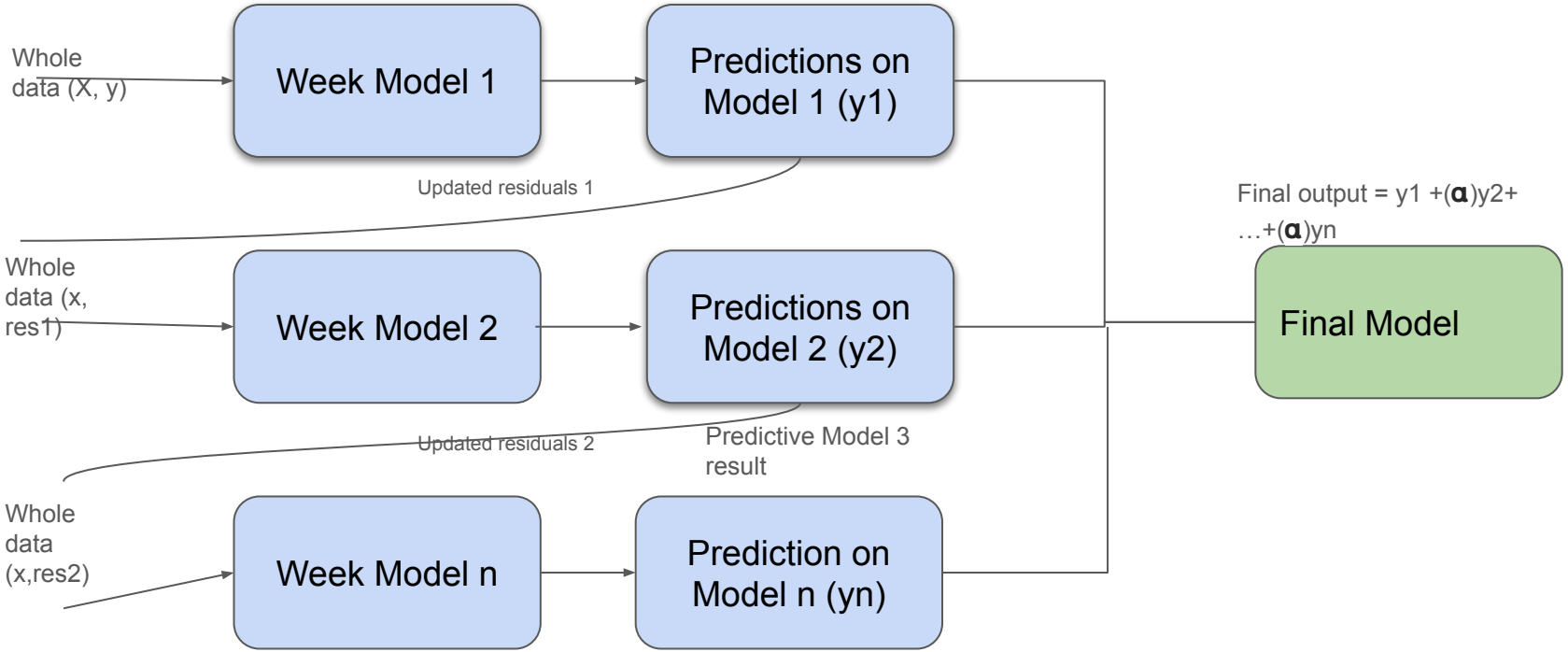
XGBoost Algorithms

Detailed presentation

What is XGBoost (Xtreme Gradient Boosting) Algorithms



Week Model 1,2,..n are individual models

Things to understand before XGBoost Detailed Steps

1. λ - Regularization parameter
2. ɣ - Threshold that defines your auto pruning or econtrols overfitting
3. η or **α** (alpha or eta)- learning rate - this helps how fast you want to converse to the next value

Example Dataset:

Exp	Gap	Salary (y)
2	yes	40k
2.5	yes	41k
3	no	52k
4	no	60k
4.5	yes	62k

Detailed Steps

1. Get output from base model by taking average of the independent variable.
   a. (40+ 41+ 52+60+62)/ 5 => 51.
2. Calculate residue based the average ie.51

Exp	Gap	Salary (y)	Residuals or errors
2	yes	40k	-11
2.5	yes	42k	-9
3	no	52k	1
4	no	60k	9
4.5	yes	62k	11

Detailed Steps

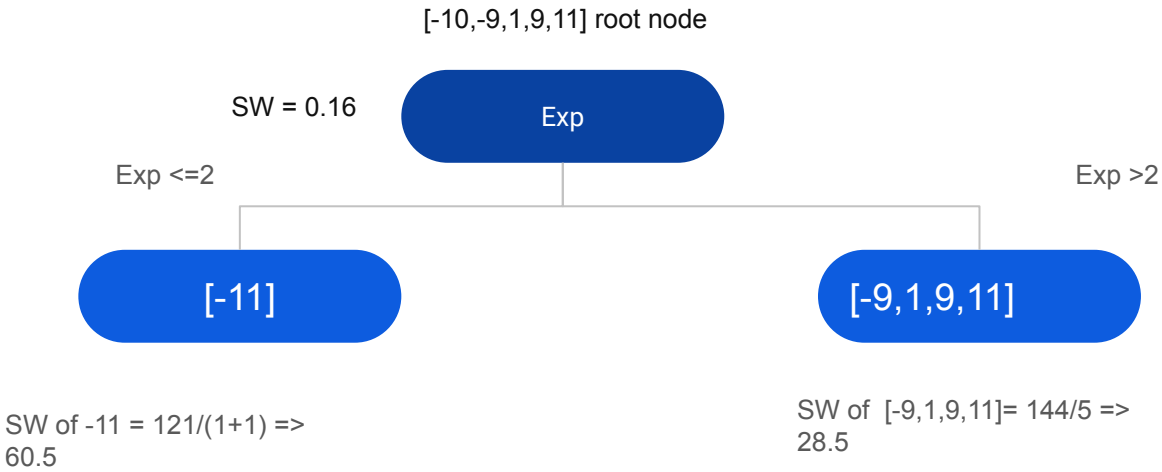
3. Calculate Similarity Score based on residual

Similarity Weights (SW) = Sum of Residuals ^{2}/ No of Residuals + λ

As per the above formule , consider λ = 1, the result is SW of root node , $(-11-9+1+9+11)^2/(5+1)$ = 1/6= 0.16 (This is the similarity score at the particular node) . Note: if λ

Value increases the similarity weight decreases.

