Classification and Regression Trees

The Classification and Regression Tree methodology, also known as the CART was introduced in 1984 by Leo Breiman, Jerome Friedman, Richard Olshen and Charles Stone.

Decision Trees

- Decision tree algorithms are nothing but if-else statements that can be used to predict a result based on data.
- A decision tree is a supervised machine learning algorithm. It has a tree-like structure with its root node at the top.

Classification Trees

- A classification tree is an algorithm where the target variable is fixed or categorical.
- The algorithm is then used to identify the “class” within which a target variable would most likely fall.
- An example of a classification-type problem
- determining who will or will not subscribe to a digital platform;
- or who will or will not graduate from high school.
- to predict among a number of different variables.
- to predict which type of smartphone a consumer may decide to purchase.

Regression Trees

- A regression tree refers to an algorithm where the target variable is continuous and the algorithm is used to predict it’s value.
- As an example of a regression type problem, you may want to predict the selling prices of a residential house, which is a continuous dependent variable.
- This will depend on both continuous factors like square footage as well as categorical factors like the style of home, area in which the property is located and so on.

Basic Differences between Classification Tree & Regression Tree

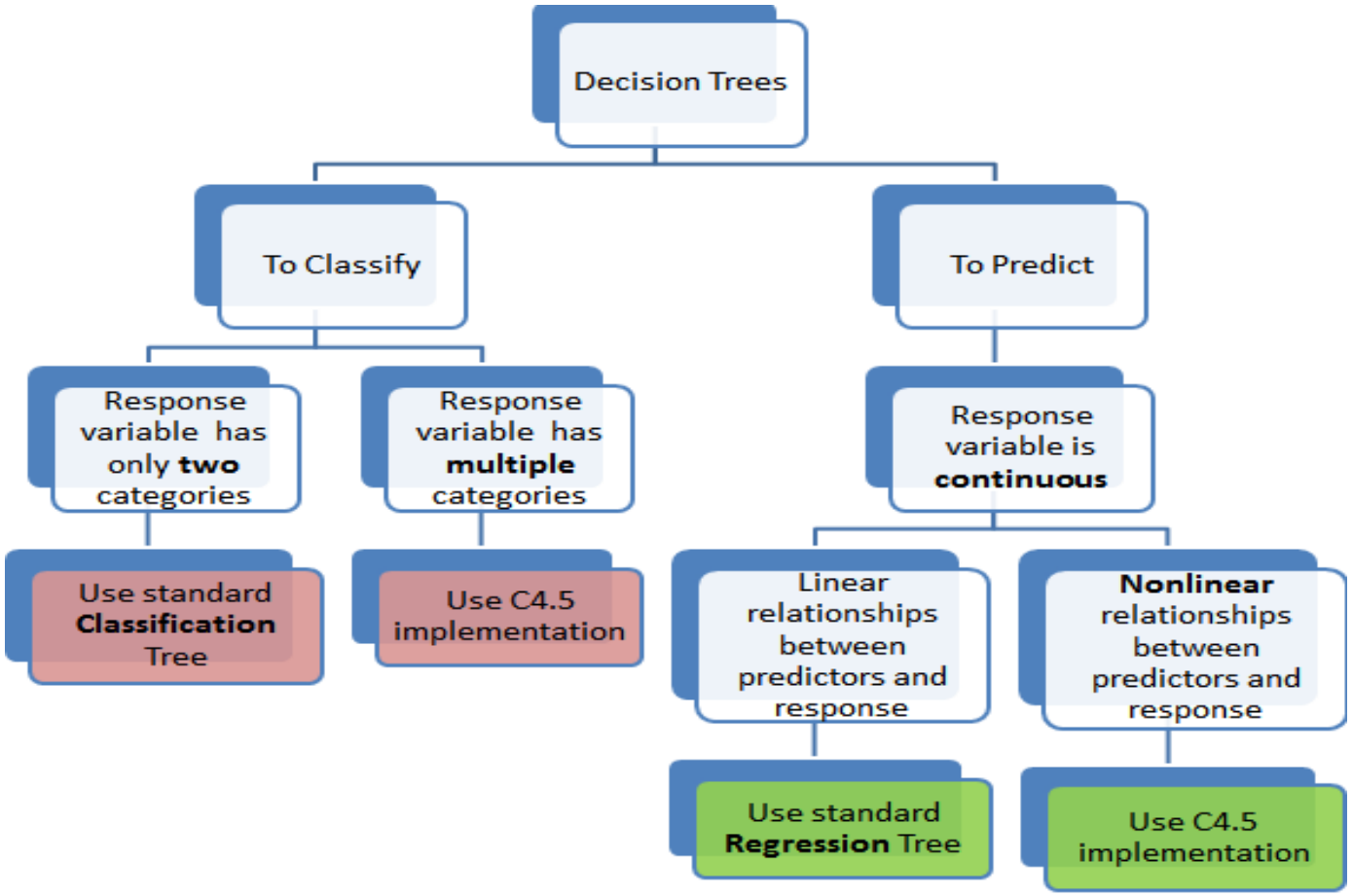
- A classification tree splits the dataset based on the homogeneity of data. Say, for instance, there are two variables; income and age; which determine whether or not a consumer will buy a particular kind of phone.
- If the training data shows that 95% of people who are older than 30 bought the phone, the data gets split there and age becomes a top node in the tree. This split makes the data “95% pure”. Measures of impurity like entropy or Gini index are used to quantify the homogeneity of the data when it comes to classification trees.

