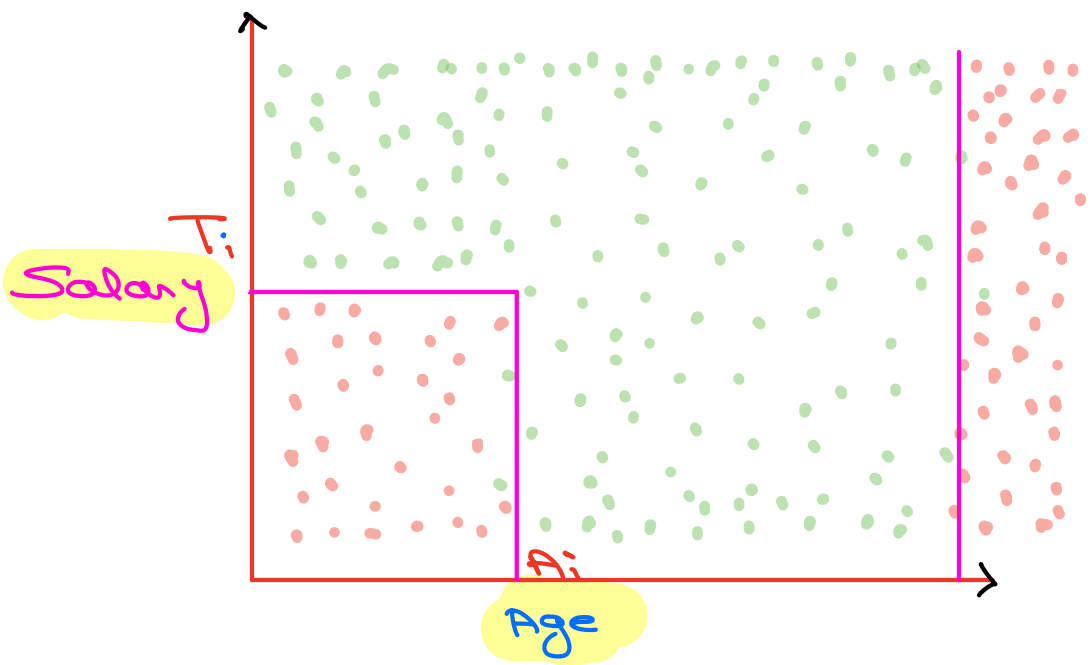
Decision Tree - 2

Agenda

- Recap
- Gini impurity
- Spliting on Numerical feature
- Imbalanced Data
- Feature Scaling
- Feature importance
- DT Regression

Recap

```
if salary < T:
    if age < A:
        P = +1
    else
        P = -1
else
    .....
```



① Decision tree splits Data into Homogeneous regions using Axis parallel Hyperplanes.

② DT is easily interpretable

1) How do we decide which feature and value to split on

2) Target: Pure / Homogenous Node

3) How do we calculate purity?

