C4.5 Decision Trees

Goals:

C4.5 decision tree algorithm

- What do we use it for
- What calculations are behind the decision making (simplified)
- What C4.5 handles that ID3 does not

Type of Problem: Classification

- Use C4.5 algorithm to predict the class in which a new instance belongs
- Supervised learning

What we need:

Training Data: things already classified

Continuous or discrete

Can have unknown values

Test Data

ID3: common decision tree algorithm cannot handle continuous data and missing values

Training Data

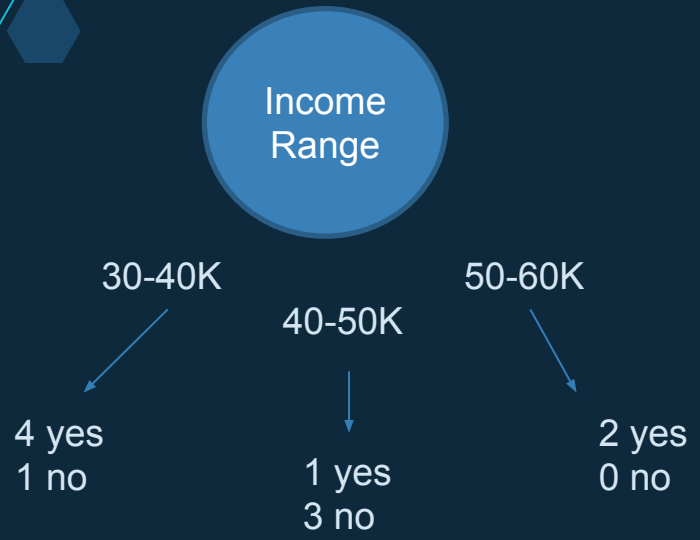
Question: which customers are most likely to take advantage of the new life insurance promotion based on past behavior

Income Range	Magazine Promo	Watch Promo	Life Ins. Promo	Credit Card Ins.	Sex	Age
40-50K	Y	N	N	N	M	45
30-40K	Y	Y	Y	N	F	40
40-50K	N	N	N	N	M	42
30-40K	Y	Y	Y	Y	M	43
50-60K	Y	N	Y	N	F	38

