Logistic Regression

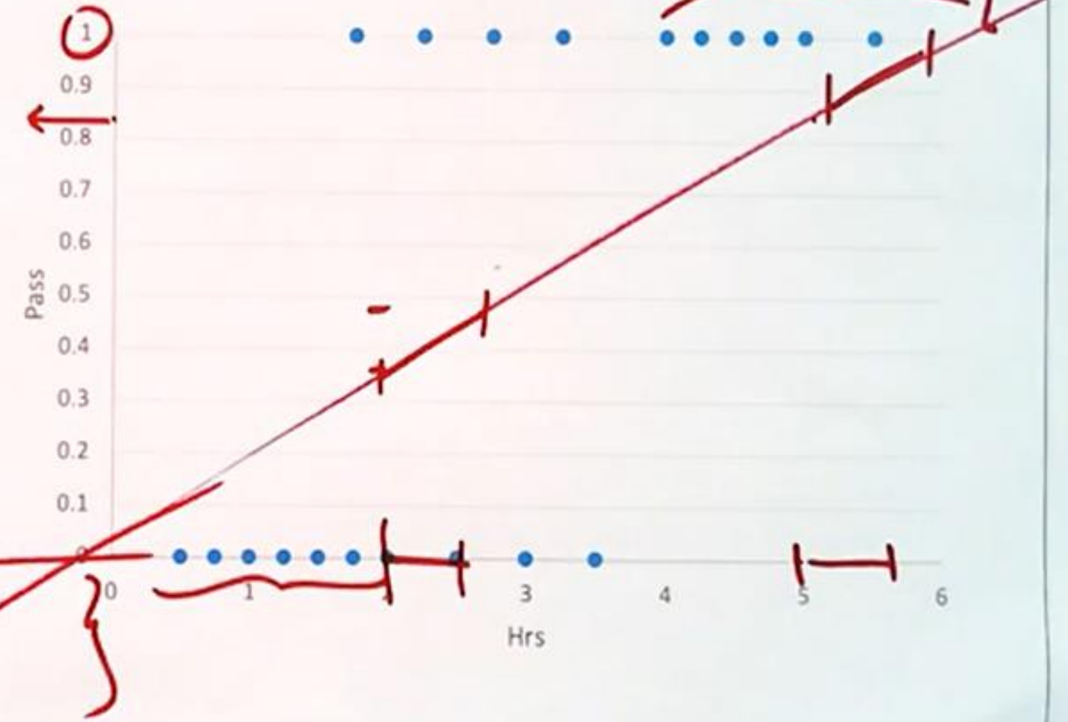
greatlearning
Learning for Life

Linear Regression for Classification?

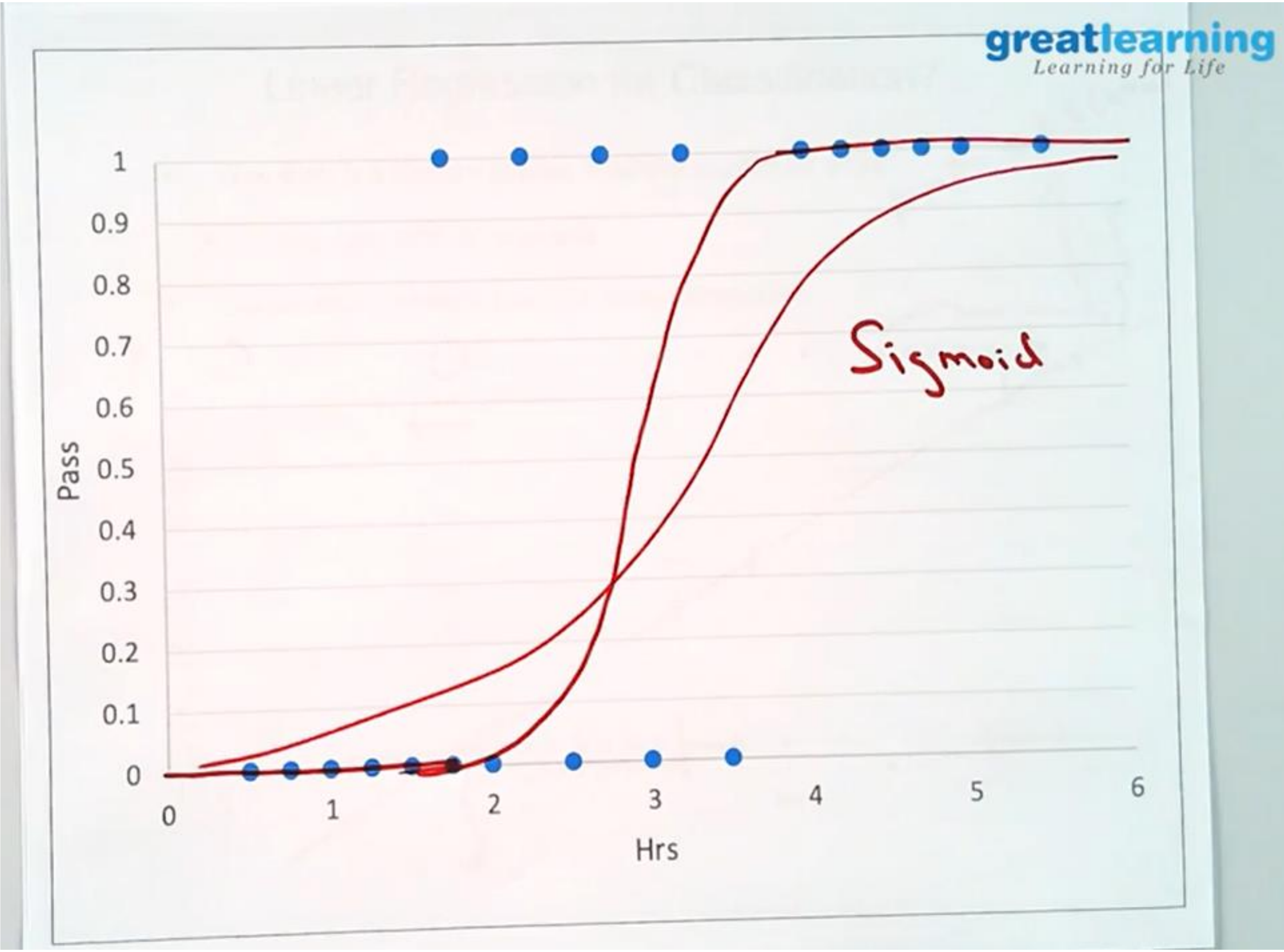
- How likely is a student to pass if he/she studies for 5 hrs?
  - Using data from 20 students
- Classification problem! Can use linear regression?

pass? = a + b(hrs)

x Hrs	y Pass?
0.5	0
0.75	0
1	0
1.25	0
1.5	0
1.75	0
1.75	1
2	0
2.25	1
2.5	0
2.75	1
3	0
3.25	1
3.5	0
4	1
4.25	1
4.5	1
4.75	1
5	1
5.5	1



Source: Wikipedia



greatlearning
Learning for Life
Sigmoid
Pass
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
6
Hrs

Instead can we fit a curve?