$$Entropy(H(y)) = -\sum_{i=1}^{k} P(y_i) \log P(y_i)$$

where k classes

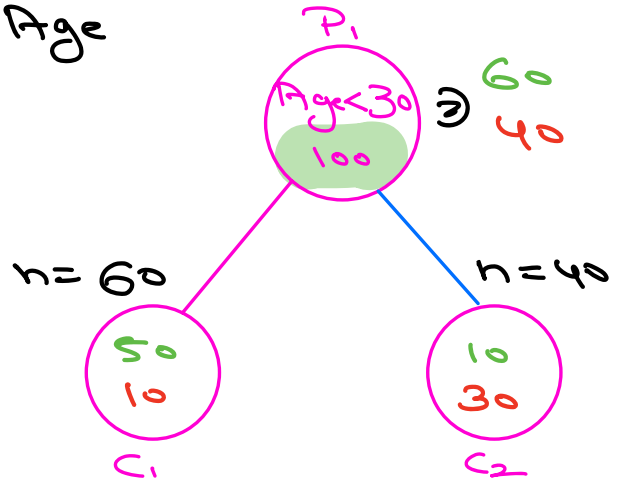
4) For Binary class : k = 2

P or 1 − P

$$H(y) = -(P \times \log P + (1-P) \times \log(1-P))$$

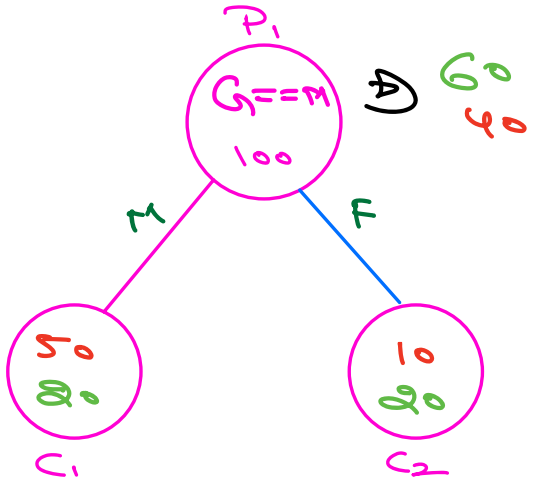
F1 = Age

Ex:



$H(C_1) = 0.65$ $H(C_2) = 0.81$

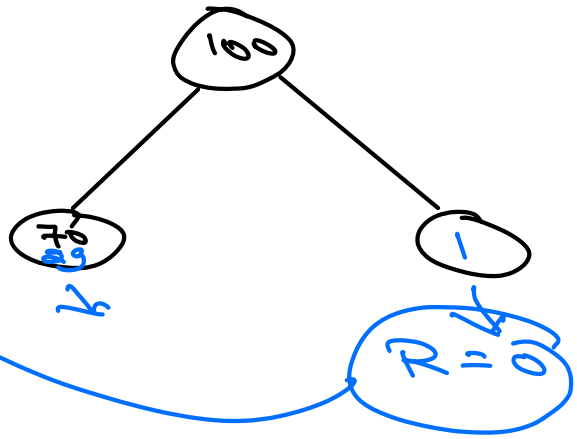
F2 = Gender



$H(C_1) = 0.86$ $H(C_2) = 0.97$

Total Entropy for Each Split

Why not $\frac{C_1 + C_2}{2}$



$$\frac{n_1}{n_{Total}} \times H(C_1) + \frac{n_2}{n_{Total}} \times H(C_2)$$

$F_1 \Rightarrow 0.81$ $\quad$ $F_2 \Rightarrow 0.67$

Information Gain
(Reduction in Entropy)

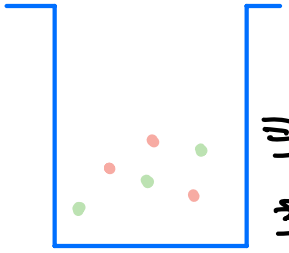
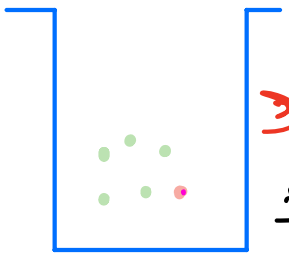
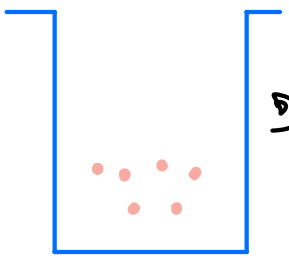
$H(Parent) - H(Child)$

