

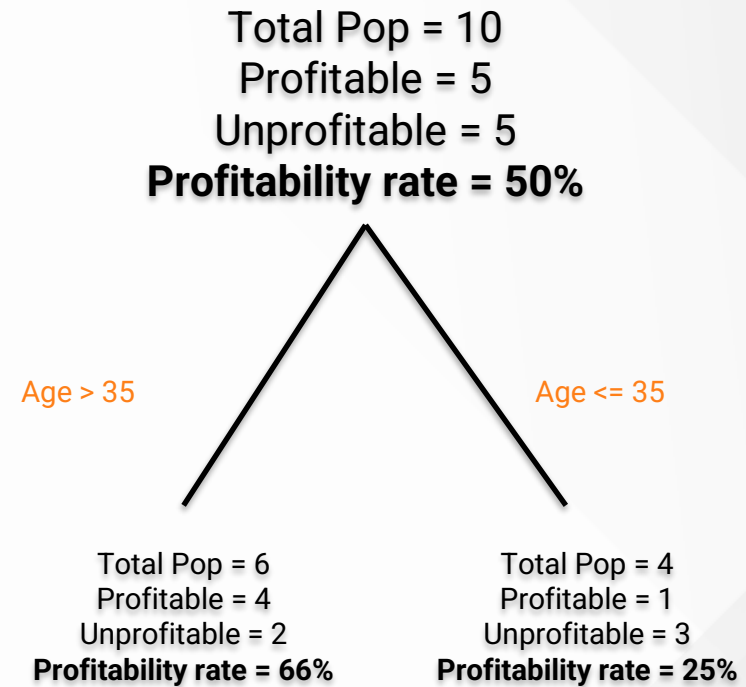
Decision Tree: Classifier, Purity Metrics

Decision Tree: Classifier, Purity Metrics

- How do we decide which variable to split on?
- Can we quantify this into a metric?
- Purity Metrics

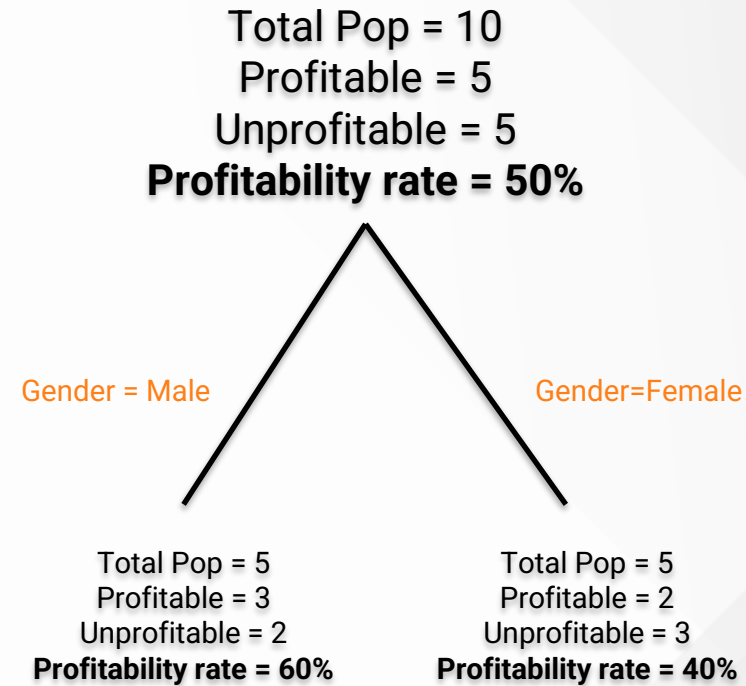
Decision Tree: Classifier, Purity Metrics

Customer	Age	Gender	Marital Status	# Cr. 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

