Today's agenda

1) Recap - Quizzes
2) Employee Attrition Dataset - graph
3) Purity of Nodes & Entropy
4) Plot for entropy
5) Weighted Entropy
6) Gini Impurity
7) Comparing Gini Impurity with Entropy
8) Code walkthrough (Time Permits)

Table :-

Splitting Categorical Variables

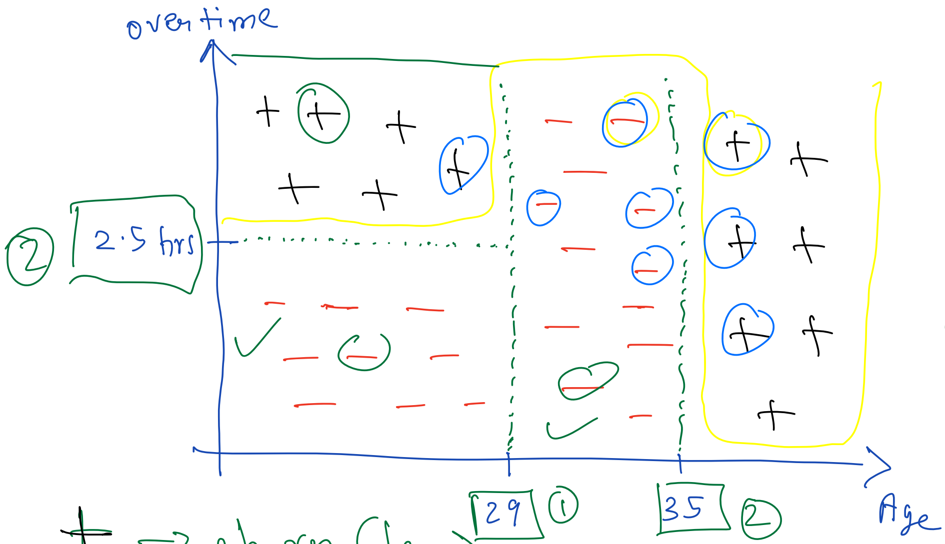
	Gender		Total
Target Variable	Male	Female	
0 (Stays)	50	10	60
1 (Churn)	20	20	40
Total	70	30	100

I)

Total 100 data points

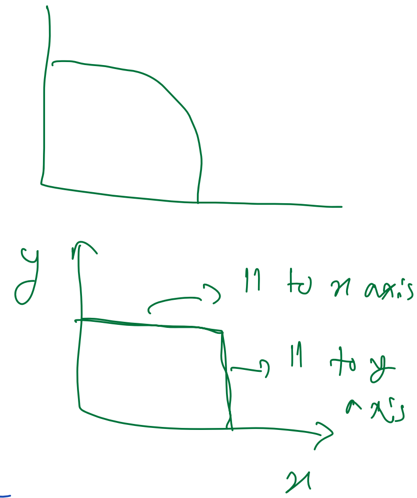
	Age		Total
Target Variable	< 35	> 35	
0 (Stays)	50	10	60
1 (Churn)	10	30	40
Total	60	40	100

