Introduction to Machine Learning

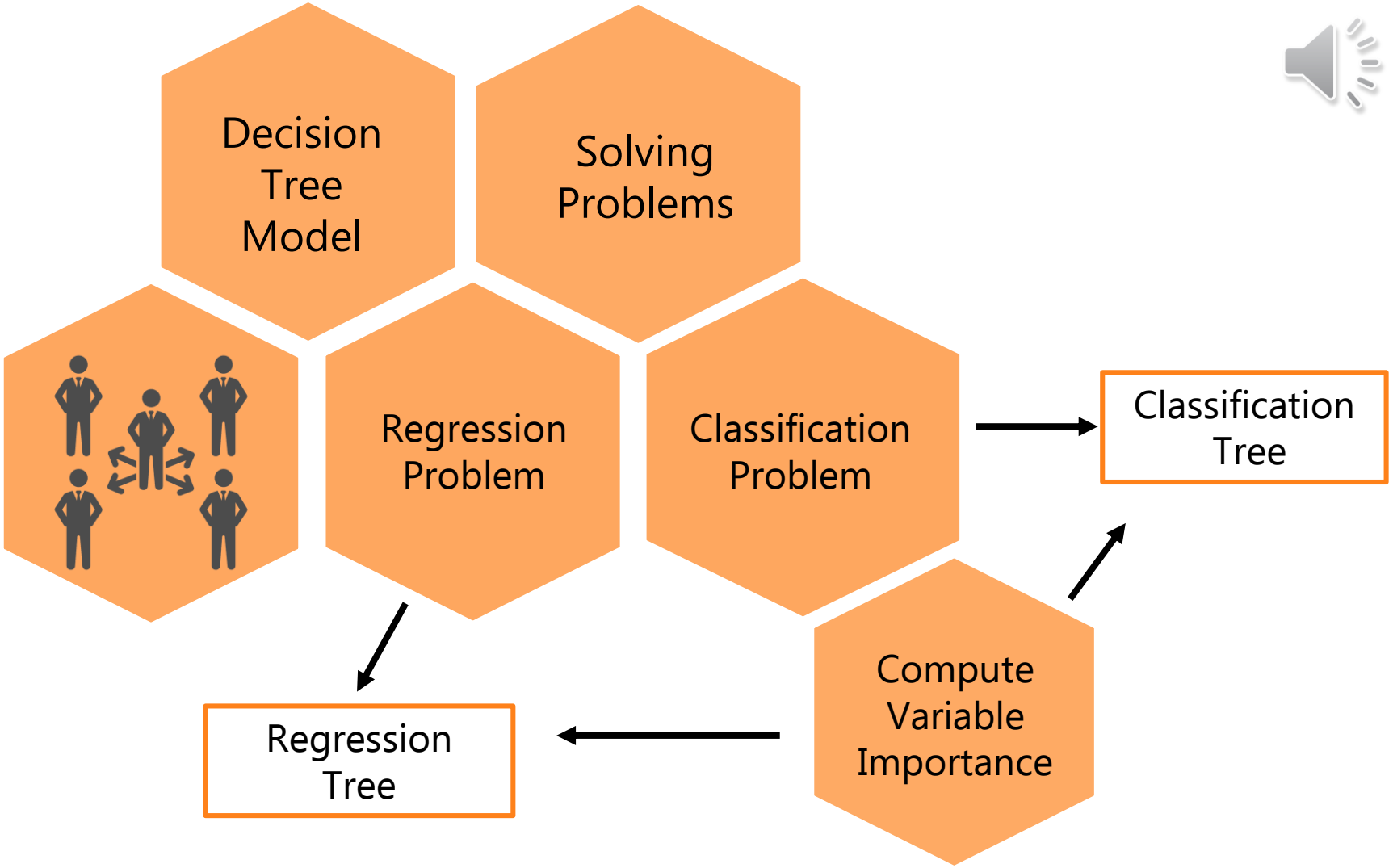
Class

Tree Based Models

Topic

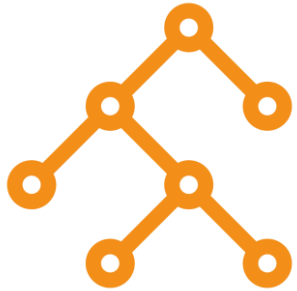
Introduction to Classification Trees

Agenda



Decision Tree: Overview

Solve both regression and classification problems



Decision Tree works is based on a branch of computer science known as **Information Theory**

