Lab 4 Report

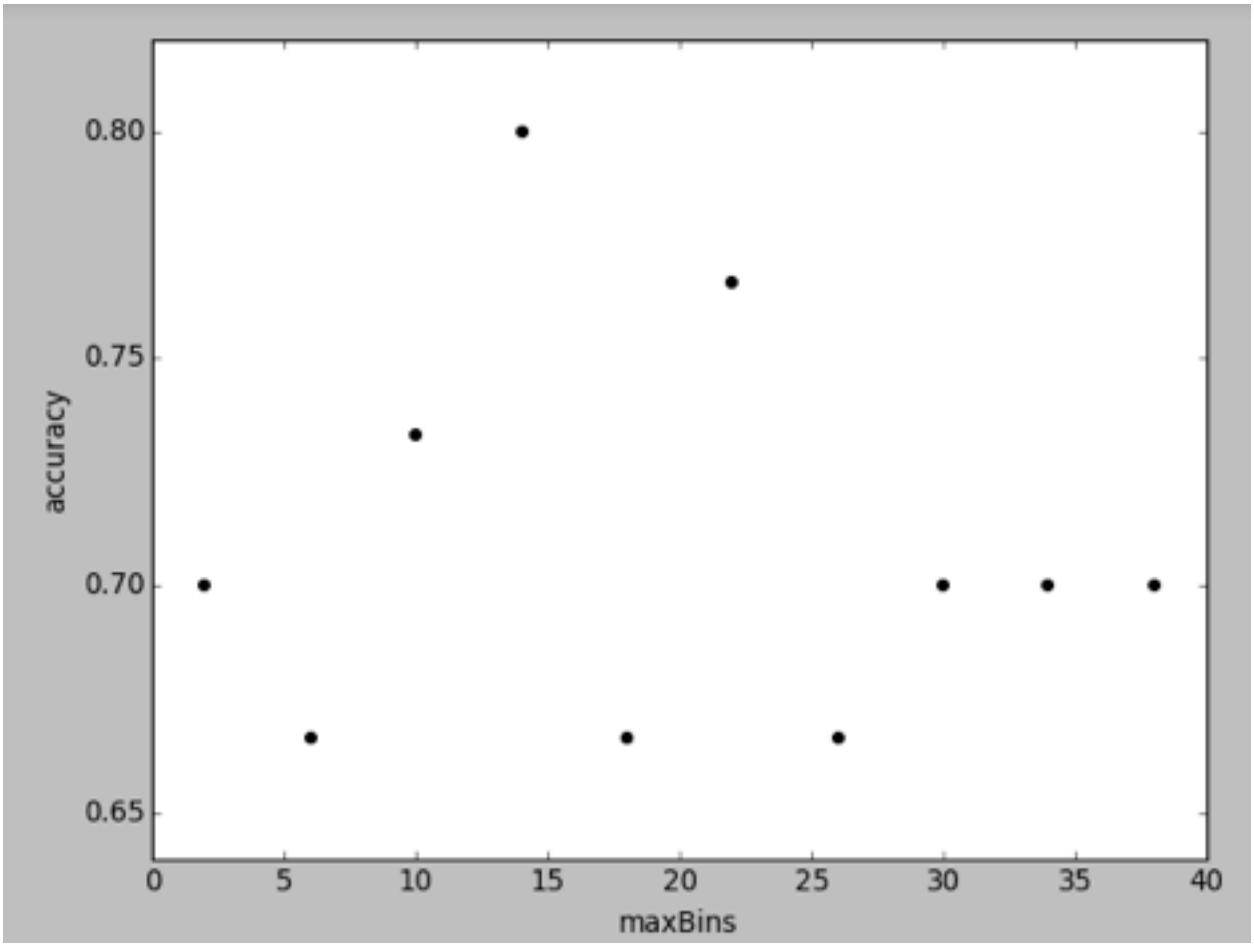
Rohan Nagar (ran679)