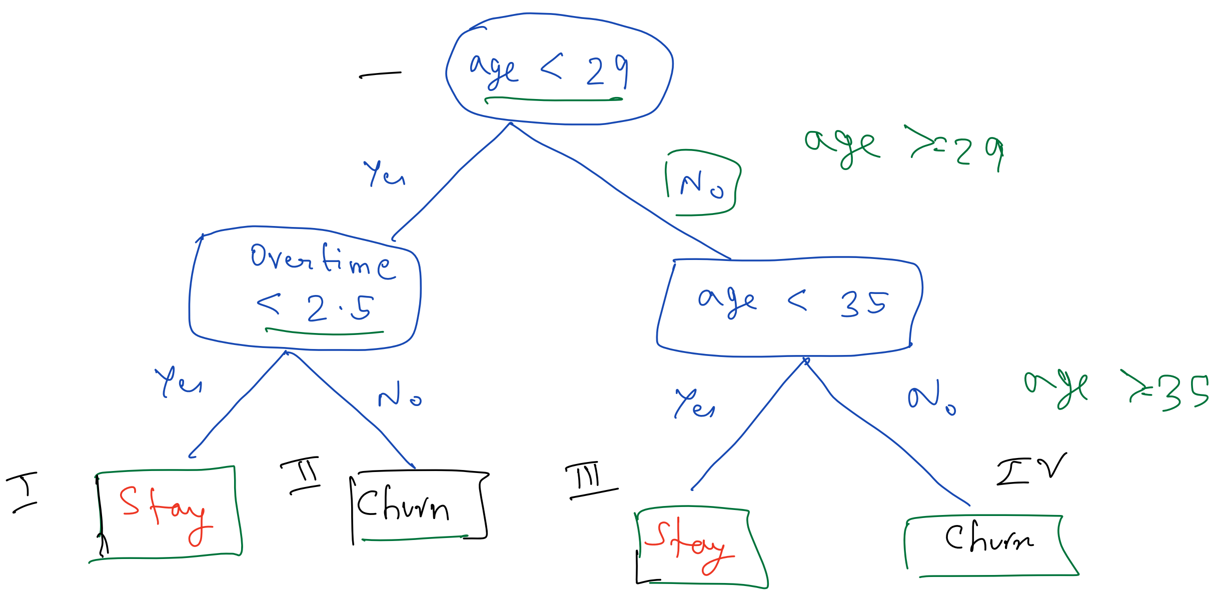
II)

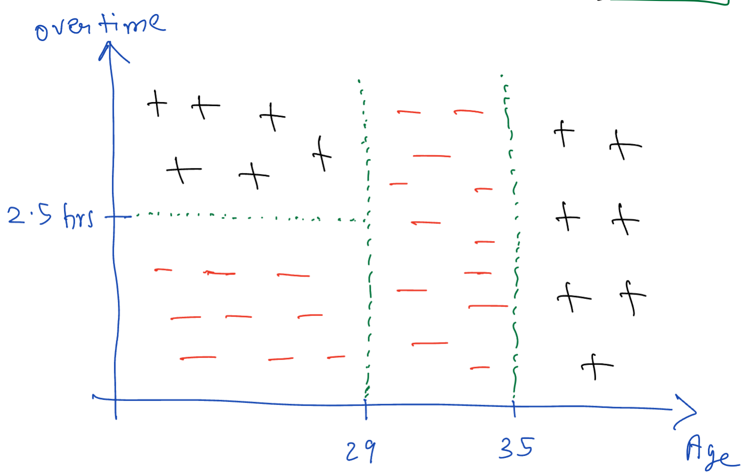


+ → churn (leave)

− → stay







If else

training data $\left[\begin{array}{l} y_+ = 40 \\ y_- = 60 \end{array} \right.$ — 100 data points

$f_1, f_2, f_3 \ldots f_d$ | y: $+$, $-$

Q) test data

↳ predict the label

100 ↳ 40 +ve, 60 -ve

majority class

$y_+ = 40$
$y_- = 60$

$\boxed{f_j < \tau_1}$

Yes — No

① Left

$\boxed{\begin{array}{l} y_+ = 10 \\ y_- = 60 \end{array}}$

$\frac{10}{70} = \frac{1}{7} \rightarrow$ +ve

$\frac{60}{70} = \frac{6}{7} \rightarrow$ -ve

② Right — node → definite decision

$y_+ = 30$
$y_- = 0$

pure node → just 1 class

$\frac{30}{30} = 1 \rightarrow$ +ve

$\frac{0}{30} = 0 \rightarrow$ -ve

impure node → more than 1 classes

Purity →

Impurity →

2 → pure

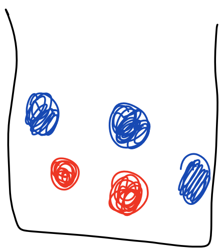
1 → impure

Q) How do we measure impurity & purity?

purity + impurity = 1

$$\boxed{\text{purity} = 1 - \text{impurity}}$$ ✓

$\boxed{\text{Entropy}}$ → measure of randomness → impurity

5  (I)

