Classification and Regression Tree (CART) is one of commonly used Decision Tree algorithms. In this post, we will explained the steps of CART algorithm using an example data.

Decision Tree is a recursive partitioning approach and CART split each of the input node into two child nodes, so CART decision tree is Binary Decision Tree. At each level of decision tree, the algorithm identify a condition - which variable and level to be used for splitting input node (data sample) into two child nodes.

CART Algorithm Steps

Decision Tree building algorithm involves a few simple steps and these are:

1. **Take Labelled Input data** - with a Target Variable and a list of Independent Variables
2. **Best Split**: Find Best Split for each of the independent variables
3. **Best Variable**: Select the Best Variable for the split
4. **Split the input data** into Left and Right Nodes
5. **Continue step 2-4** on each of the nodes until meet stopping criteria
6. **Decision Tree Pruning** : Steps to prune Decision Tree built

In the previous blog, we have explained the Gini Index calculation.

Data

We have considered a sample data for explaining the Decision Tree building process using CART algorithm. The data has a few variables and below is description of these variables.

Variable	Description
Card_Cust_ID	Credit card customer Identifier /ID Variable
Gender	Gender of Card Holder
Education_level	Education Level of Card Holder
Last_Month_spend	Spend in the latest Month
Last_3m_avg_spend	Average Spend in the last 3 Months
Spend_Drop_over50pct	Target Variable - Whether customer drop spend by over 50%

Data Sample

Parent Node has below split of Target Variable.

	100%	
T=1	125	26.3%
T=0	350	73.7%

Best Split for a Variable

We are considering Last Month Spend, Gender, Education and Last 3 month average spend for the decision tree building. Let's first consider variable "Last Month Spend" for finding out the "Best Split Cut Off" for this variables.

List of potential cut values have to be identified first. One of the common approach is to find splits /cut off point is to take middle values. For this variable, the distinct values and then finding middle points.

Below is an example of find unique values and then taking average value of two adjacent distinct values to find cut off point.

Unique	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200	210	220	230	240	250	260
Cut off Points	15	25	35	45	55	65	75	85	95	105	115	125	135	145	155	165	175	185	195	205	215	225	235	245	255	

For each of these "Cut off Point" , we need to calculate Gini Index and Gini Index for a Split. The Gini Index Calculations are explained with worked out example - here.

Cur Off (Mid Points)	15	25	35	45	55	65	75	85	95	105	115	125	135	145	155	165	175	185	195	205
Gini Index for Split	0.0005	0.0028	0.0008	0.0048	0.0113	0.0170	0.0275	0.0257	0.0247	0.0222	0.0173	0.0185	0.0146	0.0136	0.0148	0.0183	0.0151	0.0186	0.0176	0.0189

Now we need to find cut off which has the highest Gini Gain(highest drop in Gini Index for a split or highest GiniGain) and pick up the split point as "**Best Split Point**" for the variable "**Last Month Spend**".

Similarly, we need to find the best split for each of the input independent variables.

On the next page, we will explain steps to find best split for a character /categorical variable.