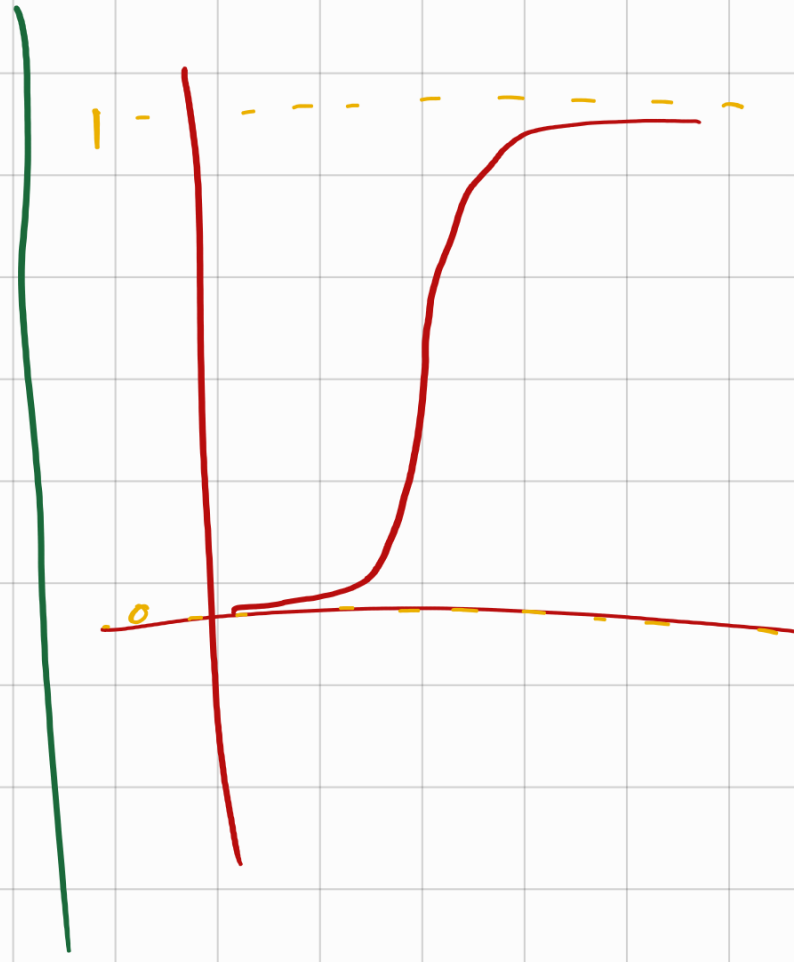
$$y = \frac{1}{1 + e^{-(a+bx)}}$$



$\Rightarrow$ We want to Minimize log loss

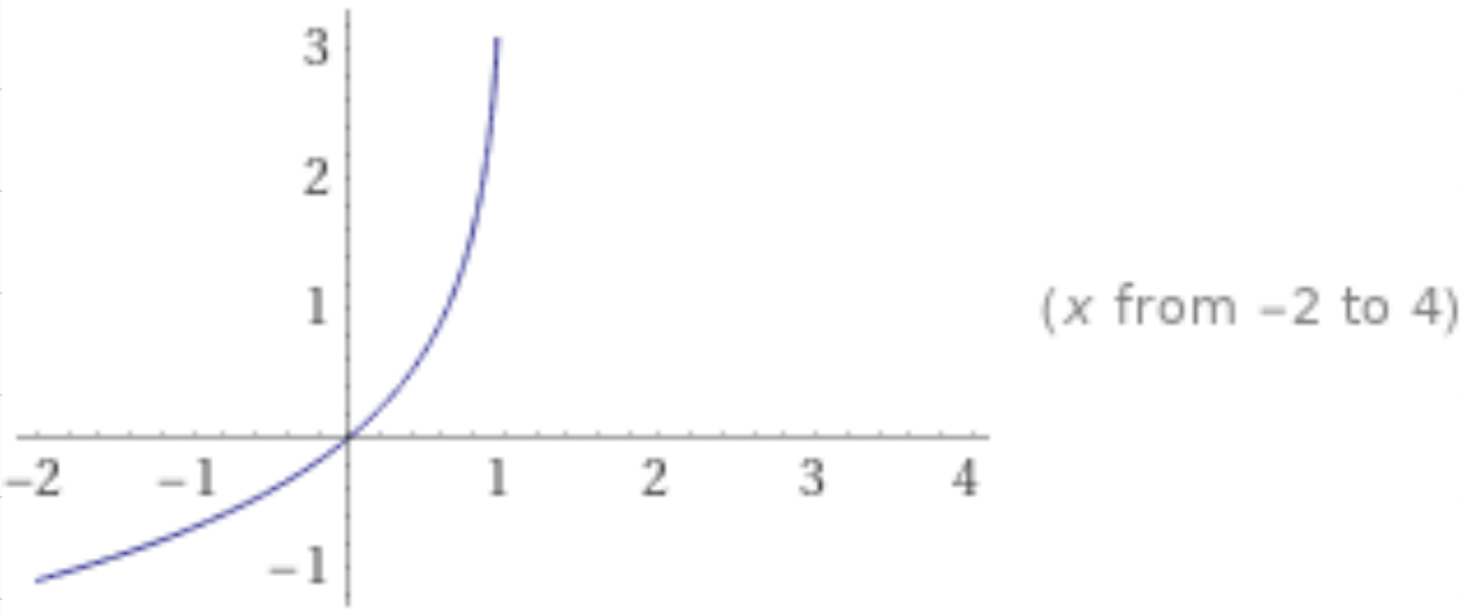
$$\text{Log loss} = \left(-y \log(\hat{y}) - (1-y) \log(1-\hat{y})\right)$$

y = observed

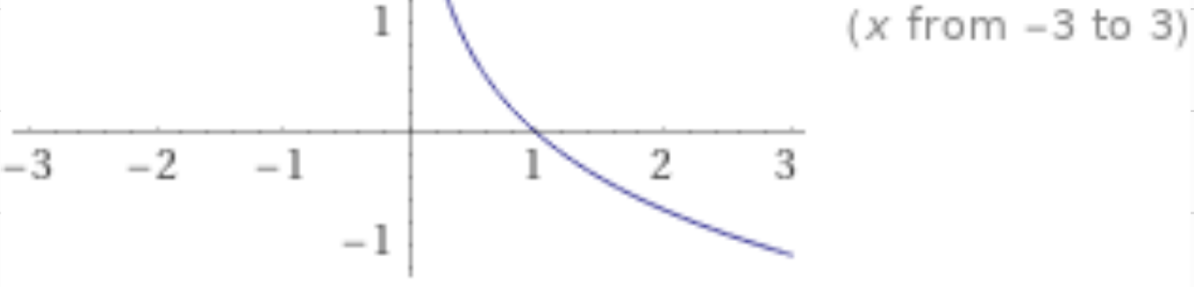
$\hat{y}$ = Predicted.

$$\begin{cases} -\log(\hat{y}) & \text{if } y=1 \\ -\log(1-\hat{y}) & \text{if } y=0 \end{cases}$$

$y = -\log(1-x)$