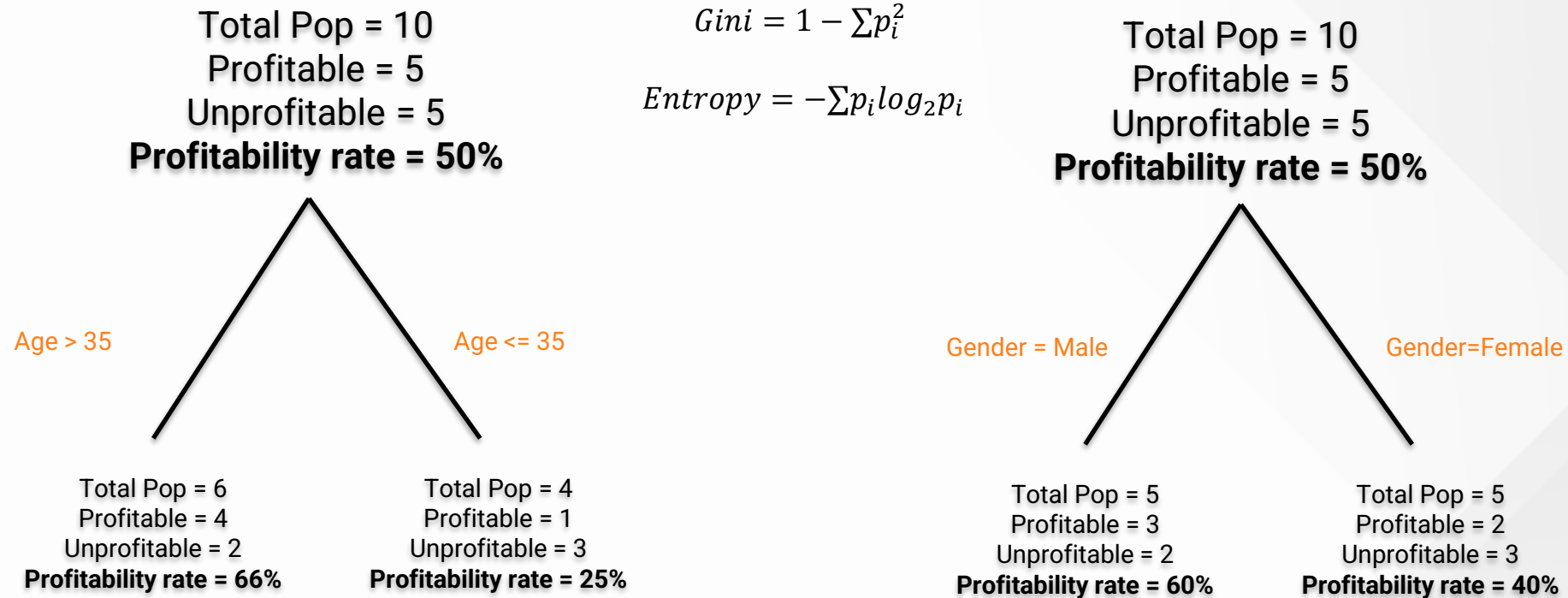


Decision Tree: Classifier, Purity Metrics

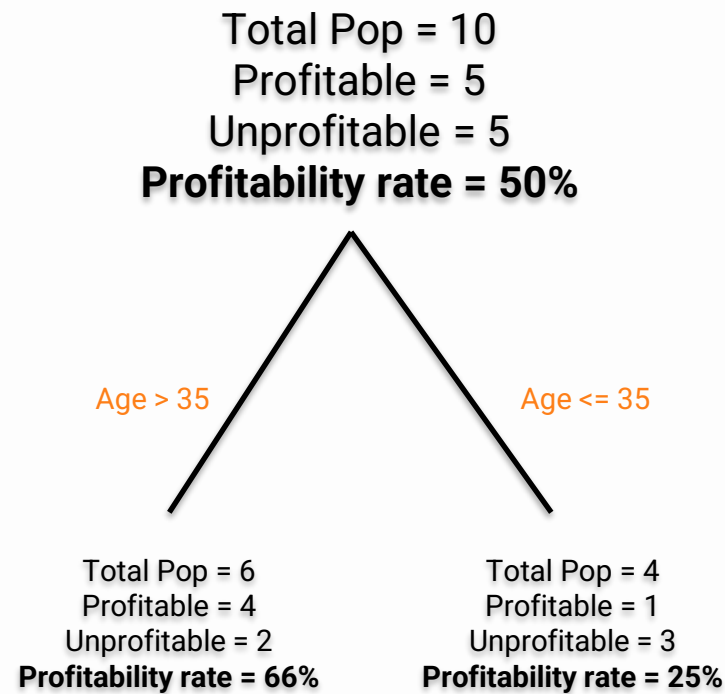
Customer	Age	Gender	Marital Status	# Cr. Cards	Profitability
1	36	M	M	1	P
2	32	M	S	3	U
3	38	M	M	2	P
4	40	M	S	1	U
5	44	M	M	0	P
6	56	F	M	0	P
7	58	F	S	1	U
8	30	F	S	2	P
9	28	F	M	1	U
10	26	F	M	0	U



