Classification or Decision Tree

CART

CLASSIFICATION AND REGRESSION TREE

Training

Customer ID	Gender	Responded to Email Marketing?
1	Male	Yes
2	Female	No

Testing

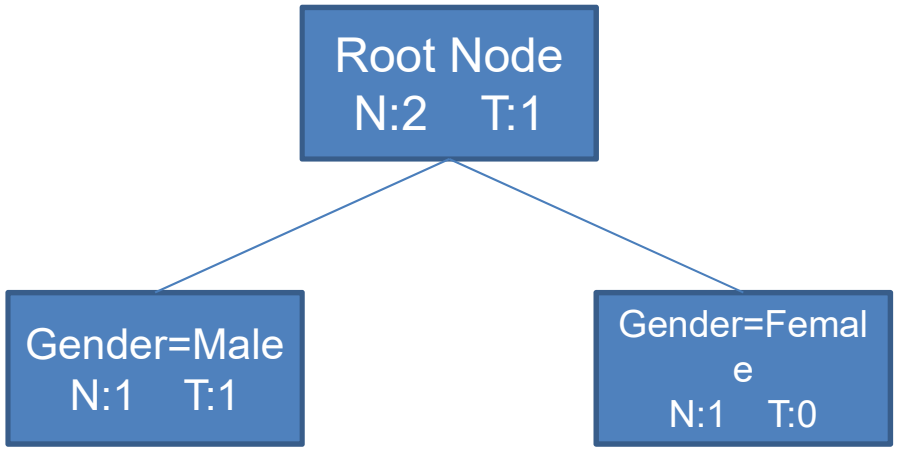
Customer ID	Gender	Likely to Respond? (Predict)
3	Male	?
4	Female	?

