Project Report

Topic:Predicting Review Sentiment Using Decision Tree

Aditya Tiwari

2016csb1029

Dataset Creation

Training Set

I have stored 1000 random reviews in a file named my_data.txt.Since I do not consider the frequency in decision tree or rating,instead I consider the presence or absence and is positive or negative rating.So data stored in my_data.txt is of first 500 positive reviews with all rating 9 or 2 and with each word having frequency 1.

TestSet:

I pick random 1000 reviews from the test reviews given.

Experiment 2:

I trained decision tree on the selected 1000 reviews and took random 1000 reviews and test accuracy of the model.

Average comes out to be **70.24** (done on 10 test data[71.2, 71.1 ,69.7, 69.6, 70.4, 70.3, 69.1, 71.0, 71.8, 68.2])

I have used pickle library to save all the input_test_data

Early Stopping:

1. Using Information Gain Threshold:

I.G Threshold	Depth of Tree created	Avg Accuracy on 10 Random Test Set(%)
0.00051	205	69.47
0.00052	205	70.01
0.00053	205	70.38

0.00054	205	69.79
0.00055	98	69.63
0.00056	**98**	**70.49**
0.00057	98	70.09
0.00058	98	69.98
0.00059	98	69.18

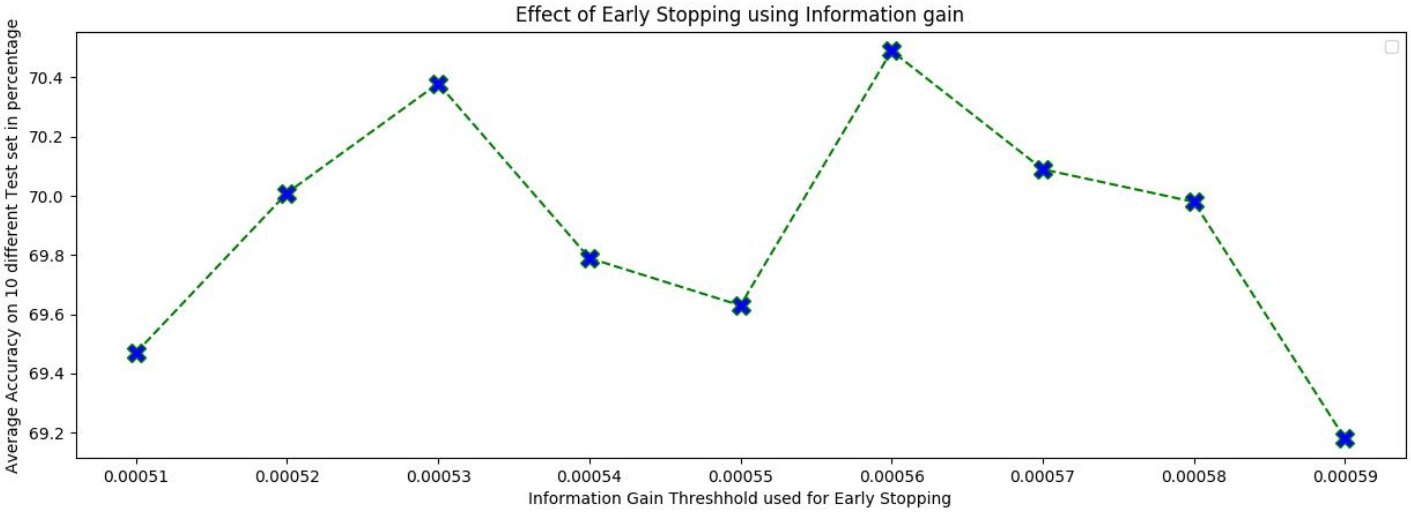
Using I.G to do early stopping didn't give good results.

