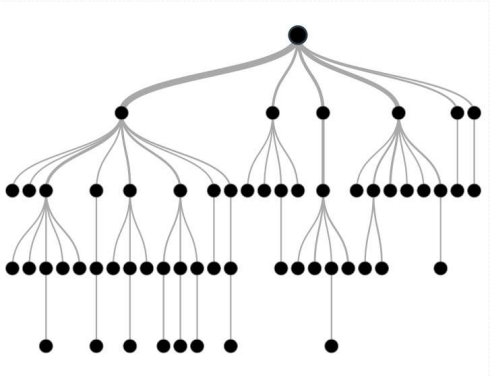
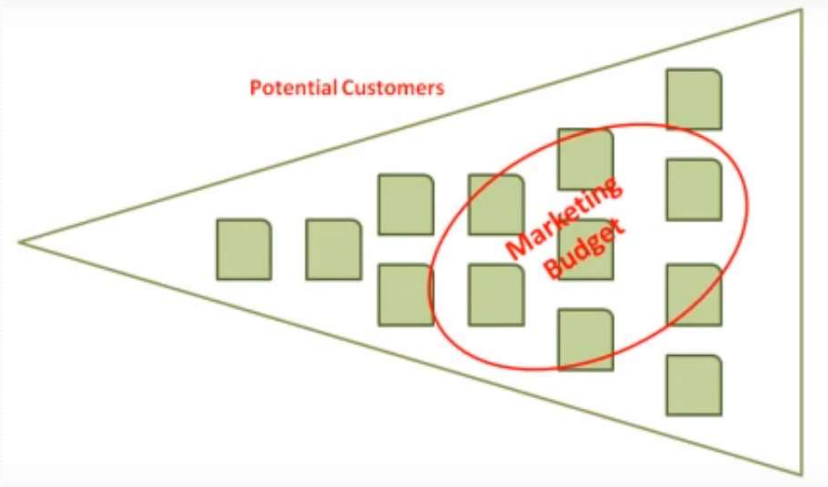
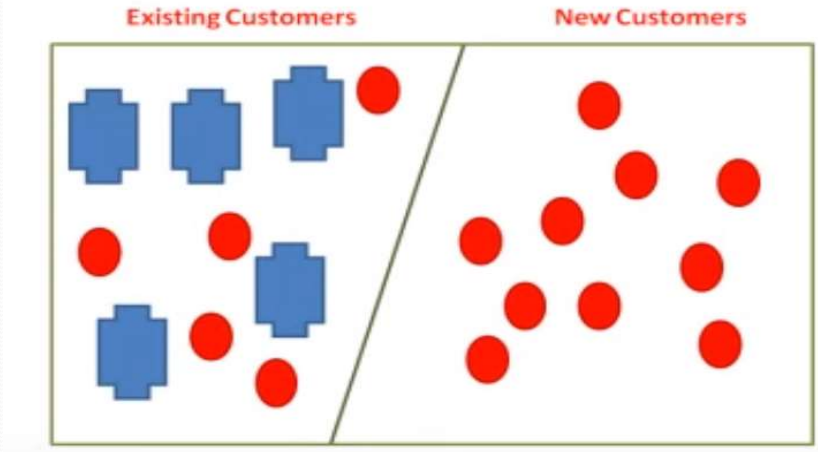
Decision Trees

