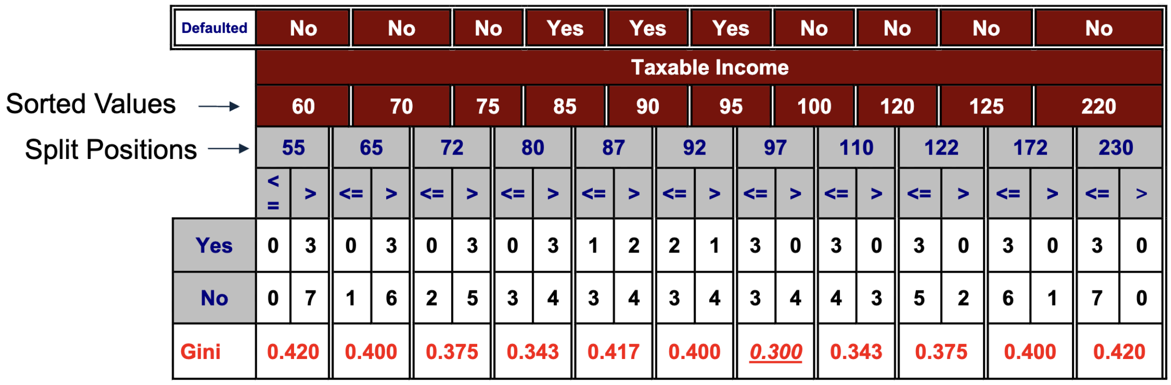
Decision Trees

Splitting of Continuous Attributes

- We will use the Gini Index to select the best splitting of a continuous attribute.
- Take the attribute, Taxable Income for example
- Use binary discretization such that it splits objects into two subsets
- Step 1: Choose a splitting point for discretization
  - Number of possible splitting points = Number of distinct values
- Step 2: Compute the Gini of the split given by each discretization
- Try each possible option
- Consider the data:

TID	Refund	Marital Status	Taxable Income	Defaulted
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

- A method for choosing splitting point
  - Sort the attribute on values
  - Take the midpoints between two adjacent sorted values as the candidate splitting point
  - Choose the split position that has the least Gini when split



Defaulted	No	No	No	Yes	Yes	Yes	No	No	No	No
	Taxable Income									
Sorted Values →	60	70	75	85	90	95	100	120	125	220

Split Positions →	55		65		72		80		87		92		97		110		122		172		230	
	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Yes	0	3	0	3	0	3	0	3	1	2	2	1	3	0	3	0	3	0	3	0	3	0
No	0	7	1	6	2	5	3	4	3	4	3	4	3	4	4	3	5	2	6	1	7	0
Gini	0.420		0.400		0.375		0.343		0.417		0.400		*0.300*		0.343		0.375		0.400		0.420	

Continuous Attributes: Exercise

- What is the best way to split on Attribute 3 on the dataset:

TID	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55k	Yes
12	Yes	Medium	80k	No
13	Yes	Large	110k	No
14	No	Small	95k	Yes
15	No	Large	67k	Yes

- Sort the attribute on values
- Take the midpoints between two adjacent sorted values as the candidate splitting point
- Choose the split position that has the least Gini when split

Yes	Yes	No	Yes	No
55k	67k	80k	95k	110k

Split	61k		73k		87k		102k	
	≤	>	≤	>	≤	>	≤	>
Yes	1	2	2	1	2	1	3	0
No	0	2	0	2	1	1	1	1
Gini	0.400		0.267		0.466		0.300	

- We would choose to split on 73k (or anything between 80k and 90k) since it has the smallest Gini Index

Tree Induction / Creating the Tree

- Greedy Strategy
  - Split the records based on an attribute test that optimizes certain criterion immediately
- Things We Need to Know
  - How to split the records
    - How to specify the attribute test condition
    - How to determine the best split
  - Determine when to stop splitting