Decision Trees

“Nothing is particularly hard if you divide it into small jobs”.

- *Henry Ford*

DECISION TREES: INTRODUCTION

Decision trees (also known as decision tree learning or classification trees) are a collection of predictive analytics techniques that use tree-like graphs for predicting the value of a response variable (or target variable) based on the values of explanatory variables (or predictors).

Decision trees use the following criteria to develop the tree

- **Splitting Criteria**
- **Merging Criteria**
- **Stopping Criteria**

Decision Trees - STEPS

- **Step1 :** Start with the **root node** in which all the data is present.
- **Step2:** Decide on a splitting criterion and stopping criteria. The root node is then split into two or more subsets leading to tree branches (called edges) using the splitting criterion. These are known as **internal nodes**. Each internal node has exactly one incoming edge.
- **Step3 :** Further divide each internal node until no further splitting is possible or the stopping criterion is met. The **terminal nodes** (aka **leaf nodes**) will not have any outgoing edges.
- **Step4 :** Terminal nodes are used for generating business rules.
- **Step 5 :** **Tree pruning** (a process for restricting the size of the tree) is used to avoid large trees and over-fitting the data. Tree pruning is achieved through different stopping criteria

CHI-SQUARE AUTOMATIC INTERACTION DETECTION (CHAID)

- CHAID is an extension of Automatic Interaction Detection (AID), which is designed to categorize the dependent variable using categorical predictors.

- CHAID trees use statistical significance of independent variables to split the subset of the data (represented by nodes of the tree).

CHAID Tree Development

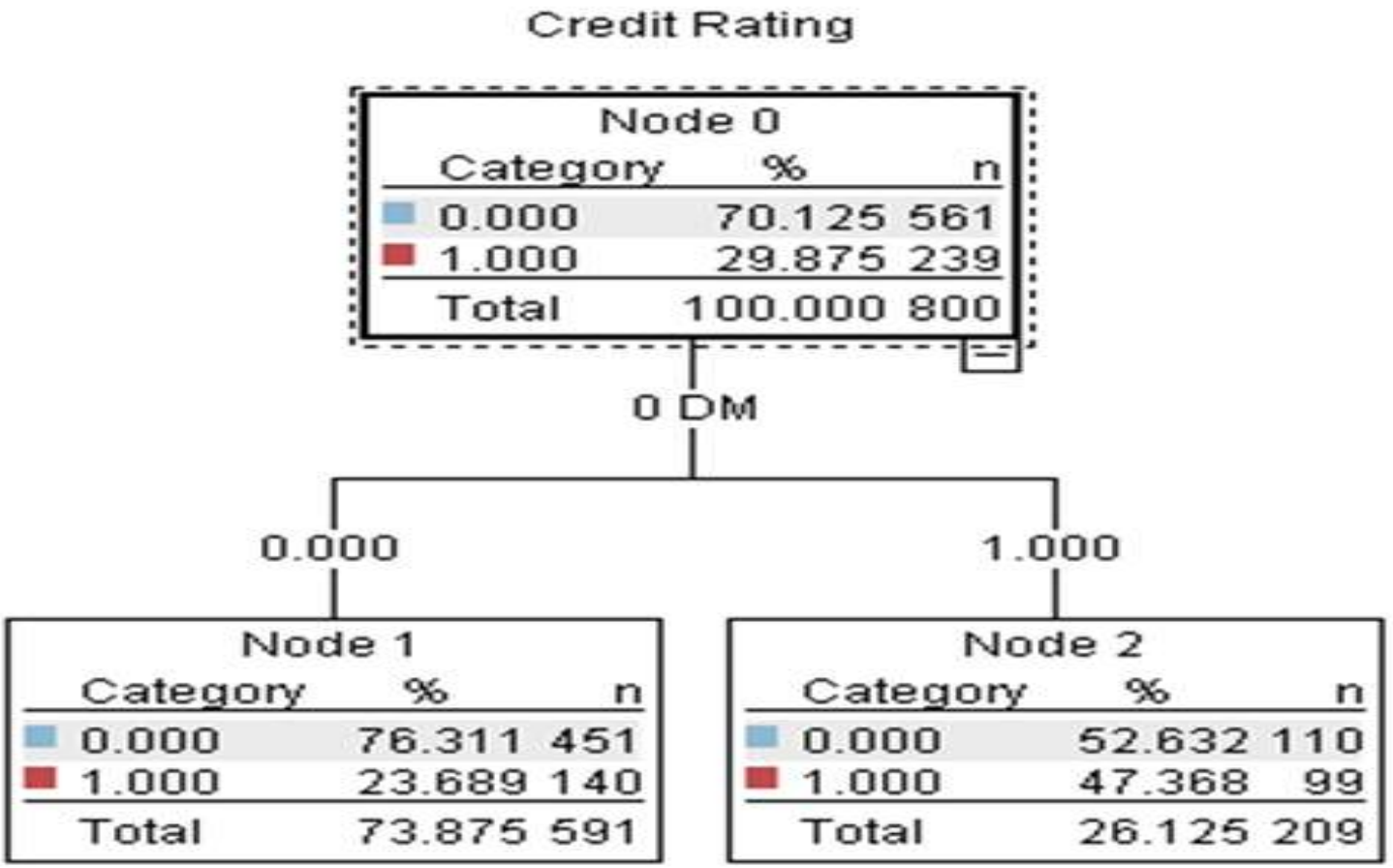
We will be using a sample of 800 observations from the German Credit Data to illustrate CHAID tree development.