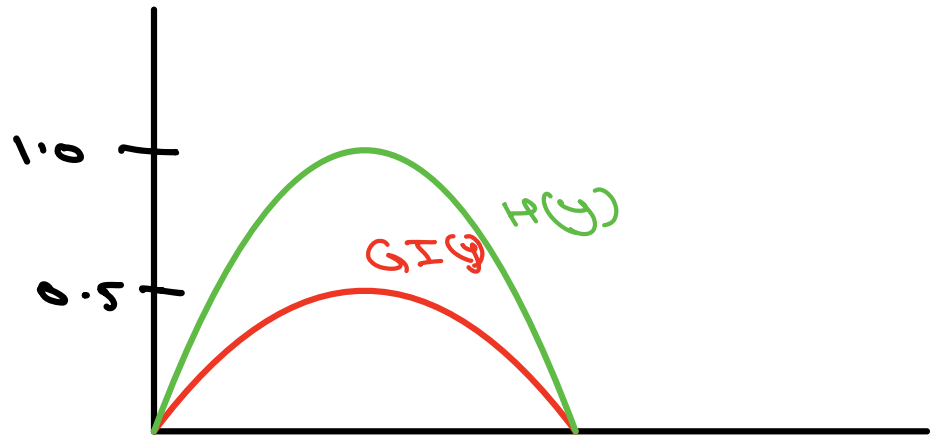
Limitation of Entropy

- Lot of Calculation involve Log

Gini Impurity

$$GI(y) = 1 - \sum_{i=1}^{k} (P(y_i))^2$$

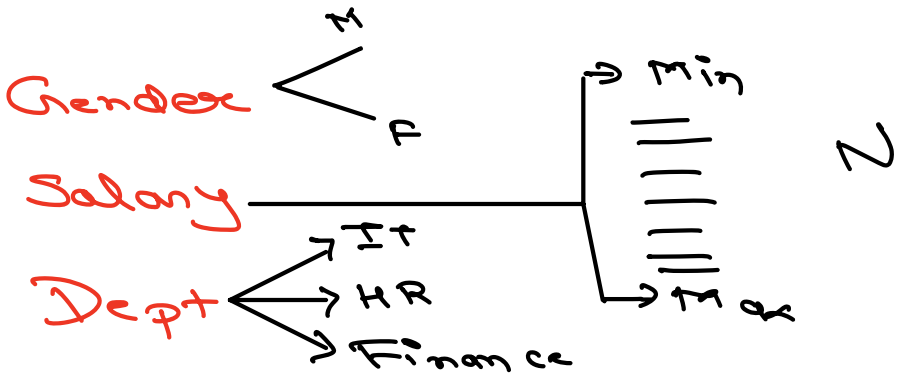
	$H(y)$	$GI(y)$
	$= 0.5 \times \log(0.5) + 0.5 \log(0.5)$ $= 1$	$1 - ((0.5)^2 + 0.5)^2$ $1 - ((0.25) + 0.25)$ $= 1 - 0.5 = 0.5$
	$= \frac{1}{6} \times \log \frac{1}{2} + \frac{5}{6} \times \log(\frac{5}{6})$ $= 0.6$	$1 - \left(\left(\frac{1}{6}\right)^2 + \left(\frac{5}{6}\right)^2\right)$ $= 0.27$
	$= 0$	$= 1 - (1^2 + 0^2)$ $= 0$



