22 Sept 2021 Data Analytics

- Finish Decision Tree lecture
- In Class Worksheet
- Discussion on interval notation used in the HW
- Discussion on using scikit-learn in the HW

In class example:

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Parent node == entire data set

Gini Index $I(t) = 1 - \sum_i p_i^2$

$$1 - \left[\frac{7}{10}\right]^2 - \left[\frac{3}{10}\right]^2 = \underline{0.42}$$

If H.O == yes

$$1 - \left[\frac{3}{3}\right]^2 - \left[\frac{0}{3}\right]^2 = 0$$

If H.O == No

$$1 - \left[\frac{4}{7}\right]^2 - \left[\frac{3}{7}\right]^2 = 1 - \left[\left(\frac{4}{7}\right)^2 + \left(\frac{3}{7}\right)^2\right]$$

$$= 0.49$$

Now Gain is

$$0.42 - \frac{3}{10}(0) - \frac{7}{10}(0.49)$$

$$= 0.077$$