$p(\text{blue}) = \frac{3}{5}$, $p(\text{red}) = \frac{2}{5}$

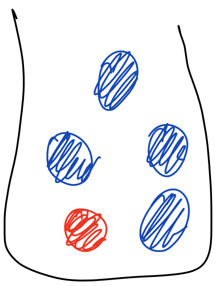
$P_1 \to p$, P_2

$P_2 = 1 - P_1$

$$\text{Entropy} = -\left[p \log_2(p) + (1-p)\log_2(1-p)\right]$$

$$= -\left[0.6 \log_2(0.6) + 0.4 \log_2(0.4)\right]$$

$$= 0.97$$

5  (II)

$p(\text{blue}) = 4/5$, $\underline{p(\text{red})} = 1/5 = p$

$$\text{Entropy} = -\left[0.8 \log_2(0.8) + 0.2 \log_2(0.2)\right]$$

$$= 0.72$$

Logistic Regression log-loss

$y = (0, 1)$

$p = (0 - 1)$

$$-\left[y \log_2(p) + (1-y)\log_2(1-p)\right]$$

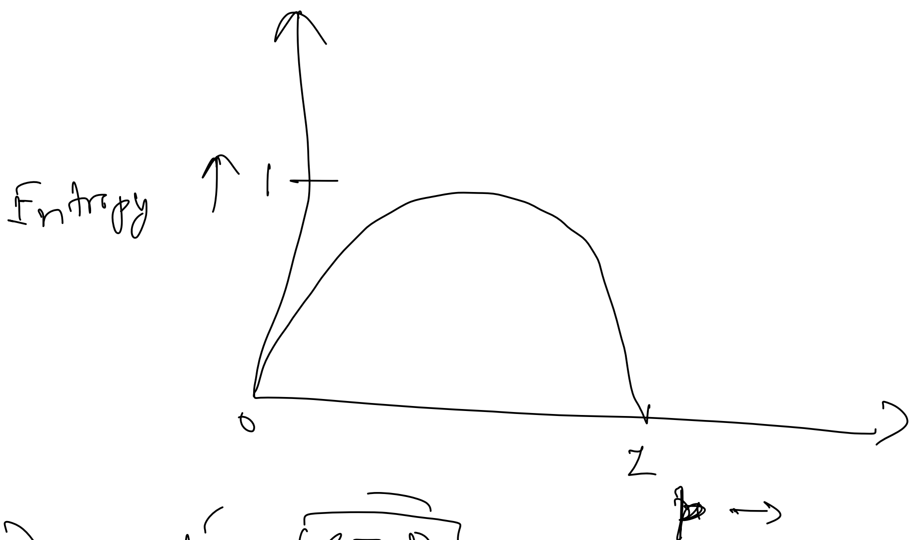
$y \to p$

KL Divergence : log-loss, entropy

$$-\left[p\log_2(p+\varepsilon)+(1-p)\log_2(1-p+\varepsilon)\right]$$

$\varepsilon = 10^{-6}$

Graph of Entropy



I) $p = 1$, ($p = 0$)