F_1 and F_2

$w(G(y)$ 0.32 $w(G(y_2)$ 0.24

Splitting on Numerical

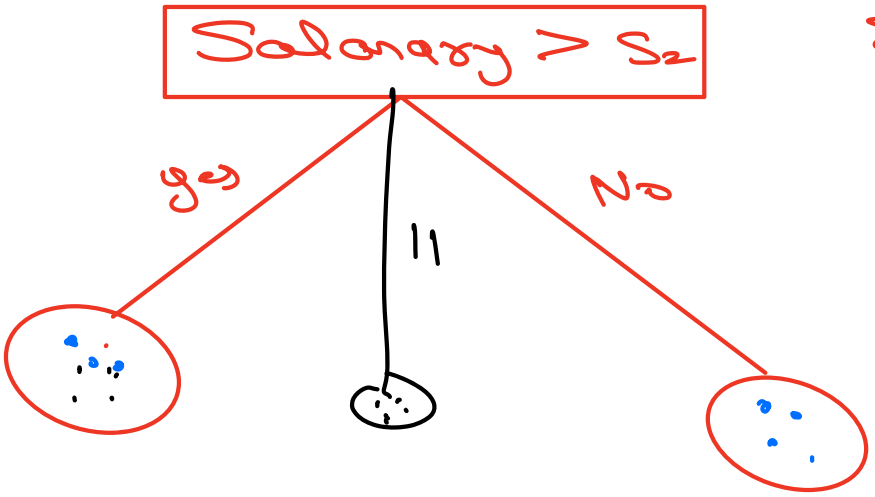


Brute Force

Step1: Get all Unique Values of Salary
$[S_1\ S_2\ S_3\ \cdots\ S_n]$

Step2: Calculate IG for each Salary Value as threshold
$[IG_1, IG_2\ IG_3\ \cdots\ IG_n]$