Checking Account Balance	Credit Rating (Observed)			Credit Rating (Expected)	
	Y = 1	Y = 0	Total	Y = 1	Y = 0
0DM = 1	99	110	209	62.44	146.56
0DM = 0	140	451	591	176.56	414.44
Total	239	561	800	239	561

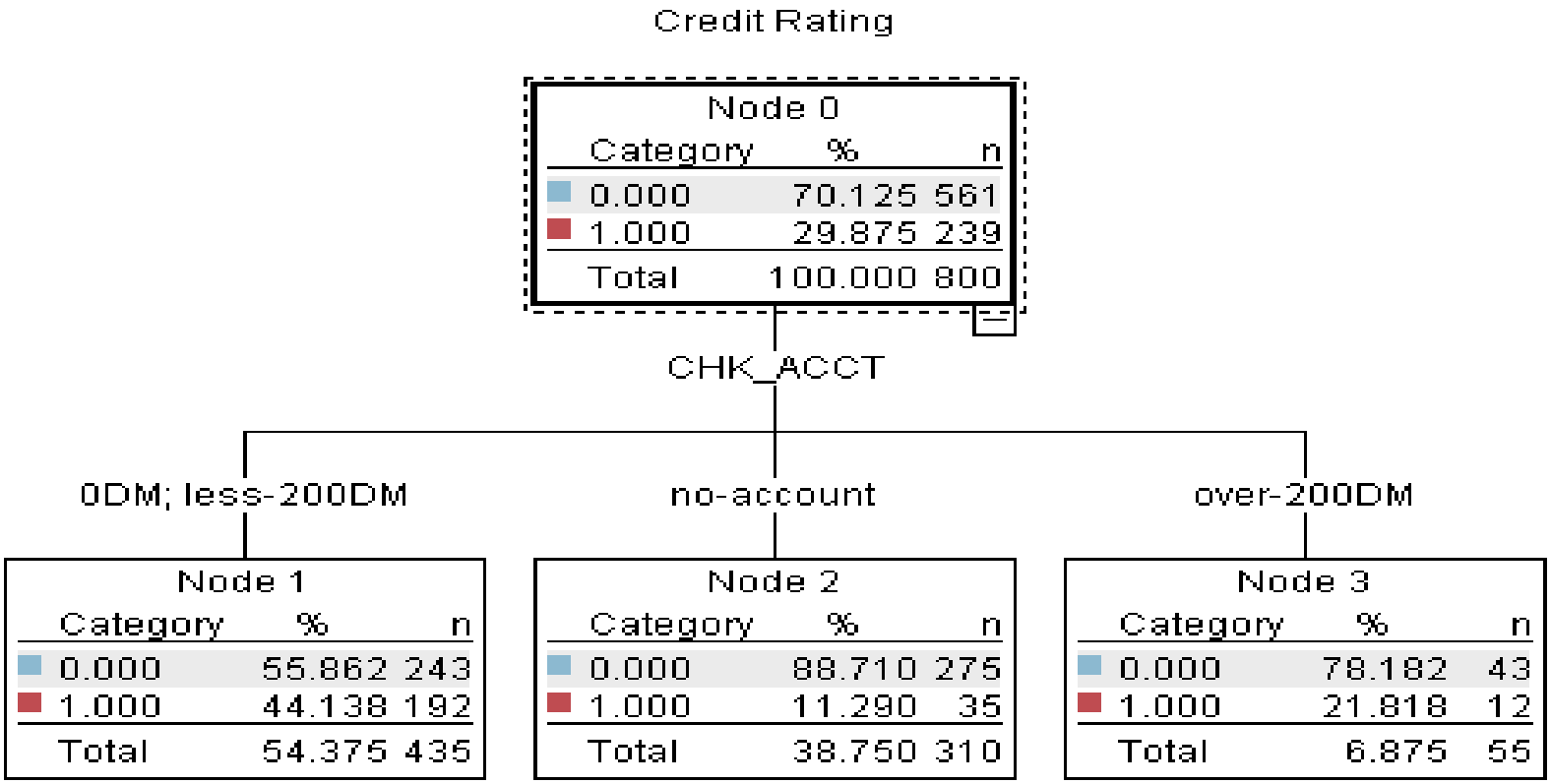
Using the values in contingency table the Chi-square statistic is given by:

$$\chi^2 = \frac{(99-62.44)^2}{62.44} + \frac{(110-146.56)^2}{146.56} + \frac{(140-176.56)^2}{176.56} + \frac{(451-414.44)^2}{414.44} =$$