The classic use case of decision trees is analysis of segments in business data

Decision Tree

Existing Data of a Bank

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

Profitable

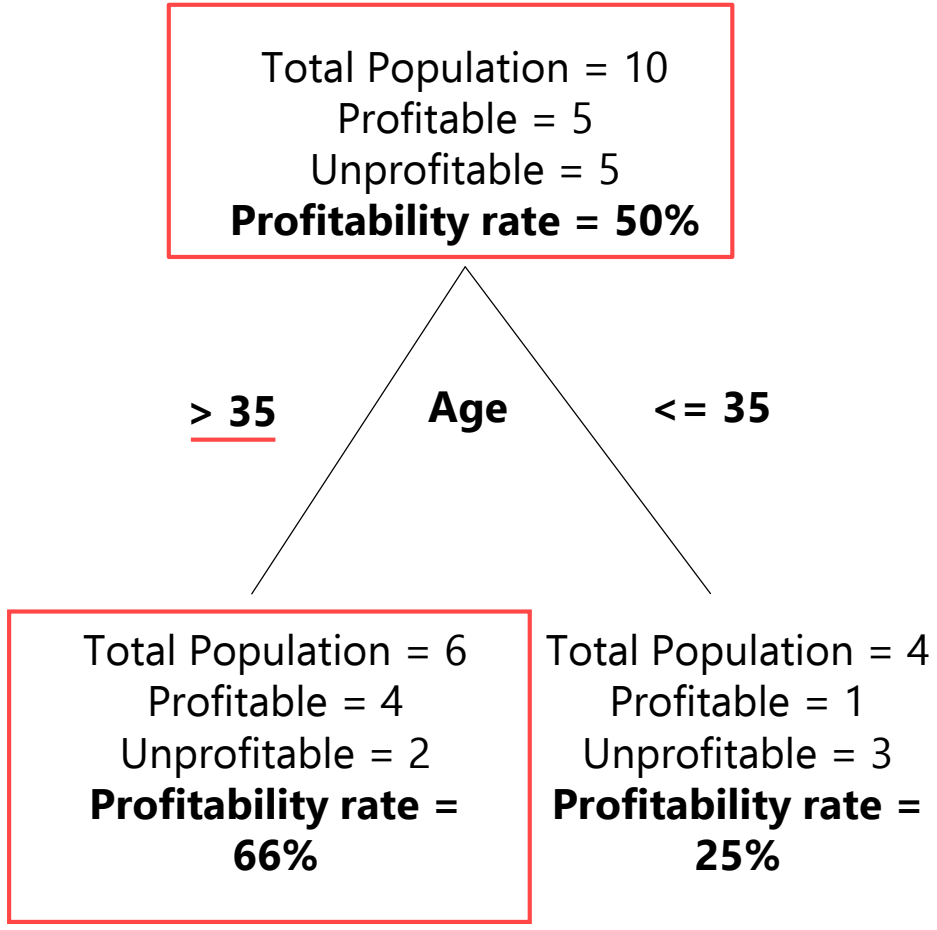
Unprofitable

To build a predictive model classifying customers logistic, Regression Classifier can be used

Decision Tree

Existing Data of a Bank

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

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

> 35 **Age** **<= 35**

Total Population = 6
Profitable = 4
Unprofitable = 2
Profitability rate = 66%

Total Population = 4
Profitable = 1
Unprofitable = 3
Profitability rate = 25%

The segment of data which is >35 has a higher chance of seeing a profitable customer

