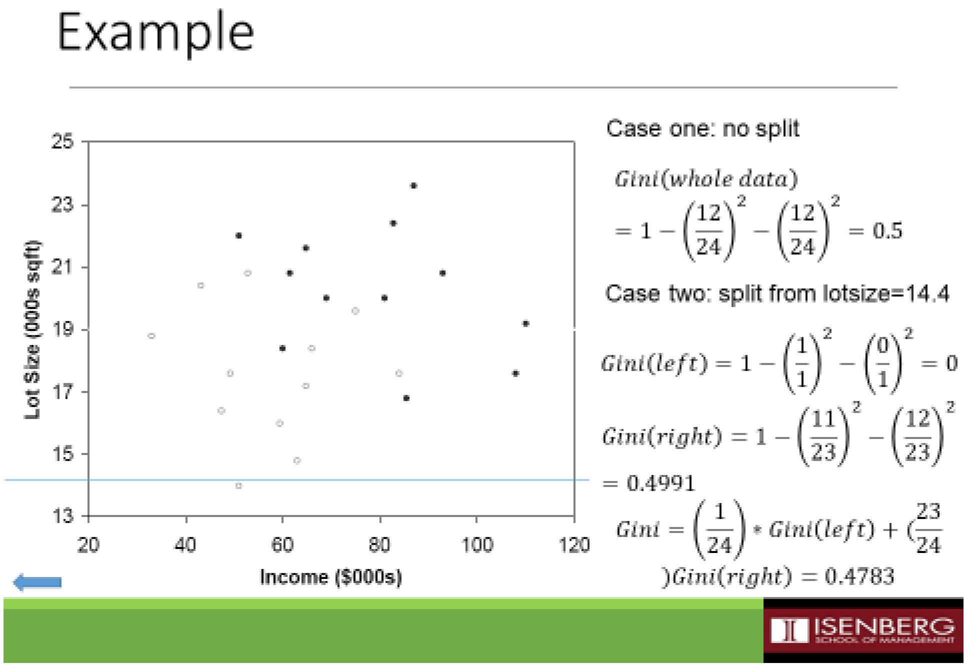
Gini index - Project

We are trying to predict whether a customer would make a claim using the following small dataset.

Gender	Driving License Years	Car category	Claim
Woman	21	Sedan	0
Woman	3	SUV	0
Woman	19	SUV	0
Woman	11	Sedan	0
Woman	10	Sport	1
Man	9	Sedan	1
Man	12	SUV	1
Woman	15	SUV	0

1. Calculate the Gini index of the full dataset.
   1-(3/8)^2-(5/8)^2=15/32=0.46



Example

Case one: no split

$$Gini(whole\ data) = 1 - \left(\frac{12}{24}\right)^2 - \left(\frac{12}{24}\right)^2 = 0.5$$

Case two: split from lotsize=14.4

$$Gini(left) = 1 - \left(\frac{1}{1}\right)^2 - \left(\frac{0}{1}\right)^2 = 0$$

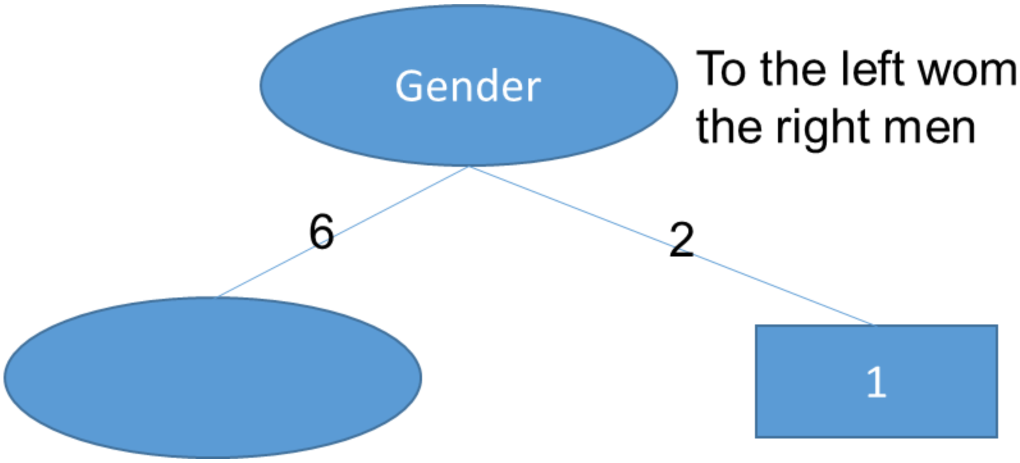
$$Gini(right) = 1 - \left(\frac{11}{23}\right)^2 - \left(\frac{12}{23}\right)^2 = 0.4991$$

$$Gini = \left(\frac{1}{24}\right) * Gini(left) + (\frac{23}{24})Gini(right) = 0.4783$$

