(Gradient) Boosted Trees Explained

Regression tree step by step example:

First, fit a single decision tree to your data using recursive binary splitting. (typically a small tree like a single split “stump”)

Second, calculate predicted values for each observation in your training data.

Third, calculate residuals for these predicted values.

Fourth, fit a new tree to the residuals of the first tree.

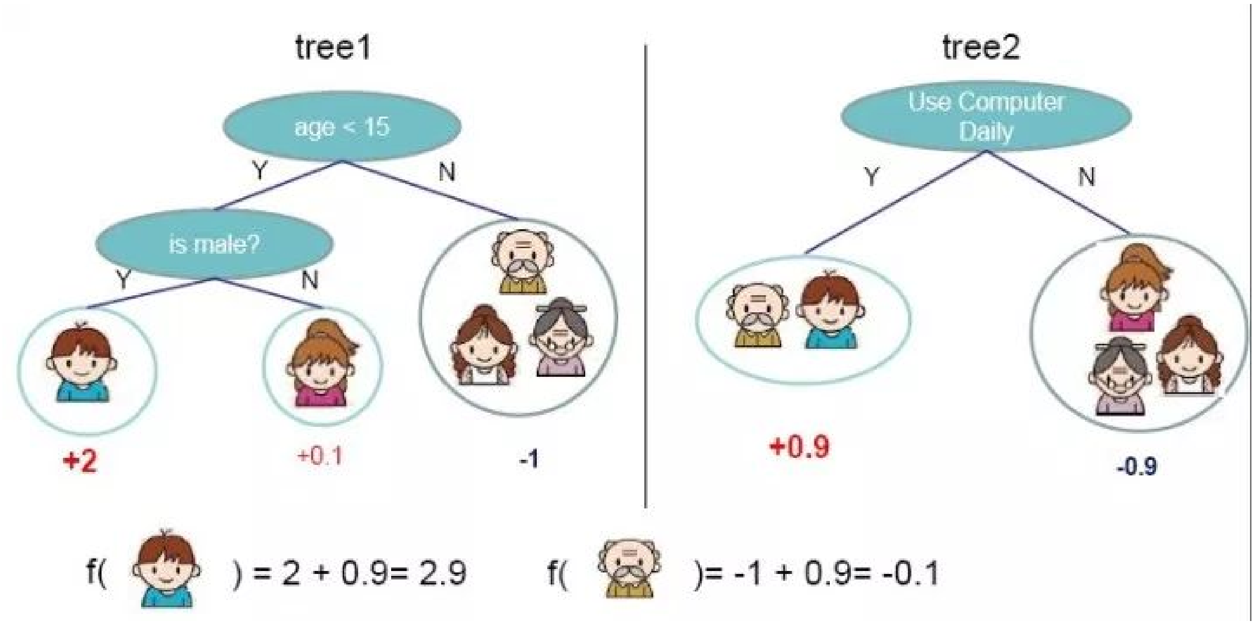
> *Note: residuals=yi for this model, so the model’s goal is to minimize rss using xi to explain the residuals. Explanatory variables that covary with the residuals will explain and therefore reduce prediction error in the previous step.*

Fifth, use predicted values from the tree fit to residuals to update predicted values from first tree.

i.e. - New Prediction= Prediction from tree #1 + prediction from tree fit to errors

(In practice we shrink each predicted error by a parameter like .01 or .001)

Link to image example thus far.
https://raw.githubusercontent.com/dmlc/web-data/master/xgboost/model/twocart.png



Next, we take the predicted values summed together from previous trees, calculate errors (i.e.-the distance from these predicted values to the original values of yi and fit a tree to these new errors

Step by step:

A. New prediction = Prediction from tree #1 + prediction from tree fit to errors
B. Original Y-New prediction= new residuals
C. Fit tree to new residuals
D. Add predicted residuals from this tree to “New prediction above” to create newest prediction

II. Repeat until:

You build a sequence of trees that predicts new data best. (use cross validation!)