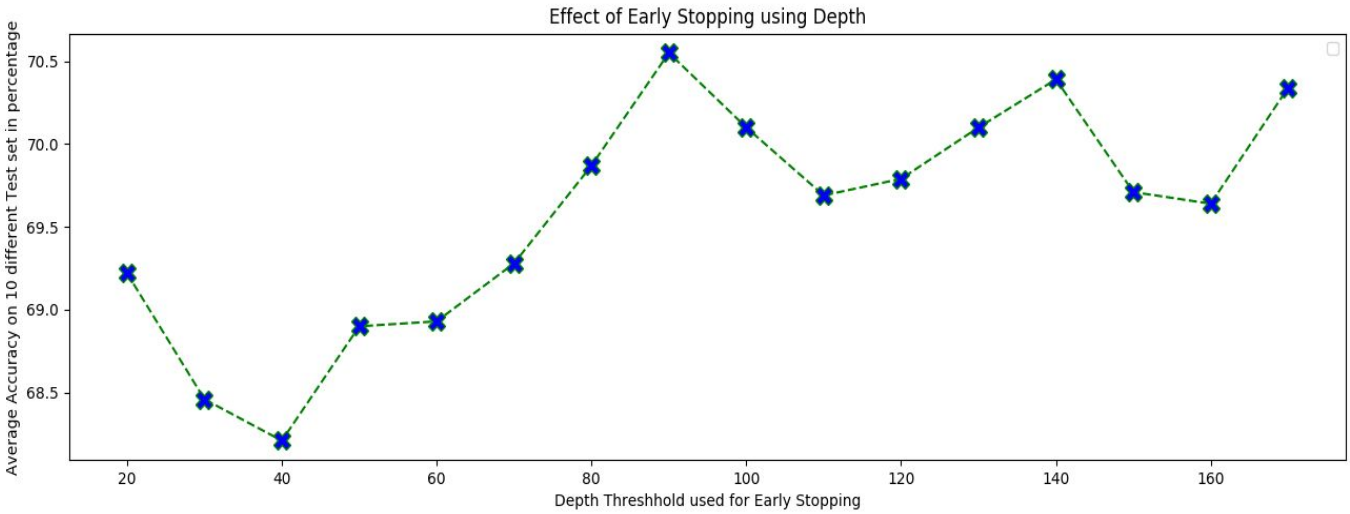
2. Using Max Depth of Tree

Depth Threshold	Depth of Tree created	Avg Accuracy on 10 Random Test Set(%)
20	21	69.22
30	31	68.46
40	41	68.21
50	51	68.9
60	61	68.93
70	71	69.28
80	81	69.87
90	**91**	**70.55**
100	101	70.1
110	111	69.69
120	121	69.79
130	131	70.1
140	**141**	**70.39**

150	151	69.71
160	161	69.64
170	**171**	**70.34**

There isn’t any considerate increase in accuracy by early stopping





Observation:

Using Depth for early stopping gives better result than Information Gain but is still negligible.

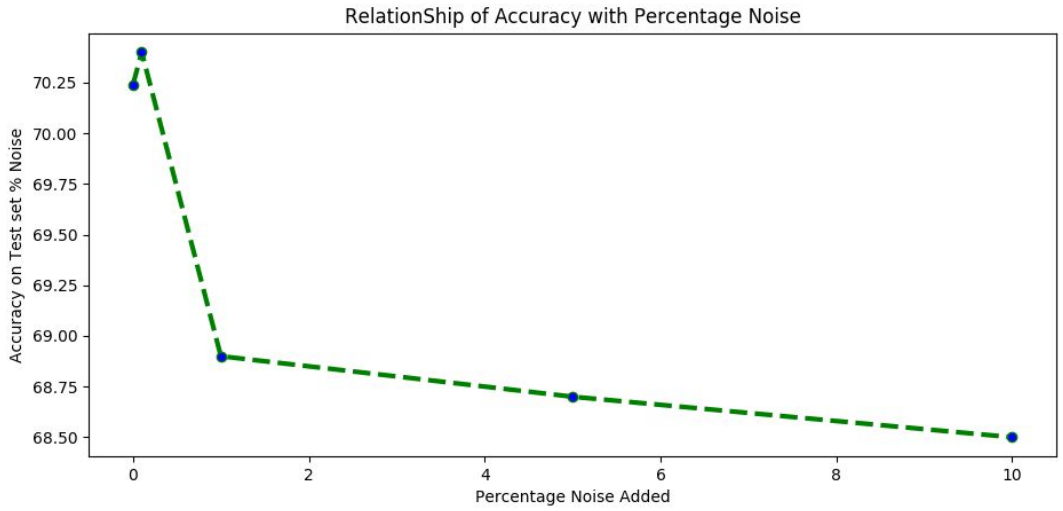
Experiment 3:

I created a total of 5 noisy data set, I changed randomly 10% of data set to opposite label in dataset.I have saved the noisy data sets as well
Following are the observations

Note:Blue Text represents original Tree data

Number of Nodes	Depth	Accuracy	Percentage Noise
847	**205**	**70.24**	**0**
849	205	70.4	0.1
863	201	68.9	1
859	213	68.7	5
819	215	68.5	10

The Accuracy Slowly decreases with the Increase in Percentage
Number of Nodes Increases at first to overfit and consider the extra rules to be made for extra noisy data points and hence require more splitting i.e more nodes.
However I'd say after a certain point these noisy points start to group together and instead increase depth than as compare to nodes.



Experiment 4:

Post-Pruning:

I have done greedy pruning, I recursively travelled through nodes and if making the current node a leaf node with label=common label in the training data that was passed for this node increases the accuracy then I make the current node leaf node else I do the same for its left and right.

Results:

I took Created Test and Validation Set from Test Data and ran

Pruning with validation set on same Training Set

Accuracy on Test Set before Pruning: 70.6

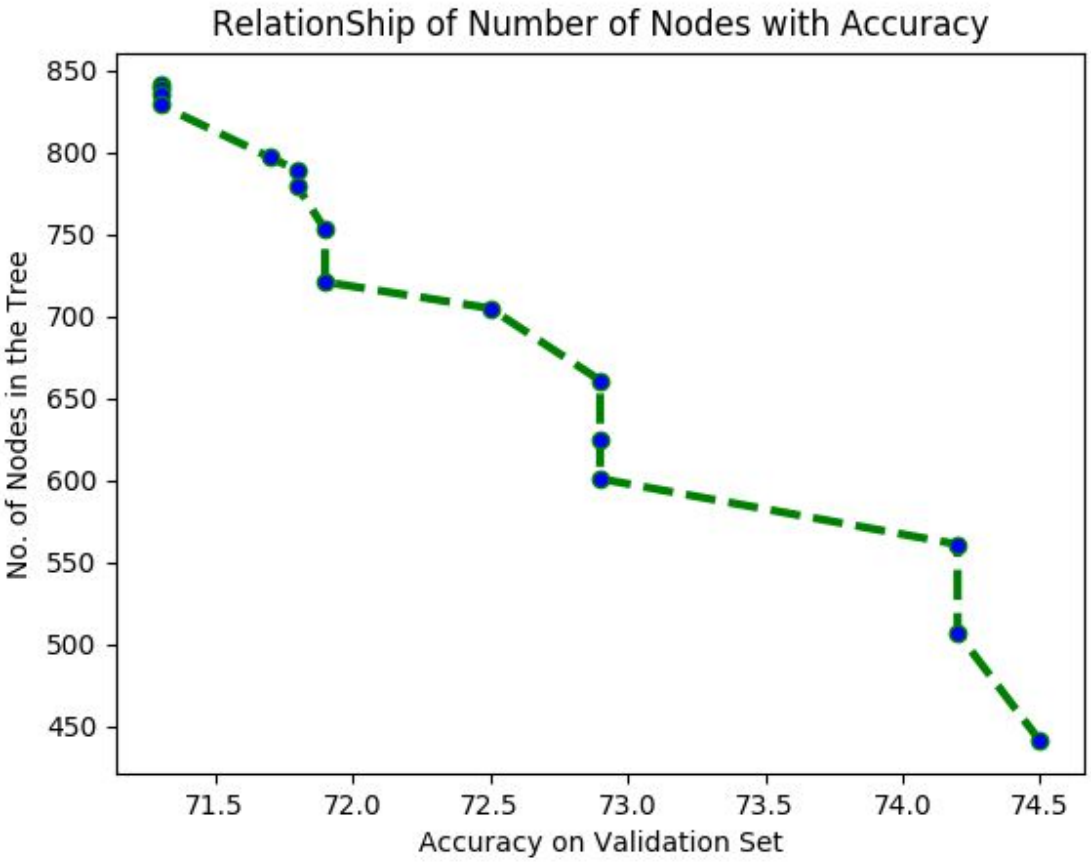
Accuracy(In Process on Validation Set)	No. of Nodes	Depth of the Tree
71.3	841	205
71.3	839	205
71.3	835	205
71.3	829	205
71.7	797	205
71.8	789	205
71.8	779	205
71.9	753	205
71.9	721	205
72.5	705	205
72.9	661	205
72.9	625	205
72.9	601	205
74.2	561	205
74.2	507	205
74.5	441	205

