Boosting: GBDT

What are Gradient Boosted Decision Trees (GBDT)?

- Fits a decision tree on the residual error of the previous tree.
- So, each new tree in the ensemble predicts the error made by the previous learner

How to build and train a GBDT?

Input: training set $\{(x_i, y_i)\}_{i=1}^{n}$, a differentiable loss function $L(y, F(x))$, number of iterations M.

Algorithm:

1. Initialize model with a constant value:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma).$$

2. For m = 1 to M:
   1. Compute so-called *pseudo-residuals*:

   $$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, n.$$

   2. Fit a base learner (or weak learner, e.g. tree) closed under scaling $h_m(x)$ to pseudo-residuals, i.e. train it using the training set $\{(x_i, r_{im})\}_{i=1}^{n}$.
   3. Compute multiplier γ_m by solving the following one-dimensional optimization problem:

   $$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right).$$

   4. Update the model:

   $$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x).$$

3. Output $F_M(x)$.

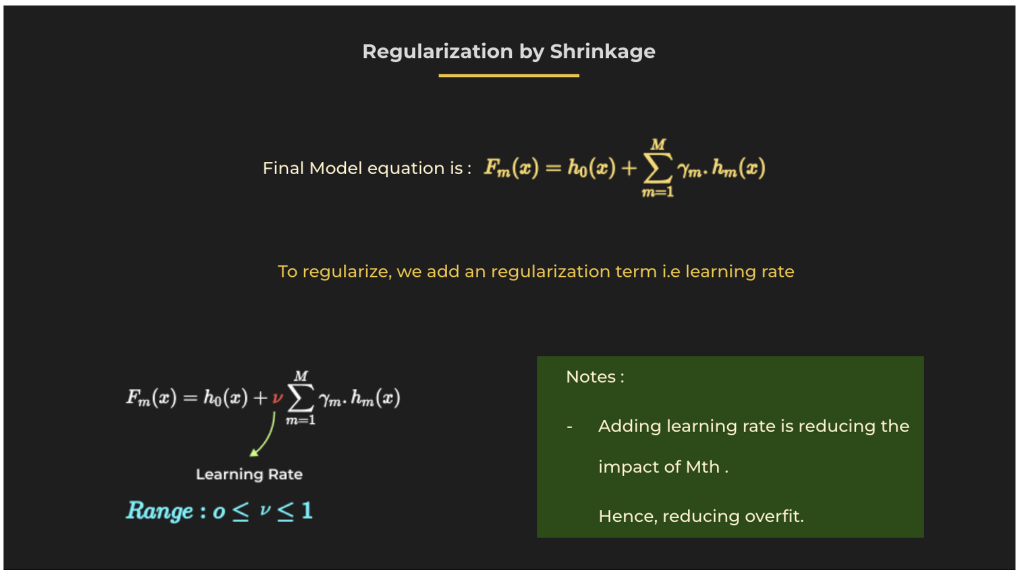
Bias Variance tradeoff

- M (number of base learners)
  - M increases → model overfits (low bias, high variance)
    - As base learners increases, GBDT starts to capture more complex relationships in data which leads to low bias
    - If data has outliers/noise, with increased base learners, GBDT starts to capture them leading to high variance
  - M decreases → model underfits (high bias, low variance)
    - GBDT Predictions becomes closer to the mean model
- Depth
  - Depth increases → model overfits (low bias, high variance)

- As deeper tree models tend to capture more complex patterns in data, it reduces the bias
- A deeper tree model can overfit on noise/outliers in data leading to high variance

- Depth decreases → model underfits (high bias, low variance)
  - As GBDT will fail to learn the basic pattern of the data

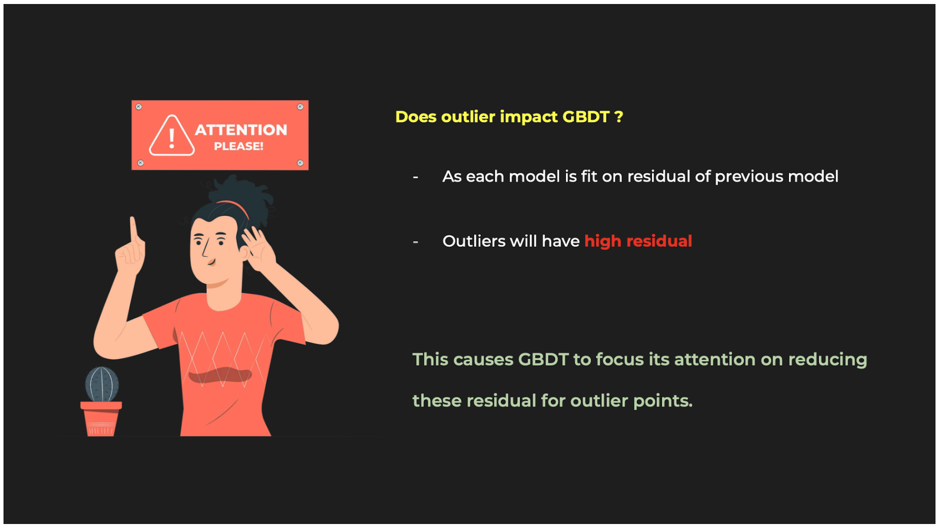
Regularization by shrinkage



Stochastic Gradient Boosting

- To reduce the GBDT overfitting issue, randomness (row and column sampling) is used
  - Sklearn provides Row sampling with **subsample** hyperparameter
  - and Column sampling using **max_features** hyperparameter

GBDT: Outlier issue



- Due to the MSE loss function, the residuals of the outlier will be exponentially high
    - Leading GBDT to overfit on these outliers

To tackle outlier issue, GBDT changes to a new loss function → **Huber Loss**