Notes on XGBoost

* Start
  + Lets understand that XGBoost was made for large and complicated data sets
  + It was also made with speed and runtime in mind
* How does XGBoost do regression
  + Step one is to make an initial prediction which by default is .5
  + The difference between the prediction and the actuals will give us an idea on how well the data was fit
  + One big difference in XGB from un-extreme gradient boost is in the regression tree
    - XGB uses a different method
  + Each tree starts out as one single leaf
    - We calculate a similarity score which is
    - Graphical user interface, text

      Description automatically generated
  + Lamda will be a regulation
  + Next is can we do better
    - This is where we will start to split things into multiple groups
  + Then it splices the data based on the x-axis based on its average and it splits that making it its own leaf
  + Diagram

    Description automatically generated with medium confidence
  + This is based on the residule or the distance from the line
  + Next we will calculate the gain
    - Gain = similarity left +similarity right – the root (score at the top of the two leafs)
    - Graphical user interface, text, application

      Description automatically generatedDefault leaf depth is 6
* Prunting a tree
  + This where gamma is selected for example we will use 130
  + A branch is removed if the root gain is less than the gamma starting at the lowest branch
* Graphical user interface, text, application, email

  Description automatically generated
* Lamda helps account for trees with a low amount of residules in it
* Lamda helps prevent overfitting training data by ensuring more trees get pruned
* Text

  Description automatically generated with low confidence
* Calculate the output values
* Timeline

  Description automatically generated
* This gives us the new prediction which is the original prediction +(learning rate\* output)
* Diagram

  Description automatically generated
* Lastly take the new resuldes and build another tree until we have small resuldes or we are done
* Classification
  + Start with a prediction like .5
  + Different formular for calculating similarity score for classification
  + The top is the same so it is still the sum of residules squared
    - The bottom is: the sum of the pervious probablty \* ( 1- previous prob) + lamda (used for overfitting)
  + Start with a leaf with all the residules
  + Split them into groups to see if we can do better for simulairty score
  + Previous prob = previous leaf (root)
  + Graphical user interface, application

    Description automatically generated
  + Gain still equals the same thing
  + Cover is the miniumn number of residules for each tree and by default is 1 in regression
    - In classification cover =
    - Graphical user interface, application

      Description automatically generated
    - Cover helps with pruning min\_child\_weight = cover
    - Output values = Diagram

      Description automatically generated
    - Next we convert to log odds after the tree is build
    - A picture containing timeline

      Description automatically generated
    - Next add the log odds of the original prediction + learning rate eta .3 \* output value – we the plug in our log prediction into a logistic regress prob
    - Graphical user interface, text, application

      Description automatically generated
  + Keep them going until you get new trees and new residules
* Math details
* The only difference between these two methids is the loss function
* Loss function
  + This tells us how good the prediction is
  + General loss function for regression
  + Diagram

    Description automatically generated
  + The loss function is stored and used to determine if our predictions are improving or now
  + General loss function for classification
  + Diagram, text

    Description automatically generated
* Goal
  + Minimize this equation
  + Diagram, text, application

    Description automatically generated
  + T is the number of terminal nodes (leaves) and gamma is a user definded penalty meant to encourage pruning
* Diagram

  Description automatically generated
* 6:24
* <https://www.youtube.com/watch?v=ZVFeW798-2I>