- Regression fits $y = a + bx$
- Instead why not fit?

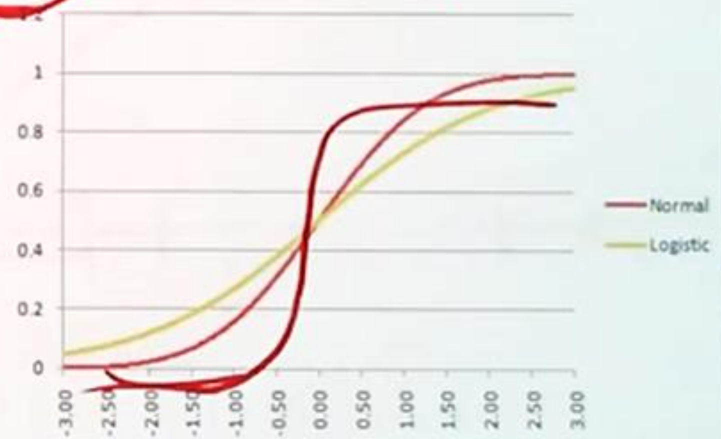
$$y = f(a + \underline{bx})$$

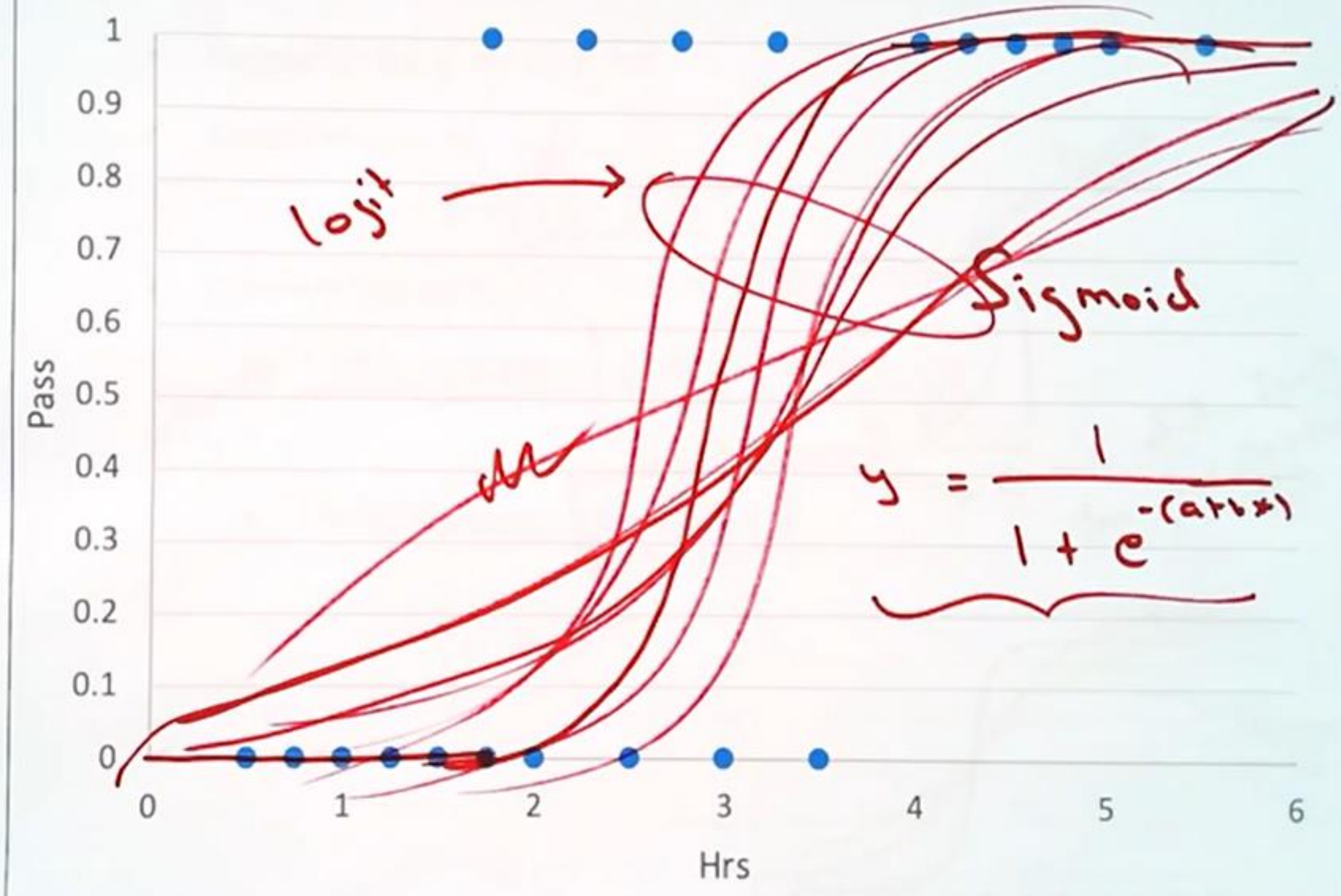
logit

- Common choices for $f()$
  - Logistic Regression: $y = \dfrac{1}{1 + e^{-(a+bx)}}$
  - Probit Regression: $\boxed{y = \Phi(x)}$

Cumm. dist. funct for the Normal dist

Probit





1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Pass
0
1
2
3
4
5
6
Hrs
logit
Sigmoid
$y = \frac{1}{1 + e^{-(a+bx)}}$