Decision Tree

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

> 35

Age

<= 35

Total Population = 6
Profitable = 4
Unprofitable = 2
Profitability rate = 66%

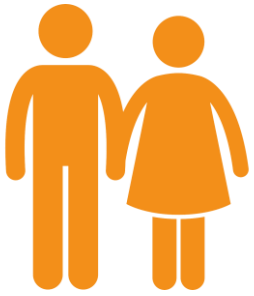
Total Population = 4
Profitable = 1
Unprofitable = 3
Profitability rate = 25%

Profitable

Married

Marital Status

Single

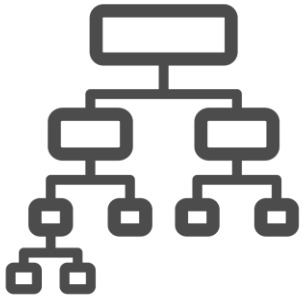


Total Population = 4
Profitable = 4
Unprofitable = 2
Profitability rate = 100%

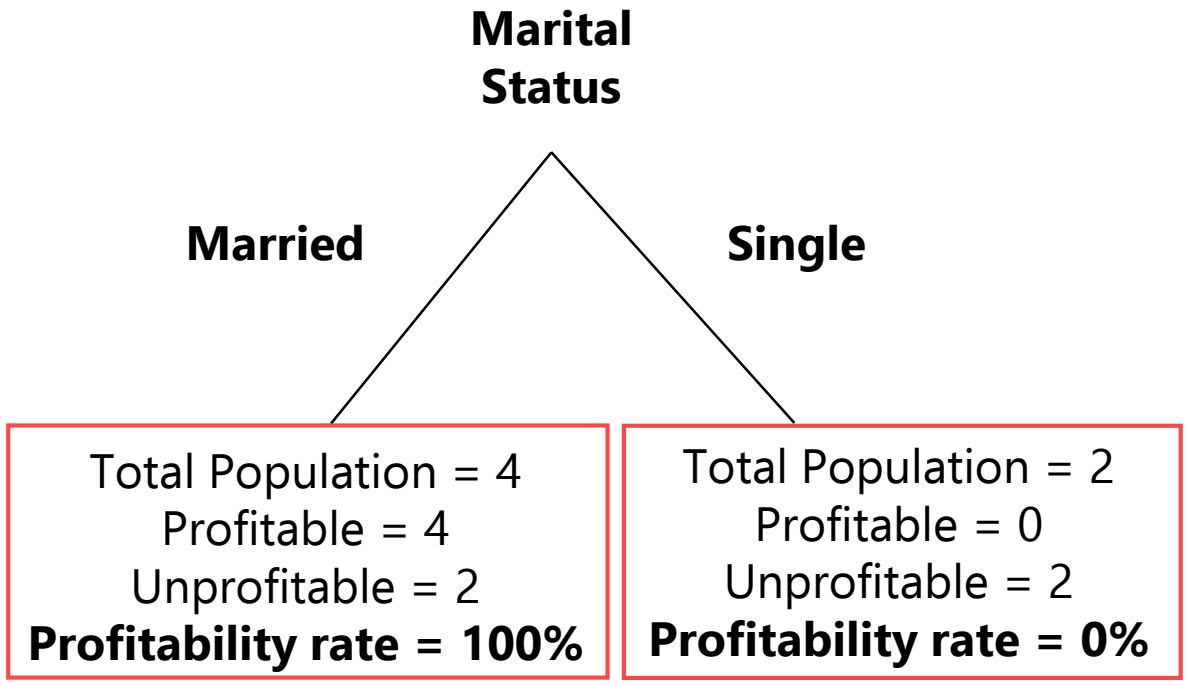
Total Population = 2
Profitable = 0
Unprofitable = 2
Profitability rate = 0%

Decision Tree

Decision tree classifier - Recursively sub-setting data can reveal interesting patterns