4. Let's split the criteria based on the input values



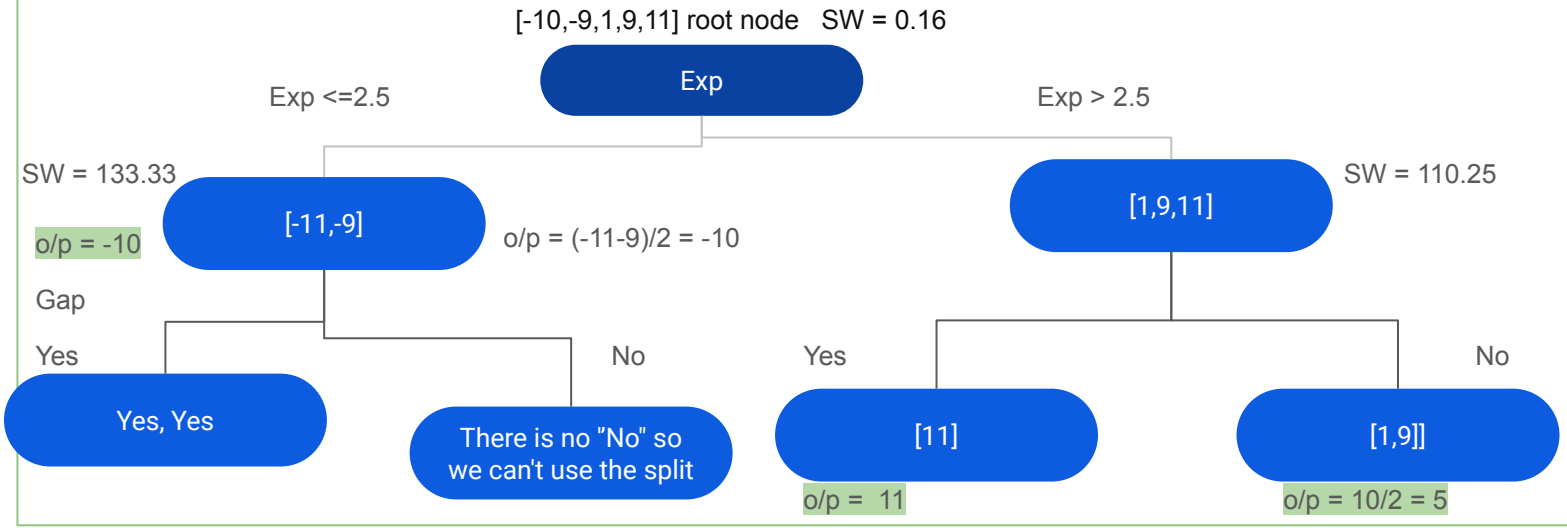
Detailed Steps

5. Calculate Gain to get max no of information by creating different decision tree and pick the trees which has highest gain for the model

Gain = SW of left and right nodes - SW of root node

Gain = (60.5+28.5) - 0.16 = **88.84** here is the total gain we get in the current split

Let's do the same for different split.Here the Gain is (133.33+ 110.25) - 0.16 = **243.2 this is greater than gain of previous split so we can pick the split which got high gain i.e 243.2. Continue splitting the tree with next categorical variable i.e Gap. Finally we got outputs from the splits. Left split = -10 and right split (11 and 5). Note the gain > gamma i.e** (ɣ) then the split will happen otherwise no split



Detailed Steps

8. Calculate output of the final split of exp variable by taking average

O/p of Left split = (-11-9)/2 = -10 and O/p of right splits = 11 => 11 and (10)/2 => 5

9. Output calculation of a given input

Output = base model or model 1 output + learning rate * model 2 output+....

Here we have one decision tree so lets say the exp is 2 then as per the previous tree the o/p is -10. ex: learning rate is 0.5

Prediction = 51 + (0.5) * -10 = 51-5 = 46 but actual output of exp 2 is 40k

Calculate o/p for all the remaining inputs

Detailed Steps

Exp	Gap	Salary (y)	Residuals 1 or errors	Outputs	Residuals 2 or error 2
2	yes	40k	-11	46	-6
2.5	yes	42k	-9	46	-4
3	no	52k	1	53.5	-1.5
4	no	60k	9	53.5	6.5
4.5	yes	62k	11	56.5	5.5

Now, we can create next decision tree based on exp and gap with residuals 2 . and goes on.. Until we get zero errors or the parameter conditions met

The final prediction = base model o/p + α1(M1 of o/p) + α2 (M2 of o/p)+ …. + αn (Mn of o/p)

Hyper Parameter Tuning

1. Gamma - it is available in booster parameter.

How to know, if we need to split the tree or not. That is decided with help of one more parameter called gamma (ɣ). This is how it helps in auto pruning and help to reduce overfitting (i.e accurate predictions on training data, but inaccurate predictions on new data)

If gamma < gain,

then split will happen

Else

No split will happen

2. Let's say λ = 2, then the gain got reduced.. Note: if you increase λ value then you are pruning the tree aggressive way . similarly if gamma value is high then also you are pruning the tree in aggressive way

Also λ value increases , the similarity score will be reduced so it helps us to reduce the effect of outliers on prediction

3. Other parameters are max_depth, learning_rate, _estimators, objective