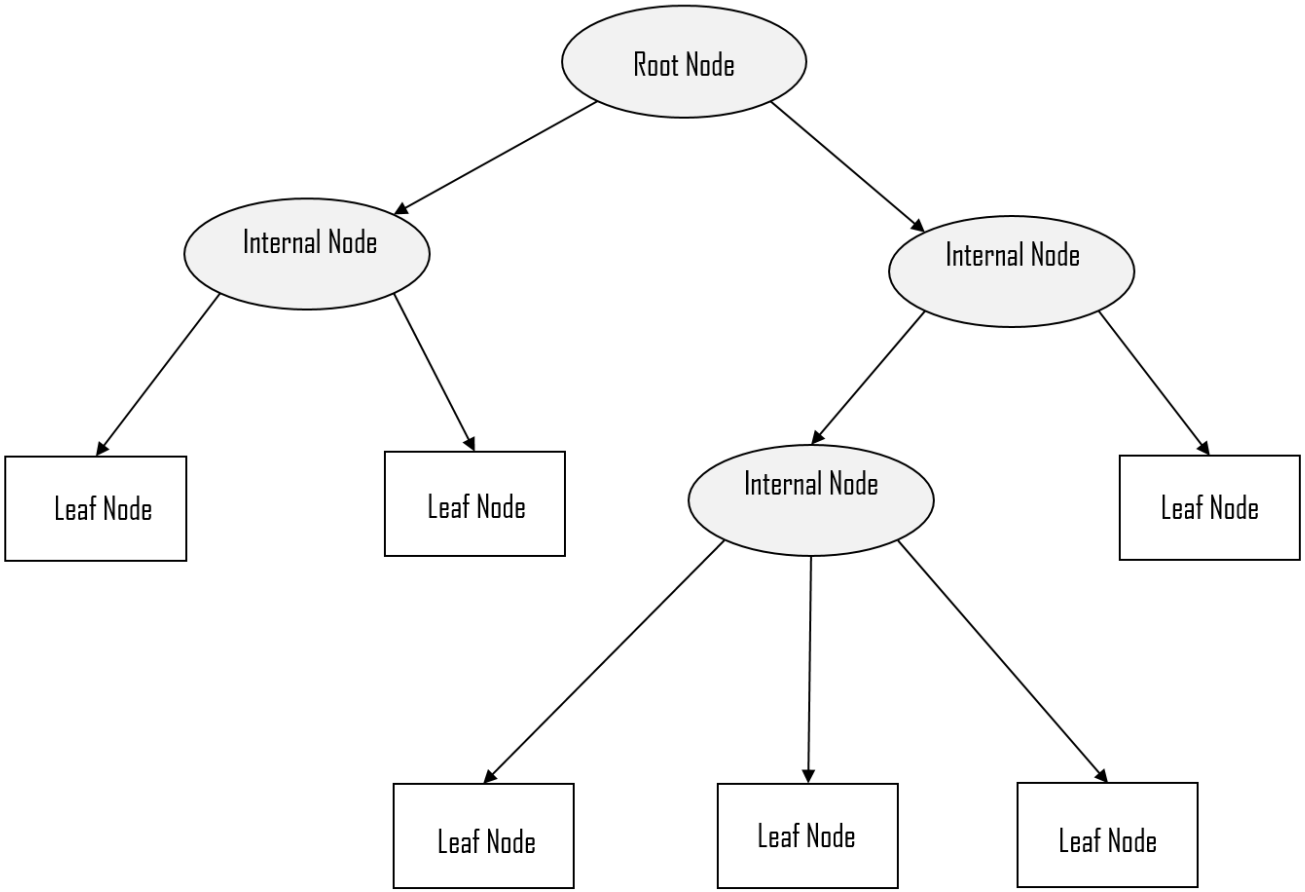
Regression Trees

Regression Trees

Linear regression is the most common probability model for regression settings, even though it can be considered a non-probability model. The most common non-probability model is a regression tree.