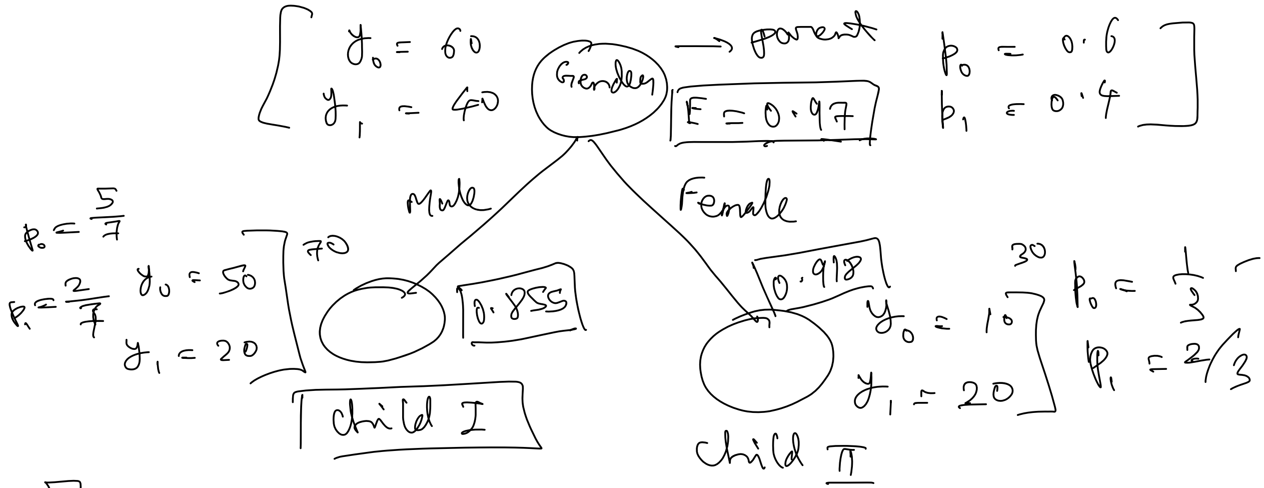
$1-p=0$

$$\text{Entropy} = -\left[1\times\log_2(\underset{0}{1}) + \underset{0}{0\times\log_2(0)}\right]$$
$$= 0$$

II) $p = 0.5$

$1-p=0.5$

$$E = -\left[0.5\times\log_2(0.5) + 0.5\times\log_2(0.5)\right]$$
$$= -\left[1\times\log_2(0.5)\right]$$
$$= -\left[1\times\log_2(2^{-1})\right]$$
$$= -\left[1\times(-1)\right] = 1$$



$$E_{parent} = -[0.6 \log_2(0.6) + 0.4 \log_2(0.4)]$$
$$= \underline{0.97}$$

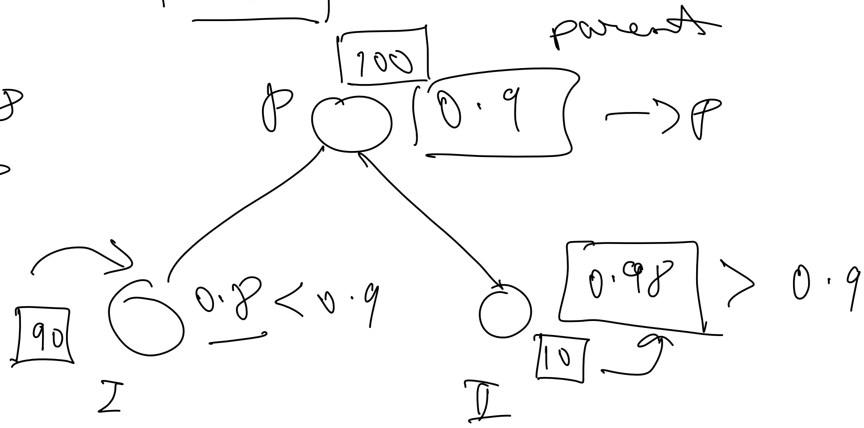
$$E_{child\ I} = -[0.72 \log_2(0.72) + 0.28 \log_2(0.28)]$$
$$= \boxed{0.855}$$

$$E_{child\ II} = -[1/3 \log_2(1/3) + 2/3 \log_2(2/3)]$$
$$= \boxed{0.918}$$

$90 \rightarrow 0.8$

$10 \rightarrow 0.98$

$p = 0.9$



$$\text{weighted entropy} = \frac{90}{100} \times 0.8 + \frac{10}{100} \times 0.98$$

$$WE = \boxed{0.818} < \boxed{0.9} \rightarrow P$$

$$WE_{children} = \frac{70}{100} \times 0.855 + \frac{30}{100} \times 0.918$$

$$= \boxed{0.874} < \boxed{0.97} \rightarrow P$$

Information Gain = Reduction in Entropy
(IG)

