Cheng-Ruei Chiu

DSCI - 552: Machine Learning for Data Science
Assignment - 4
Decision Tree

Using the following dataset:

	Holiday	Discount	Purchase
1	No	Yes 1	Yes 1
2	No	Yes 2	Yes 2
3	No	No	No
4	Yes 1	Yes 3	Yes 3
5	Yes 2	Yes 4	Yes 4
6	Yes 3	No	No
7	Yes 4	Yes 5	Yes 5
8	No	Yes 6	Yes 6
9	Yes 5	Yes 7	Yes 7
10	Yes 6	Yes 8	Yes 8
11	Yes 7	No (1)	Yes 9
12	Yes 8	No	No
13	Yes 9	Yes 9	Yes 10
14	Yes 10	Yes 10	Yes 11
15	Yes 11	Yes 11	Yes 12
16	No	Yes 12	Yes 13
17	Yes 12	No (2)	Yes 14
18	No	Yes 13	Yes 15
19	Yes 13	No (3)	Yes 16
20	Yes 14	No (4)	Yes 17

21	No	Yes 14	Yes 18
22	Yes 15	Yes 13	No
23	Yes 16	No (5)	Yes 19
24	No	Yes 16	Yes 20
25	yes 17	No	No
26	No	No	No
27	No	Yes 17	Yes 21
28	No	Yes 18	Yes 22
29	yes 18	Yes 19	Yes 23
30	yes 19	Yes 20	Yes 24

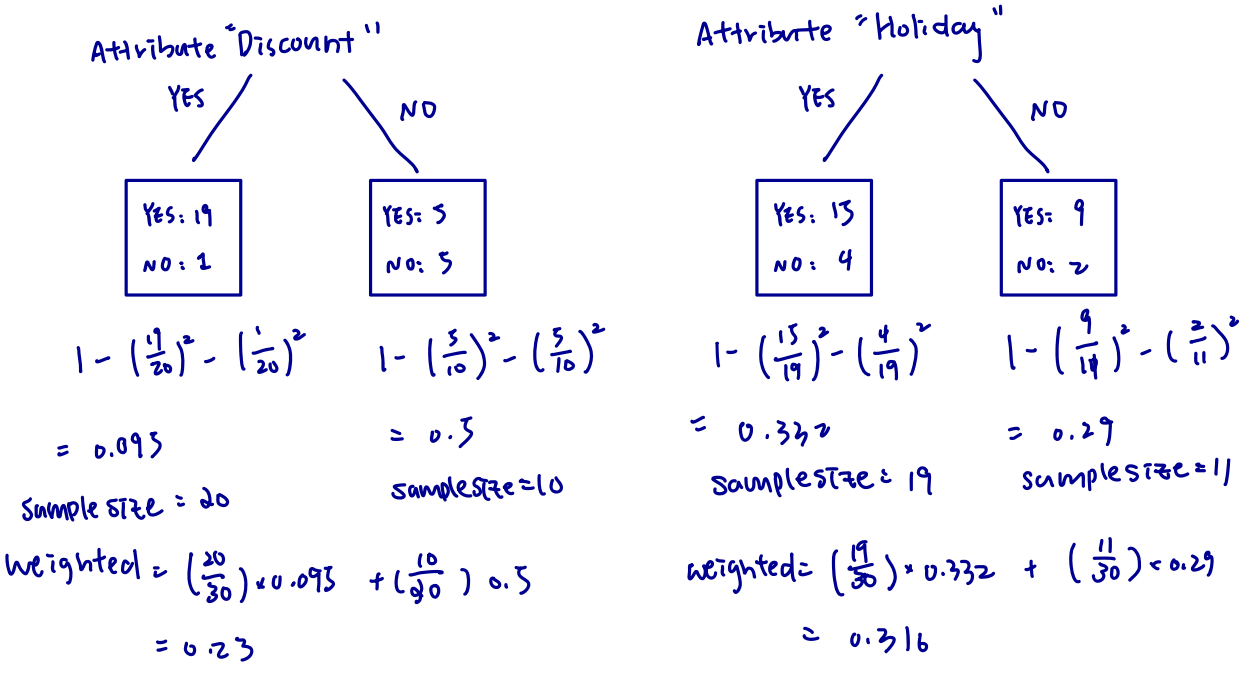
Create a decision tree based on the "Discount" and "Holiday" variables to predict the value of "Purchase". Using weighted gini index as the splitting criteria and show the resulting leaf nodes for each branch of the tree. For the root and each node of the decision tree, calculate the gini index, sample size, and sample distribution.

Hint: Sample distribution = [a,b] where a is the number of “yes” of target value in the current sample, b is the number of “no” of target value in the current sample.

purchase = [24, 6]

gini of root = $1 - (\frac{24}{30})^2 - (\frac{6}{30})^2$

$= \frac{8}{25}$



$1 - (\frac{19}{20})^2 - (\frac{1}{20})^2$ $= 0.095$

Sample size = 20

$1 - (\frac{5}{10})^2 - (\frac{5}{10})^2$ $= 0.5$

samplesize = 10

weighted = $(\frac{20}{30}) \times 0.095 + (\frac{10}{30}) 0.5$

$= 0.23$

$1 - (\frac{15}{19})^2 - (\frac{4}{19})^2$ $= 0.332$

samplesize = 19

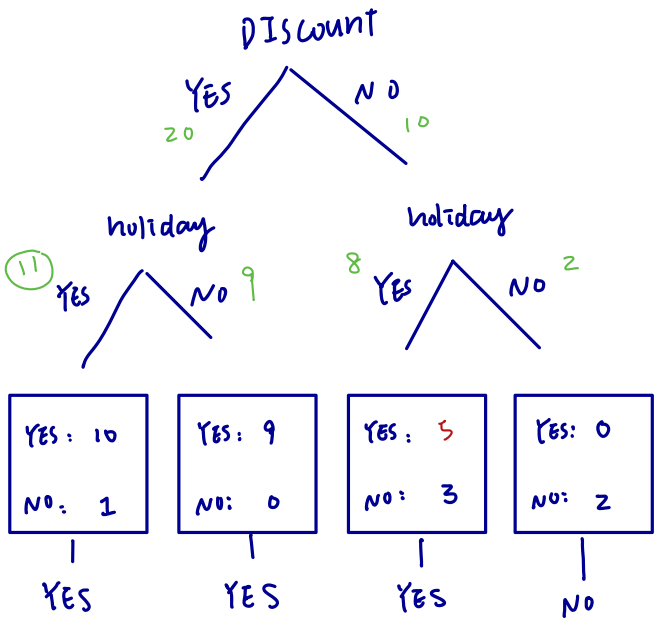
$1 - (\frac{9}{11})^2 - (\frac{2}{11})^2$ $= 0.29$

samplesize = 11

weighted = $(\frac{19}{30}) \times 0.332 + (\frac{11}{30}) \times 0.29$

$= 0.316$

Thus the Decision tree should be



* picking Majority

⇒