2. To find the first split node, corresponding Gini index. And please draw the current tree.

driving license	gini left	gini right	gini	
6	0	24/49	3/7	0.43
9.5	0.5	4/9	11/24	0.46
10.5	4/9	8/25	11/30	0.37

11.5	0.5	3/8	7/16	0.44
13.5	12/25	0	0.3	0.3
17	0.5	0	3/8	0.375
19	24/49	0	3/7	0.43
	gini woman	gini man	gini	
gender	5/18	0	5/24	0.21
	gini sedan	gini others	gini	
car|sedan	4/9	12/25	7/15	0.47
	gini suv	gini others	gini	
car|suv	3/8	0.5	7/16	0.44
	gini sport	gini others		
car|sport	0	24/49	3/7	0.43



To the left women, to the right men

3. Now we find the first node, please list the dataset will be used on the left after splitting and dataset on the right after splitting.

Left

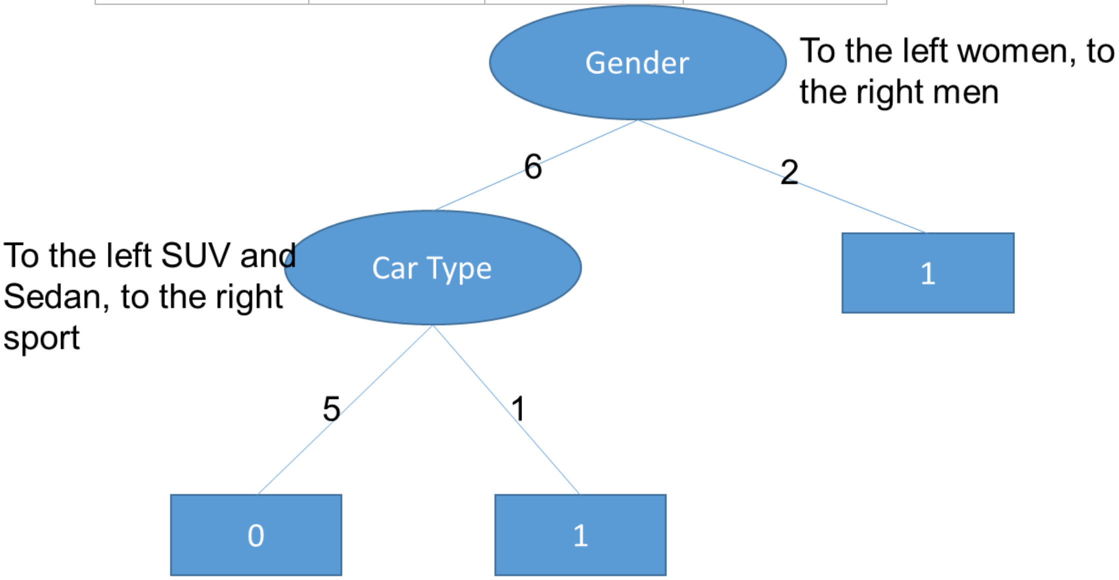
Gender	**Driving License Years**	**Car category**	**Claim**
Woman	3	SUV	0
Woman	10	Sport	1
Woman	11	Sedan	0
Woman	15	SUV	0
Woman	19	SUV	0
Woman	21	Sedan	0

Right

Gender	**Driving License Years**	**Car category**	**Claim**
Man	9	Sedan	1
Man	12	SUV	1

4. Now we trying to find further nodes. Based on the dataset on the left after splitting, please find the combinations of splitting variables and values, and calculate their corresponding Gini index like what I did in part 2. Based on the results, which combination of variable and splitting value could be the next node on the left? And please draw the current tree.

driving license	gini left	gini right	gini
6	0	0.32	0.27
10.5	0.5	0	0.17
13	0.44	0	0.22
17	0.375	0	0.25
20	0.32	0	0.27
	gini sedan	gini others	gini
car|sedan	0	0.375	0.25
	gini suv	gini others	Gini
car|suv	0	0.44	0.22
	gini sport	gini others	
car|sport	0	0	0



To the left women, to the right men

To the left SUV and Sedan, to the right sport

5. Similarly, based on the dataset on the right after splitting, please find the combinations of splitting variables and values, and calculate their corresponding Gini index like what I did in part 2. Based on the results, which combination of variable and splitting value could be the next node on the left? And please draw the current tree.
All the same classification, no need to split