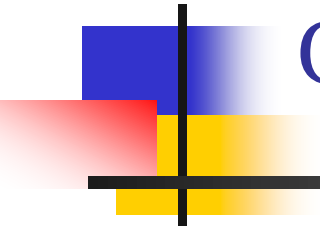


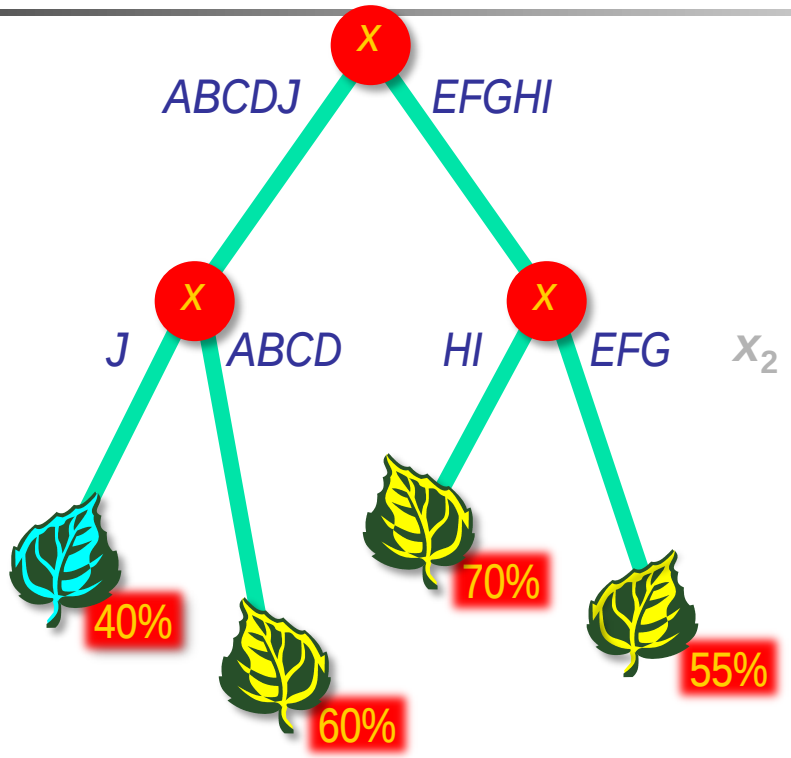
Consolidating Categorical Inputs

Dr. Goutam Chakraborty

What to do With Categorical Variables with Large Number of Categories?

- Use as it is (adds too many flag variables increasing dimensionality)
- Consolidate levels:
  - Using domain expertise
  - Using a decision tree
  - Using Weight-of-evidence (WOE) approach

Categorical Input Consolidation



Combine categorical input levels that have similar primary outcome proportions.

Demo

- Follow handout titled “Consolidating categorical variable_handout”
- Connect a **Decision Tree** node to the **Impute** node and rename it **Consolidation Tree.**
- Make these changes in the Train property group.
  - Under the Subtree section, select **Assessment Measure** ⇨ **Average Squared Error**. This optimizes the tree for prediction estimates.
  - Under the P-Value Adjustment section, select **Bonferroni Adjustment** ⇨ **No**.
- Make these changes in the Score property group.
  - Select **Variable Selection** ⇨ **No**. This prevents the tree from rejecting inputs in subsequent nodes.
  - Select **Leaf Role** ⇨ **Input**. This adds a new input (**_NODE_**) to the training data**.**
- Right-click tree > Edit variables > select **Use** ⇨ **Yes** for **DemCluster** and **TARGET_B** and **No** for all other variables
- Then use the Interactive Tree tool to cluster **DemCluster** values into related groups
  - Right-click the root node and select **Train Node** from the option menu.