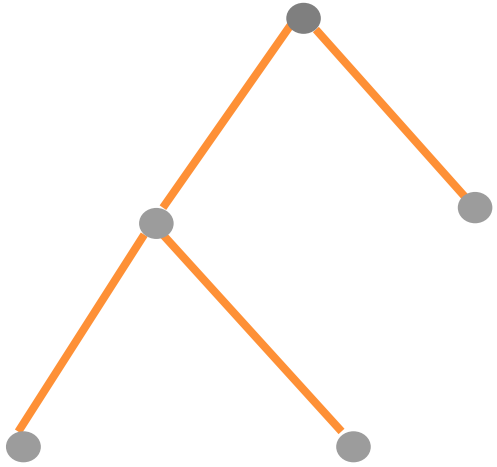


Data needs to be split in such a way so that the subsets of data end up being dominated by one class of the target variable



Decision Tree

Decision Tree splits into 2 parts at each node



Most implementations of a decision trees produce binary splits

Binary Tree

Decision Tree: Algorithm

How to decide which variable should be used to create splits?

Understand the intuition behind creating splits

The intuition will be formalized by introducing purity metrics

Decision Tree: Algorithm

Previous Example

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

> 35 **Age** **<= 35**

Total Population = 6
Profitable = 4
Unprofitable = 2
Profitability rate = 66%

Total Population = 4
Profitable = 1
Unprofitable = 3
Profitability rate = 25%

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

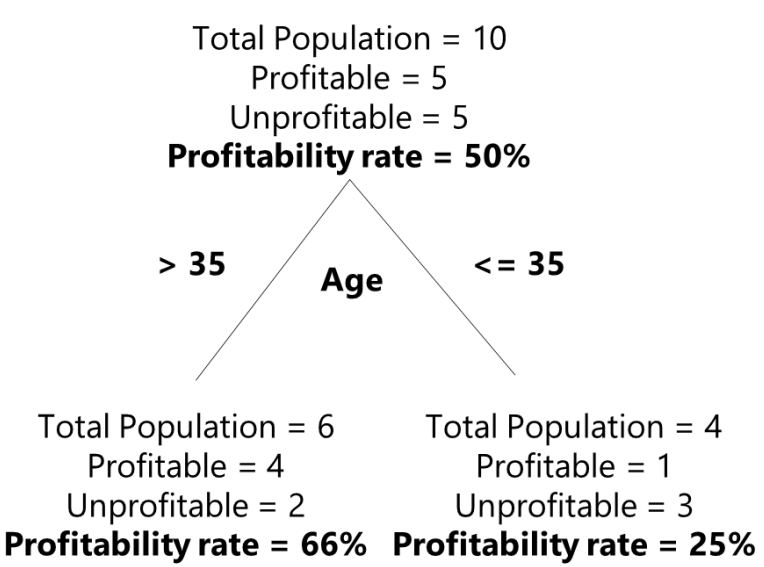
Male **Gender** **Female**

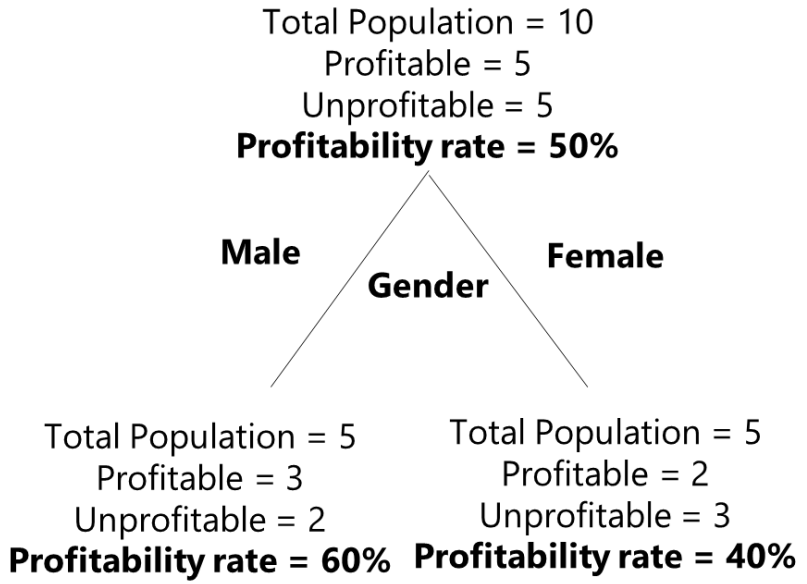
Total Population = 5
Profitable = 3
Unprofitable = 2
Profitability rate = 60%