Decision Tree: Classifier, Purity Metrics



Decision Tree: Classifier, Purity Metrics



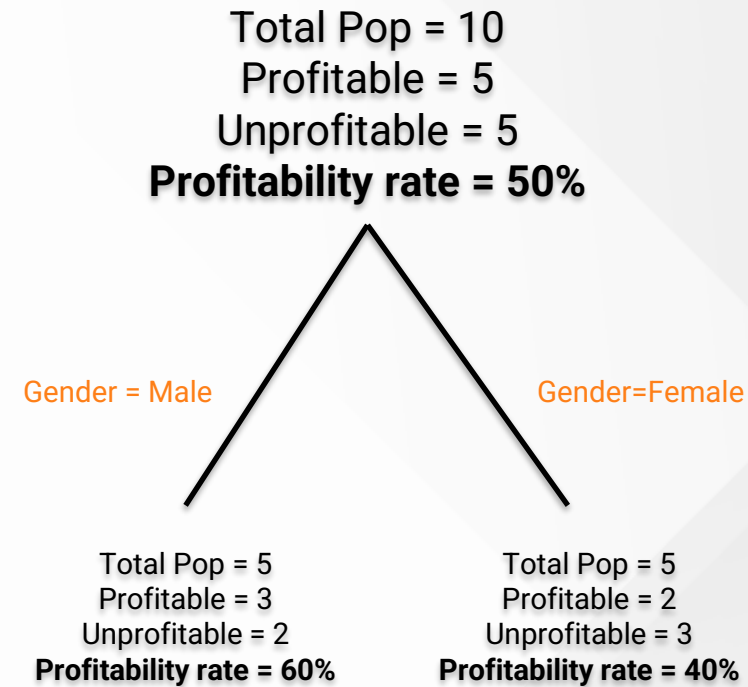
$$1 - \left[\left(\frac{4}{6} \right)^2 + \left(\frac{2}{6} \right)^2 \right]$$

0.44

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

0.375

$$Gini = 1 - \sum p_i^2$$



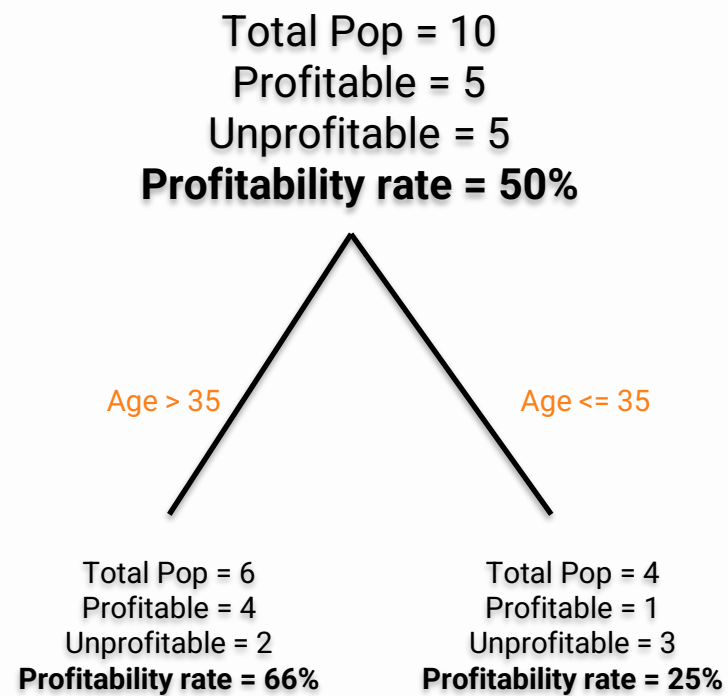
$$1 - \left[\left(\frac{3}{5} \right)^2 + \left(\frac{2}{5} \right)^2 \right]$$