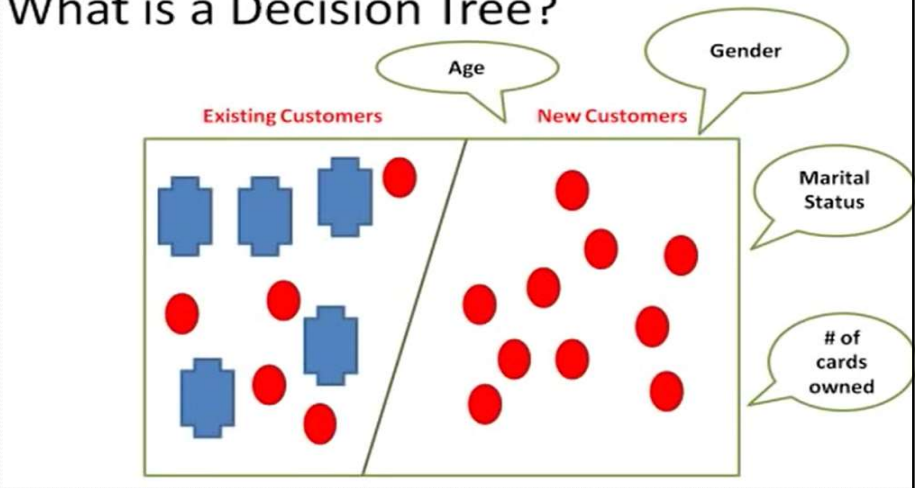
Nonlinear Classification and Regression with Decision Trees