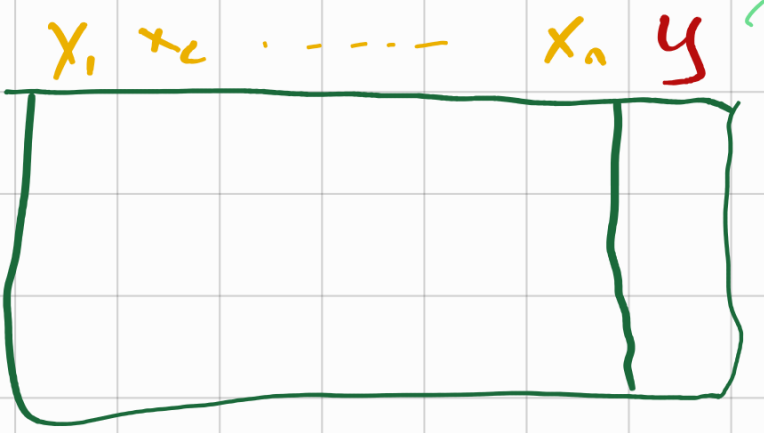
$-\log(x)$



final Logit function

$$y = \frac{1}{e^{-(a + b_1x_1 + b_2x_2 \cdots + b_nx_n)}}$$

Thresholds



→ $\hat{y}$ is a probability

=> When deciding on the threshold for logit functions,

Put it where outliers or mis-predictions will have a very low cost

Confusion Matrix

Pred \ Truth	P	N
P	True positives TP	false positives FP
N	false negatives FN	True negative TN

