End Notes

Random forest gives much more accurate predictions when compared to simple CART/CHAID or regression models in many scenarios. These cases generally have high number of predictive variables and huge sample size. This is because it captures the variance of several input variables at the same time and enables high number of observations to participate in the prediction. In some of the coming articles, we will talk more about the algorithm in more detail and talk about how to build a simple random forest on R.

What is Random forest algorithm?

- Random forest algorithm is a supervised classification algorithm. As the name suggest, this algorithm creates the forest with a number of trees.
- In general, the more trees in the forest the more robust the forest looks like. In the same way in the random forest classifier, the higher the number of trees in the forest gives the high accuracy results.
- If you know the decision tree algorithm. You might be thinking are we creating more number of decision trees and how can we create more number of decision trees. As all the calculation of nodes selection will be same for the same dataset.
- Yes. You are correct. To model more number of decision trees to create the forest you are not going to use the same apache of constructing the decision with **information gain** or **gini index** approach.
- If you are not aware of the concepts of decision tree classifier, Please spend some time on the below articles, As you need to know how the decision tree classifier works before you learning the working nature of the random forest algorithm.

Basic Decision Tree Concept

- Decision tree concept is more to the rule based system. Given the training dataset with targets and features, the decision tree algorithm will come up with some set of rules. The same set rules can be used to perform the prediction on the test dataset.
- Suppose you would like to predict that your daughter will like the newly released animation movie or not. To model the decision tree you will use the training dataset like the animated cartoon characters your daughter liked in the past movies.
- So once you pass the dataset with the target as your daughter will like the movie or not to the decision tree classifier. The decision tree will start building the rules with the characters your daughter like as nodes and the targets like or not as the leaf nodes. By considering the path from the root node to the leaf node. You can get the rules.
- The simple rule could be if some x character is playing the leading role then your daughter will like the movie. You can think few more rule based on this example.

- Then to predict whether your daughter will like the movie or not. You just need to check the rules which are created by the decision tree to predict whether your daughter will like the newly released movie or not.
- In decision tree algorithm calculating these nodes and forming the rules will happen using the information gain and gini index calculations.
- In random forest algorithm, Instead of using information gain or gini index for calculating the root node, the process of finding the root node and splitting the feature nodes will happen randomly. Will look about in detail in the coming section.
- Next, you are going to learn why random forest algorithm? When we are having other classification algorithms to play with.