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

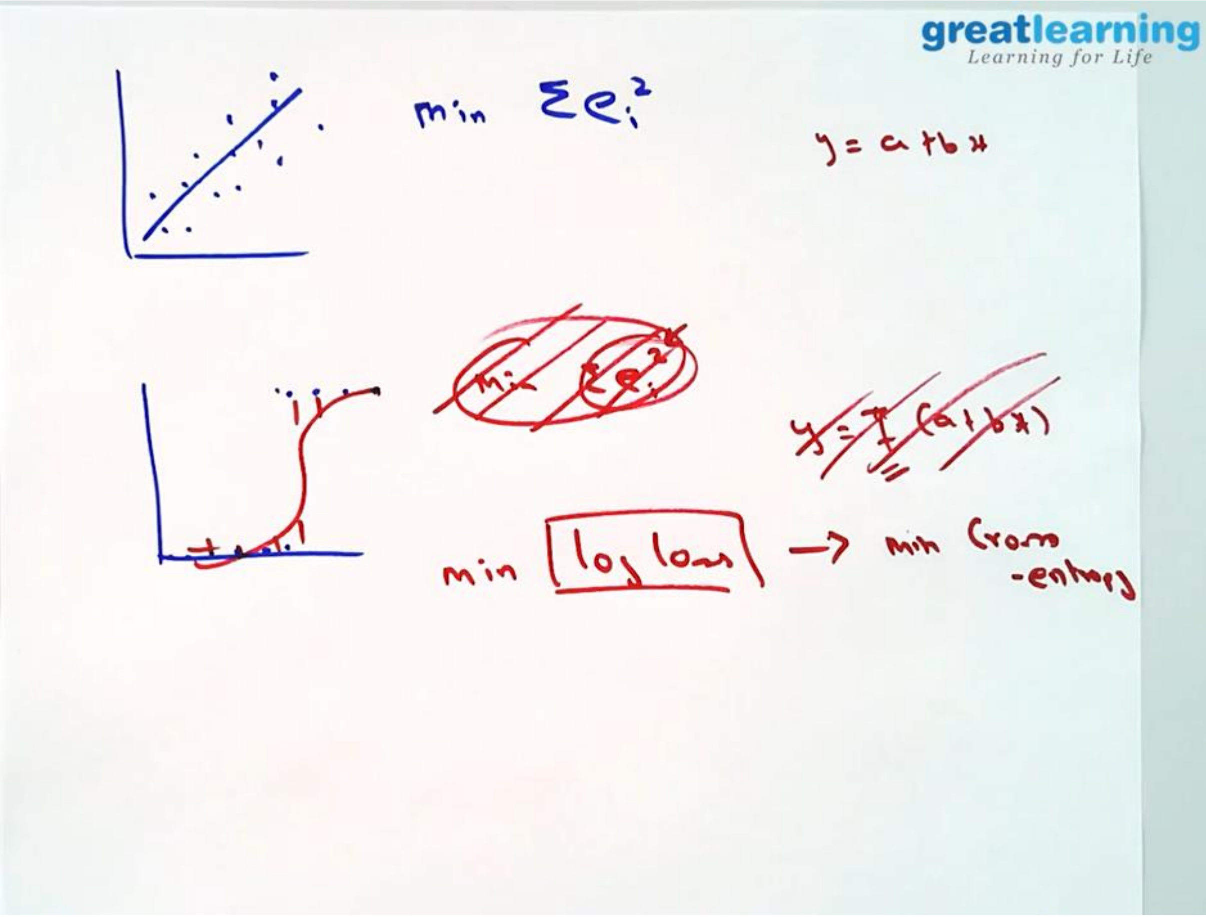
between
$\boxed{0 \text{ and } 1}$

$$\log\left(\underbrace{\frac{y}{1-y}}_{\text{odds ratio}}\right) = a+bx$$

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

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

Log Loss


greatlearning
Learning for Life
min Σe_i^2
y = a + bx
min log loss → min cross-entropy

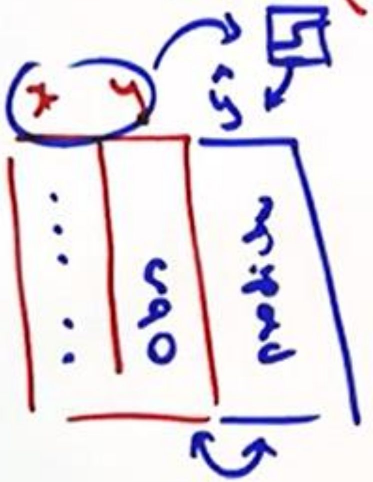
Finding the best fit logic curve?

$\cancel{\sum e_i^2}$

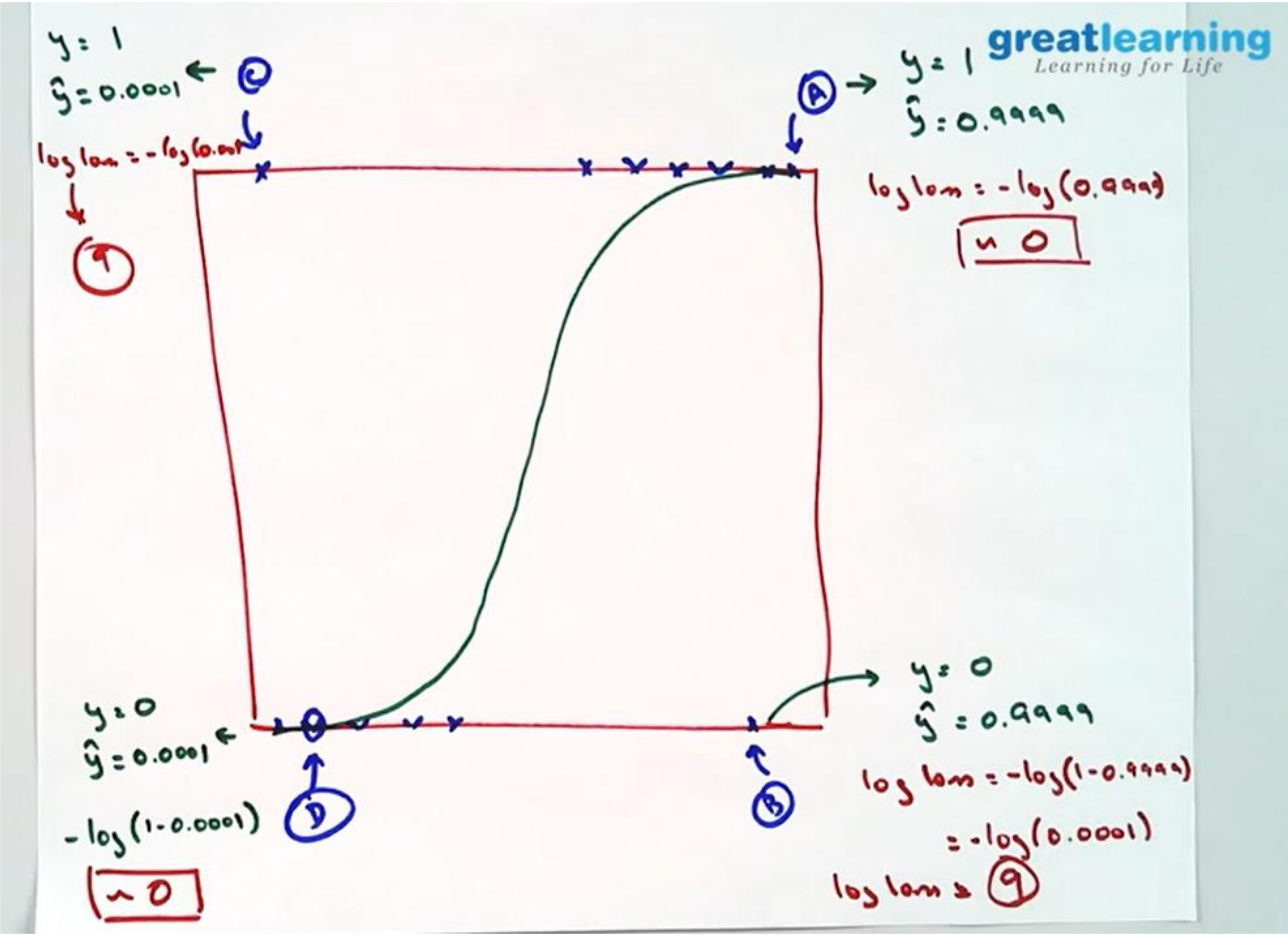
- Linear regression minimized sum of squared residuals. This unfortunately will not work in logistic regression!
- Instead we choose to minimize the "Log Loss" or "Cross-Entropy"