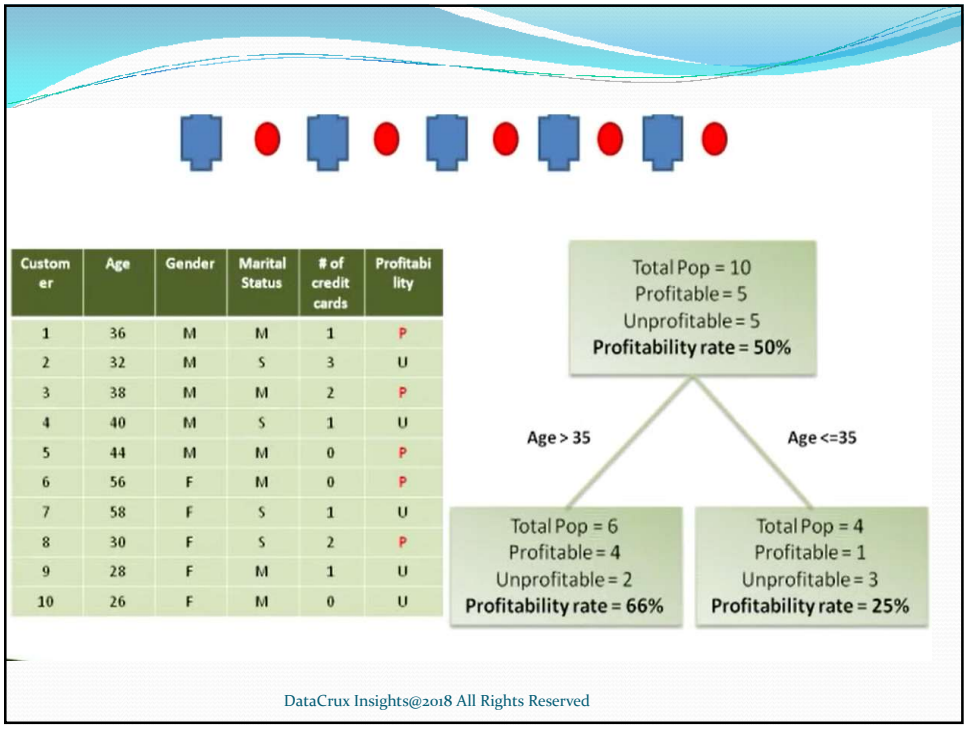




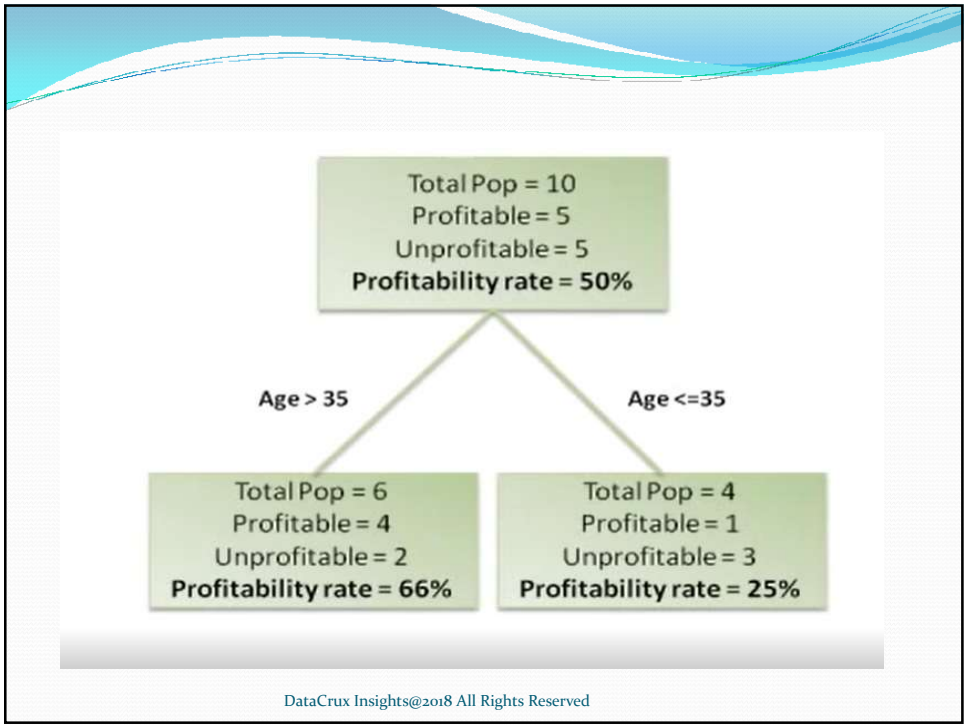


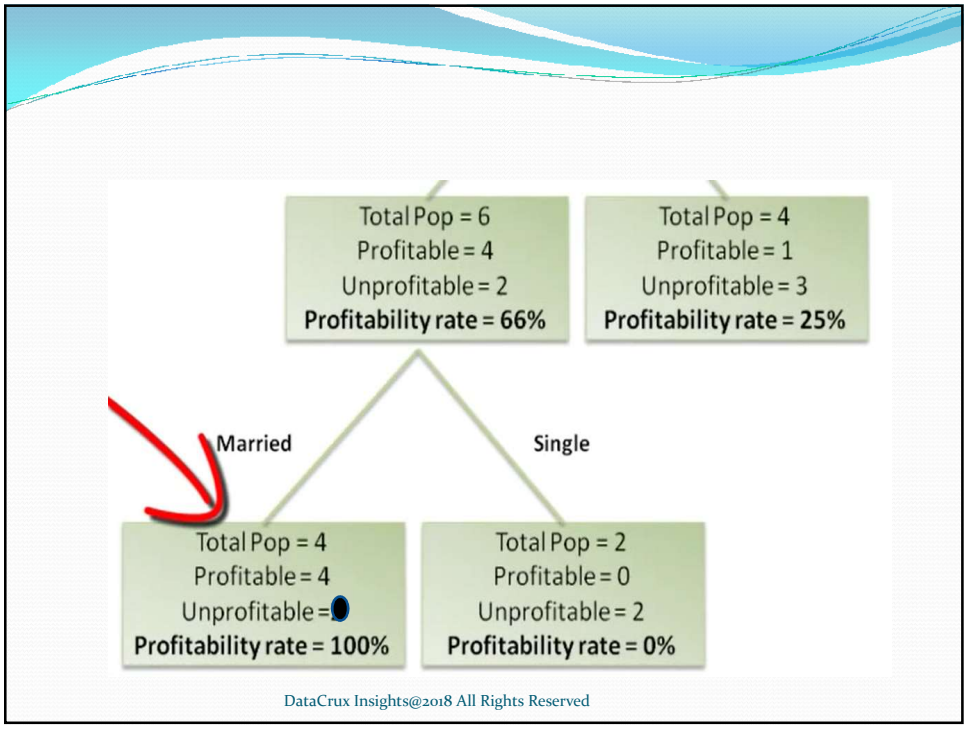
What is a Decision Tree?