A **regression tree** is a series of if-then decision rules that results in binning the data into several terminal groups. The estimate for that group is just the mean of the group.

Regression Trees



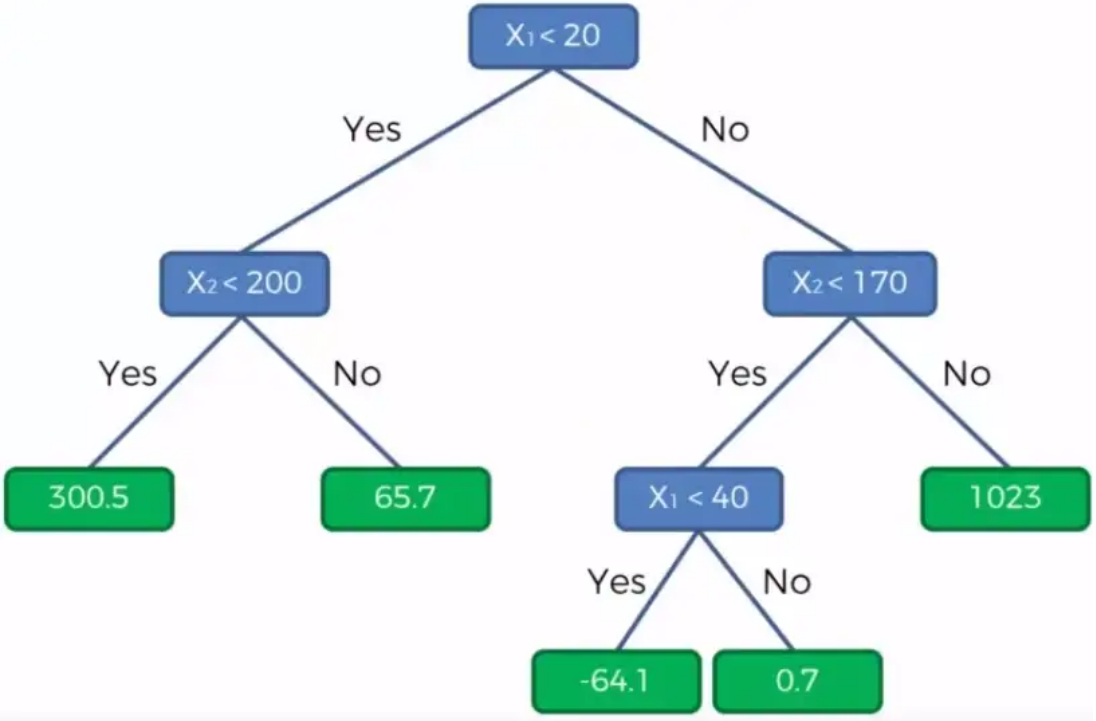
Regression Trees

Terms explained:

- At the **root node**, all the data is in one large group. The first decision rule is made and the data is split. If it meets the decision rule it is split to the left, if not it goes to the right.
- An **internal node** only contains some subset of the data but it will also split the data like the root node did.
- A **leaf node** is a node where no splits are made on the data that resides there. Also called a **terminal node**

Regression Trees

There is a target variable Y and predictors, $X_1, X_2, \ldots$. Where the splits happen depend on the predictors and will try to split in such a way that every leaf node has data where the target is similar.



Estimating the splits

1. On the full data, order the values of the predictor from least to greatest
2. For every possible split, meaning between every successively increasing predictor value, we split the data. This creates two groups for the data.
3. For every split, calculate the squared error loss from the two groups, using the mean of the group as the estimate. For example, the whole group might have a squared error loss of 100, but by splitting, group 1 has a squared error loss of 90 and group 2 has a squared error loss of 75
4. Take the weighted average of the squared error loss for the two groups.
5. Choose the split with the smallest squared error loss.
6. Repeat for all internal nodes

Estimating the splits

How does it know when to stop? You can dictate the stopping criteria in a few simple ways:

- Specify a max depth of the tree
- Specify the minimum number of observations in a node to be able to split

There is one more we will talk about later in class.