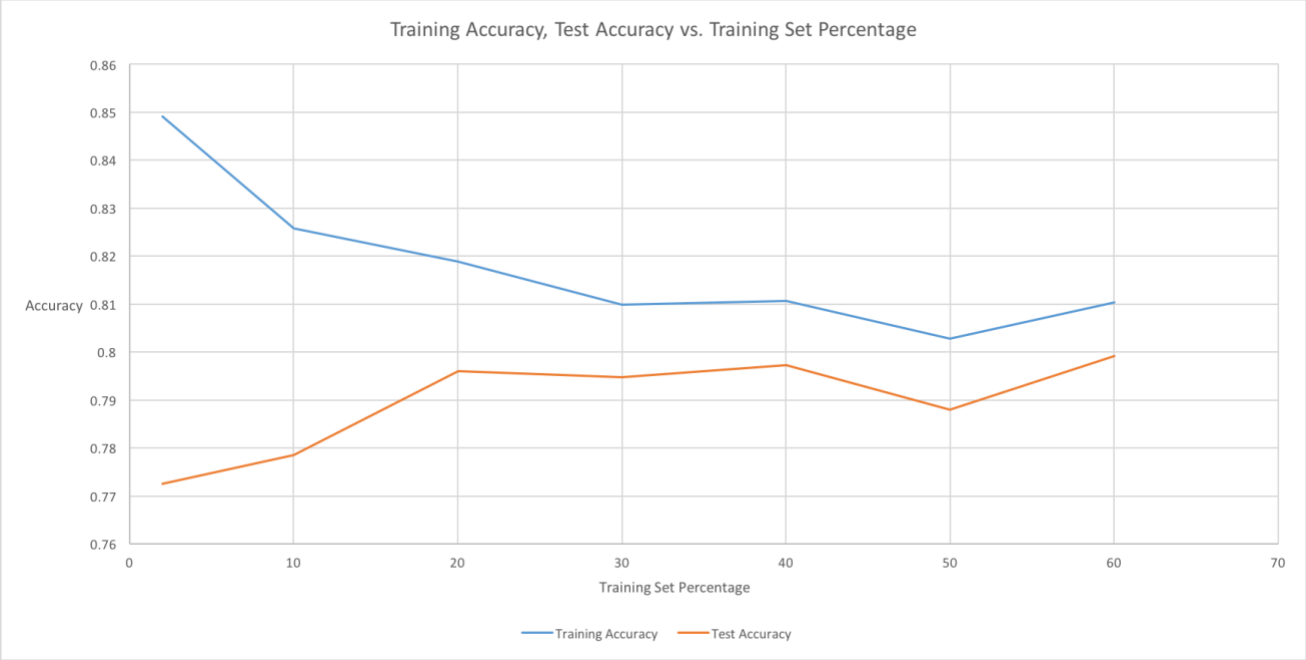
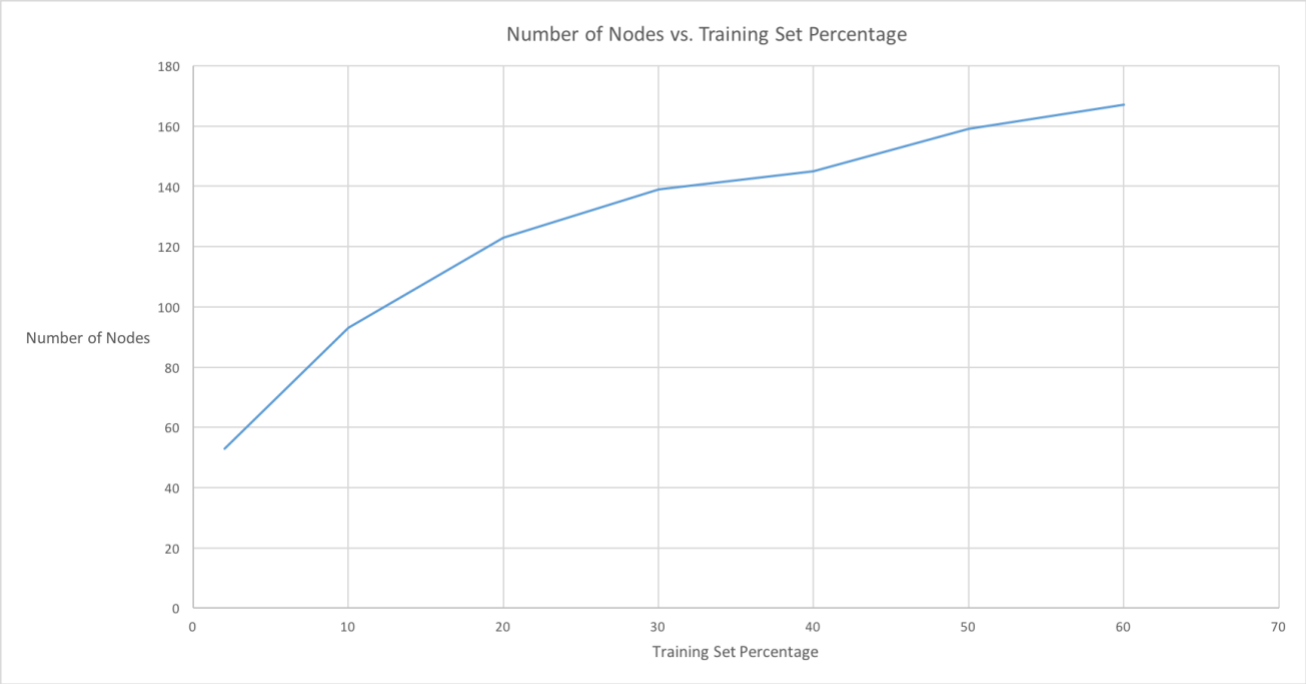
Haoran Wang
00274-00605