Part 1 - Decision Trees

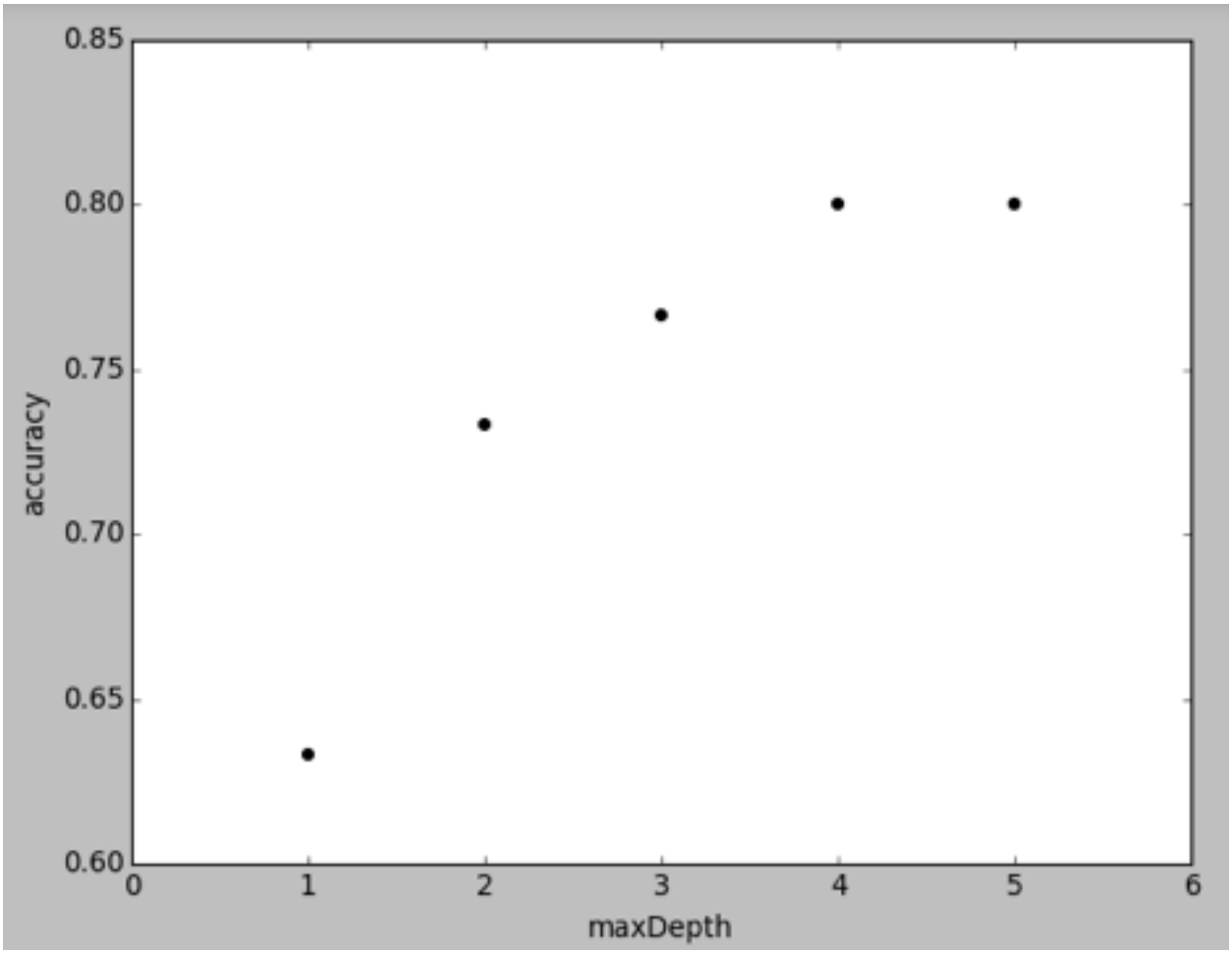
Note: I ran all of the Decision Tree code using 2 threads for the num_threads Spark argument.

1. Keeping maxBins set the 32 and the maxDepth at a default of 5, the accuracy for different impurity measures were:
    - Gini: Accuracy = 0.767
    - Entropy: Accuracy = 0.700

2. Varying the number of bins. I kept the impurity measure fixed as gini and kept maxDepth at the default of 5. I used bins of [2, 6, 10, 14, 18, 22, 26, 30, 34, 38]. The highest accuracy was when maxBins=14, so I used that when training future models. The plot is below.



3. Varying the max depth from 1 to 5. I kept impurity as gini and the max bins as 14, since that produced the best accuracy from part 2. From the plot, we can see that accuracy improved as the max depth got higher, until it leveled off at 4 and 5. The plot is below.