0.48

$$1 - \left[\left(\frac{2}{5} \right)^2 + \left(\frac{3}{5} \right)^2 \right]$$

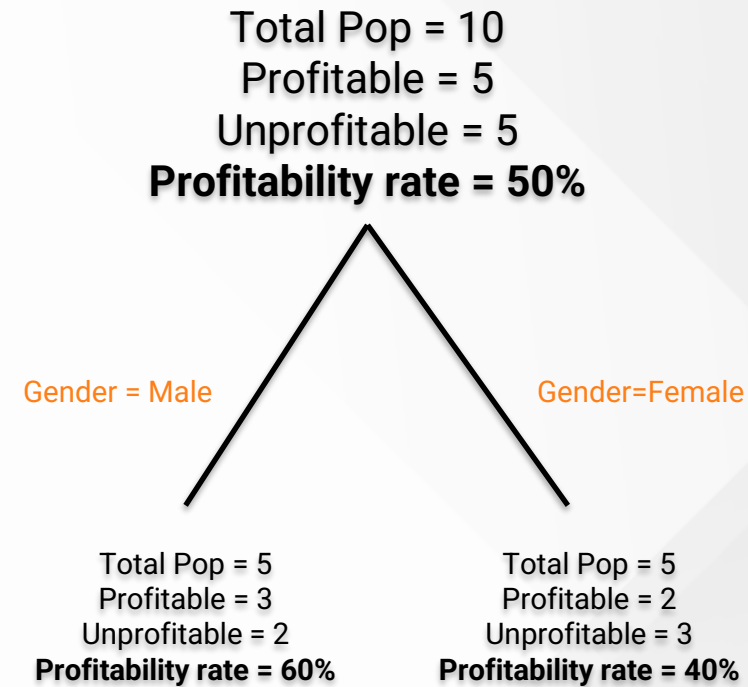
0.48

Decision Tree: Classifier, Purity Metrics



$$\left(\frac{6}{10}\right) * 0.44 + \left(\frac{4}{10}\right) * 0.375 = 0.41$$

$$Gini = 1 - \sum p_i^2$$



$$\left(\frac{5}{10}\right) * 0.48 + \left(\frac{5}{10}\right) * 0.48 = 0.48$$

Decision Tree: Classifier, Purity Metrics

$$Entropy = -\sum p_i \log_2 p_i$$

Total Pop = 10
Profitable = 5
Unprofitable = 5
Profitability rate = 50%

Age > 35

Age <= 35

Total Pop = 6
Profitable = 4
Unprofitable = 2

Profitability rate = 66%

$$-\left[\left(\frac{4}{6}\right) * \log_2 \left(\frac{4}{6}\right) + \left(\frac{2}{6}\right) * \log_2 \left(\frac{2}{6}\right)\right]$$

0.91

Total Pop = 4
Profitable = 1
Unprofitable = 3

Profitability rate = 25%

$$-\left[\left(\frac{1}{4}\right) * \log_2 \left(\frac{1}{4}\right) + \left(\frac{3}{4}\right) * \log_2 \left(\frac{3}{4}\right)\right]$$

0.81

Total Pop = 10
Profitable = 5
Unprofitable = 5
Profitability rate = 50%

Gender = Male

Gender=Female

Total Pop = 5
Profitable = 3
Unprofitable = 2

Profitability rate = 60%

$$-\left[\left(\frac{3}{5}\right) * \log_2 \left(\frac{3}{5}\right) + \left(\frac{2}{5}\right) * \log_2 \left(\frac{2}{5}\right)\right]$$