Check relevance

Output

For each branch being used, choose the most frequently occurring decision

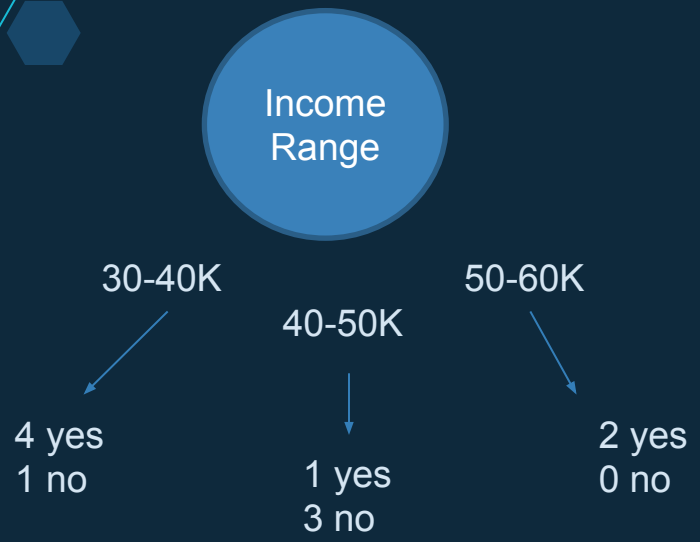


Cost: 3 branches

Accuracy: (4+3+2) / (5+4+2) = 9/11 = 82%

Goodness Score: 9/11/3 = .273

For each branch being used, choose the most frequently occurring decision



Cost: 3 branches

Accuracy: (4+3+2) / (5+4+2) = 9/11 = 82%

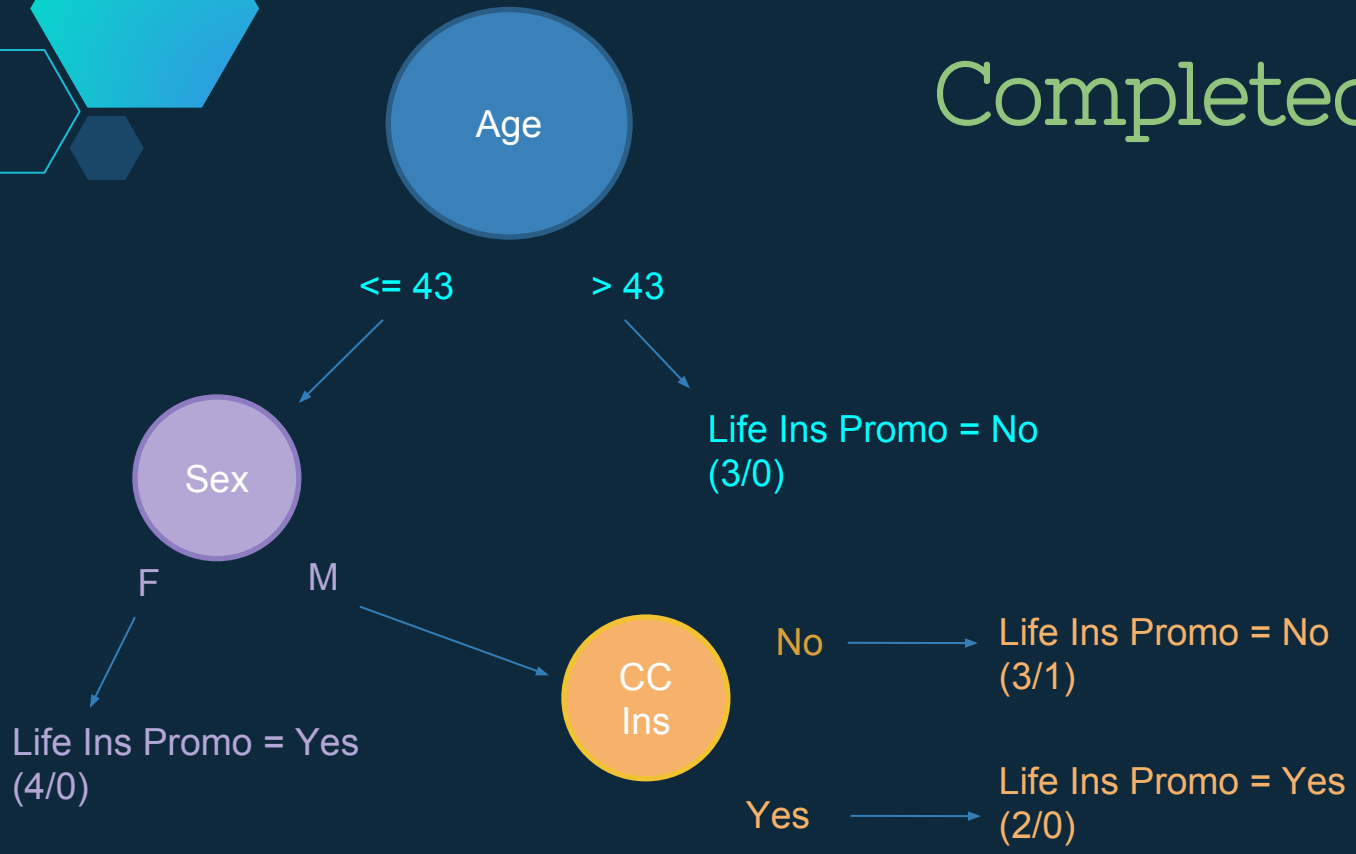
Goodness Score: 9/11/3 = .273

Repeat for all

	Cost	Accuracy	Goodness
CC ins	2 branches	60%	.3
Age	2 branches	80%	.4
Sex	2 branches	73%	.37

Consider: are there any subtrees we could terminate?

Completed Tree



Handling Continuous Data

- A binary split is performed

Handling Missing Values

- Missing data treated as separate class (“?”) and not used in calculation of gain

Pruning

- Reduce overfitting
- ID3 grows a tree until it makes no errors over the set of training data