Measures

(1) Accuracy

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

(2) Recall (Sensitivity / True Positive rate)

$$\text{Recall} = \frac{TP}{TP + FN}$$

Proportion of total positives that were correctly identified

(3) Precision

$$\text{Precision} = \frac{TP}{TP + FP}$$

④ Classification Error rates

$$\text{Error Rate} = \frac{FP + FN}{TP + TN + FP + FN}$$

$$\text{Accuracy} = 1 - \text{Error rate}$$

⑤ Specificity

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Propotion of total negatives that were correctly identified

ROC Curves, Gini Coefficien

x x_2 x_3 ..... x_n	P	A	$\overline{A}$
	0.9	1	1
	0.51	0	1
	0.3	1	0

A = Actual

P = Probability output / Prediction

$\overline{A}$ = Value after Applying threshold.

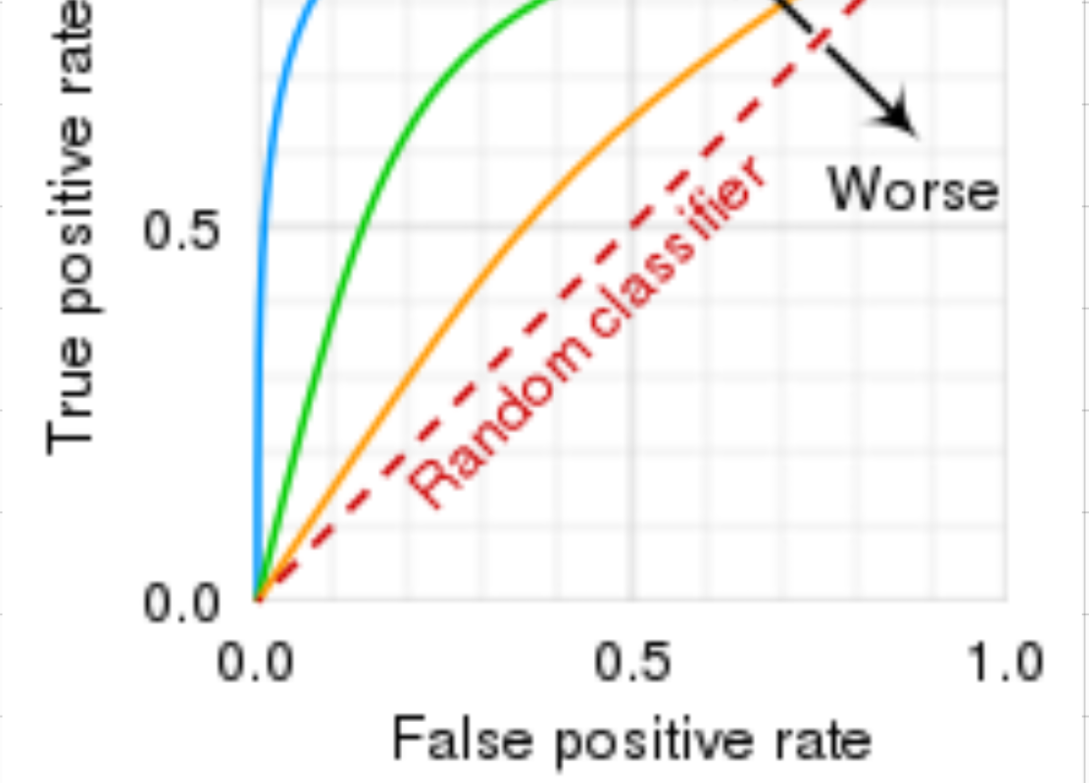
threshold = 0.5

1) ROC Curve

⇒ A plot of the True Pos Rates V.S True Neg rates for each possible threshold

TP

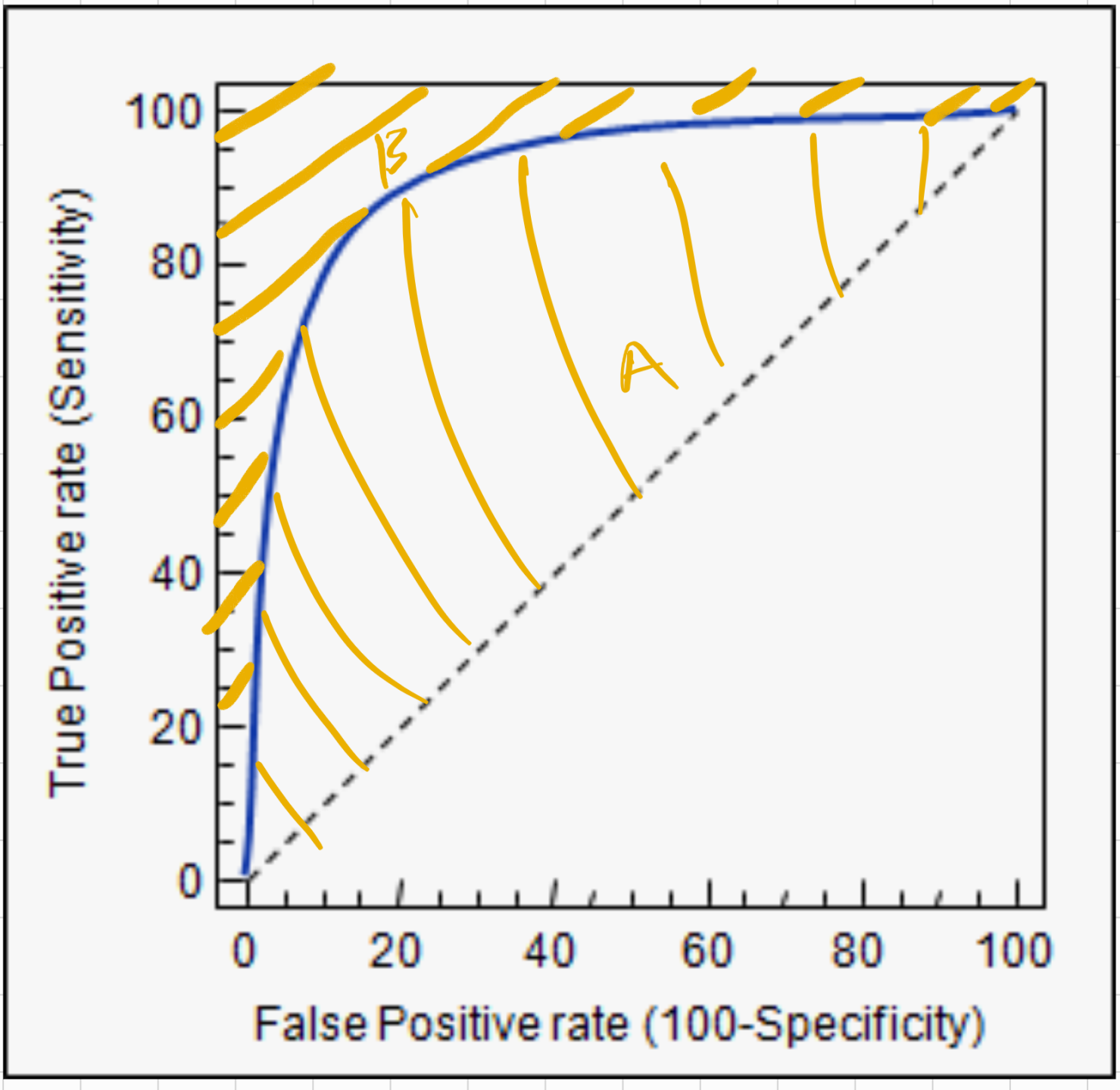
TP + FN



$$\Rightarrow \frac{FP}{FP + TN}$$

Pred \ Truth	P	N
P	True TP Positives	false FP Positives
N	false FN negatives	True negative TN

2) Gini Coefficient



⇒ the large the Area A, the more the

inequality???

$$\text{Gini Coefficient} = \frac{A}{A+B}$$