Total Population = 5
Profitable = 2
Unprofitable = 3
Profitability rate = 40%

Both splits can be compared to understand which split is better

Decision Tree: Algorithm

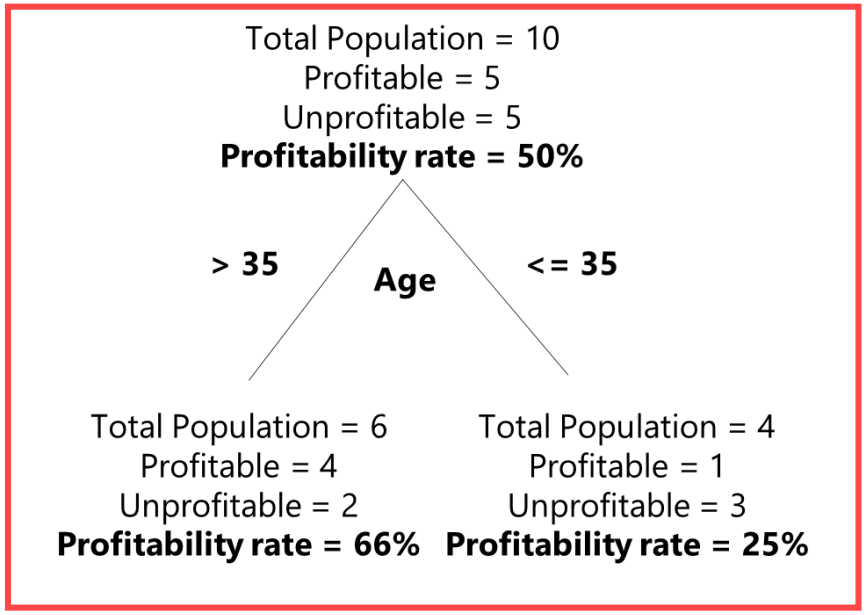


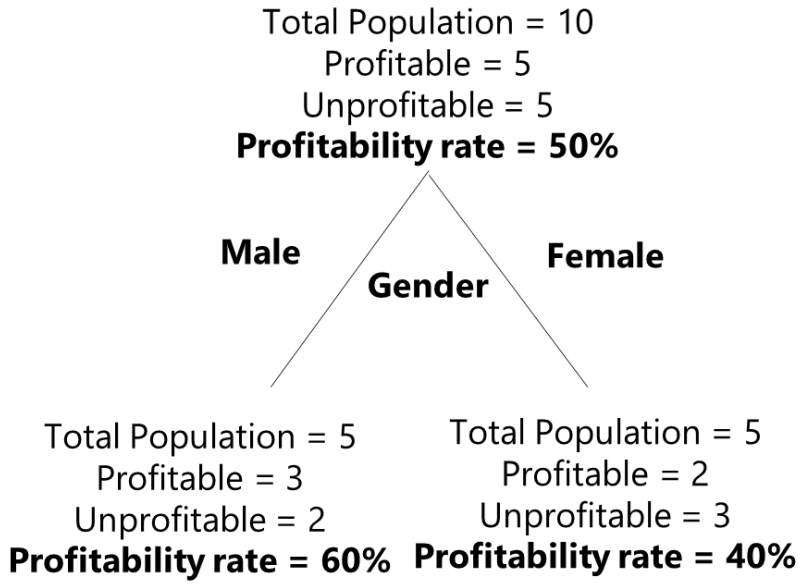


Which variable produces better splits?

Age or Gender?