$$= 0.97 - 0.874$$

$$= \boxed{\sim 0.1} \rightarrow IG$$

Split objective

① is maximize IG

Gini Impurity

$$GI = 1 - [p(0)^2 + p(1)^2]$$

$$E = -[p \log(p) + (1-p)\log(1-p)]$$

$$\boxed{p = p(0)}$$

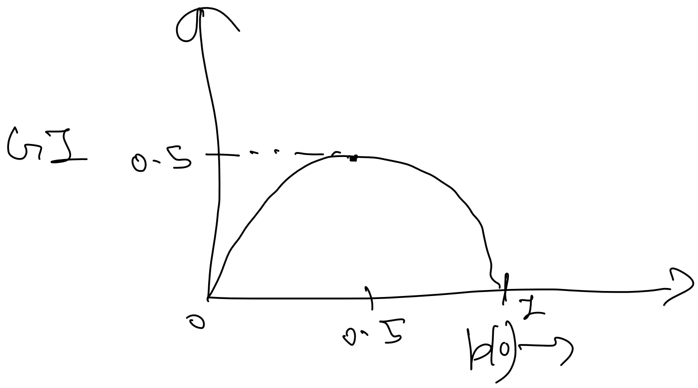
Compute intensive

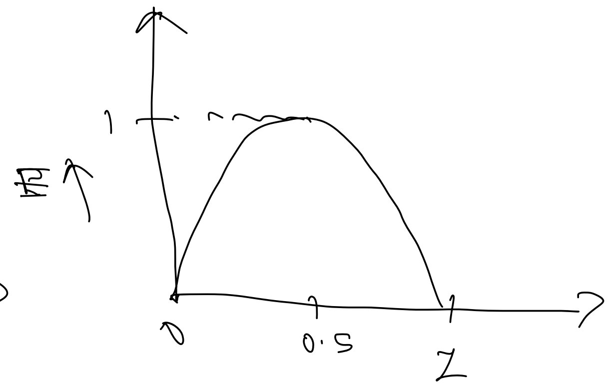
$\boxed{GI}$ I: $p(0) = 1, \quad p(1) = 0$

$$GI = 1 - [1^2 + 0^2] = 0$$

II: $p(0) = 0.5, \quad p(1) = 0.5$

$$GI = 1 - [0.5^2 + 0.5^2] = 0.5$$

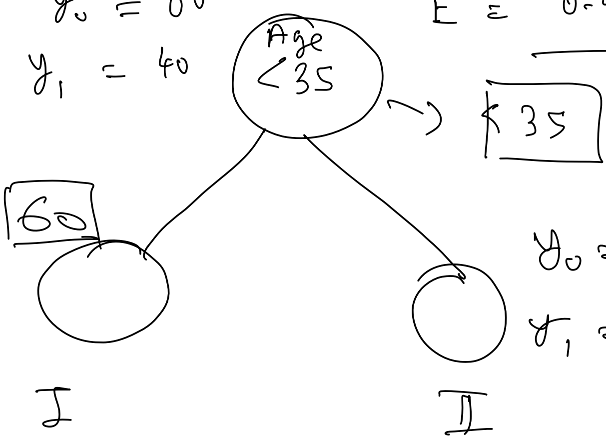




$y_0 = 60$

$y_1 = 40$

$E = 0.97$



$y_0 = 50$

$y_1 = 10$

$p_0 = 5/6$

$p_1 = 1/6$

$y_0 = 10$

$y_1 = 30$

$p_0 = 1/4$

$p_1 = 3/4$

$$E_{I} = -\left[\frac{5}{6} \log_2\left(\frac{5}{6}\right) + \frac{1}{6} \log_2\left(\frac{1}{6}\right) \right]$$

$$= 0.65$$

$$E_{II} = -\left[\frac{1}{4} \log_2\left(\frac{1}{4}\right) + \frac{3}{4} \log_2\left(\frac{3}{4}\right) \right]$$

$$= 0.811$$

$$WE_{I \& II} = \frac{60}{100} \times 0.65 + \frac{40}{100} \times 0.811$$

$$= 0.7144$$

$$I.G = 0.97 - 0.7144 = \boxed{0.2566} \rightarrow ?$$

$\checkmark$ $0.2566 > \boxed{0.1}$

GI for Gender

$GI_p = 1 - [0.6^2 + 0.4^2]$
$= \boxed{0.48}$

$GI_I = 1 - [(5/7)^2 + (2/7)^2]$
$= 0.408$

$GI_{II} = 1 - [(1/3)^2 + (2/3)^2]$
$= 0.444$

$WGI_{I\&II} = \frac{70}{100} \times 0.408 + \frac{30}{100} \times 0.444$
$= \boxed{0.4188}$