min $\boxed{-y\log(\hat{y}) - (1-y)\log(1-\hat{y})}$

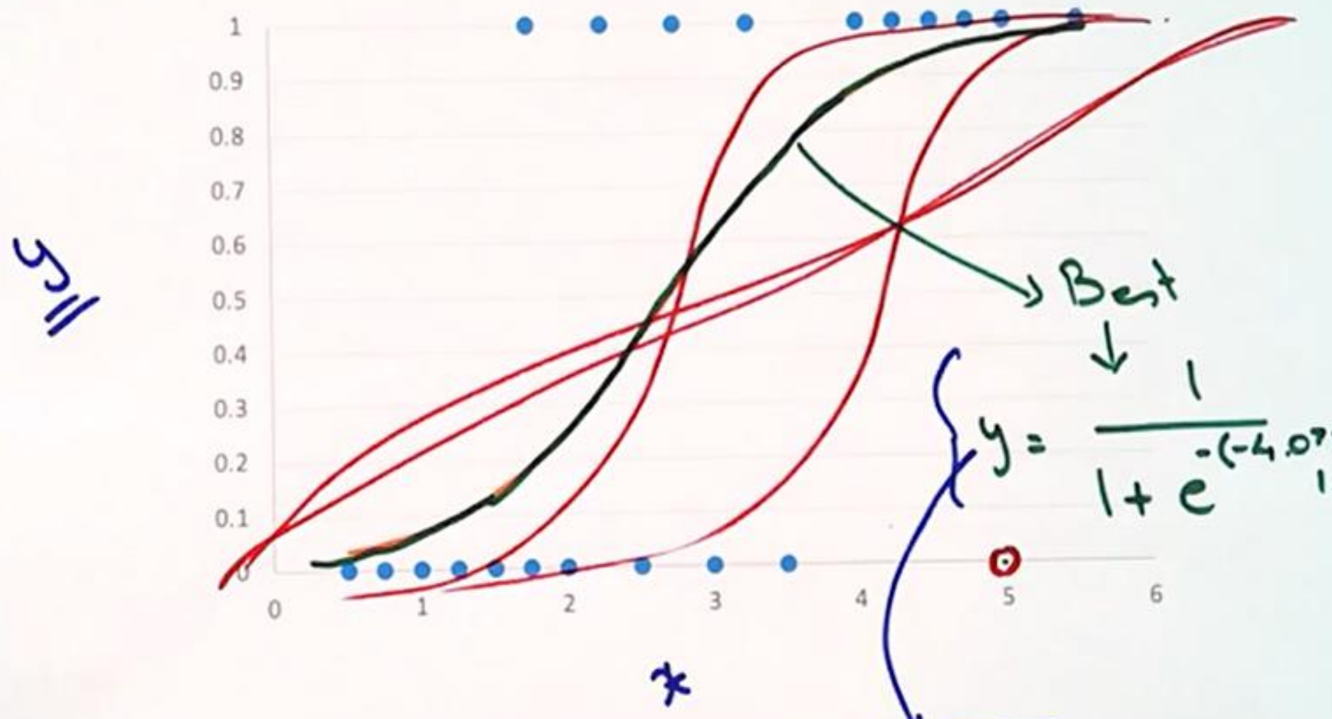
$y \rightarrow$ obs
$\hat{y} \rightarrow$ predicted



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


greatlearning
Learning for Life
y = 1
ŷ = 0.0001
C
log loss = -log(0.0001)
9
A
y = 1
ŷ = 0.9999
log loss = -log(0.9999)
~ 0
y = 0
ŷ = 0.0001
D
-log(1-0.0001)
~ 0
B
y = 0
ŷ = 0.9999
log loss = -log(1-0.9999)
= -log(0.0001)
log loss = 9

- How likely is a student to pass if he/she studies for 5 hrs?