Step3: Argmax of IG



* Problem with this approach

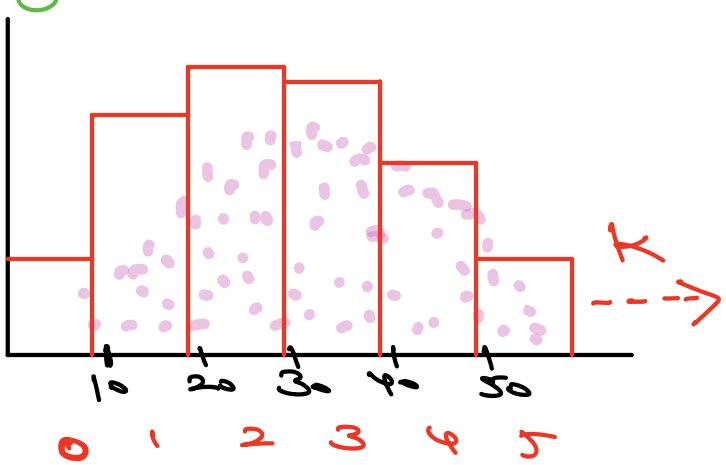
d numerical feature

n rows

$O(n \times d)$

How do we solve this

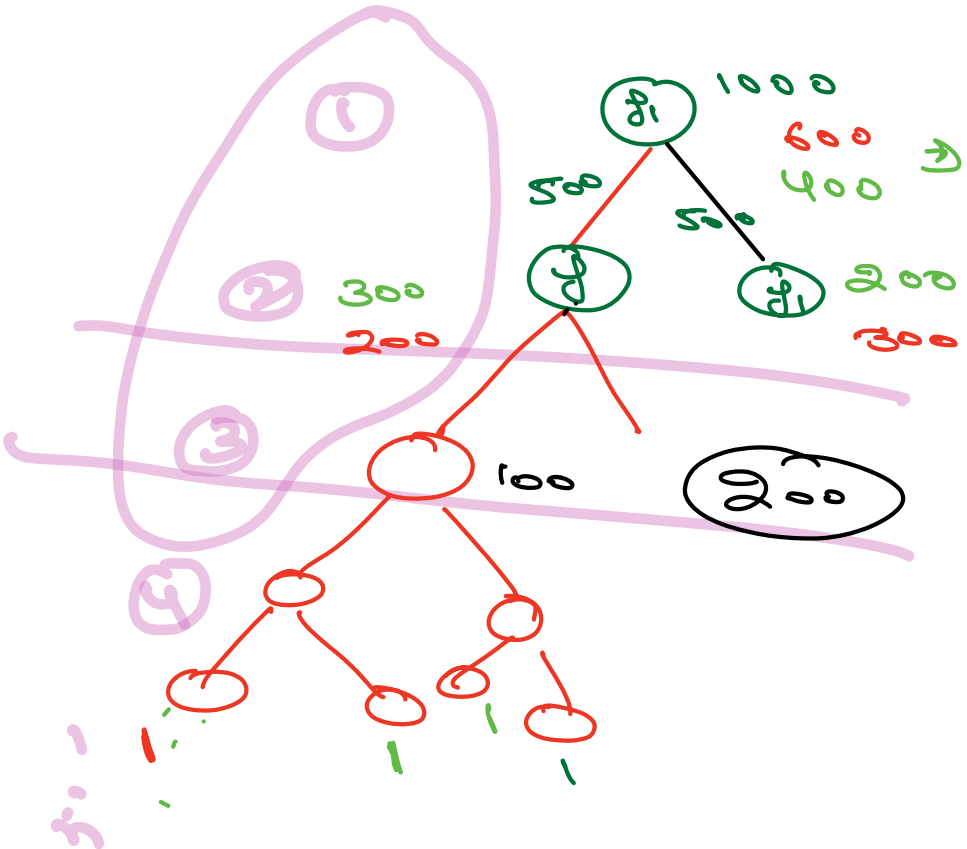
Binning



$n \leq 10000$ Unique Salary & n entropy

Binning

k entropy



→ Entropy

If left Unchecked

* DT Goes very deep and fits every Single Point including OUTLIER

Total accuracy = 100% TrainAcc.

Hence:

Decision Tree is very prone to Overfitting

↓

Low BIAS + HIGH VARIANCE
(Complex Model)

How to Handle This:

* Pruning: Cutting Un-necessary Branches

① Max-depth → Max depth of Tree

② Min-sample-Split : Min nodes for Split Consider

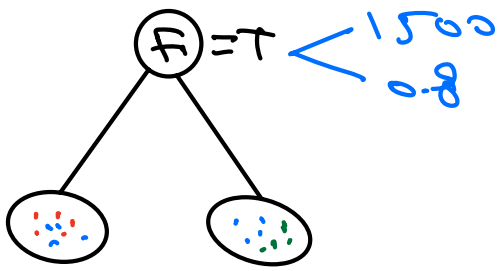
③ Min-sample-leaf → Min no every Node Should Have

④ Max-leaf-Nodes → Max number of Nodes Tree Can have

Q- Will feature scaling have any impact on DT?