$IG = 0.48 - 0.4188 = \boxed{0.0612}$

GI for Age < 35

$GI_p = 0.48$

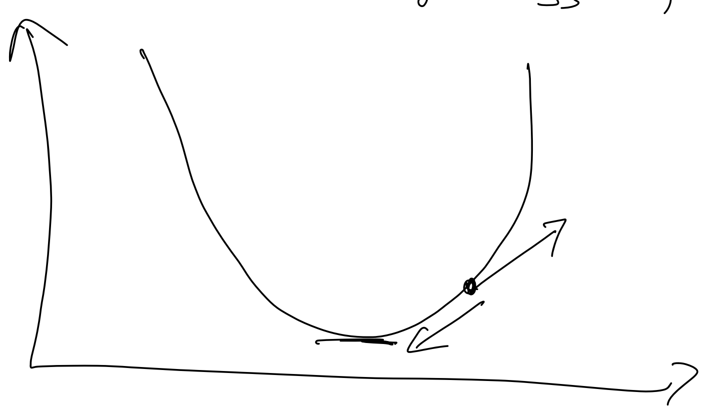
$GI_I = 1 - [(5/6)^2 + (1/6)^2]$
$= 0.277$

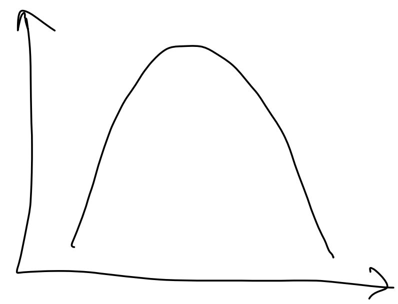
$GI_{II} = 1 - [(1/4)^2 + (3/4)^2]$
$= 0.375$

$WGI_{I\&II} = 6/10 \times 0.277 + \frac{4}{10} \times 0.375 = 0.3162$

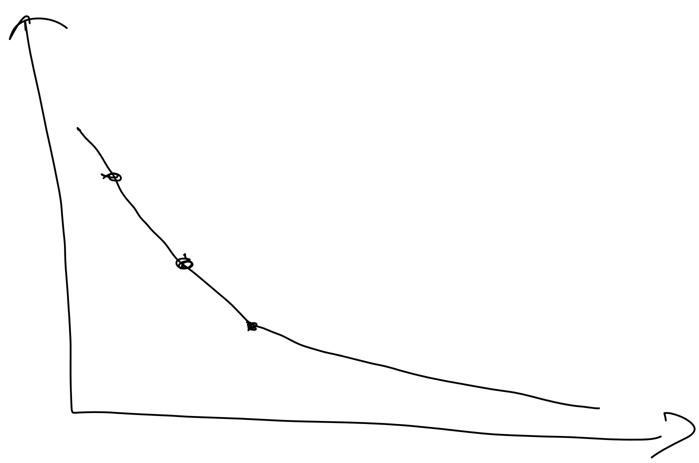
$IG = 0.48 - 0.3162 = \boxed{0.1638}$

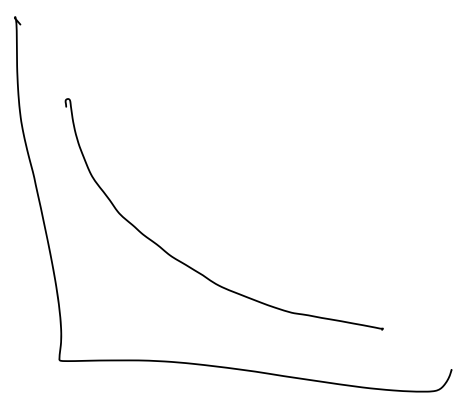
age < 35 → better split





WCSS





k →

plot WCSS vs $\boxed{\text{initialization}}$

d dimensional vector

$\boxed{\text{WCSS}}$ vs d_1, d_2, d_3 . . .

WCSS

d →