Decision Tree: Algorithm





Good split in context of classification problem

Split produced by variable age are better than the splits produced by variable gender

Greater the **class imbalance**, better the split

Decision Tree: Algorithm

Class imbalance can be measured by computing Gini or Entropy

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

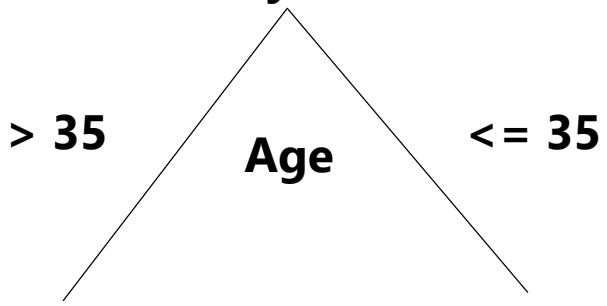
$$Entropy = -\sum p_i log_2 p_i$$

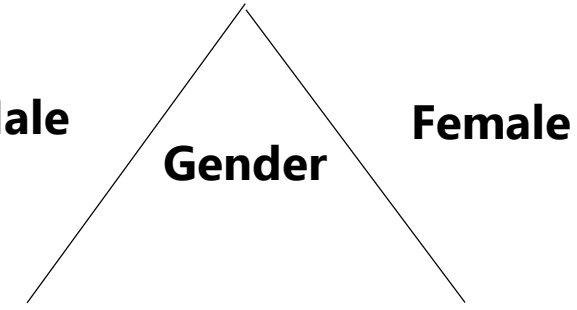
Decision Tree: Algorithm

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

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

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





Total Population = 6
Profitable = 4
Unprofitable = 2
Profitability rate = 66%

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

0.44

Total Population = 4
Profitable = 1
Unprofitable = 3
Profitability rate = 25%

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

0.375

Total Population = 5
Profitable = 3
Unprofitable = 2
Profitability rate = 60%

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

0.48

Total Population = 5
Profitable = 2
Unprofitable = 3
Profitability rate = 40%

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

0.48

Decision Tree: Algorithm

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

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

> 35 — **Age** — **<= 35**

Total Population = 6
Profitable = 4
Unprofitable = 2
Profitability rate = 66%

Total Population = 4
Profitable = 1
Unprofitable = 3
Profitability rate = 25%

$(\frac{6}{10}) * 0.44 + (\frac{4}{10}) * 0.375$

0.41

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

Male — **Gender** — **Female**

Total Population = 5
Profitable = 3
Unprofitable = 2
Profitability rate = 60%

Total Population = 5
Profitable = 2
Unprofitable = 3
Profitability rate = 40%

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

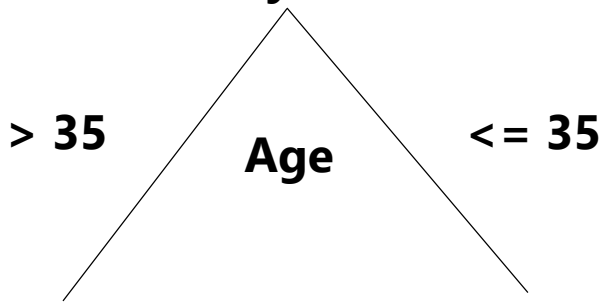
0.48

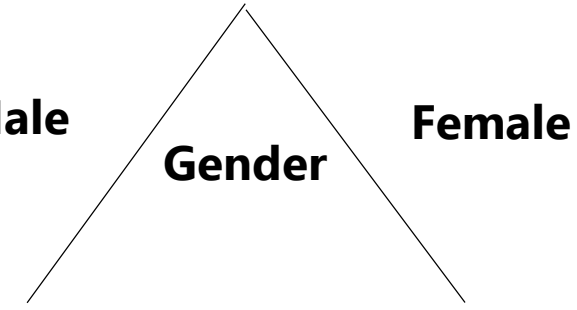