4. The final parameters I used for my model were:

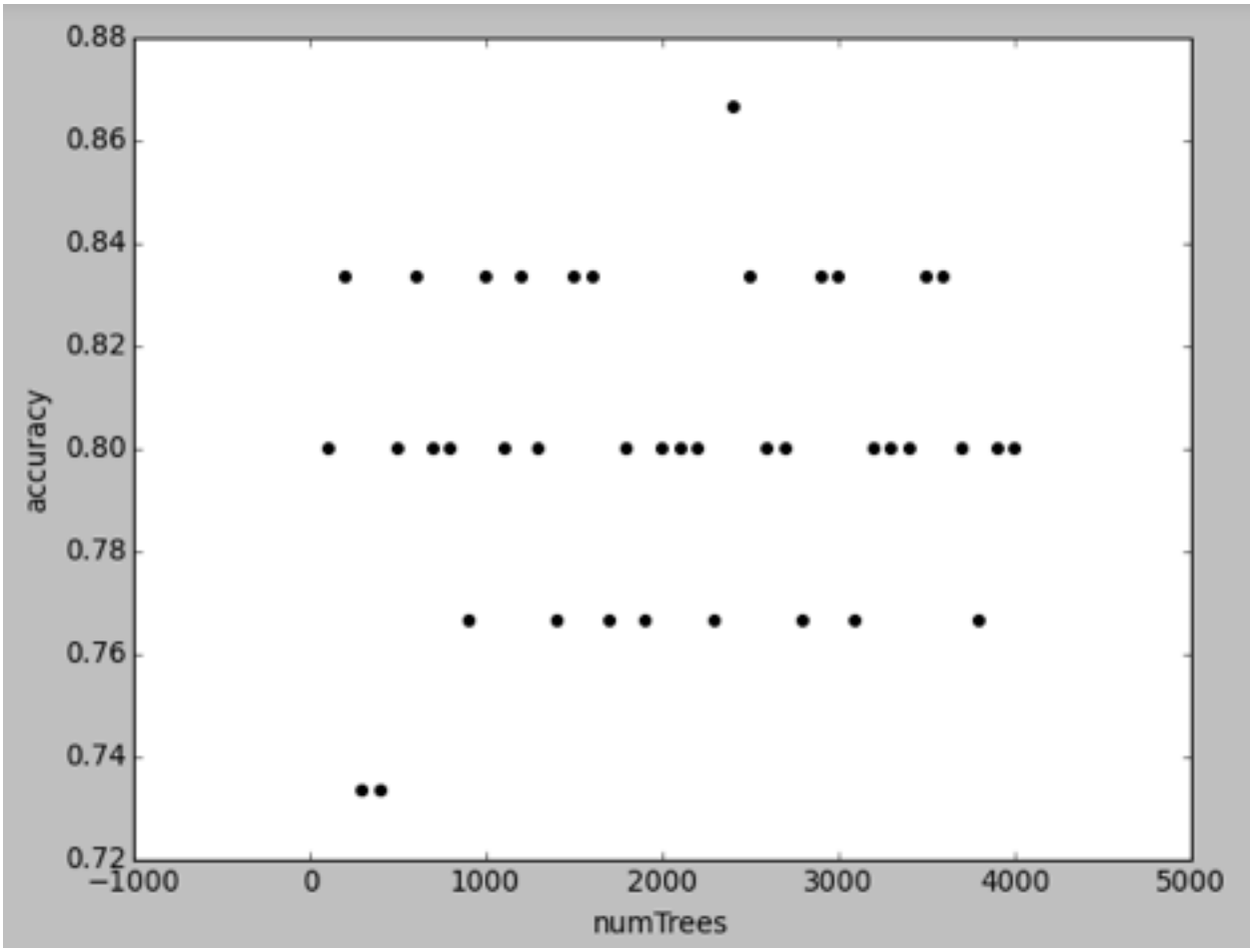
- Impurity: Gini
- maxBins: 14
- maxDepth: 5
- minInfoGain: 0.05