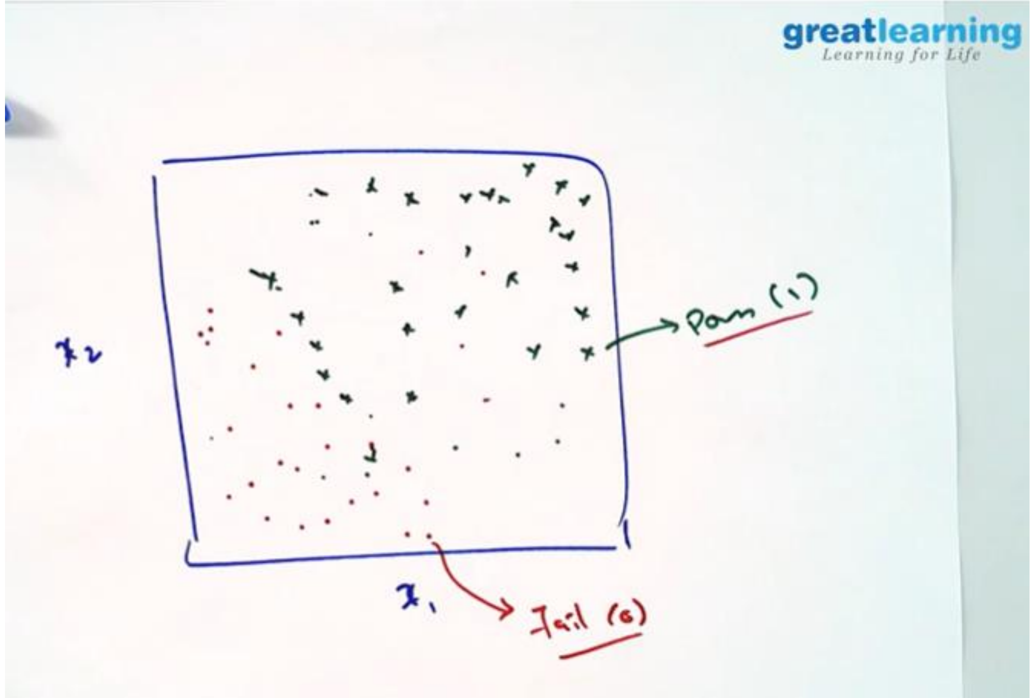


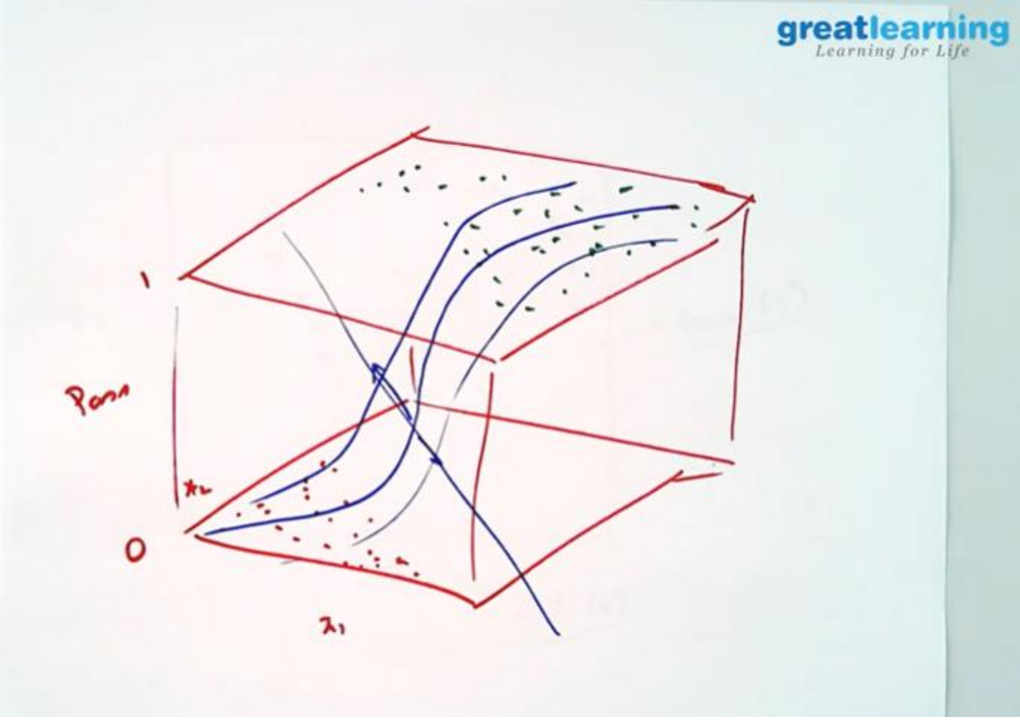
$$y = \frac{1}{1 + e^{-(-4.078 + 1.505x)}}$$

5

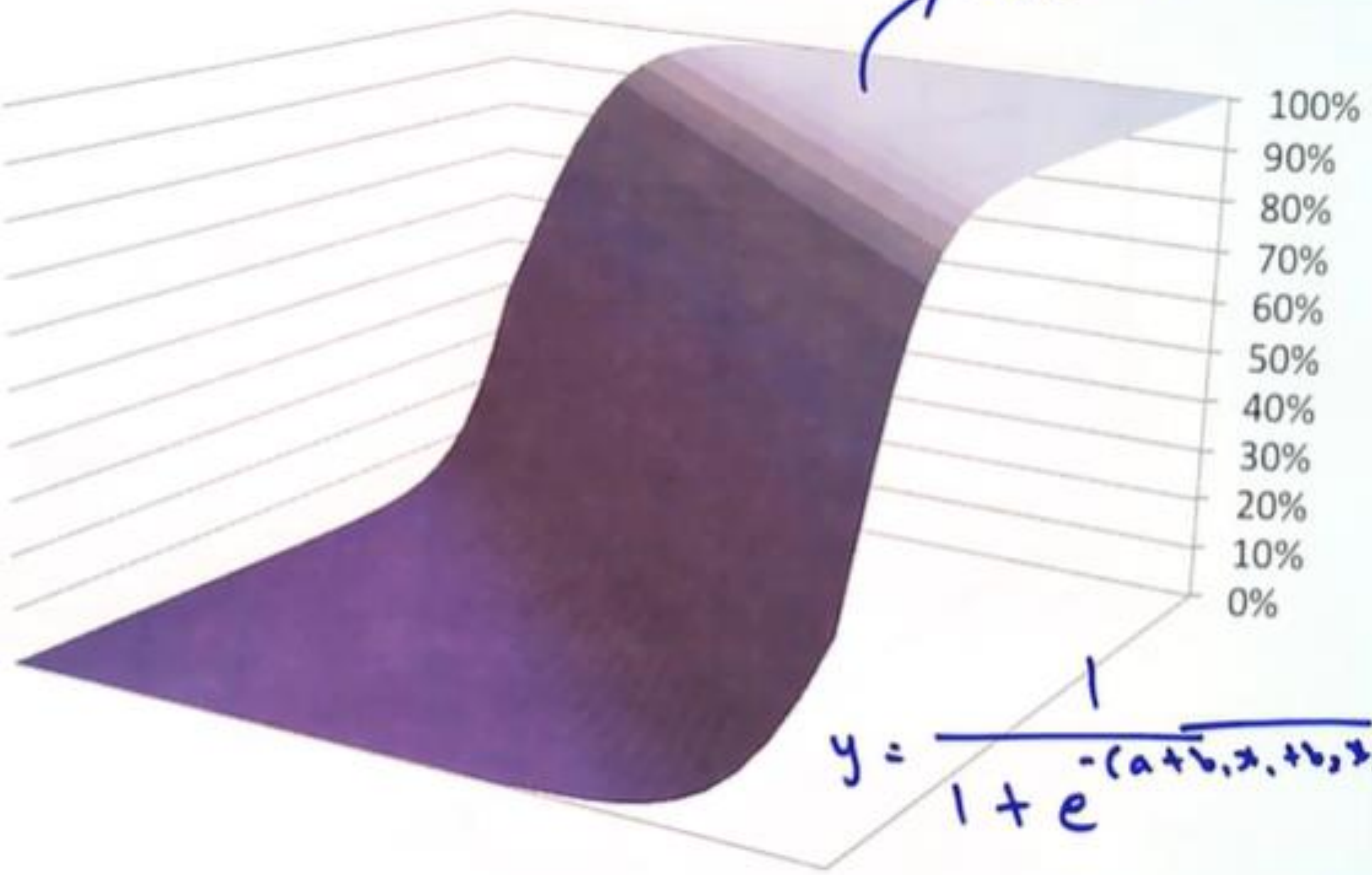
$\boxed{0.97}$

$$\rightarrow y = f(x_1, x_2 \ldots x_k)$$


greatlearning
Learning for Life
x_2
x_1
Pass (1)
Fail (0)


