GINI Index: Worked out Example

- The Gini Index measures the impurity of data set (D) as: -

$$\text{Gini(D)} = 1 - \sum_1^n p_i^2$$ Where, n = Number of classes, pi = Probability of ith class.

- It considers binary split for each attribute.
- When D is partition into D1 and D2 then Gini(D) = D1/D Gini(D1) + D2/D Gini(D2)
- The attribute that maximize the reduction in impurity is selected as splitting attribute.

Consider an example where target variable is binary, the summary table for such example will be similar to below table.

Target Variable Value	Count	%
Yes	346	74%
No	124	26%
Ovarall	470	100%

Gini Index = $1 - 0.74^2 - 0.26^2 = 1-0.5476 -0.0676 = 0.3848$

Nominal Target Variable: Worked out Example

When target variable is nominal variable with different levels, we can calculate Gini Index in a similar way. When target variable is nominal the summary table looks similar to the below table. In India, there are number of different cuisines available and people have different food preferences. We have considered a few options and the table below shows proportion of people with their food preferences. This is an illustrative example.

Target Variable Value	Count	%
South Indian	20	8%
North Indian	35	13%
Continental	75	29%
Other Cuisine	130	50%
Overall	260	100%

Gini Index $= 1 – (20/260)^2 - (35/260)^2 – (75/260)^2 –(130/260)^2$
$= 1 - 0.006 - 0.018 - 0.083 - 0.250$
$= 0.643$

GINI of a split

GINI (s, t) = GINI (t) – P_L GINI (t_L) – P_R GINI (t_R)

Where

- s : split
- t : node
- GINI (t) : Gini Index of input node t
- P_L : Proportion of observation in Left Node after split, s
- GINI (t_L) : Gini of Left Node after split, s
- P_R : Proportion of observation in Right Node after split, s
- GINI (t_R) : Gini of Right Node after split, s

GINI Index: Worked out Example

Example

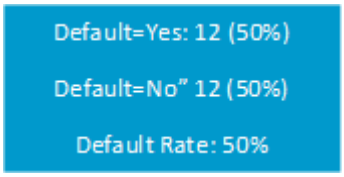
Example, banks and financial institutions grant credit facility after evaluating credit risk involved. Credit risk involved in credit decisions is evaluated using Credit Scorecard [Credit Score: What is it and how is it developed?]. Also, there are a few additional decisions involved in credit underwriting [Credit Underwriting: Minimize credit risk losses using Data Science and Analytics].

Decision Tree: Non Technical Explanation

The last 2 years of customer performance on meeting credit obligations is available with us. We want to understand the variable(s) explains high risk of customers who defaulted on a credit facility given to them.

The sample has 24 customers. And for making it simple, only customer age and gender are considered. Age is a continuous variable and Gender is nominal variable.

Input sample has 12 customers who have defaulted on the credit facility. So, default rate is 50%.



We want to understand if the customers with certain age group has higher chance of defaulting, or one gender has higher default rate than that of the other gender.

We have an example in which input node, parent node, has equal number of Target variable values- "Yes" and "No". Overall number of observations are 24.

Gender variable is considered to split the node. Gini Split value is calculated as below.

	Target Value		
Gender	**No**	**Yes**	**Total**
Female	6	2	8
Male	6	10	16
Total	**12**	**12**	**24**

Gini index for this node will be

GINI (t) = $1- (1/2)^2–(1/2)^2$

= 1- 0.25 -0.25

= 0.5

Now we want to split the code based on Gender Variable. After the split we will have following summary.
Now, let's calculate GINI index of the split using Gender variable.