greatlearning

Training

Customer ID	Gender	Responded to Email Marketing?
1	Male	Yes
2	Female	No

Trained Model

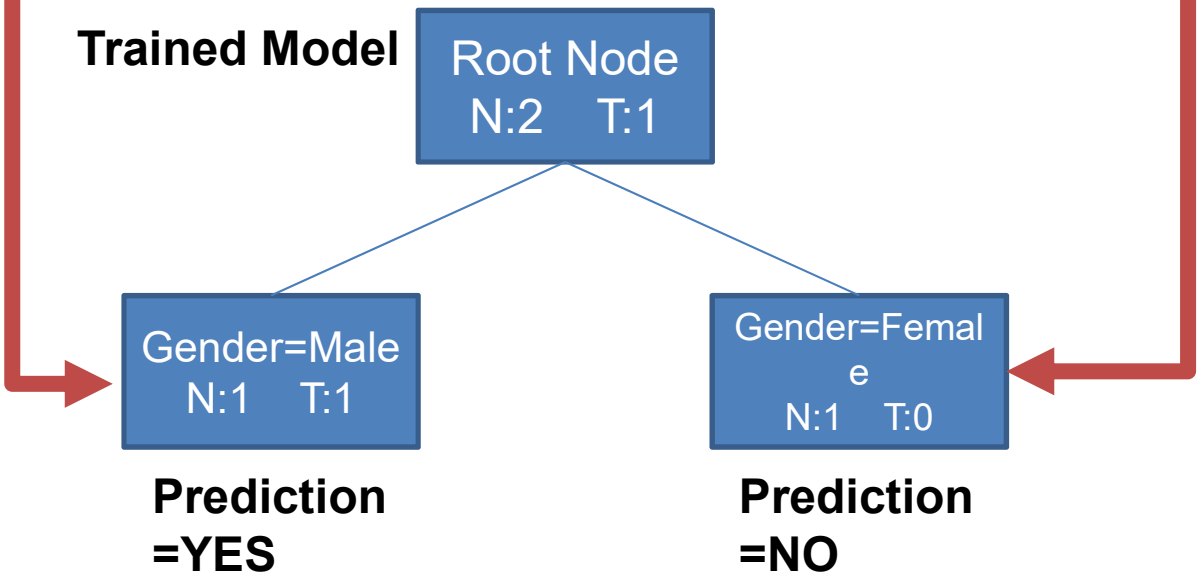


greatlearning

Testing

Customer ID	Gender	Likely to Respond? (Predict)
3	Male	**YES**
4	Female	**NO**

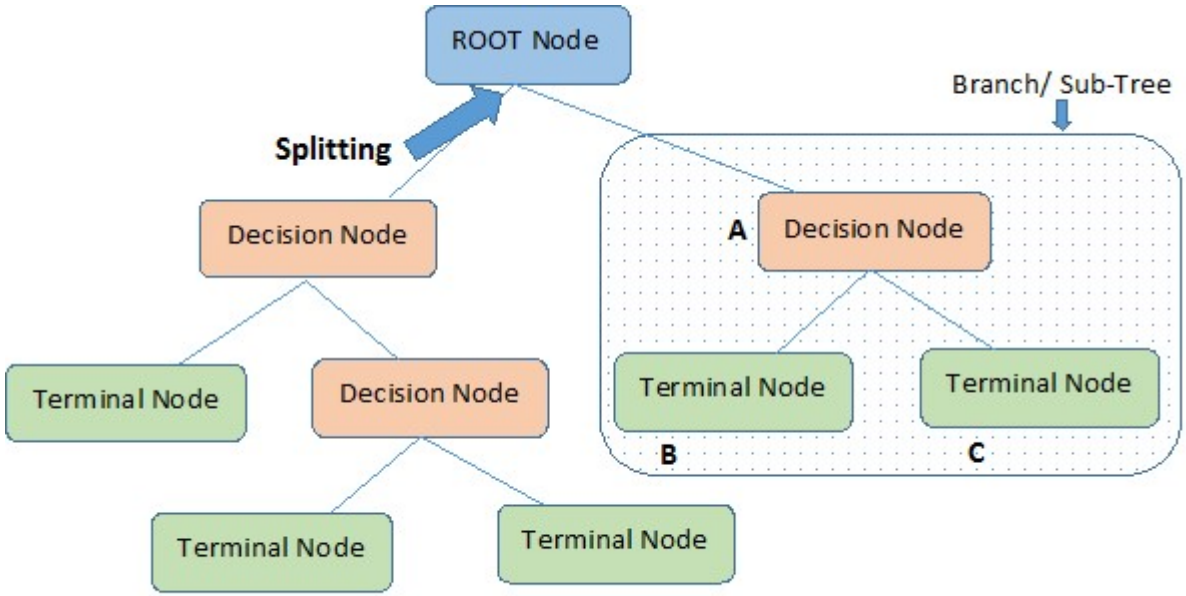
Trained Model



Extending the logic of Decision Trees

- What if there are more Independent Variables? – Choose the Best Variable using a splitting criterion
- What are the splitting criteria available? - Gini
- How many splits to perform for a data? – Depends on Purity of a Node
- Should we always perform a Binary Split? – Binary & Multiway Split Models are available

Decision Tree Terminology



Note:- A is parent node of B and C.

Decision Trees

greatlearning

- Supervised Machine Learning algorithm
- Mainly used for classification
- Works for both categorical and continuous output variables
- Terminology:
  - Root Node: Represents entire Training Dataset
  - Decision Node/Parent Node: Node that is split into sub-nodes
  - Splitting: Dividing Decision node into sub-nodes
  - Leaf/Terminal Node: Nodes that can no longer split
  - Branch: Sub-section of entire tree
  - Child node – The resulting Nodes after splitting a decision node

Decision Trees

greatlearning

- Advantages
  - Easy to interpret
  - Automated field selection
  - No data processing required
    - Variable transformation not required
    - Can handle outliers
    - Missing value tolerant
- Disadvantages
  - They are unstable
  - Often inaccurate and poor compared to other models (Solution – Random Forest)
  - Generally not preferred for continuous prediction