greatlearning
Learning for Life
1
Pass
x_2
0
x_1

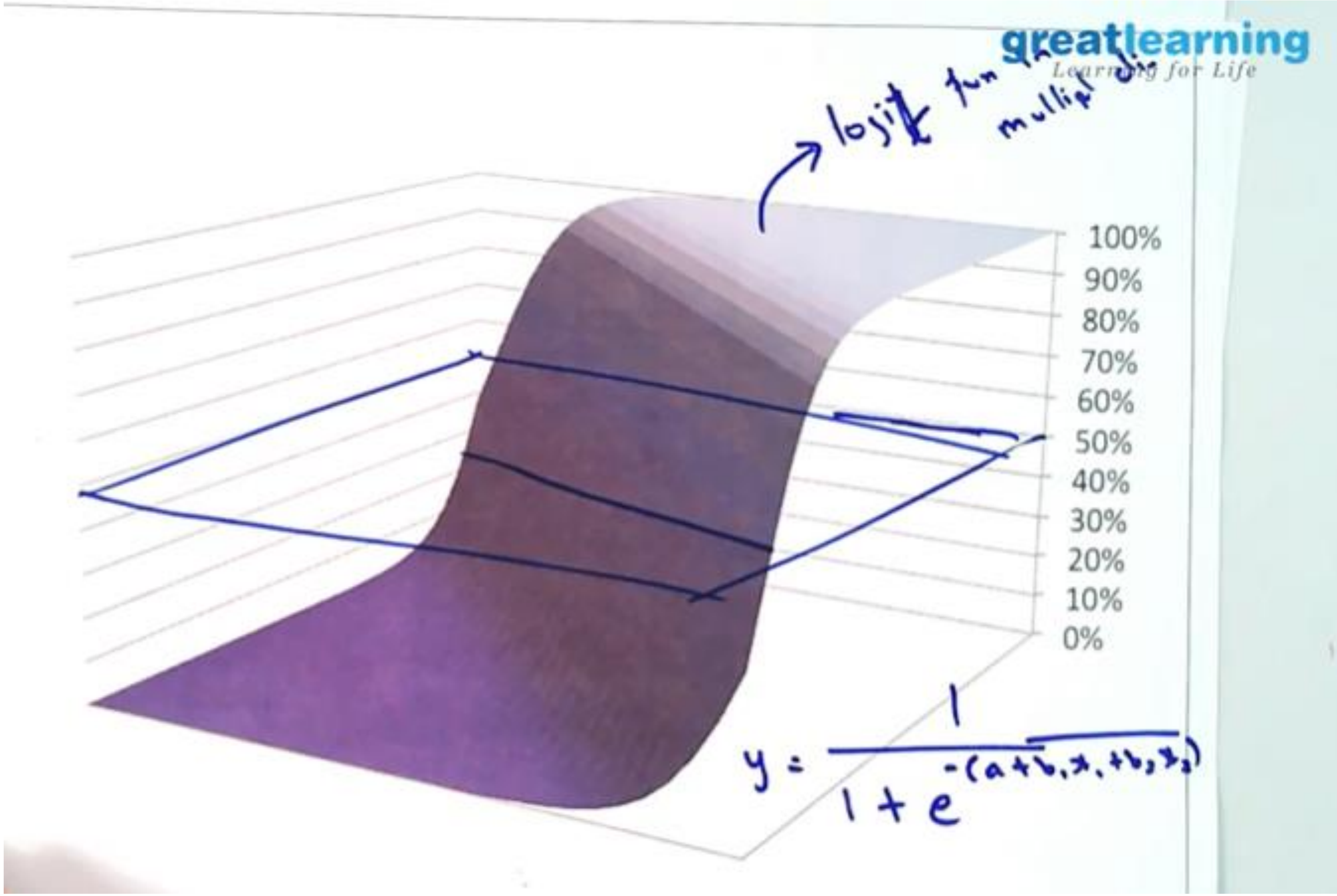
logit fun in multiple dim



$$y = \frac{1}{1 + e^{-(a + b_1 x_1 + b_2 x_2)}}$$

Logistic Reg - Pros and Cons

- Advantages
  - A classification model that does give probabilities
  - Easily extended to multiple classes (multinomial regression)
  - Quick to train and very fast at classifying unknown records
- Disadvantages
  - Constructs linear boundaries
  - Assumes that variables are independent (eg. does not include interaction terms)
  - Interpretation of coefficients is difficult


greatlearning
Learning for Life
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
$y = \frac{1}{1+e^{-(a+b_1x_1+b_2x_2)}}$