0.97

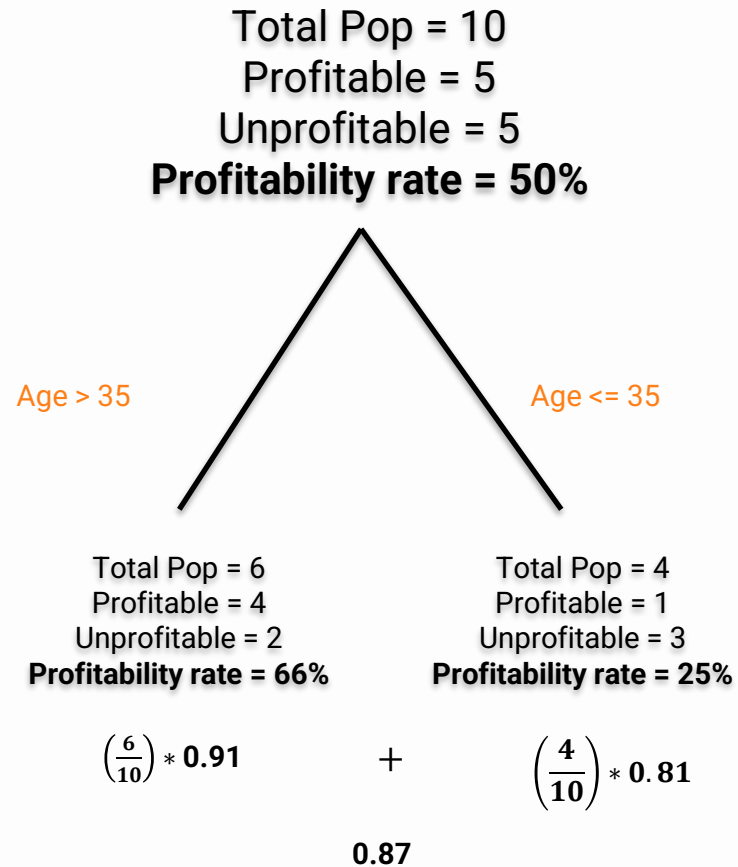
Total Pop = 5
Profitable = 2
Unprofitable = 3

Profitability rate = 40%

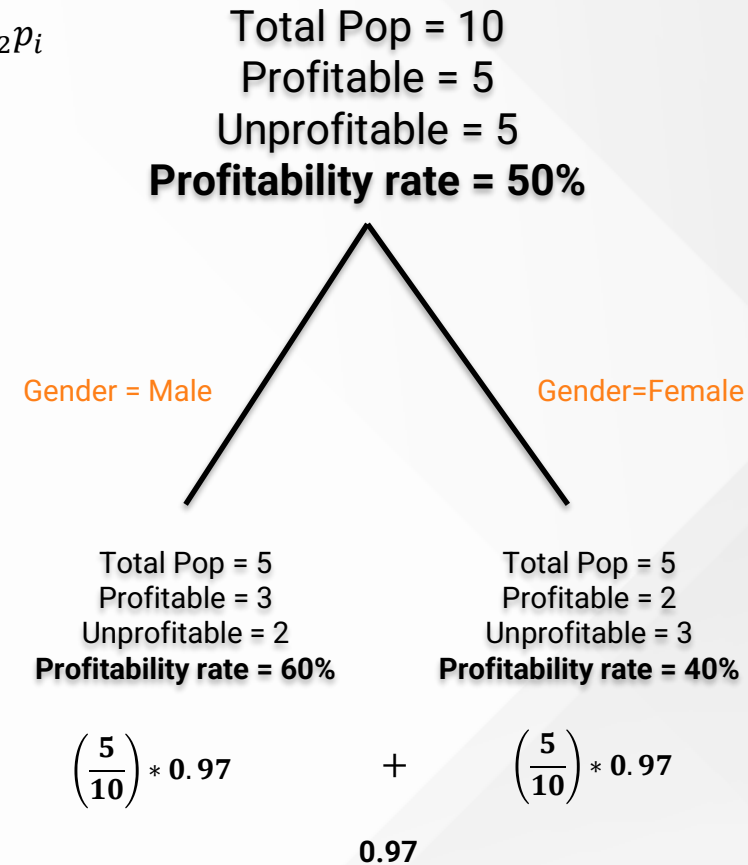
$$-\left[\left(\frac{2}{5}\right) * \log_2 \left(\frac{2}{5}\right) + \left(\frac{3}{5}\right) * \log_2 \left(\frac{3}{5}\right)\right]$$

0.97

Decision Tree: Classifier, Purity Metrics



$$Entropy = -\sum p_i \log_2 p_i$$



Decision Tree Algorithm: Overview

Algorithm can be summarised as:

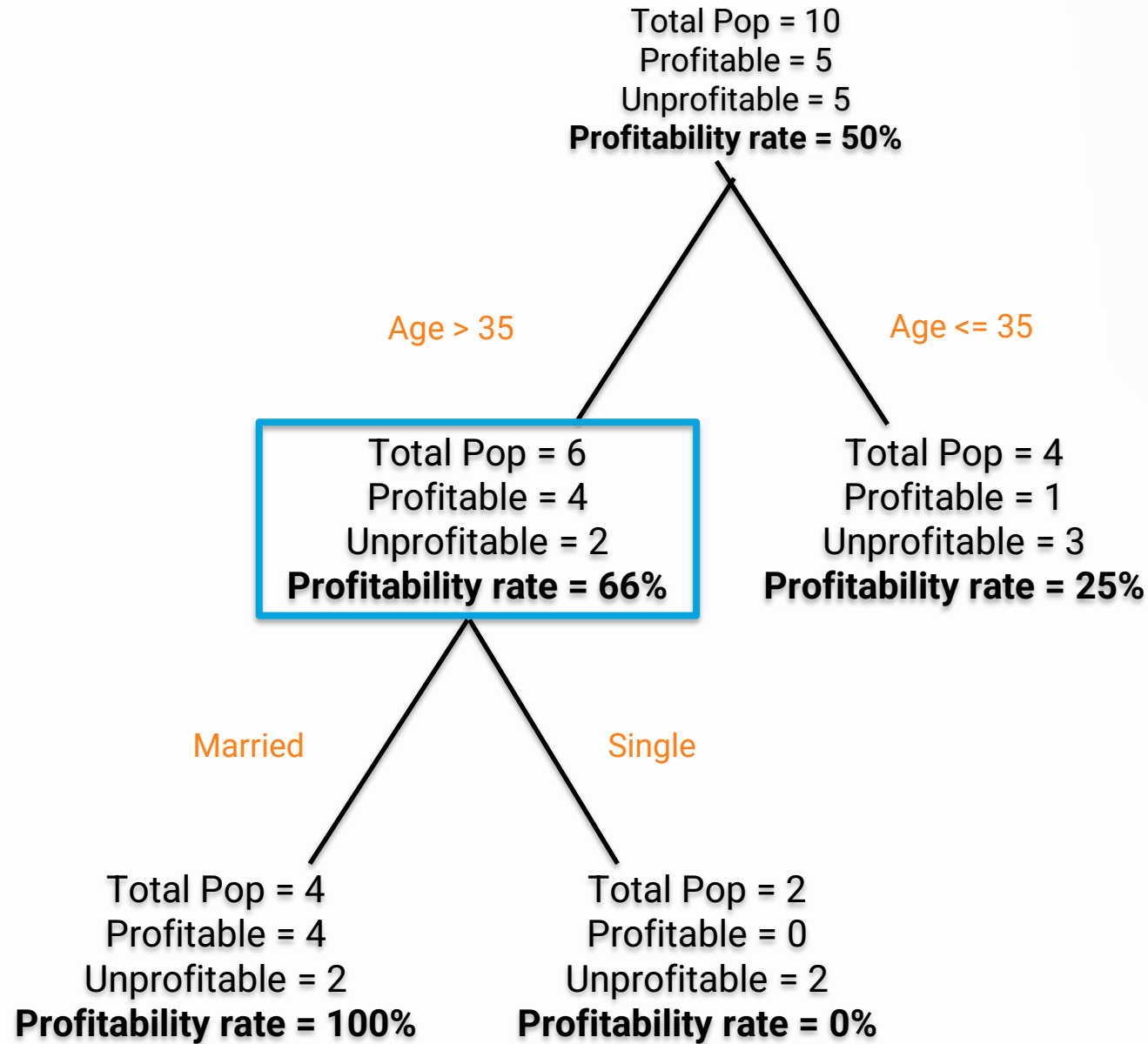
- For each split:
 - Compute purity metric (Gini or Entropy) for each variable
 - Choose the variable which results in lowest value of purity metric
- Continue doing 1 until some stopping criteria is met