Accuracy on Test Set after Pruning: 75

The Above stats may vary a little but The increase in Accuracy is roughly 5% or comes to near about 75%

I ran it a couple of times, but Depth doesn't changes.

It would seem the node that has the max depth , that subtree has the max Information Gain Split



and hence its removal would only decrease accuracy.

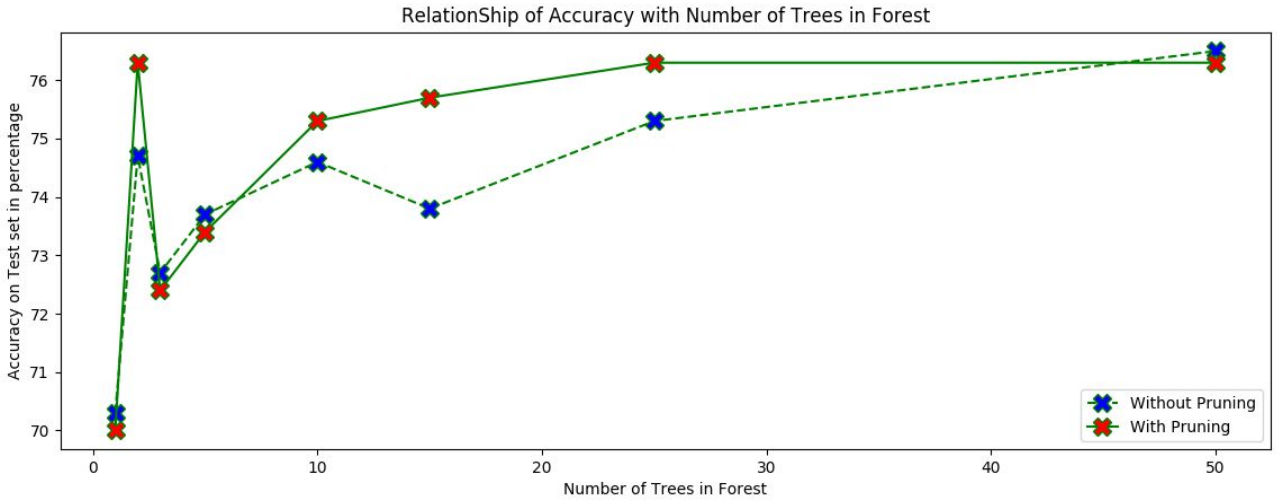
Experiment 5:

I first select top 5000 Features and randomly Select 2000 features out of them and create a tree and add it to forest.
Following is the Data:

Forest Length	Forest Accuracy
1	70.3
2	74.7
3	72.7
5	73.7
10	74.6
15	73.8
25	75.3
50	76.5

I
experimented if The Trees in the forest are pruned as well,then what is the accuracy
The Results are:

Forest Length	Forest Accuracy
1	70.0
2	76.3
3	72.4
5	73.4
10	75.3
15	75.7
25	76.3
50	76.3



Observation:

Accuracy vibrates in start but increases later with increase in number of trees in forest and finally becomes constant.

It seems Pruning Achieves Stability faster than without pruning forest as each tree is maximised to have more accuracy and hence becomes constant faster.

So one could use pruning and less trees in random forest instead of increasing the number of trees simply in the forest