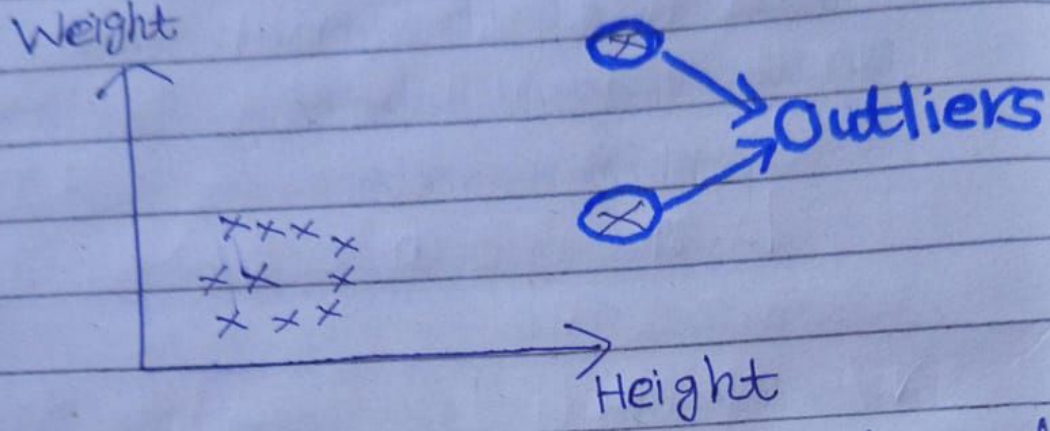
⇒ Anomaly Detection (To detect outliers)



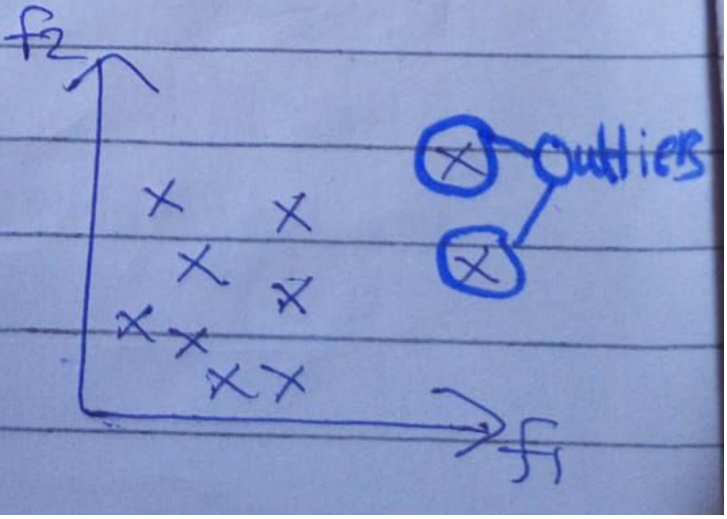
⇒ When outliers are very important for a specific problem statement then we use Anomaly Detection
e.g from IPs list which IPs are hacking IPs

⇒ Techniques of Anomaly Detection

① Isolation Forest:

- It uses Decision Trees

f_1	f_2	f_3	f_4
—	—	—	—
—	—	—	—
—	—	—	—
—	—	—	—



$E(h(x)) \Rightarrow$ Average search depth for x from all the isolated trees.

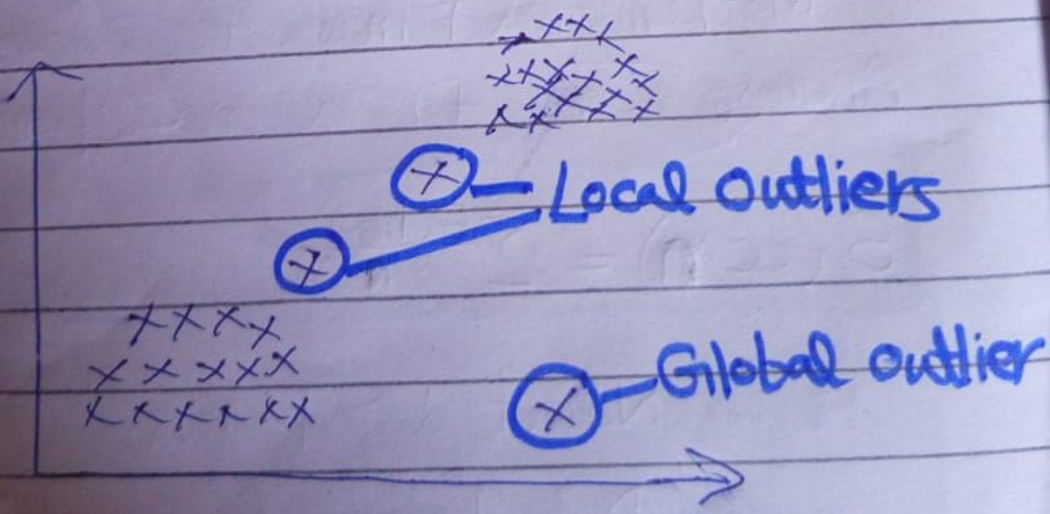
$c(n) \Rightarrow$ Theoretical Average of [illegible] path length in a BST with n points