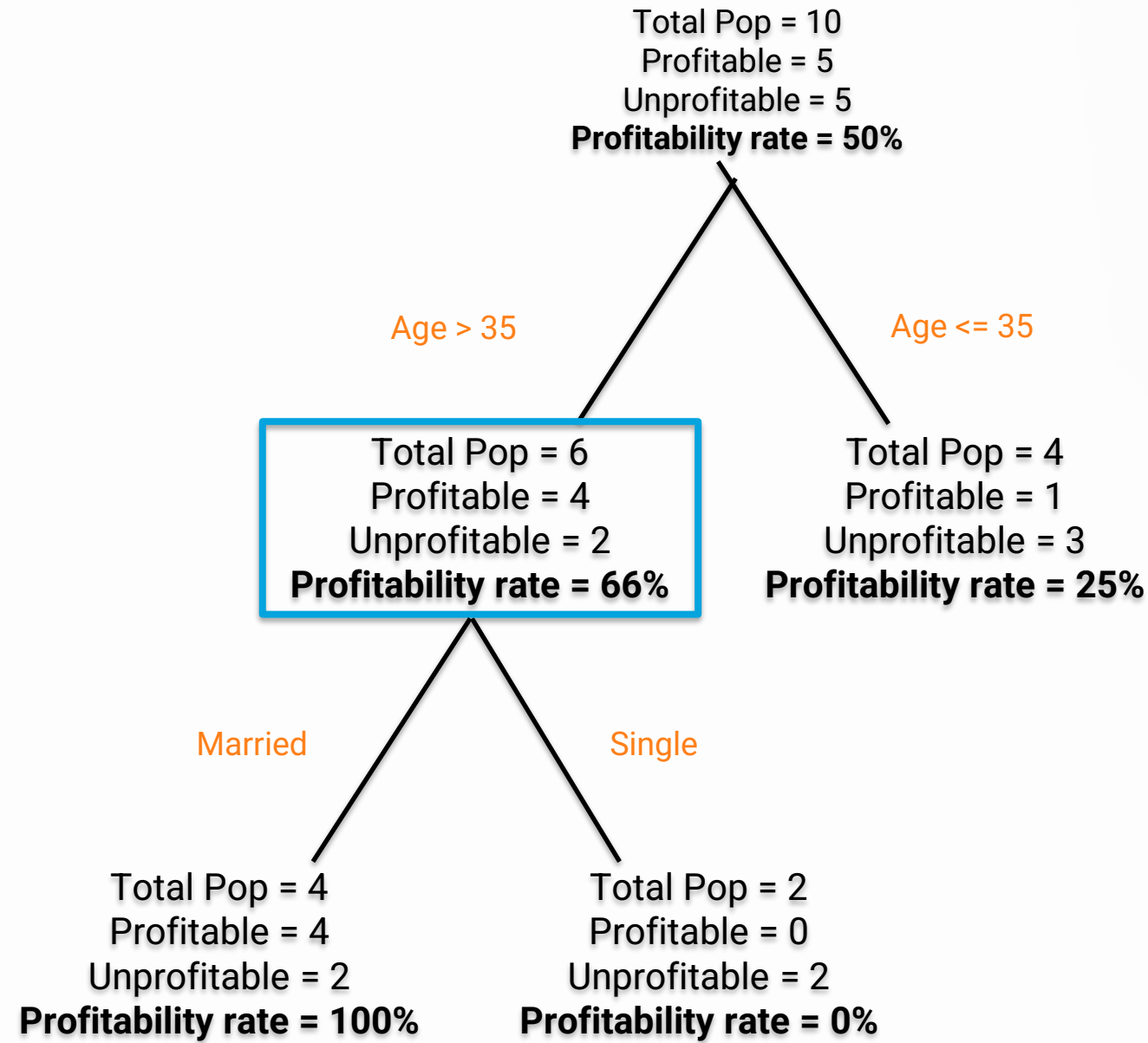
Commonly used stopping criteria are:

- Depth of tree
- Improvement in purity metric
- Number of observations in terminal nodes

In Video Question:

A person is 20 years old, what will the model predict?





A person is 20 years old

Prediction is 25% chance he will be profitable

Decision Tree: Classifier Performance Metrics

- A decision tree classifier can output probabilities.
- Hence, one can use confusion matrices, ROC curves, and AUC.
- For multiclass problems, usually, people use accuracy as a performance measure.

Decision Tree: Classifier Hyperparameters

- What are the parameters of a decision tree model?
- What inputs does a user have to specify while building a tree model?
- Gini/Entropy?
- Depth of tree: Should it be 1 level deep, 2 levels deep?
- These parameters are obtained by: Cross Validation.
- Usually, we want trees with less rules.

Decision Tree: Classifier

Code Demo



Thank You!

© 2023 Hero Private Limited. All rights reserved. This session is the proprietary of Hero Vired and/or its licensor. Your use/access or download shall be governed as per our IPR Policy.