Decision Tree: Algorithm

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

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

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





Total Population = 6
Profitable = 4
Unprofitable = 2
Profitability rate = 66%

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

0.91

Total Population = 4
Profitable = 1
Unprofitable = 3
Profitability rate = 25%

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

0.81

Total Population = 5
Profitable = 3
Unprofitable = 2
Profitability rate = 60%

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

0.97

Total Population = 5
Profitable = 2
Unprofitable = 3
Profitability rate = 40%

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

0.97

Decision Tree: Algorithm

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

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

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

> 35 **Age** **<= 35**

Male **Gender** **Female**

Total Population = 6
Profitable = 4
Unprofitable = 2
Profitability rate = 66%

Total Population = 4
Profitable = 1
Unprofitable = 3
Profitability rate = 25%

Total Population = 5
Profitable = 3
Unprofitable = 2
Profitability rate = 60%

Total Population = 5
Profitable = 2
Unprofitable = 3
Profitability rate = 40%

$\left(\frac{6}{10}\right) * 0.91 + \left(\frac{4}{10}\right) * 0.81$

0.87

$\left(\frac{5}{10}\right) * 0.97 + \left(\frac{5}{10}\right) * 0.97$

0.97

Decision Tree: Algorithm Overview

For each split the purity metric is computed

Choose the lowest variable which results in lowest value of purity metric

Continue doing these till some **stopping criteria** is met

Decision Tree: Algorithm Overview

Stopping Criteria

Depth of tree

Specifying the levels of the tree

Improvement in purity metric

Specifying the minimum change in purity metric from one split to another

Value in terminal node

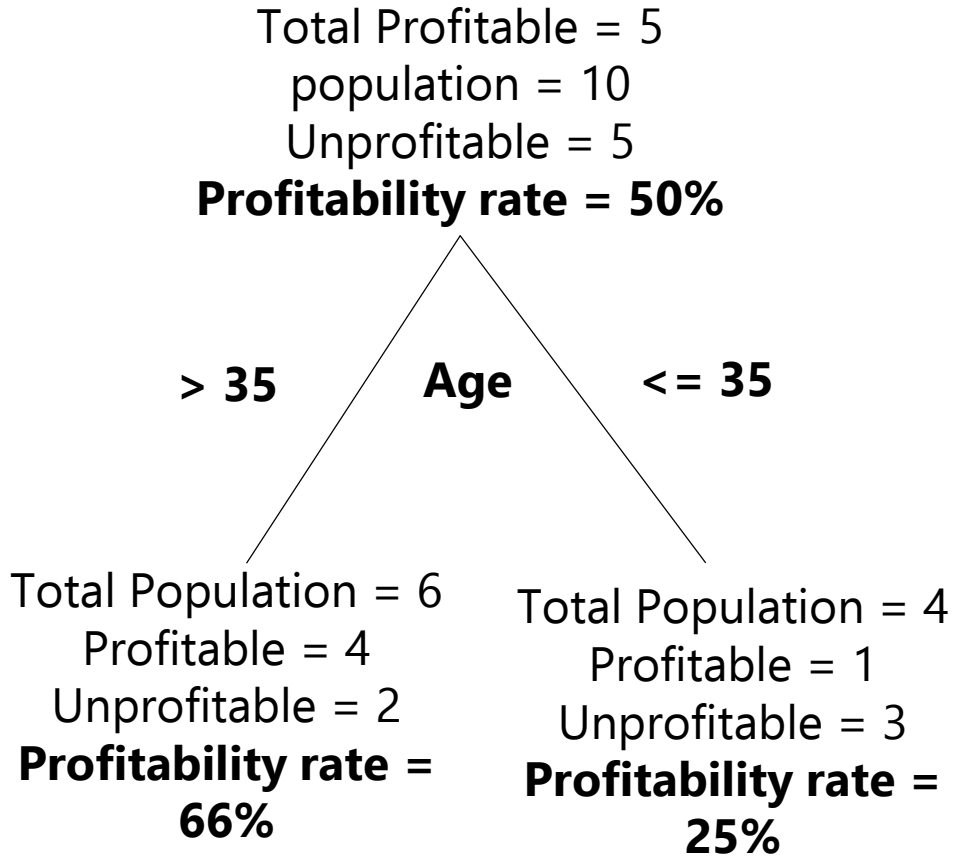
Specifying the number of value in the terminal node