CHAID Tree German Credit Data



CHAID Tree with all categories within checking account balance



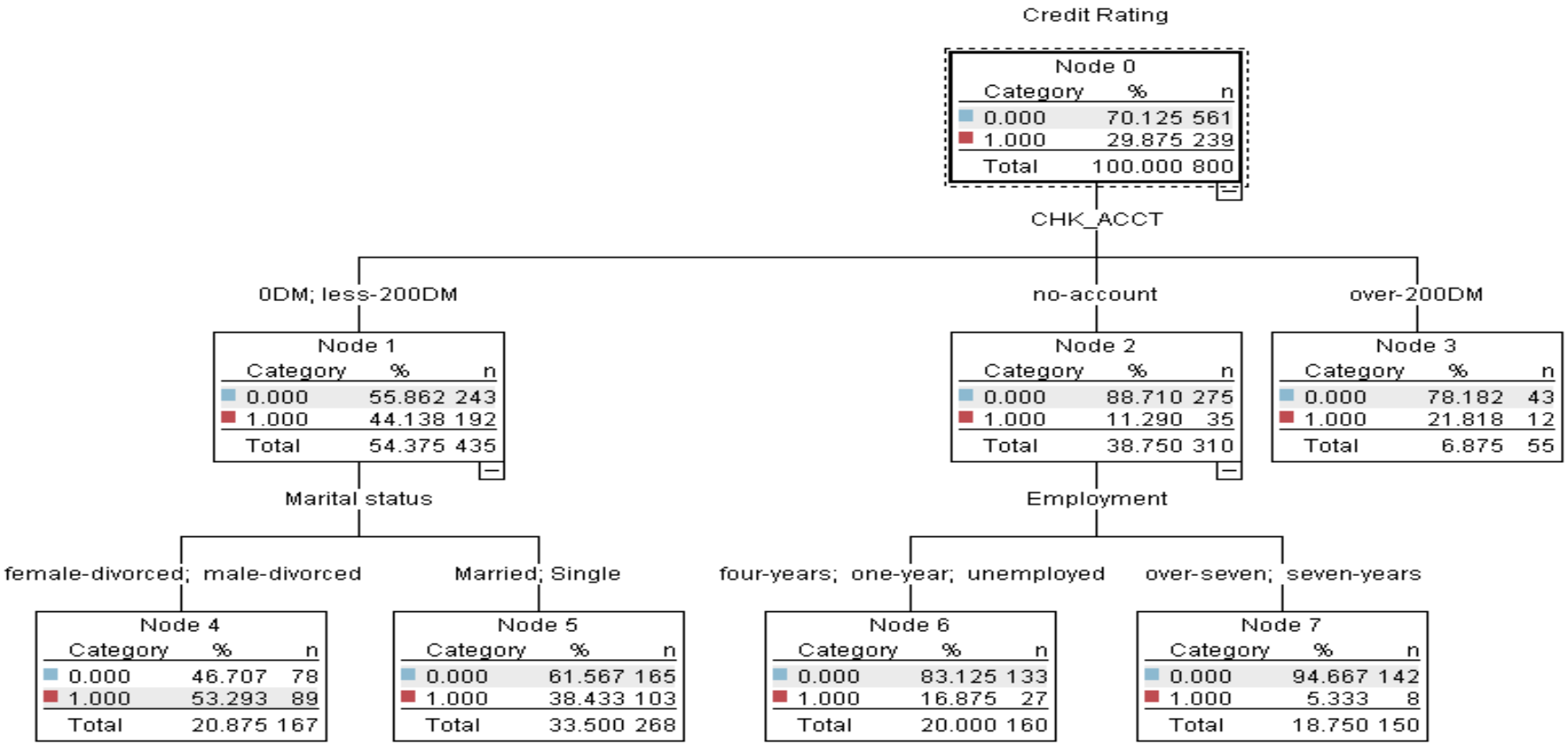
Bonferroni Correction

The Bonferroni Correction sets the significant cut-off for individual test at α/n instead of α when n hypothesis tests are conducted simultaneously (Armstrong, 2014).

That is, we set a lower type I error (α/n) at individual tests, but this may increase the type II error which is one of the criticisms of the Bonferroni Correction.

Generating Business Rules using CHAID Tree

CHAID tree for German credit rating sample data.



Business Rules and Support

Node	Business Rule	Support
3	Checking account balance is more than 200DM, classify the outcome as Y = 0. Classification accuracy is 78.18%.	6.875%
4	Checking account balance is either 0DM or less than 200DM AND the marital status is male divorced or female divorced, classify outcome as Y = 1. Classification accuracy is 53.293%.	20.875%
5	Checking account balance is either 0DM or less than 200DM AND the marital status is married or single, classify outcome as Y = 0. Classification accuracy is 61.56%.	33.500%
6	No checking account AND the employment is 1. Unemployed or 2. One year or 3. Four years Then classify the outcome as Y = 0. Classification accuracy is 83.125%.	20.00%
7	No checking account AND the employment is either seven years or over seven years then classify the outcome as Y = 0. The classification accuracy is 94.667%.	18.750%

CLASSIFICATION AND REGRESSION TREE

- Classification and Regression Tree (CART) is a common terminology that is used for a **Classification Tree** (used when the dependent variable is discrete) and a **Regression Tree** (used when the dependent variable is continuous).

- Classification tree uses various impurity measures such as the Gini Impurity Index and Entropy to split the nodes.

- Regression Tree, on the other hand, splits the node that minimizes the Sum of Squared Errors

STEPS

The following steps are used to generate a classification and a regression tree (Breiman *et al.* 1984)

- **Step 1 :** Start with the complete training data in the **root node**.
- **Step 2 :** Decide on the **measure of impurity** (usually Gini impurity index or Entropy). Choose a predictor variable that minimizes the impurity when the parent node is split into **children nodes** [see Eq. (12.4)].

This happens when the original data is divided into two subsets using a predictor variable such that it results in the maximum reduction in the impurity in the case of discrete dependent variable or the maximum reduction in SSE in the case of a continuous dependent variable.

- **Step 3 :** Repeat step 2 for each subset of the data (for each **internal node**) using the independent variables until:

  - ✓ All the dependent variables are exhausted .
  - ✓ The stopping criteria is met. Few stopping criteria used are number of levels of tree from the root node, minimum number of observations in parent/child node (eg. 10% of the training data) and minimum reduction in impurity index.

- **Step 4 :** Generate business rules for the **leaf (terminal) nodes** of the tree.