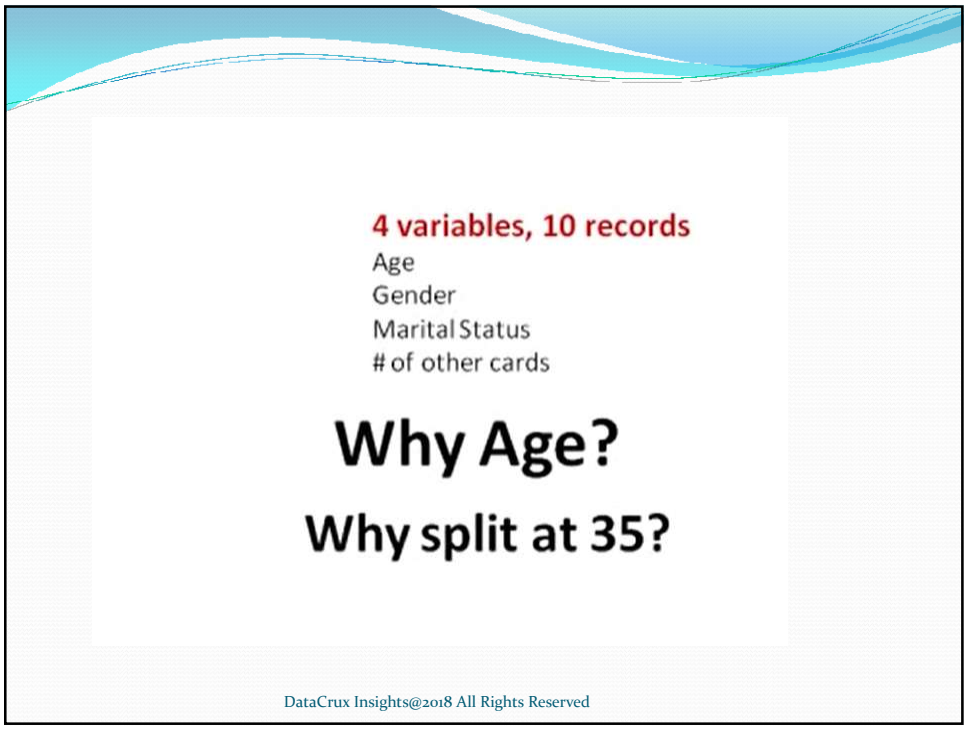




Customer	Age	Gender	Marital Status	# of credit cards	Profitability
1	36	M	M	1	P
2	32	M	S	3	U
3	38	M	M	2	P
4	40	M	S	1	U
5	44	M	M	0	P
6	56	F	M	0	P
7	58	F	S	1	U
8	30	F	S	2	P
9	28	F	M	1	U
10	26	F	M	0	U







Decision Trees

Gini

Entropy

Chi-square

Reduction of variance

Decision Tree Algorithm

Basic algorithm (a greedy algorithm)

- → Tree is constructed in a top-down recursive divide-and-conquer manner
- → At start, all the training examples are at the root
- → Attributes are categorical (if continuous-valued, they are discretized in advance)
- → Input data is partitioned recursively based on selected attributes
- → Test attributes at each node are selected on the basis of a heuristic or statistical measure (e.g., information gain)