Decision Tree: Prediction

Use decision tree classifier as prediction

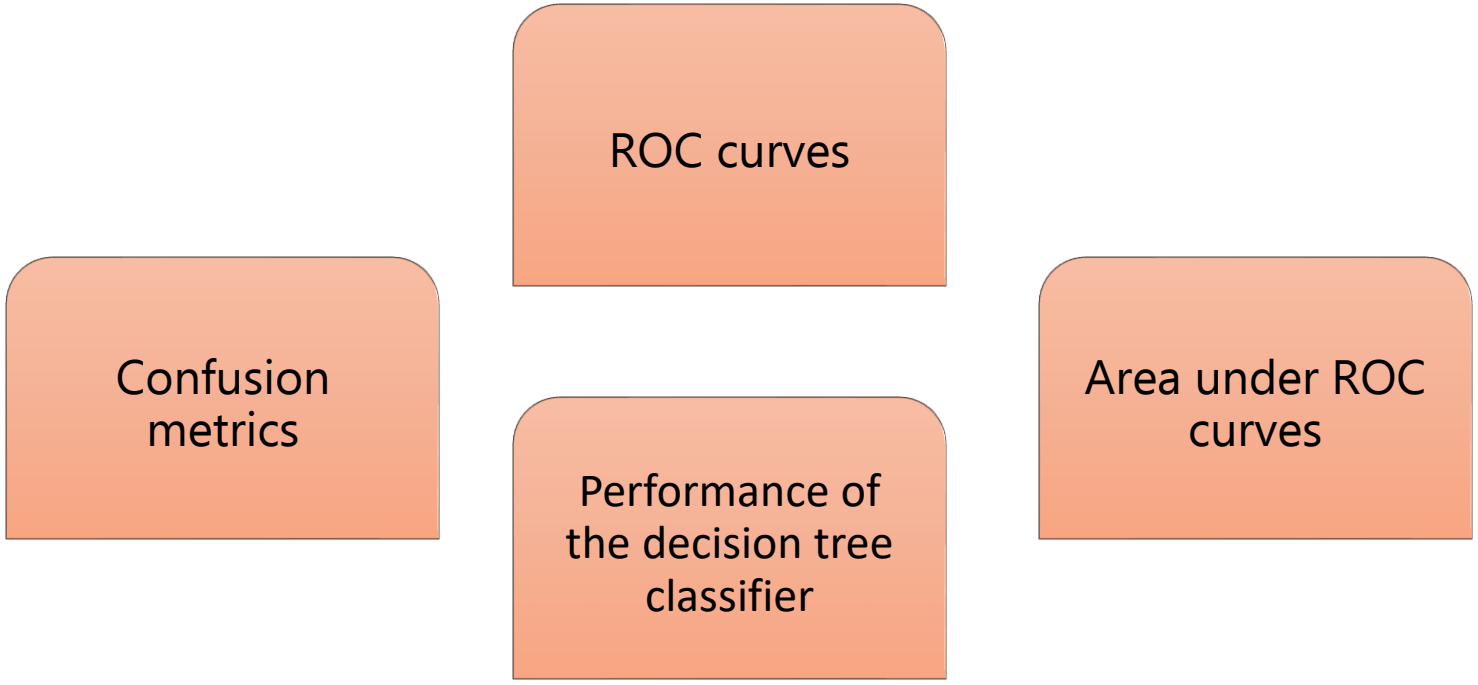
Available data – 20 year old person

Prediction – 25% Chance of him being profitable



Decision Tree: Performance Metrics

Decision tree classifier output probabilities



For multiclass problems, accuracy is used as a performance measure

Decision Tree: Parameters and Hyperparameters

Parameters of a decision tree

Data

Purity metric- Gini or Entropy?

Depth of the tree

These parameters are estimated using cross validation

At the model level of decision tree rules are decided for predicting probabilities or classes

Recap

- Decision Tree Overview
- Decision Tree Algorithms – Gini and Entropy
- Decision Tree Performance Metrics
- Decision Tree Parameter and Hyperparameter