Gini Impurity Index

Gini impurity index is one of the measures of impurity that is used by classification trees to split the nodes.

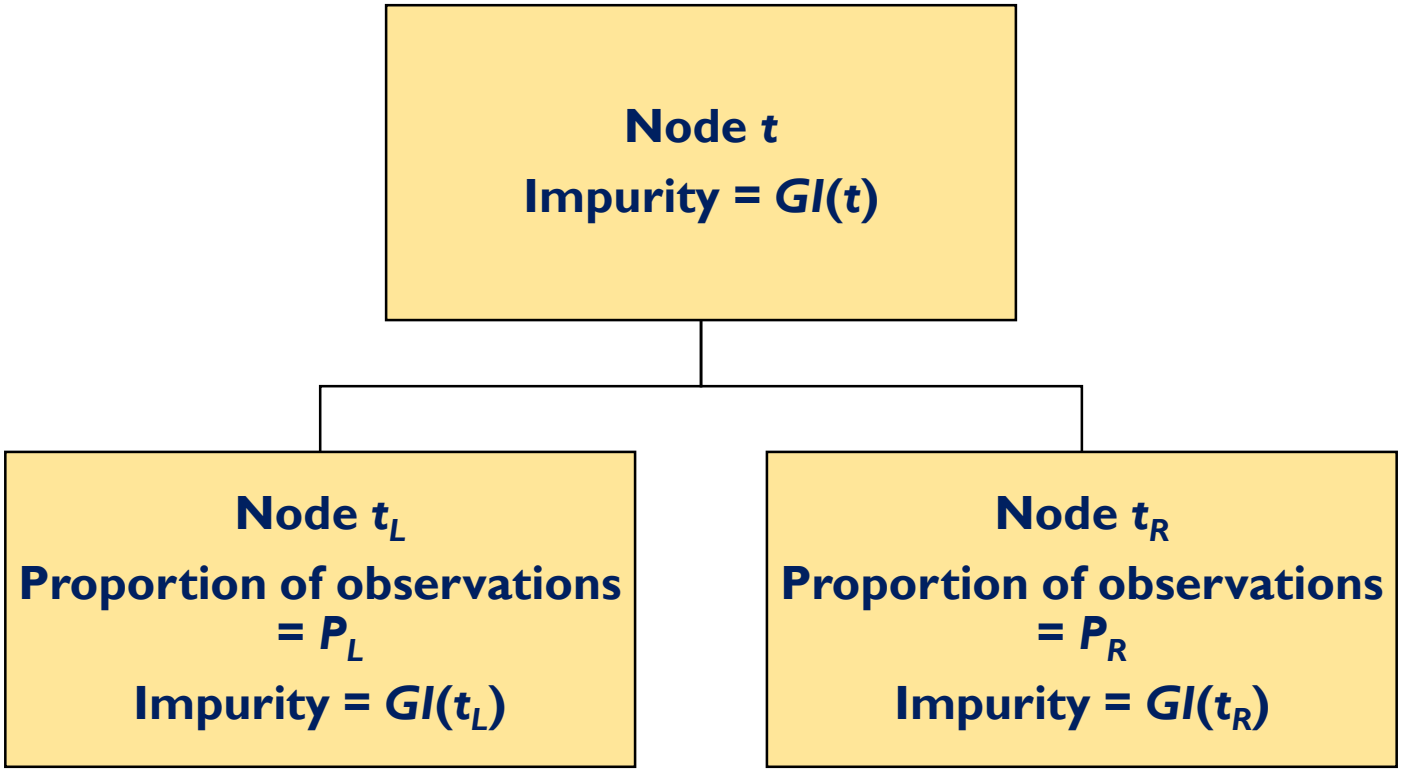
$$GI(t) = \sum_{i=1}^{K} \sum_{j=1, j\neq i}^{K} P(C_i|t)P(C_j|t) = \sum_{i=1}^{K} P(C_i \mid t)(1 - P(C_i \mid t) = 1 - \sum_{i=1}^{K} [P(C_i \mid t)]^2$$

