depth is 4.

c. What is the most important feature for making a prediction? How can we find it? Report the name of it (see data dictionary) as well as the value it takes for the split.

The first one, which is "Population, 2014 estimate". The reason we think it is this one is because it has the highest gain from the split.

2.

a. For *max_depth* = 7, *n* = 50 trees, and *max_features* ∈ [1, 2, 5, 8, 10, 20, 25, 35, 50], plot the train and testing accuracy of the forest vs. the number of trees in the forest 'n'. Include train and testing of F1_score vs. number of trees as well.

b. What effect does adding more trees into a forest have on the train/testing performance? Why?

c. Repeat above experiments for *max_depth* = 7, *n* = 50 trees, and *max_features* of [1, 2, 5, 8, 10, 20, 25, 35, 50]. How does *max_features* change the train/validation accuracy? Why?

d. Run 10 trails of different seeds. Overall, how do you think randomness affects performance?

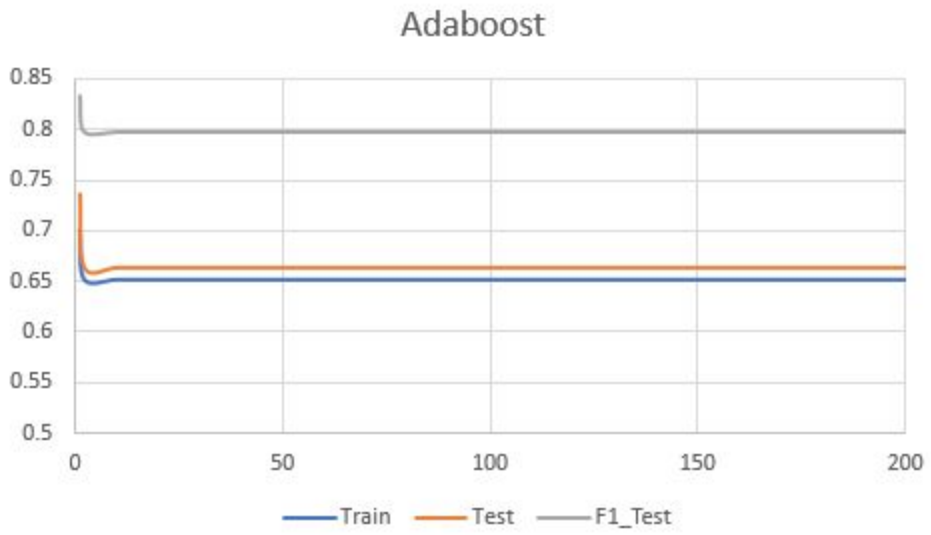
3.

a. Report the train and validation accuracy for L ∈ [10,20,...,200]

All of them are train =0.652, Test = 0.789 F1 Test = 0.797

b. Plot the train/test accuracy and train/test F1 scores of AdaBoost against the parameter L

Yunqiao Cai
Ghaith Shan
Sarun Pitaksuteephong
May 11, 2020



4. Article Summary & Our Ideas for Handling Imbalanced Dataset:

The article references that in machine learning the most ideal data sets are ones that are balanced with an even number of each type of data leaning toward a 50/50 distribution. However, this isn't the case for real-world data where cases like modeling minorities vs majorities gets you a highly imbalanced data due to one side having a larger population than the other.

The author then briefly discusses some methods to circumvent this. Some of these methods include oversampling and undersampling but these methods require you to collect new data or alter the data set. Another method is changing the weight for each class such that one has greater effect than the other. This exaggerates the minority data to create a more balanced dataset. Alternatively, changing the decision gives a similar result but makes it so that you need to reach either a higher or lower threshold depending on the minority data.

In our case, our data has a base accuracy of 66% as opposed to the ideal 50%. To lessen the imbalance, we decided to adjust our weights so the accuracy is closer to 50% although it is hard to completely circumvent it. An idea we could try to balance the distribution is by adding more feature(s) as a right one can help further distinguish the data and therefore improve the accuracy. Another possible method is to train the classifier in batches. This method can help reduce the overall proportion variance within the entire data set, increasing stability and balance.