5. Using these parameters, the accuracy I got on the validation set was 0.80.

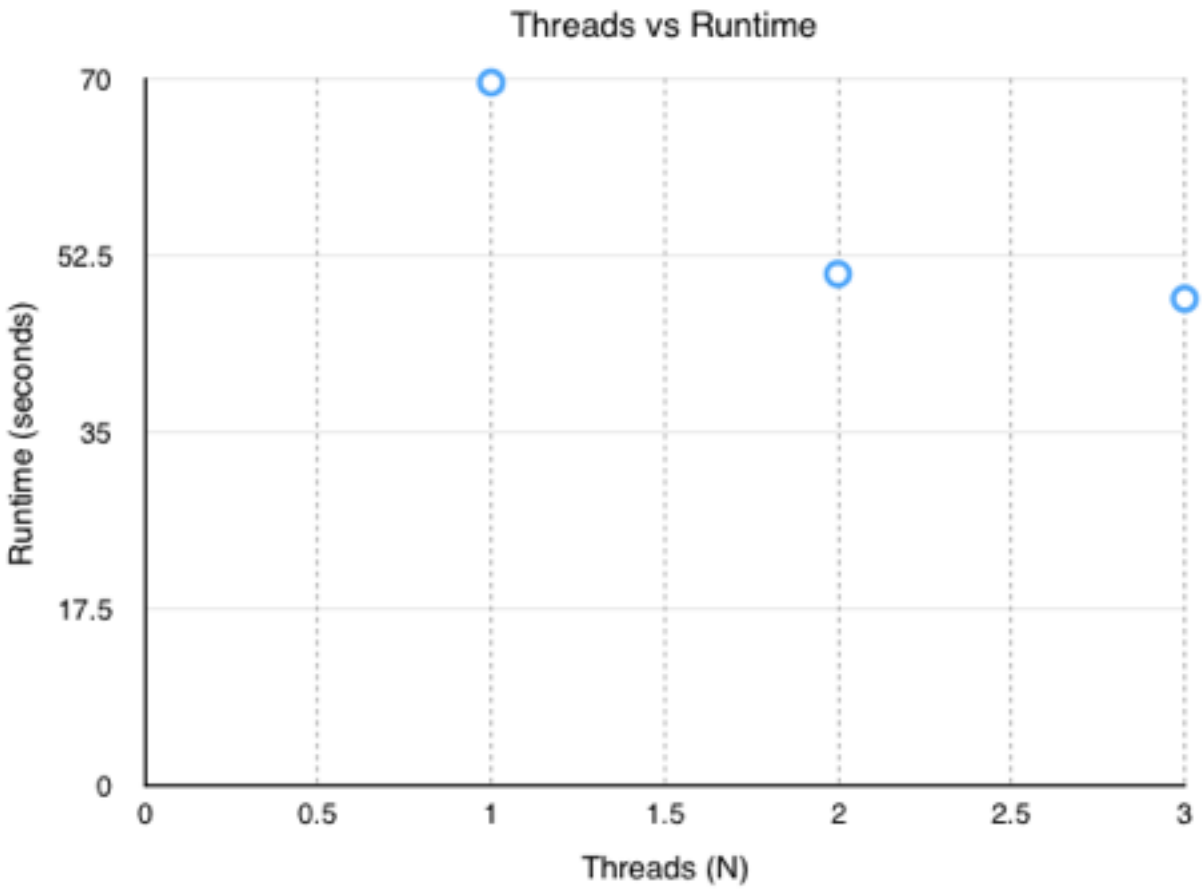
6. Using these parameters, the accuracy I got on the holdout test set was 0.66. I didn't do as well on the test set as I hoped, especially considering that the validation set accuracy for that model was quite high, but I tried tweaking even more to make it better and it wouldn't get any higher.

Part 2 - Random Forests

1. I trained in intervals of 100 trees so that it wouldn't take hours to run. I could only go up to 4000 trees, any more than that caused an OutOfMemoryError. The highest accuracy was with 2400 trees. The plot is below.



2. I was only able to go up to N=3 threads, because that is the max that I could do on my virtual machine. The plot is below, but it is clear that more threads allow the model to train faster.



3. The final parameters I used for my model were:
    - Impurity: Gini
    - maxBins: 14
    - maxDepth: 5
    - numTrees: 2400

4. Using these parameters, the accuracy I got on the validation set was 0.80.

5. Using these parameters, the accuracy I got on the test set was 0.77. This is much improved from the decision tree model, which is expected.