where

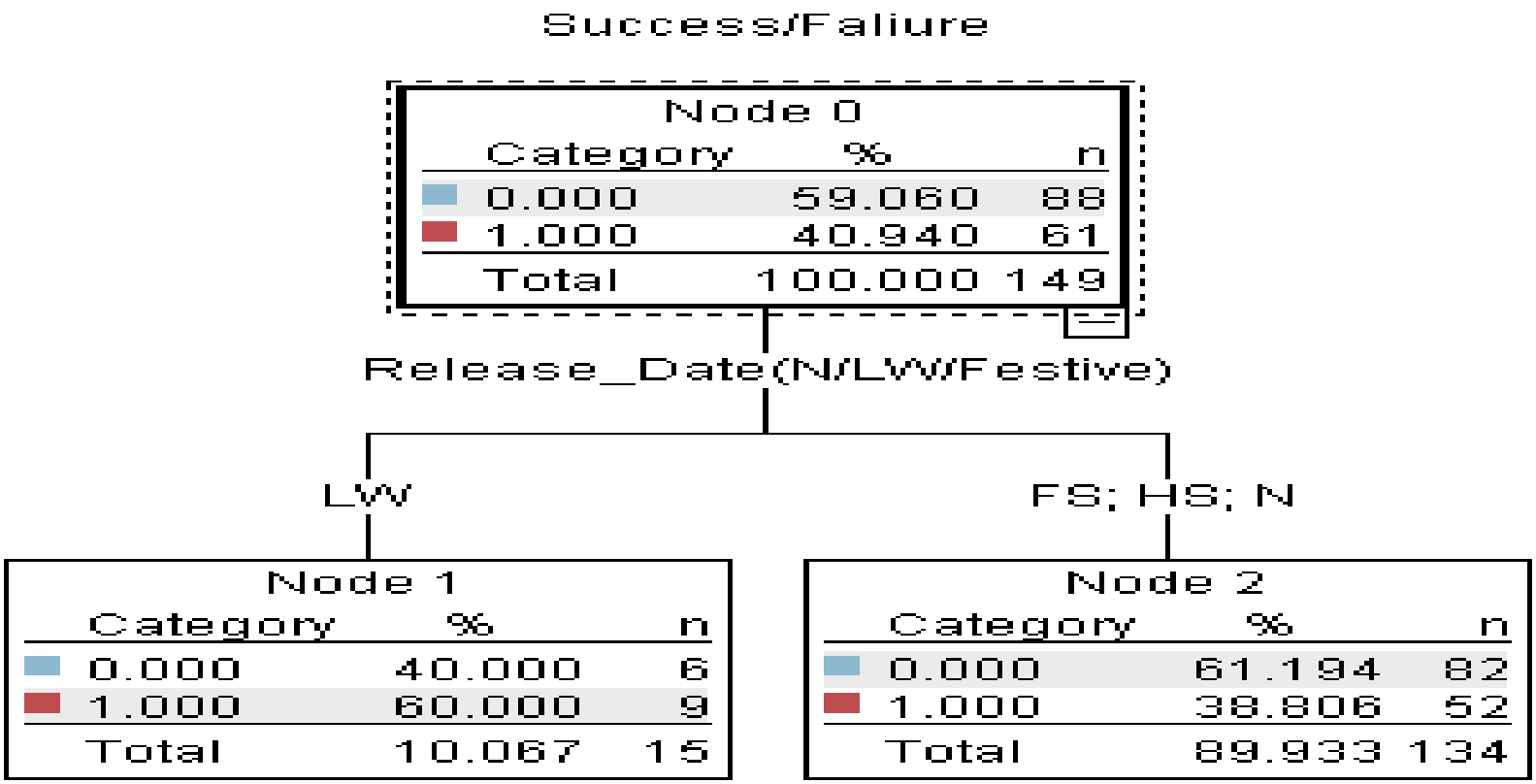
$GI(t)$ = Gini index at node t

$P(C_i|t)$ = Proportion of observations belonging to class C_i in node t

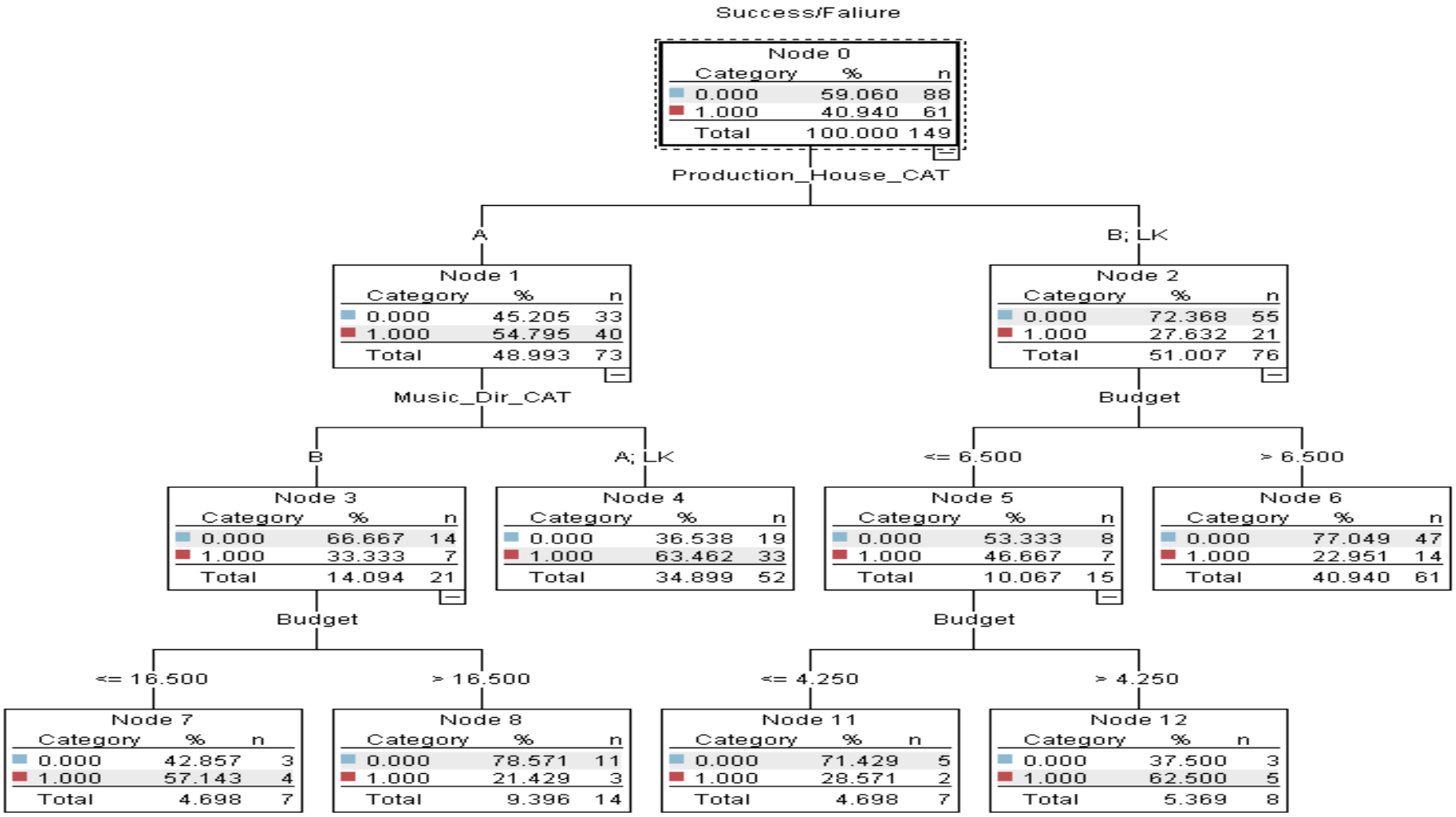
Splitting strategy in CART



Classification tree for success/failure of a Bollywood movie based on release



Classification tree for success/failure of a Bollywood movie.

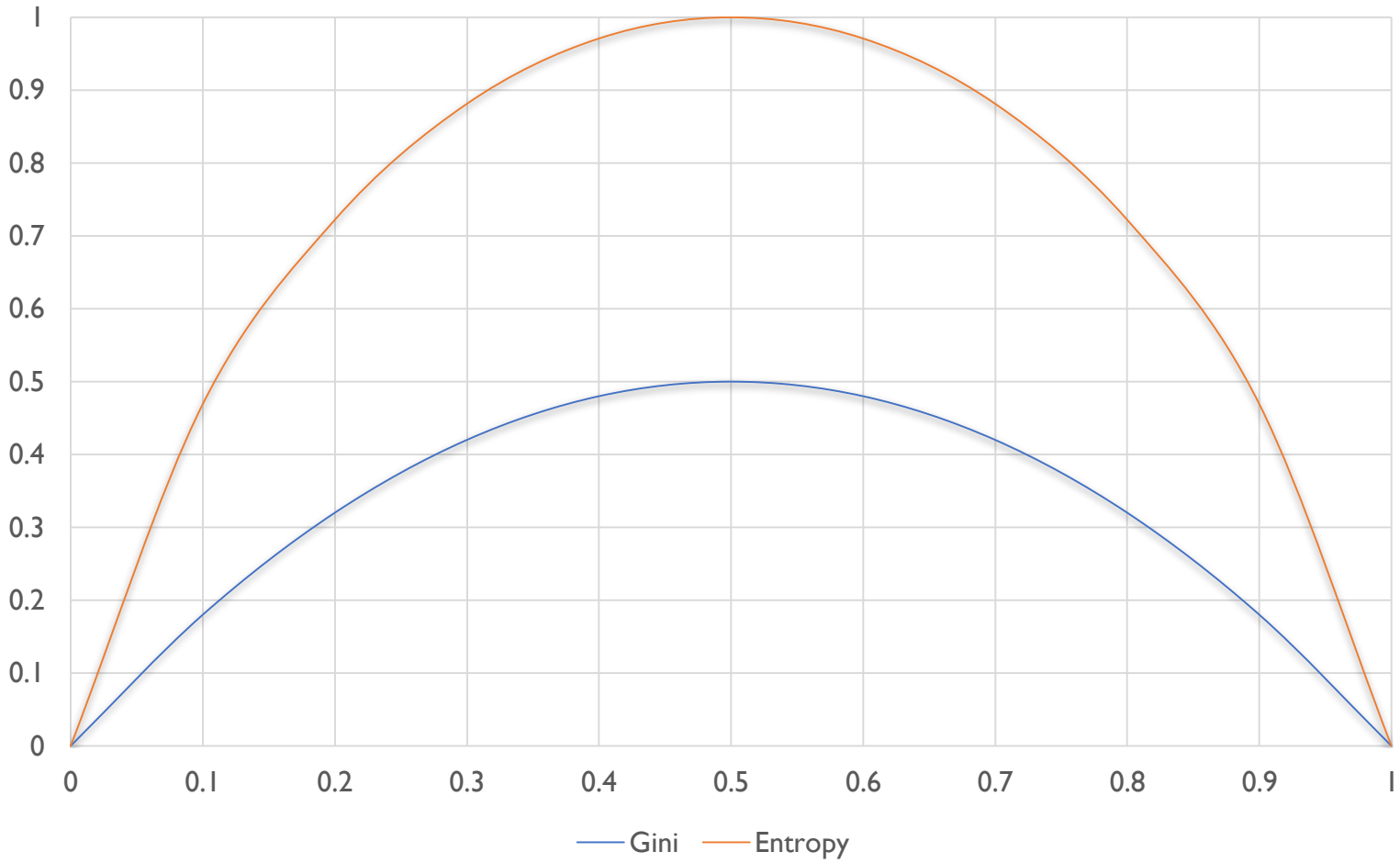


Entropy

Entropy is another popular measure of impurity that is used in classification trees to split a node.

$$Entropy(t) = -\sum_{i=1}^{K} P(C_i \mid t) \times \log_2 P(C_i \mid t)$$

Comparison of Gini index and entropy



Cost Based Splitting Criteria

- Other than impurity measures such as Gini impurity index and entropy, decision makers may also use **Cost of Misclassification** to split the data.