Decision Tree – Model Design

- Data should have both 0 (Bad) and 1 (Good) data
- Remove indeterminate values (NAs)
- Look for categorical variables
- Look for meaningful trend. Eg: Height should increase with Age.
- Look for default values like -999. Convert them to missing values. Maybe remove it.
- Look for capping/floor values – Age > 100
- Reasons to create meaningful group – Group all small Northeastern states

Classification Techniques

- **Classification and Regression Tree (CART):**
  - Binary Decision Tree
  - Classification (Categorical output variable)
  - Regression (Continuous output variable)
  - Uses Gini Index
- **CHAID – CHI-squared Automatic Interaction Detector:**
  - Non-Binary Decision Tree
  - Use statistical significance of proportions

LET US BUILD A DECISION TREE!!!

Gini Index Calculations

greatlearning

$$Gini\,(D) = 1 - \sum_{i=1}^{m} p_i^{\,2}$$

m : Number of Classes

p : Probability that a record in D belongs to class Ci

Gini Index consists of binary split ($\mathbf{D_1}$ and $\mathbf{D_2}$) for each attribute **A**

$$Gini_{\,A}(D) = \frac{D_1}{D} Gini\,(D_1) + \frac{D_2}{D} Gini\,(D_2)$$

Reduction in Impurity is given by:

$$\Delta Gini\,(A) = Gini\,(D) - Gini_{\,A}(D)$$

The attribute that Maximizes the reduction in impurity is chosen as the Splitting Attribute