If
$E(h(x)) << c(n) \Rightarrow S(x,n) \approx 1 \Rightarrow$ Anomaly Score $\Rightarrow$ outlier

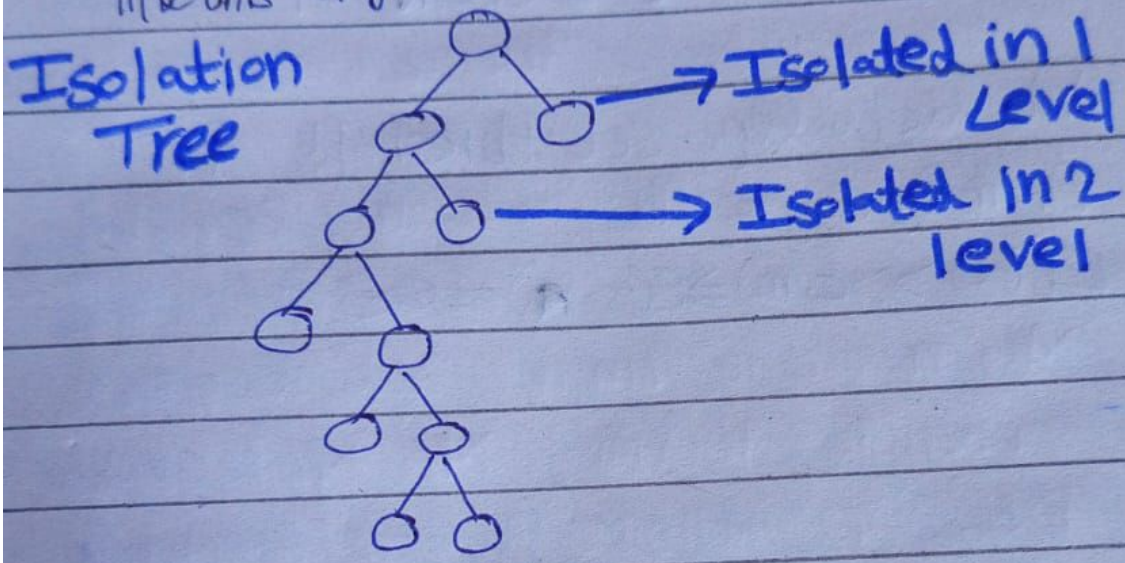
We actually set threshold for Anomaly score
$\gtrsim >= 0.7$ etc.

$Eh(x) >> c(n) \Rightarrow S(x,n) \approx 0.5 \Rightarrow$ Anomaly Score $\Rightarrow$ Not outlier

$\Rightarrow$ Local Outlier Factor Anomaly Detection:



- In Isolation Forest, we internally make split and make decision trees and as early as a datapoint isolated means become leaf node they are Outliers. We create multiple Isolation trees like this using different features



Anomaly Score:

Mathematical Formula for calculating Anomaly score for a new point

$$S(x,n) = 2^{-\frac{E(h(x))}{c(n)}}$$

$n \Rightarrow$ No. of data points

$x \Rightarrow$ Data point

① **Local Outlier:** When datapoints are somewhere near to clusters

② **Global Outlier:** When datapoints are not near to clusters

It uses **K Nearest Neighbour** internally

⇓

Local Density

- If we have to check that a particular point is Local outlier we will spcify value of K i.e K=5 we will find K-nearest neighbours and find average distance (inverse to local density) and then we move to those neighbours and find average distance (local density inverse). If the succeeding average distance is less i.e local density is high then we will consider that point as outlier (Local)

This is something called LOF (Local Outlier Factor) score