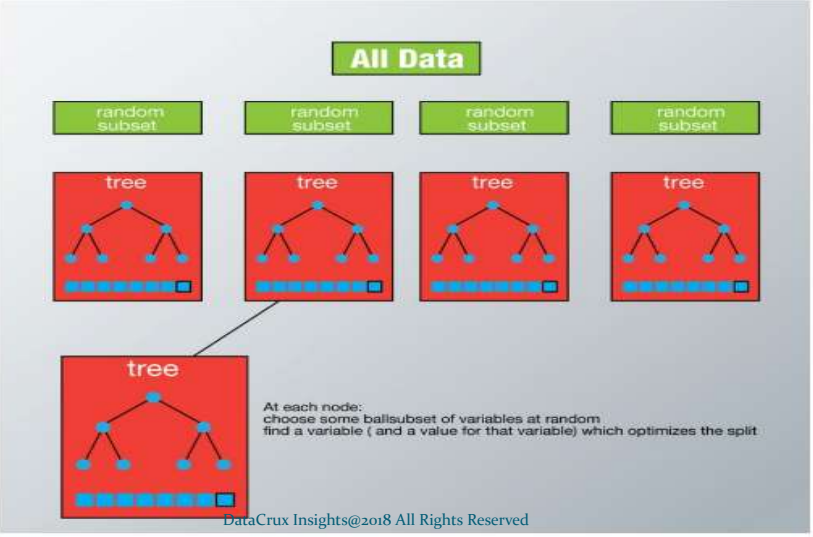
Conditions for stopping partitioning

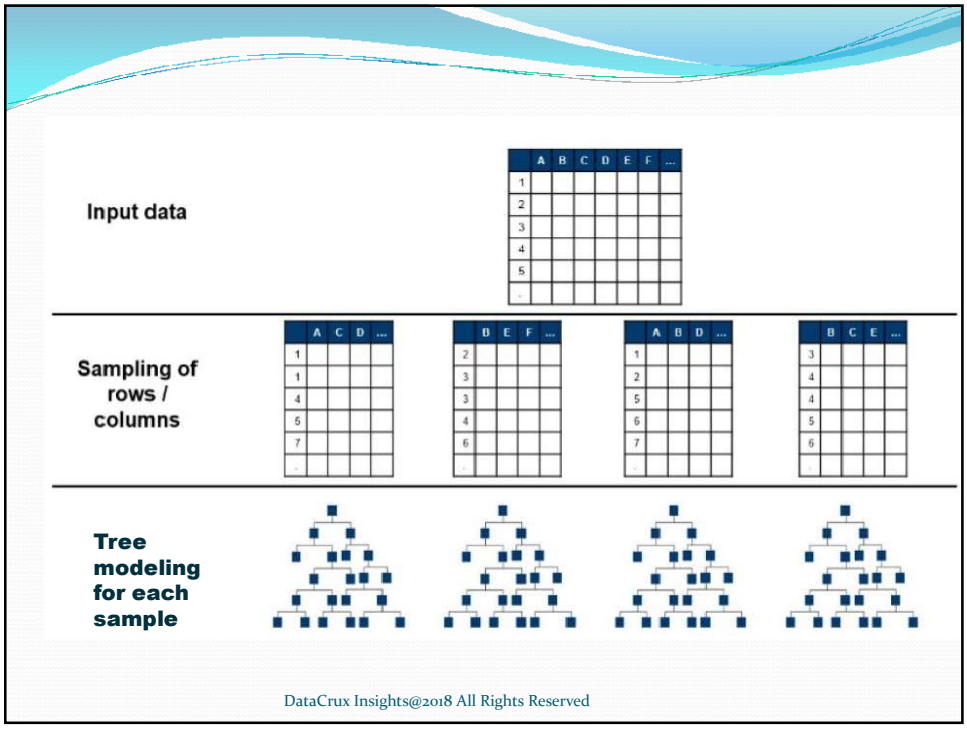
- → All samples for a given node belong to the same class
- → There are no remaining attributes for further partitioning – majority voting is employed for classifying the leaf
- → There are no samples left

Random split

- The tree can grow huge
- These trees are hard to understand.
- Larger trees are typically less accurate than smaller trees.

Random Forest





Bagging