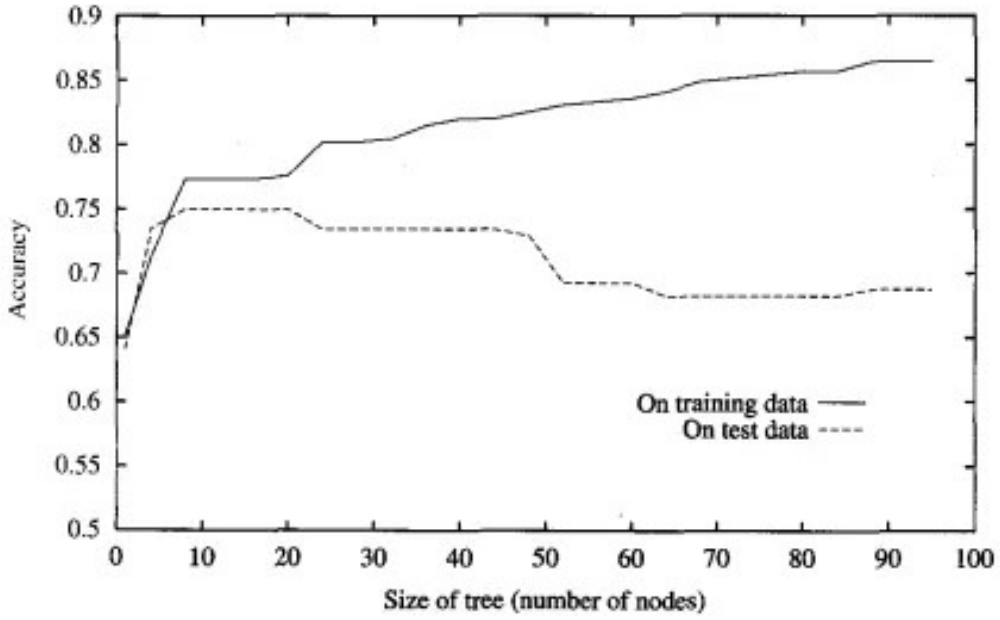
Limitations of Decision Trees

- Vulnerable to over-fitting
  Solution – Pruning

- Greedy Algorithm
  Solution – Cross Validation

Over Fitting

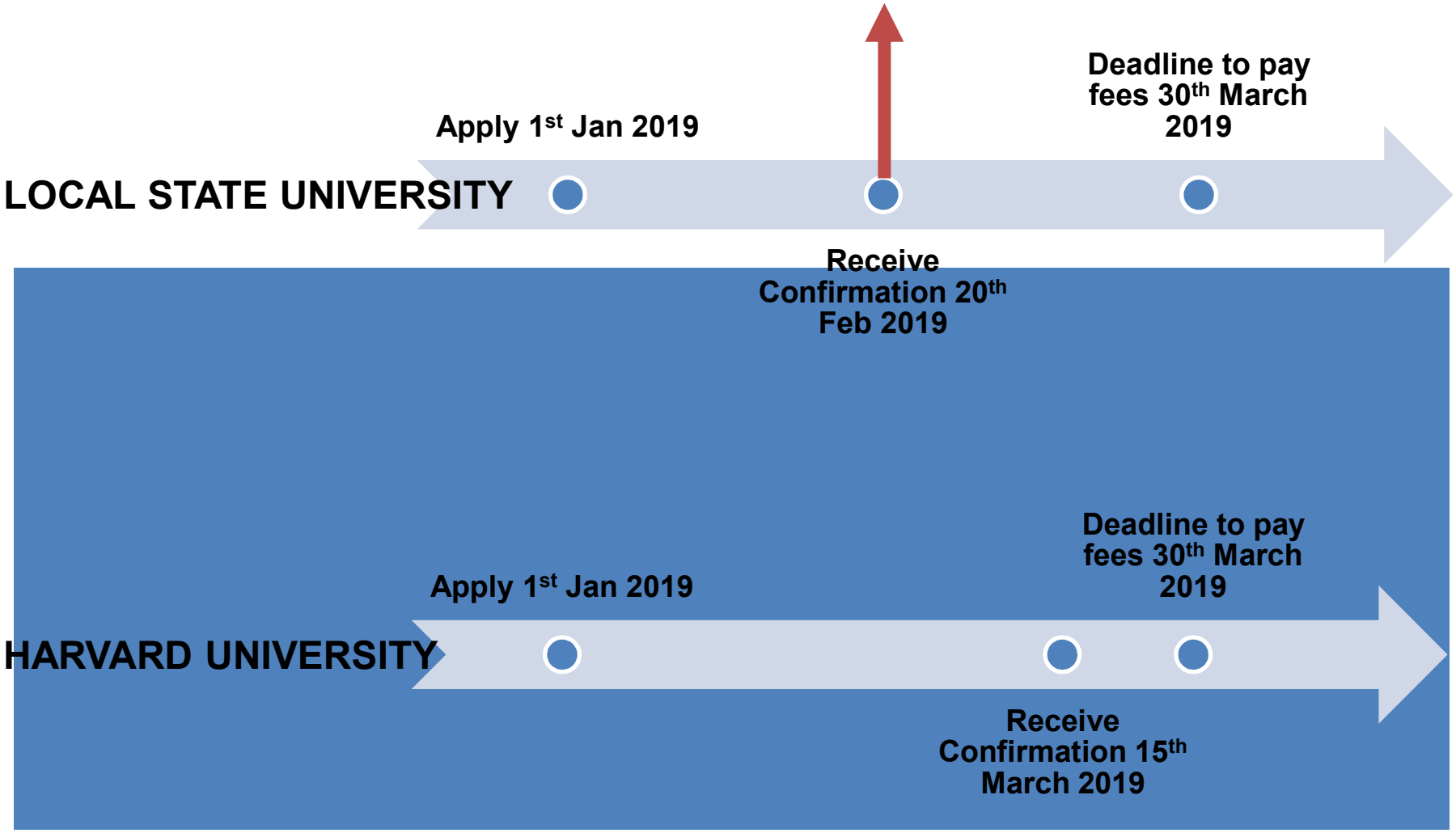
- Works extremely well on Training dataset
- Performs poorly on unseen dataset



Greedy Algorithm

greatlearning

Deciding to pay fees on the same day (Greedy Decision, Not Optimal)