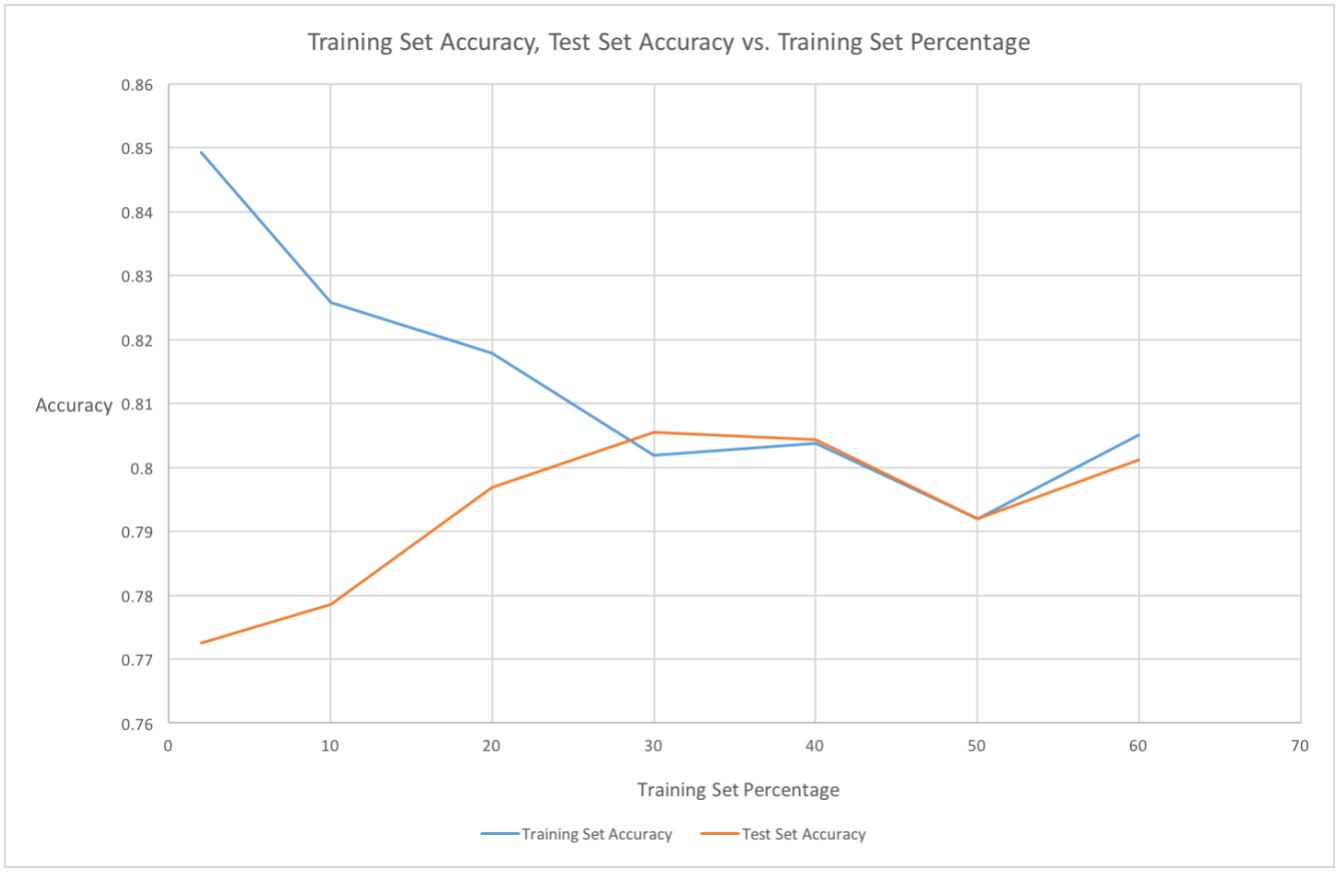
Part 2

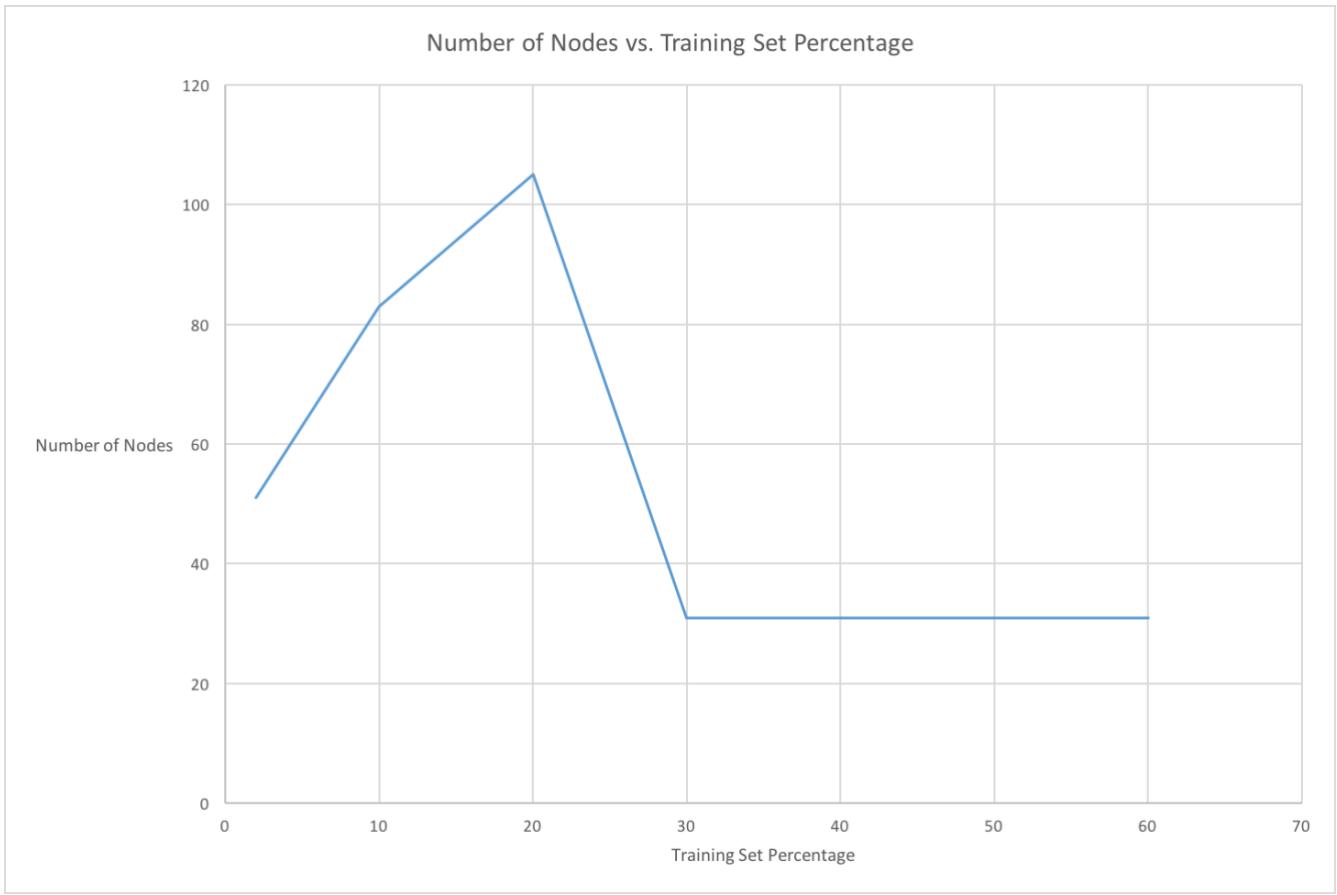
Q1. Vanilla Tree

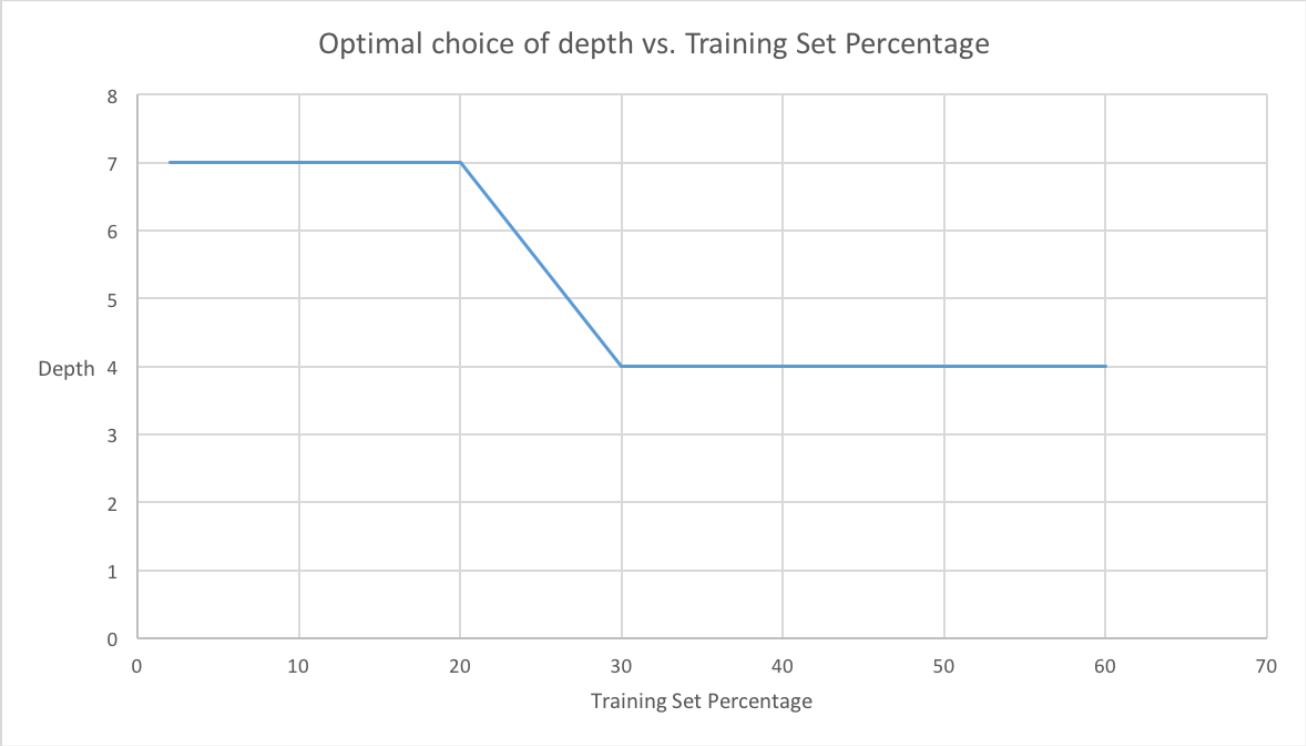




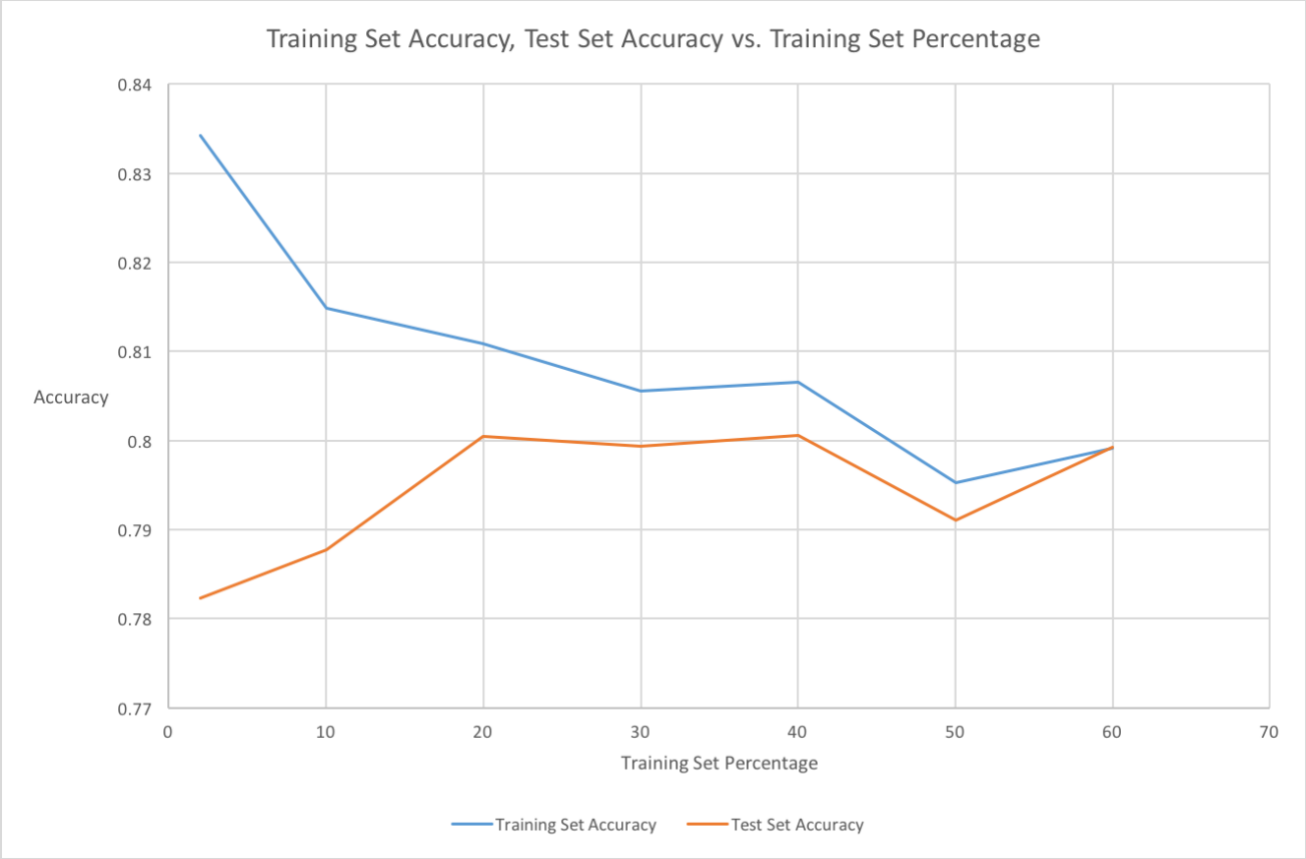
Q2. Depth Tree

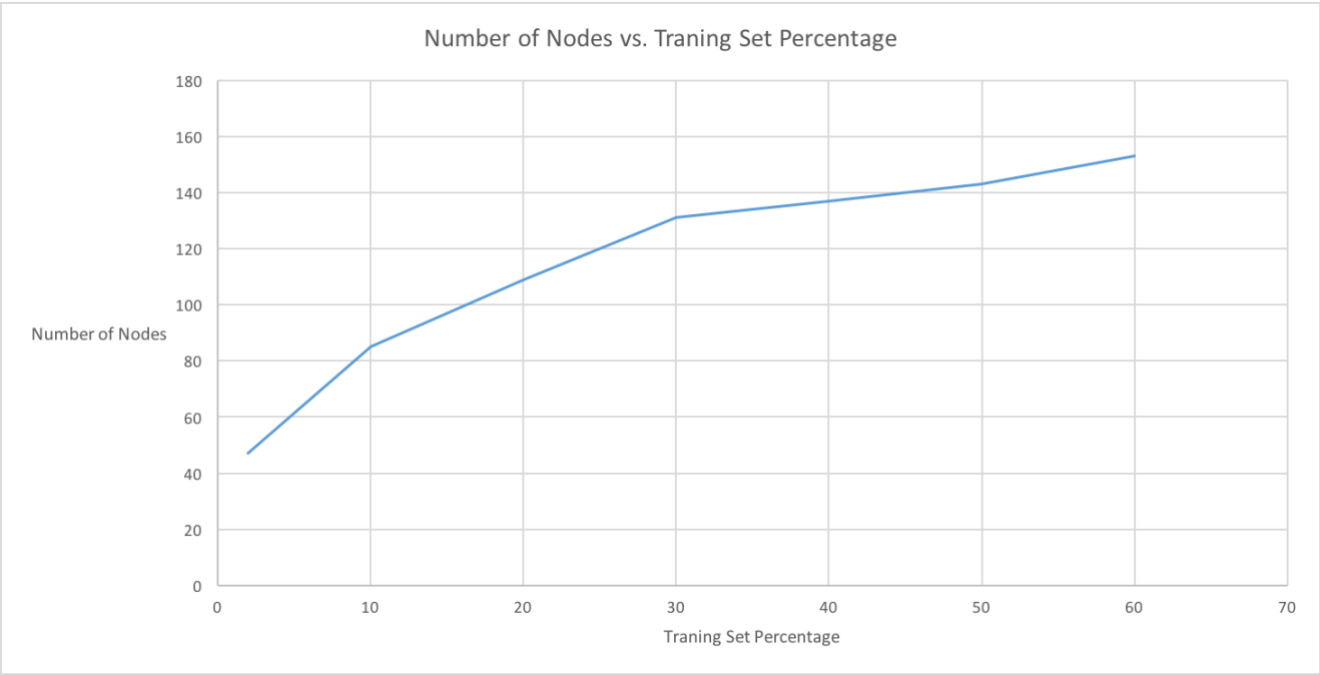

Training Set Accuracy, Test Set Accuracy vs. Training Set Percentage
Accuracy
0.86
0.85
0.84
0.83
0.82
0.81
0.8
0.79
0.78
0.77
0.76
0
10
20
30
40
50
60
70
Training Set Percentage
Training Set Accuracy
Test Set Accuracy


Number of Nodes vs. Training Set Percentage
Number of Nodes
120
100
80
60
40
20
0
0
10
20
30
40
50
60
70
Training Set Percentage



Q3. Prune Tree





Q4.

The purpose of pruning is to solve overfitting problem. Overfitting is caused by the model learns the details and noise in the training data set. Overfitting decision tree will perform better on training data set than the test data set. Therefore, if we prune on the test set, we cannot solve the overfitting problem because we did not change training set.

We use a separate validation set to provide an unbiased evaluation of a model fitting on the training set, tune the hyperparameters of the classifier and to avoid overfitting problem.

Q5.

We can replace “>50k” and “<=50k” labels with continuous variables such as the person’s exact income as our new labels. Then we can divide income into several intervals and make the intervals discrete. When we use the decision tree to predict, the leaves are the intervals and therefore we completed ranking.