10/25/2023

Name: Nikhila Bommareddy
netid: nib22003
student id: 3053527

OPIM:5512 Midterm Assignment - FALL 2023

After performing the feature engineering on variables in the dataset

- → house_median_age
- → total_rooms
- → median_income
- → median_house_value (target variable)

By using seed = 3053527 (student id) the values

	housing median-age	total rooms	median_income	median house_value
0	0	1	1	1
1	0	1	0	1
2	0	1	1	0
3	0	1	0	0
4	0	1	0	0
5	0	1	0	0
6	0	0	0	1
7	0	1	1	1
8	0	1	0	0
9	0	1	0	0
10	1	0	1	1
11	1	0	1	1

Name: Nilchila
netid: nib22003

Entropy of the Entire system

Formula :- $\boxed{E = - \sum_{i=1}^{n} P_i * \log_2 (P_i)}$

probability of the high median house value (1)

$= 6/12$

probability of Low median house value (0)

$= 6/12$

Entropy of the entire system $E =$

$= -\left(\frac{6}{12}\right) * \log_2 \left(\frac{6}{12}\right) - \left(\frac{6}{12}\right) * \log_2 \left(\frac{6}{12}\right)$

$= -0.5 * (-0.30) - 0.5 * (-0.30)$

$= -0.15 - 0.15$

$\boxed{E = 0.3}$

Name : Nikhila
netid : nib2200

Selected features for first split in the entire dataset are

- housing_median_age
- total rooms
- median_income

To get the best decision tree classifier model we have to maximize the information gain and decrease the randomness. To calculate that formula is

$$\boxed{\text{Information gain} = \text{Entropy (parent)} - \text{Entropy(child)}}$$

Above we calculated the entropy of the parent node. Entropy of the child node is the statistical mean of each of the sub nodes.

Entropy of housing_median_age variable:-

Let us assume that older homes values (i.e = 1) i.e greater than the median value

older homes	housing median_age	median house_value
11	1	1
12	1	1

Name: Nikhila
netid: nib22003

Entropy of older_home

$$= -\left(\frac{2}{2}\right) * \log_2\left(\frac{2}{2}\right) - \left(\frac{0}{2}\right) * \log_2\left(\frac{0}{2}\right)$$

$$\boxed{\text{Entropy} = 0}$$

For Newer homes (ie = 0)

	house median_age	median house_value
0	0	1
1	0	1
2	0	0
3	0	0
4	0	0
5	0	0
6	0	1
7	0	1
8	0	0
9	0	0

Entropy of newer_home

$$= -\left(\frac{6}{10}\right) \log_2\left(\frac{6}{10}\right) - \left(\frac{4}{10}\right) \log_2\left(\frac{4}{10}\right)$$

$$= -0.6 * -0.22 - (0.4)(-0.3)$$

$$= 0.132 + 0.12 = 0.252$$

$\boxed{\text{Entropy} = 0.252}$

Name: Nikhila
netid: nib22003

older homes $(E=0,\ n=2)$ Newer-homes $(E=0.252,\ n=10)$

Entropy housing median age

$$= 0 * \left(\frac{2}{12}\right) + 0.252\left(\frac{10}{12}\right)$$

$$= 0 + 0.21$$

Entropy $= 0.21$

Information-gain $= 0.3 - 0.21$

$= 0.09$

Entropy of Total-rooms variable:-

For more-rooms:- $(i=1)$

	~~total~~-rooms	median_house-values
0	1	1
1	1	1
2	1	0
3	1	0
4	1	0
5	1	0
7	1	1
8	1	0
9	1	0

Name: Nichila
vtidi nib22003

Entropy of more rooms.

$$= -\left(\frac{3}{9}\right)\log_2\left(\frac{3}{9}\right) - \left(\frac{6}{9}\right)\log_2\left(\frac{6}{9}\right)$$

$$= -0.3 * (-0.47) - 0.6\,(-0.17)$$

$$= 0.141 + 0.10$$

$$\boxed{\text{Entropy} = 0.241}$$

For less rooms:- (i=0)

	~~total~~ rooms	median house value
		1
6	0	
	0	1
10		
	0	1
11		

Entropy of less rooms

$$= -\left(\frac{3}{3}\right)\log_2\left(\frac{3}{3}\right) - \left(\frac{0}{3}\right)\log_2\left(\frac{0}{3}\right)$$

$$\boxed{\text{Entropy} = 0}$$

more rooms ($\epsilon = 0.241$, $n = 9$), less rooms ($\epsilon = 0$, $n = 3$)

Entropy of Total rooms:

$$= 0.241\left(\frac{9}{12}\right) + 0\left(\frac{3}{12}\right)$$

Name: Nikhila
netid: nib22003

Entropy of total rooms

$= 0.18$

Information gain $= 0.3 - 0.18$

$\boxed{\text{Information gain} = 0.12}$

Entropy of median_income variable:-

For low income (i.e = 0):-

	median_income	median house_value
1	0	1
3	0	0
4	0	0
5	0	0
6	0	1
8	0	0
9	0	0

Entropy of low income =

$= -\left(\frac{5}{7}\right) * \log_2\left(\frac{5}{7}\right) - \left(\frac{2}{7}\right) \log_2\left(\frac{2}{7}\right)$

$= -(0.71) * (-0.14) - (0.28) * (-0.54)$

$= 0.\cancel{85}10 + 0.\cancel{82}1512$

$\boxed{\text{Entropy} = \cancel{1.00}\ 0.25}$

Name: Nikhila
netid: nib22003

For high income (ie = 1):-

	median_income	median house-value
0	1	1
2	1	0
7	1	1
10	1	1
11	1	1

Entropy of high income =

$= -\left(\frac{4}{5}\right) * \log_2\left(\frac{4}{5}\right) - \left(\frac{1}{5}\right) * \log_2\left(\frac{1}{5}\right)$

$= -0.8 * (-0.09) - (0.2) * (-0.6)$

$= 0.072 + 0.12$

$= 0.192$

Entropy of median-income:-

more-income ($\epsilon = 0.192, n = 5$) less-income ($\epsilon = 1.67, n = 7$)

$= 0.192\left(\frac{5}{12}\right) + \cancel{1.00}\ 0.25\left(\frac{7}{12}\right)$

$= 0.08 + \cancel{0.97}\ 0.\cancel{5}814$

$= \cancel{1.05}\ 0.\cancel{99}$

Name: Nilchila
netid: nib22003.

Information gain = 0.3 − 0.22

$\boxed{\text{Information gain} = 0.08}$

Summary :-

house-median-age = 0.09

total-rooms = 0.12

median-income = 0.08

Since, ~~total~~-rooms is having the high information gain, We can choose the total rooms as the first split.

Decision-tree :- First split starts from total rooms to calculate median-house value.

total-rooms ≤ median[total-rooms]
entropy = 0.3
samples = 12
value = [9, 3]

False

entropy = 0.241
samples = 9
value = [3, 6]

True

entropy = 0
samples = 3
value = [3, 0]