- The total penalty is $C_{01}P_{01} + C_{10}P_{10}$

  Where

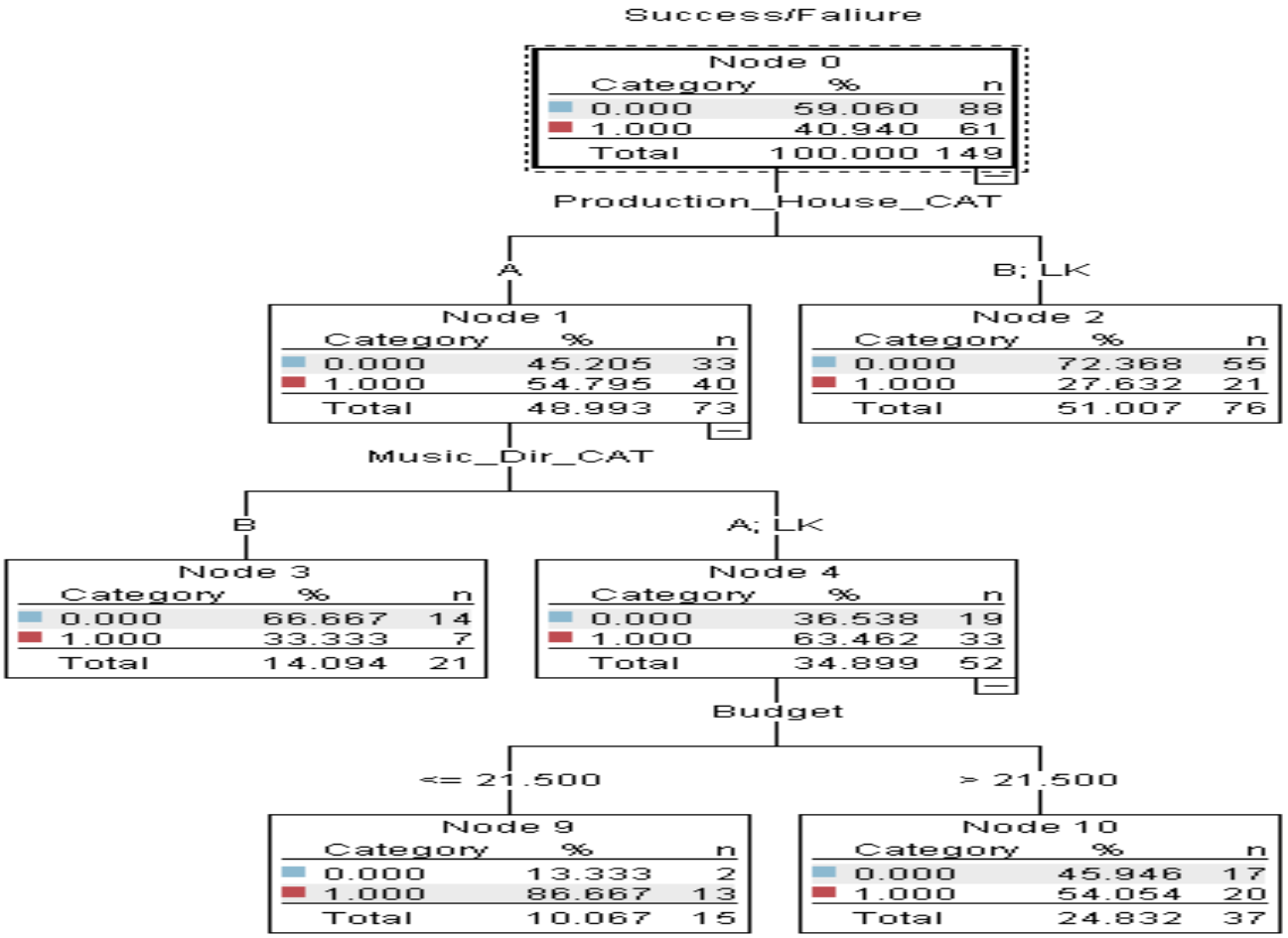
  P_{01} = Proportion of 0 classified as 1

  P_{10} = Proportion of 1 classified as 0

  C01 = Penalty for classifying 0 as 1

  C10 = Penalty for classifying 1 as 0

Classification tree based on misclassification cost (C_{01} = 5 and C_{10} = 1



ENSEMBLE METHOD

The ensemble method is a machine-learning-algorithm that generates several classifiers (a classifier means a classification model) using different sampling strategies such as bootstrap aggregating.

RANDOM FOREST

- Random Forest is one of the popular ensemble method in which several trees (thus the name "forest") are developed using different sampling strategies. One of the most frequently used sampling strategy is the **Bootstrap Aggregating** (or **Bagging**).

- Bagging is a random sampling with replacement.

STEPS

Random forests are developed using the following steps:

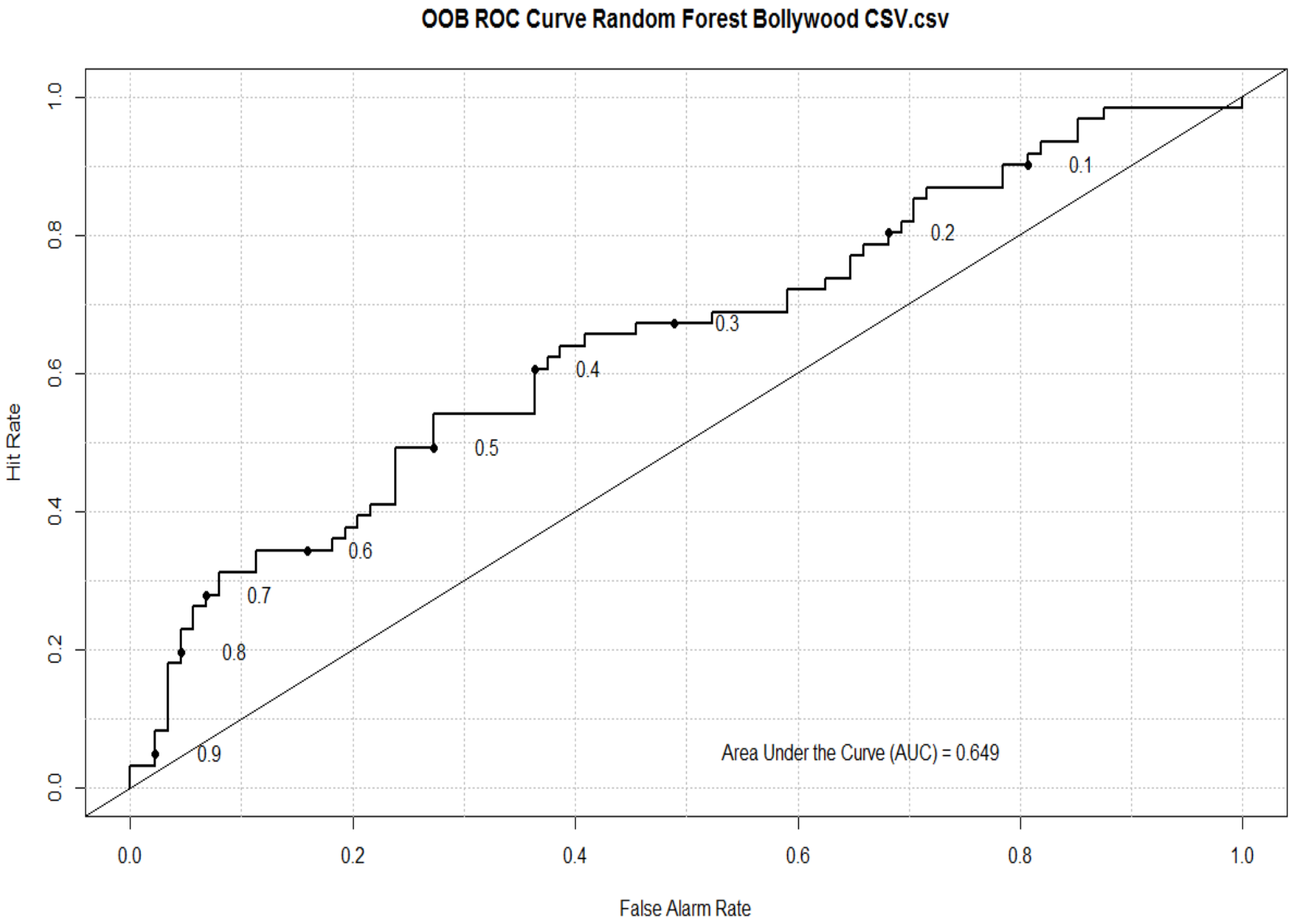
- Step 1 : Assume that the training data has *N* observations. One needs to generate several samples of size *M* (*M* < *N*) with replacement (called **Bagging**). Let the number of samples based on sampling of the training dataset be S_1.
- Step 2 : If the data has *n* predictors, sample *m* predictors (*m* < *n*).
- Step 3 : Develop trees for each of the samples generated in steps 1 using the sample of predictors from step 2 using CART.
- Step 4 : Repeat step 3 for all the samples generated in step 1.
- Step 5 : Predict the class of a new observation using majority voting based on all trees.

- One has to be aware of possible overfitting while using random forest. The model is validated using the validation data, known as **Out-of-Bag (OOB)** data.

- All such cases that are not part of training data of a tree can be used as a validation data and such cases are called Out-Of-Bag data.

- The random forest model using 500 trees and sampling 2 out of 5 variables in table 12.3 is developed using Rattle Package

- Rattle is one of the packages in open source software R


OOB ROC Curve Random Forest Bollywood CSV.csv
Hit Rate
False Alarm Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Area Under the Curve (AUC) = 0.649

Summary

- ✓ Decision trees are important set of techniques used in predictive analytics to solve problems associated with continuous as well as discrete dependent variables. However, decision trees are mostly used for solving classification problems and such trees are called classification trees.
- ✓ Decision trees are supervised learning algorithms. They are developed using different splitting, merging and stopping criteria.
- ✓ There are several decision tree learning algorithms and they differ mainly in the splitting criteria.
- ✓ Chi-square automatic interaction detection (CHAID) uses chi-square test of independence when the dependent variable is categorical; F-test when the dependent variable is continuous and likelihood ratio test when the dependent variable is ordinal as splitting strategy.

- In classification and regression tree (CART) impurity measures such as Gini impurity index or entropy are used as splitting criteria when the dependent variable is categorical and sum of squared errors (SSE) is used when the dependent variable is continuous.

- Decision trees are integral part of random forest algorithm. Random forest technique uses sampling with replacement (bagging) to create several trees and the class of a new observation is decided on the basis of majority voting.

- One of the major advantages of decision tree learning algorithms is that it can be used for generating business rules from the model directly.

References

- Armstrong R A (2014), "When to Use Bonferroni Correction", *Ophthalmic and Physiological Optics*, **34**, 502-508.

- Breiman L, Friedman, J H, Olshen R A and Stone C J (1984), " Classification and Regression Trees", Chapman and Hall, USA

- Kass G V (1980), "An Exploratory Technique for Investigating Large Quantities of Categorical Data", *Applied Statistics*, **20**(2), 119-127.