# CSE537 - Decision Tree for Clickstream

Group Member: {Yangchun Li: 109819462, Anke Li: }

## Training of Decision Tree

We use a class named “Node” to represent every node in the decision tree. The “Node” class is used in class ID3Tree which represents the decision tree itself. The “Node” class stores basic information for a Node in the decision tree, such as child nodes, predict label of this node (if this is leaf node), etc. And the construction of the decision tree is the process of generating these nodes and build connection of them.

For every step of generating the tree, we first test if the training samples all have the same label, if this is the case, we shall just create a Node with according label and return this Node. Or if we have exhausted all the features, we should use the majority of the label in the training sample and return a Node with this label. Another thing to clarify, positive samples and negative samples are of no equal number. So we can’t just compare the absolute number of positive samples and negative samples, but rather use the ratio to judge the majority of label. Then if above two are not the case, meaning that we need to continue build up the tree. Then we should choose a best feature to split the dataset based on maximize gain of entropy. After we find the best feature, we need to use chi-square test to check if this feature is irrelevant to our target. If so, we need to do early stopping, and return the majority label as above cases.

Then we can recursively build the tree given the above strategy. In order to do the chi-square test, we use the python library: “SCIPY”, which gives convenient way to compute probability of chi-square distribution.

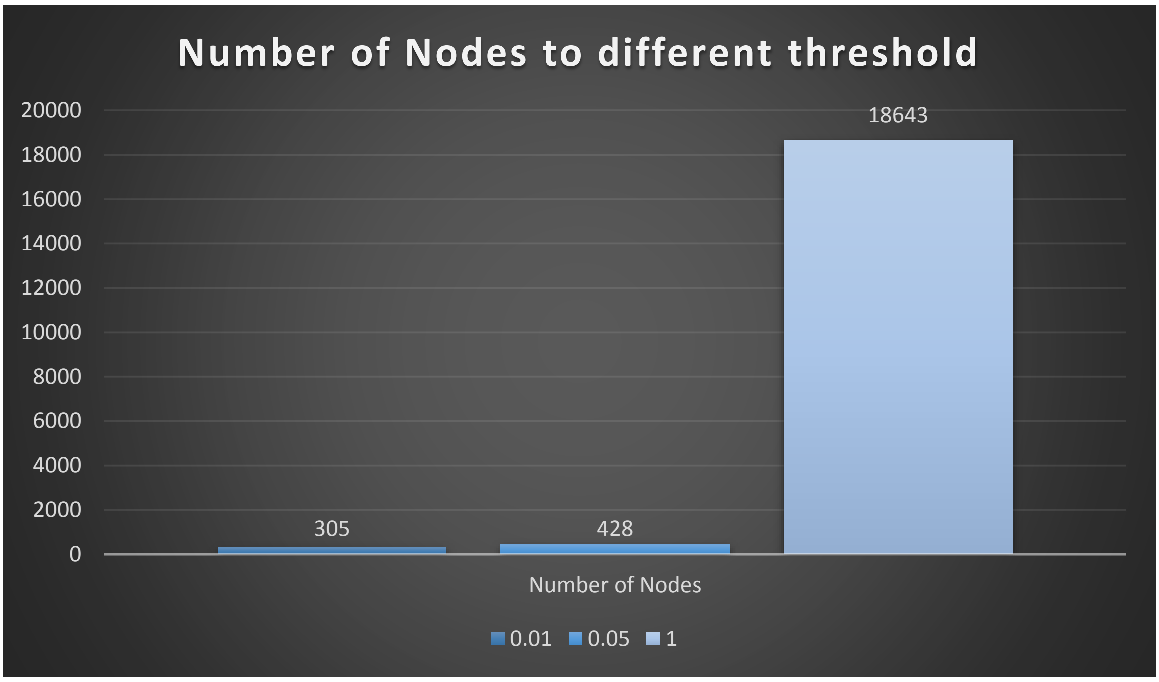
## Test Results

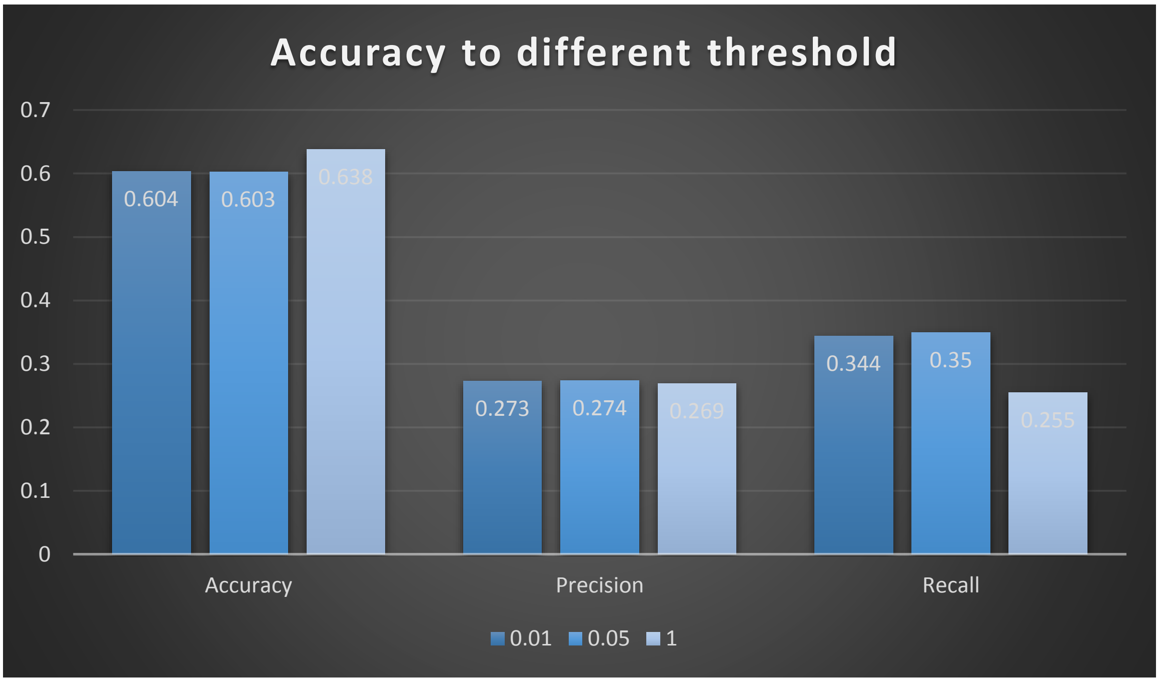
We run the test for different threshold on the whole test set.

Firstly, we try out the full tree, threshold to be 1. The results are as follows:
Node in the decision tree is: 18643
Precision is: 0.269
Recall is: 0.255
Accuracy is: 0.638

Then we try out threshold being 0.05.
Node in the decision tree is: 428
Precision is: 0.274
Recall is: 0.350
Accuracy is: 0.603

The results with threshold being 0.01.
Node in the decision tree is: 305
Precision is: 0.273
Recall is: 0.344
Accuracy is: 0.604





## Analysis:

The result shows that full tree gives the best result. But full tree uses 60 times more nodes than tree with threshold of 0.05. And accuracy is slight lower than full tree, about 5% decrease in accuracy. And another thing to mention is that negative samples in the training set is much more than positive samples. Thus the precision and recall is not high.