- In a regression tree, a regression model is fit to the target variable using each of the independent variables. After this, the data is split at several points for each independent variable.
- At each such point, the error between the predicted values and actual values is squared to get “A Sum of Squared Errors”(SSE). The SSE is compared across the variables and the variable or point which has the lowest SSE is chosen as the split point. This process is continued recursively.

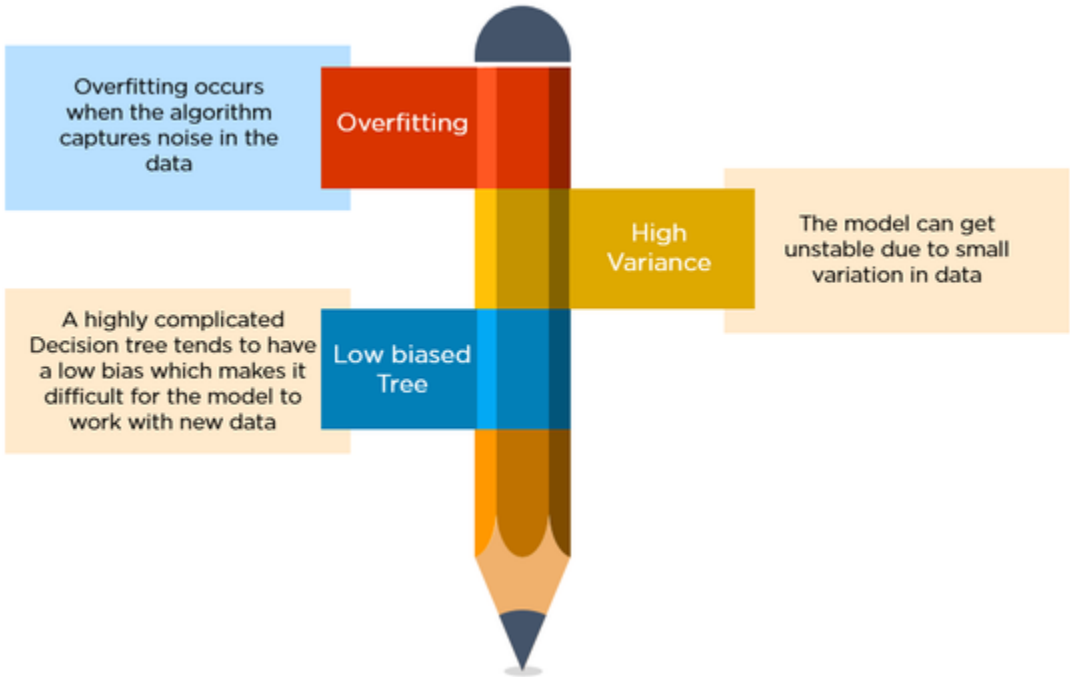
Advantages of Classification and Regression Trees

- The purpose of the analysis conducted by any classification or regression tree is to create a set of if-else conditions that allow for the accurate prediction or classification of a case
- The results are simplistic
- Classification and Regression Trees are Nonparametric & Nonlinear
- Classification and Regression Trees Implicitly Perform Feature Selection

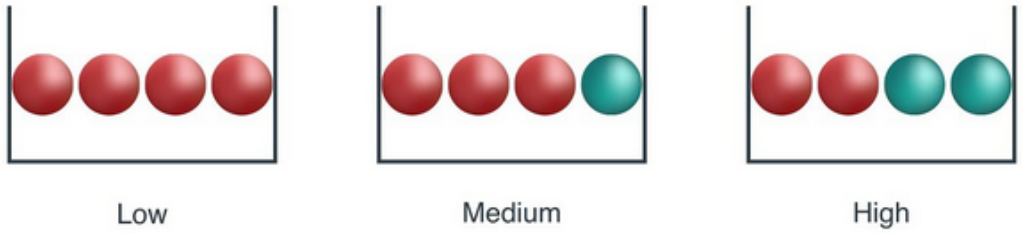
CART


Decision Trees
To Classify
To Predict
Response variable has only **two** categories
Response variable has **multiple** categories
Response variable is **continuous**
Use standard **Classification** Tree
Use C4.5 implementation
Linear relationships between predictors and response
Nonlinear relationships between predictors and response
Use standard **Regression** Tree
Use C4.5 implementation

Limitations of the CART



Types of Impurity



- In this image, a ball is randomly drawn from each bowl.
- So how much information you needed to accurately tells the color of ball.
- So, left bowl needed less information as all of the ball is red colored, central bowl needed more information than left bowl to tell it accurately, and right bowl needed maximum information as both number of both color ball are same.
- As information is measure of purity, so we can say that left bowl is pure node, middle is less impure and right is more impure.

Measures of impurity

- Entropy
- Gini index/ Gini impurity

- https://medium.com/@ankitnitjsr13/math-behind-decision-tree-algorithm-2aa398561d6d

Entropy & Gini Index

- Entropy is amount of information is needed to accurately describe the some sample.
- So if sample is homogeneous, means all the element are similar than Entropy is 0, else if sample is equally divided than entropy is maximum 1.
- So, left bowl has lowest entropy, middle bowl has more entropy and right bowl has highest entropy.

- Gini index is measure of inequality in sample.
- It has value between 0 and 1.
- Gini index of value 0 means sample are perfectly homogeneous and all element are similar, whereas, Gini index of value 1 means maximal inequality among elements.
- It is sum of the square of the probabilities of each class.

CART

- It is used for generating both classification tree and regression tree.
- It uses Gini index as metric/cost function to evaluate split in feature selection in case of classification tree.
- It is used for binary classification.
- It use least square as a metric to select features in case of Regression tree.