f_1	f_1-scaled
2000	1.5
250	1.8
1500	0.8

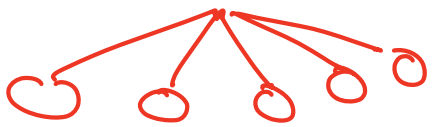
No impact



No need to Scale for DT

Q Categorical feature

→ If Unique Categorical : OHE



→ 10000 Unique Categories

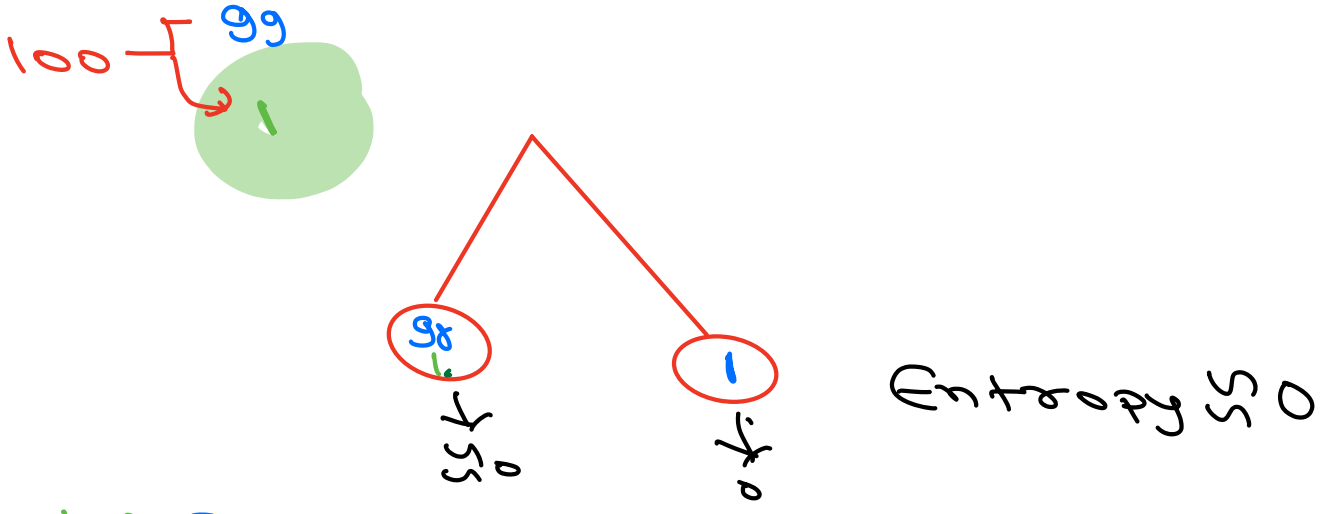
Target Encoding

XOOOOOOOOOO

Impact of Outliers? Yes

↳ prune Tree

Data Imbalance Impact DT?



1 → 98/99 %

How do we Solve this

↳ Class-Weights
↳ SMOTE
↳ Undersampling / Oversampling

For Tmrw

Feature Importance ✓

Decision Tree for Regression ?

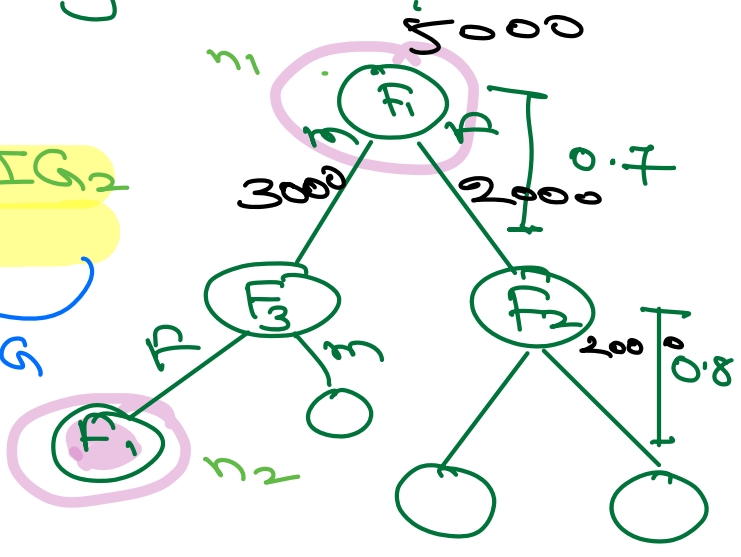
Q What is feature Importance

- How did you determine feature Importance in Linear
- Linear Model
  - ↳ weights (coef) [-10, 1]
- Decision Tree
  - No weights ✗

- Higher the information Gain, more important is the feature ?