Greedy Algorithm

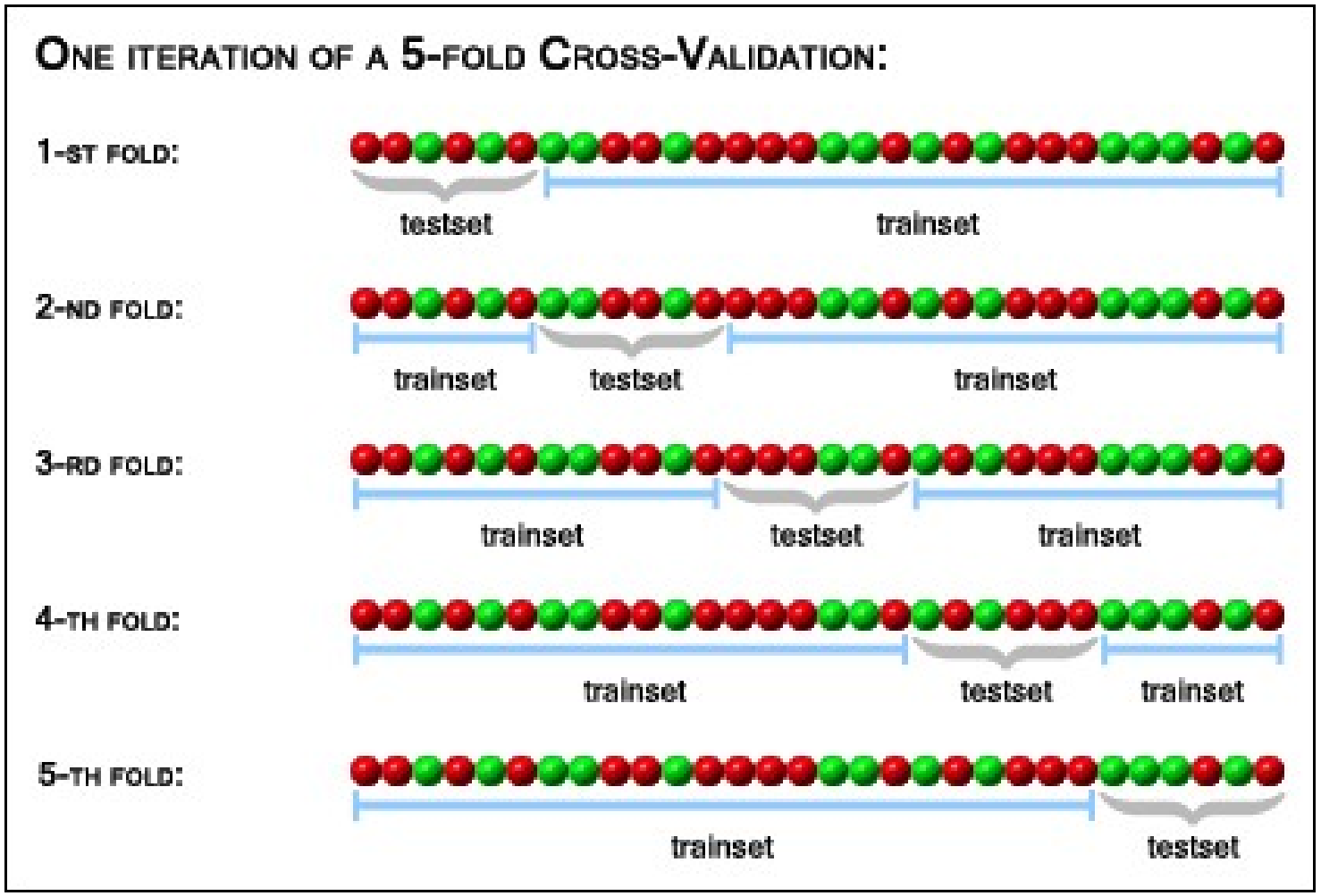
- When a Split happens using the best independent variable , the model does not consider the future states

- What-if the model has higher accuracy if a different independent variable is chosen instead of the best one?

Cross Validation

greatlearning

Source: https://stats.stackexchange.com/questions/1826/cross-validation-in-plain-english



Cross-Validation

- Helps overcome Greedy Algorithm problem
- How good is the model with unseen data?
- Also helps address ‘Over Fitting’

greatlearning

Thank You