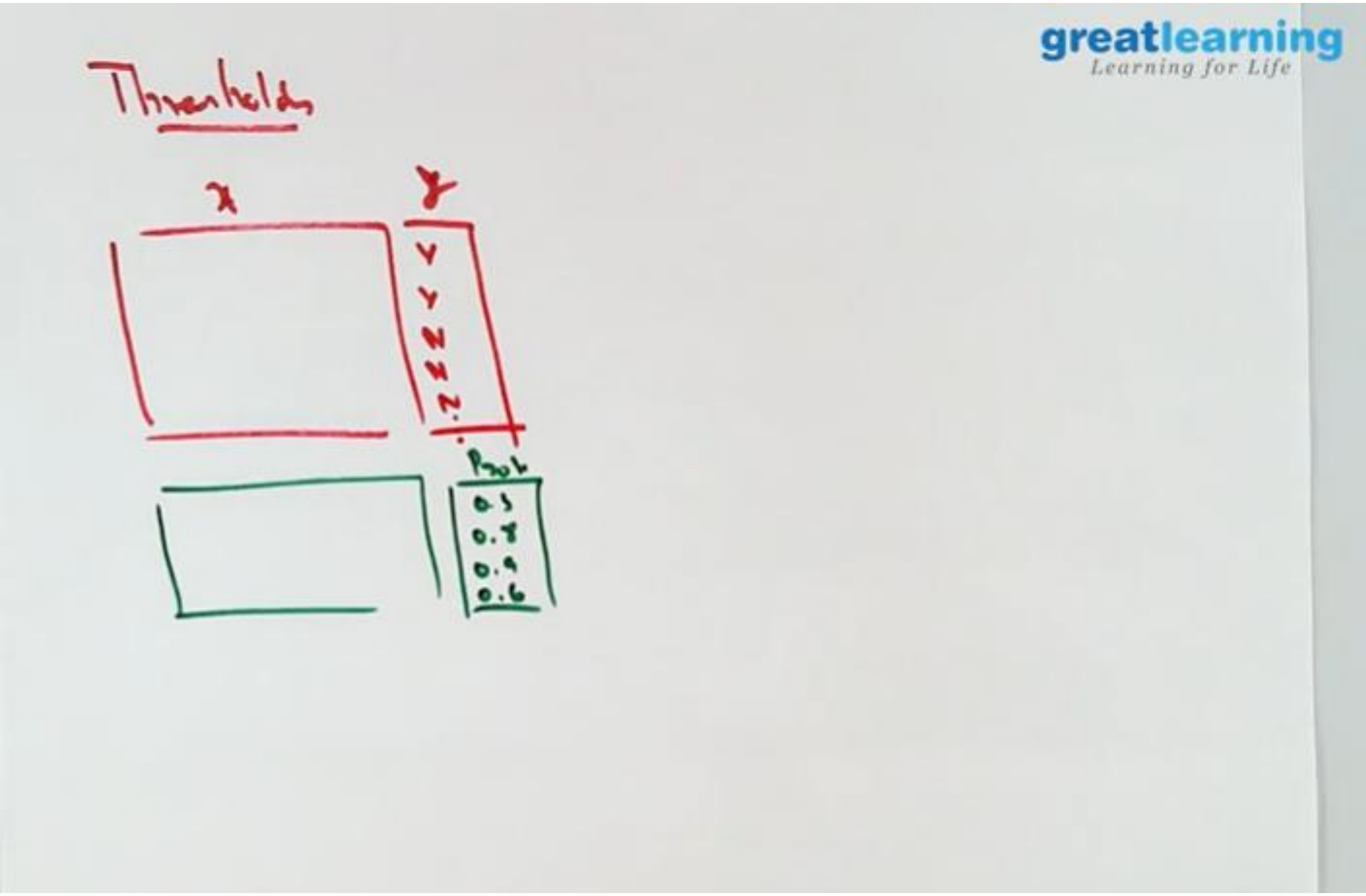
Logistic Reg - Pros and Cons

$y = a + b_1 x_1 + b_2 x_2$

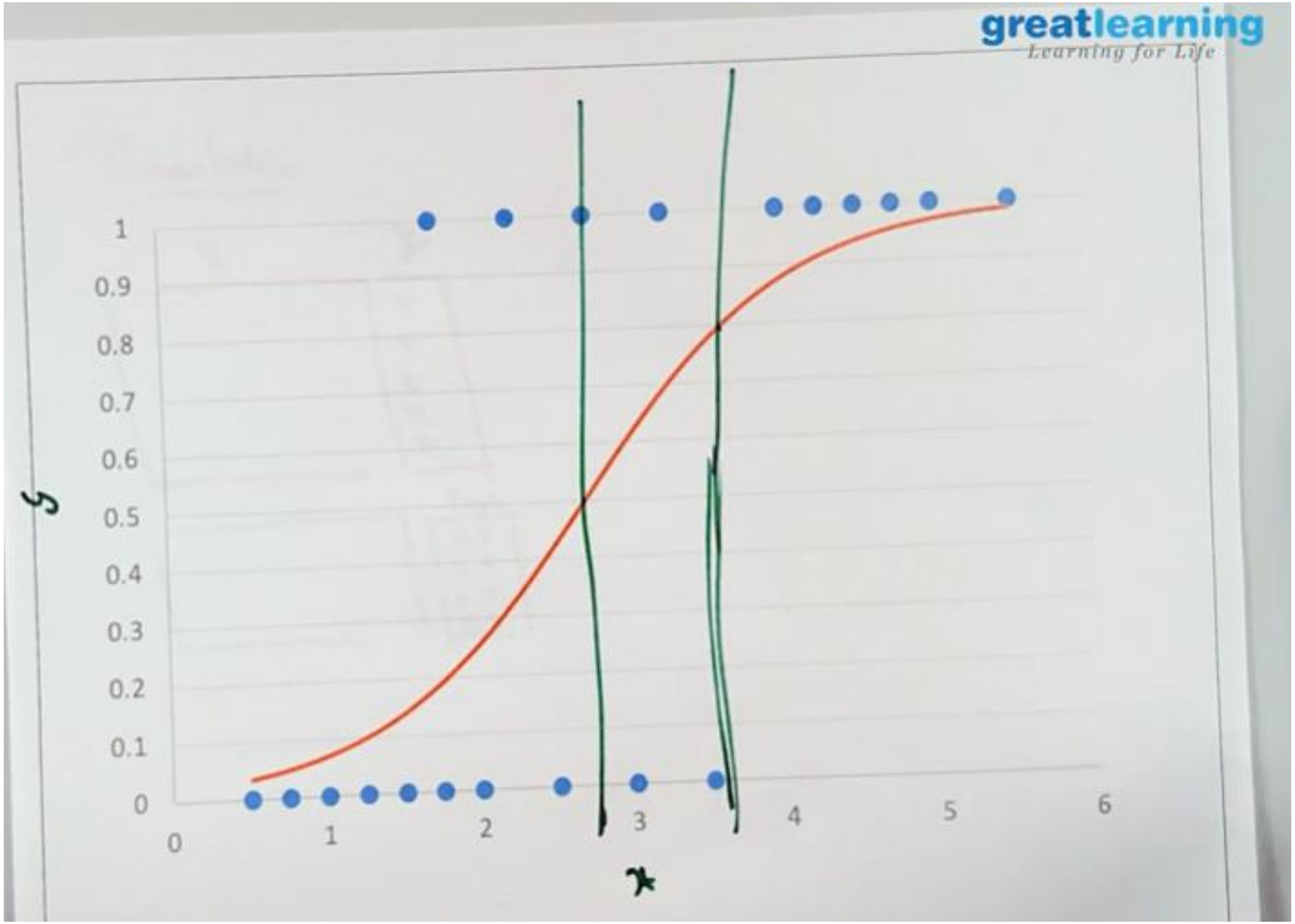
- Advantages
  - A classification model that does give probabilities
  - Easily extended to multiple classes (multinomial regression)
  - Quick to train and very fast at classifying unknown records
- Disadvantages
  - Constructs linear boundaries
  - Assumes that variables are independent (eg. does not include interaction terms)
  - Interpretation of coefficients is difficult