$$FI(G_i) = \frac{n_1}{n} IG_1 + \frac{n_2}{n} * IG_2$$

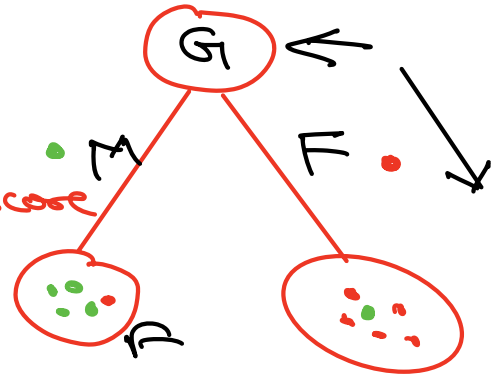
normalised IG



How Can we Use DT for Regression

DataPoint ⇒ F

What will be label and prob_score



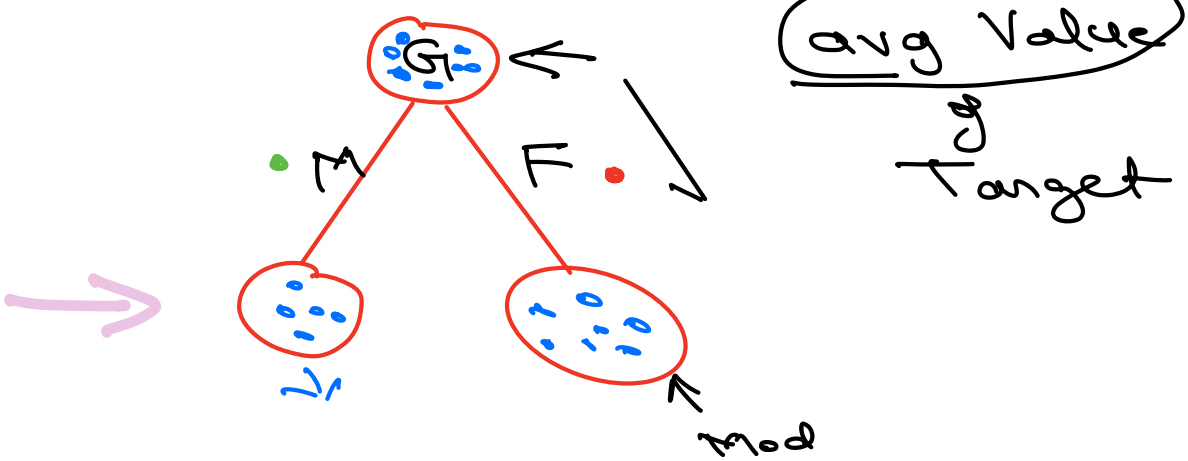
label ⇒ Red

Prob_score ⇒ 5/6

inference ⇒ O(1) ⇒ O(d)

* DT is very fast for Inference

Regression



Inference ⇒ Mean or Average ①

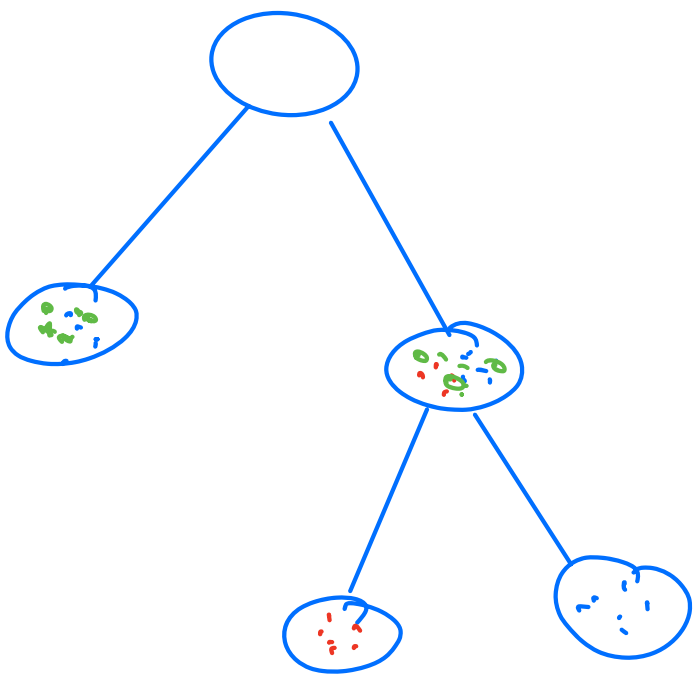
Split ⇒ MSE or RMSE ②

IG

SmoteD To create Artificial D.P

k - points

$$P'_1 = \lambda_1 P_1 + \lambda_2 P_2 + \lambda_3 P_3$$

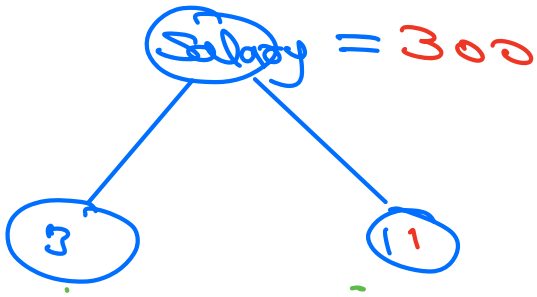


Min - leaf

Max-dep

1,1	1,2	1,3
2,1	2,2	2,3
10,1	10,2	10,3

Salary	Churn
5000	1
3000	1
2000	0
5000	1
1000	1



Salary	IG
5000	x_1
3000	x_2
2000	x_3
	1
1000	

x_3 → IG is Max