GINI (s, t) = **GINI (t) – P_L GINI (t_L) – P_R GINI (t_R)**

GINI (t_L) = $1- (6/8)^2 – (2/8)^2$

= 1- 0.5625 -0.0625

= 0.375

GINI (t_R) = $1- (6/16)^2 – (10/16)^2$

= 1 - 0.140625 -0.390625

= 0.469

GINI (s, t) = 0.5 – (8/24)*0.375 – (16/24)*0.469

= 0.5 - 0.125 -0.313

= 0.0625

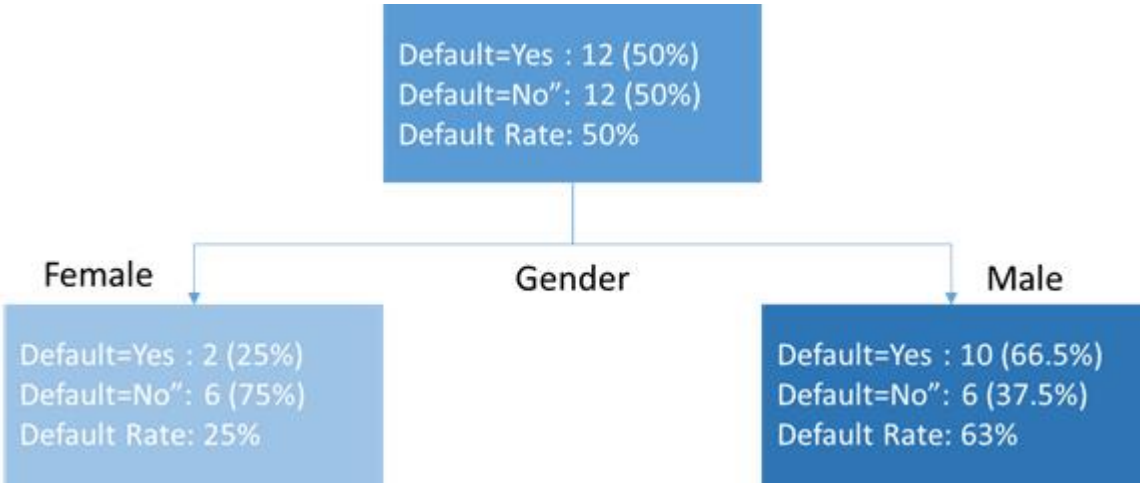
GINI Index: Worked out Example

Similarly, we need to find GINI index value for all the split points and select the best split for a variable. Also, the best split points are calculated for all the variables. The best variable and the split is selected to split the input node.

Decision Tree is one of the techniques which can help us answer these questions. Decision Tree process has to find the variable and cut off (for numeric and group values for nominal variables) to be considered for the split. The aim of the split will be to improve impurity (default rate) of the child nodes.

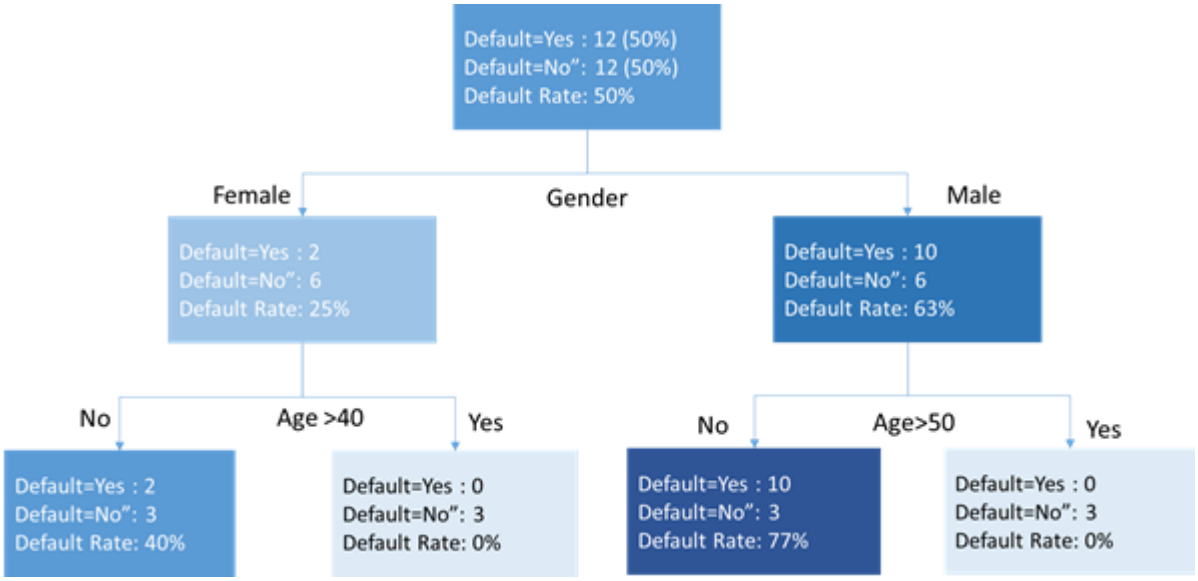
Based on exploratory analysis, we can see that Male group has higher default rate of 63% whereas Female group has 25%. Average age of default customers is around 39 years as compare to non-defaulting customers has average age of 47.

Decision Tree can help in find the cut of Age variable and interaction effect between Age and Gender. Also, if there are more number of variables, the efficacy exploratory data analysis in selecting the variables or finding association with target variable could be low.



In this example, Gender variable is selected for partitioning the input data sample. After split, there are two samples (or child nodes) – one for each Gender.

Impurity (or default rate) has increased for one child node to 63%. Now, each of these child nodes are further partitioned to improve the impurity. Since each child node undergoes same process of partitioning as their parent node, the process is called recursive partitioning.



GINI Index: Worked out Example

Left Node (Gender=Female) is partitioned based on Age>40 condition and Right Node (Gender=Male) is partitioned using Age >50 condition. In this example, Age is only variable so left and right node are partitioned using same variable but in reality all the input variables considered for each node, and the best variable and split point will be selected for each of the nodes.

Default rate for Male Customers who are aged below 50 is 77% compared to that of the customers who have 50% default rate.