$$y = \frac{1}{1 + e^{-(a + b_1 x_1 + b_2 x_2)}}$$

Threshold

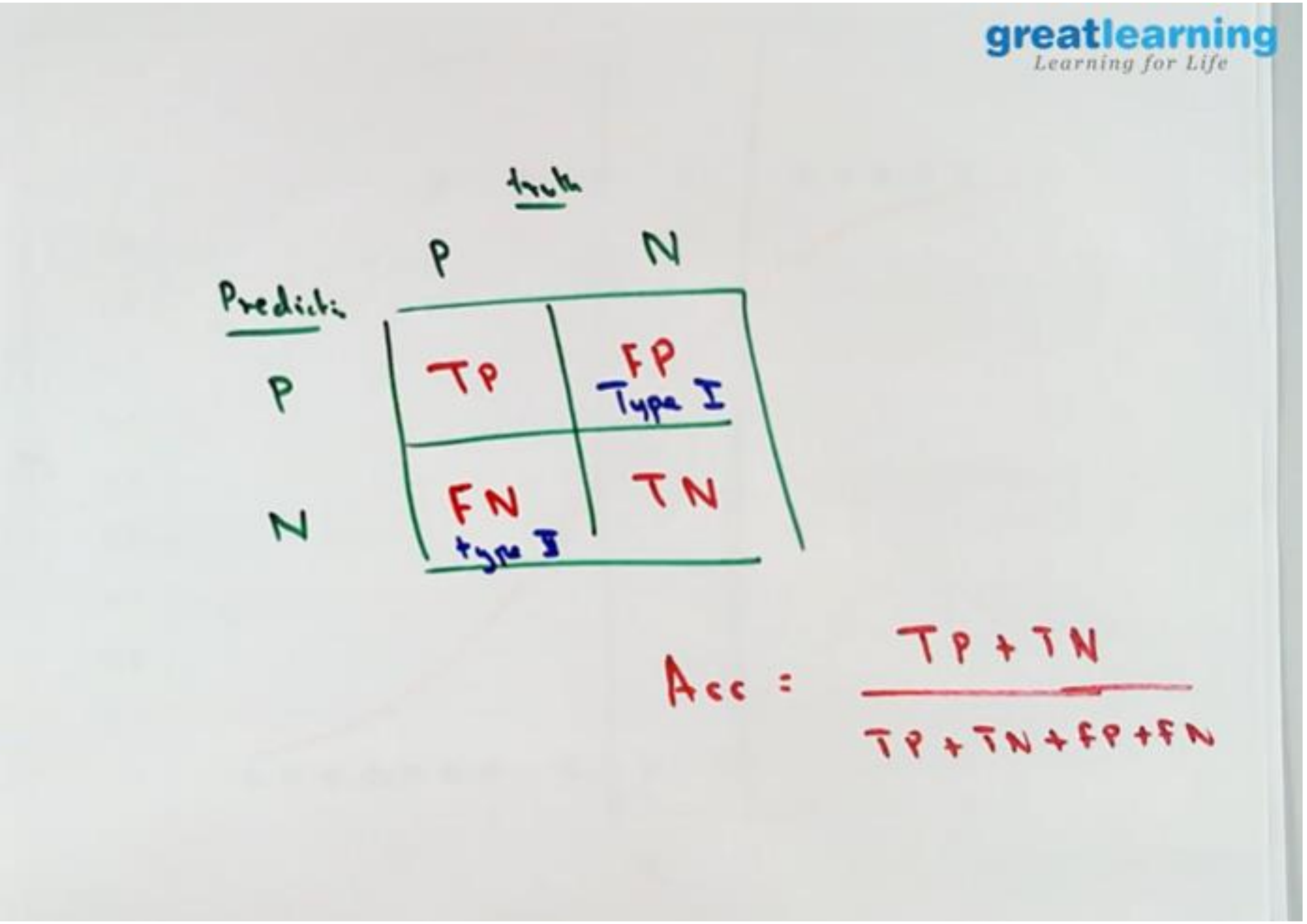


greatlearning
Learning for Life
Thresholds
x
y
Y
Y
N
N
N
0.8
0.9
0.6



greatlearning
Learning for Life
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
6
y
x

Confusion Metrix & performance measures



1 Billion   10 terr.

	P	N
P	0	0
N	10	$10^9 - 10$

$$\text{Acc} = \frac{10^9 - 10}{10^9}$$

$$= 1 - 10^{-8}$$

$$= 0.99999999$$

or $\boxed{99.999999\%}$

Recall (Sensitivity or TPR)

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

out of all terr. what fraction did you identify label all as Not a

1 Billion   10 terr.

	P	N
P	0	0
N	10	$10^9 - 10$

label all as terr

	P	N
P	10	$10^9 - 10$
N	0	0

$$Acc = \frac{10^9 - 10}{10^9}$$

$$= 1 - 10^{-8}$$

$$= 0.99999999$$

$$\text{or } \boxed{99.999999\%}$$

Recall (Sensitivity or TPR)

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

out of all terr. what fraction did you identify

Precision

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

out of all predicted terr. what frac. were really terr.

	Acc	Recall	Precision
all as Not terr.	1	0	0
all as terr.		1	low
predict the top terr. only		0	1

$$F_1 = \frac{2 \times P \times R}{P + R}$$

Predictive Models

- Machine Learning fundamentally differentiates itself from classical statistics by not assuming that the data comes from a specific model
- Hence ML, justifiably, can try different models on a given dataset to eventually pick the **best** one.
- Obviously, to do this one needs to define what it means to be **best.**
- Different "model performance measures" exist and any of these can be used to compare models - largely depending on the context and the kind of output:
  - Regression outputs are continuous numbers
  - Classification outputs are either
    - Class output (from algorithms like SVM and KNN that usually give a classification) or
    - Probability output (from algorithms like Logistic Regression, Random Forest that can give probability outputs)

- For classification problem with a class output, the confusion matrix gives the counts of correct and erroneous predictions:

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

Actual

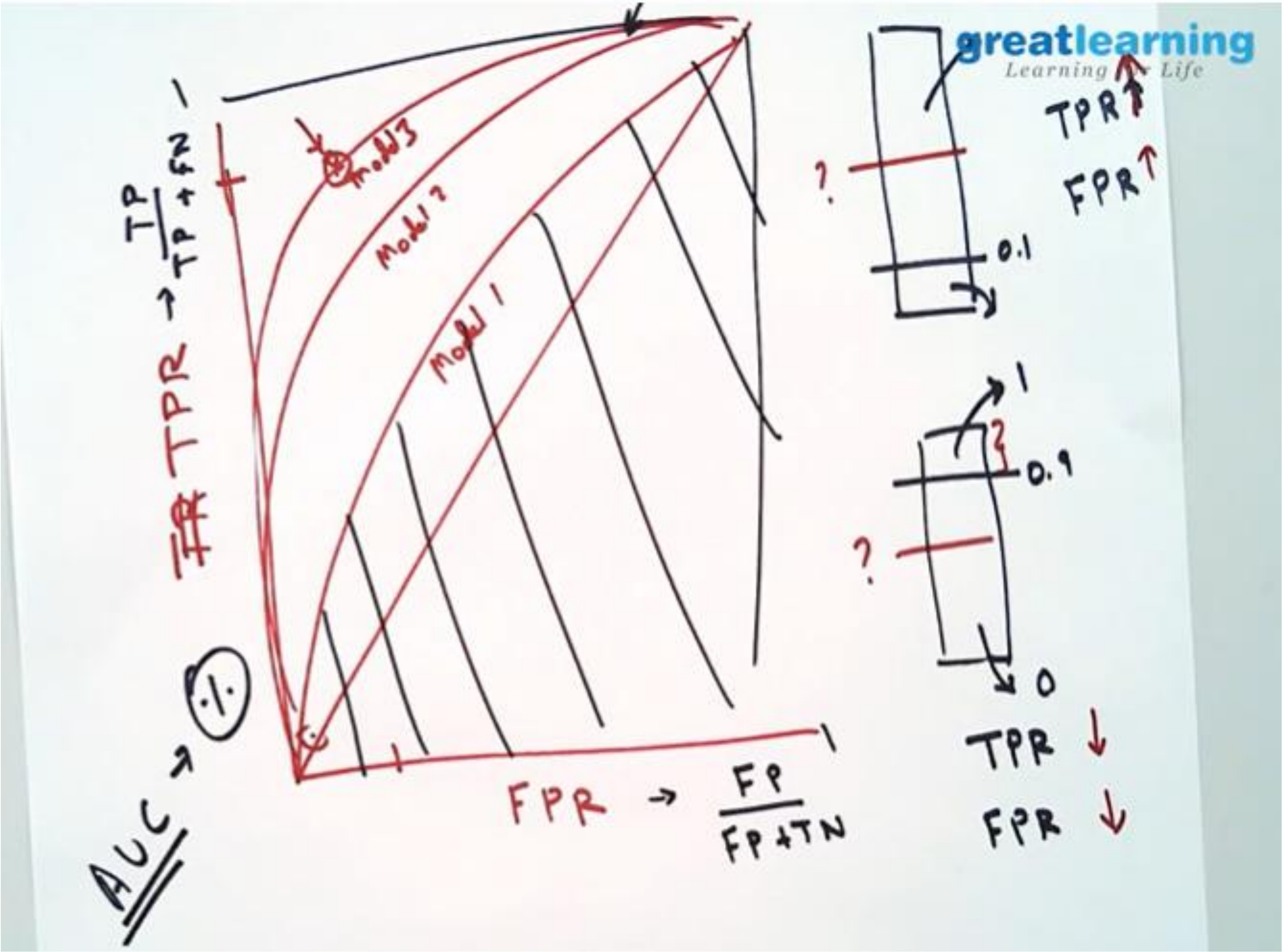
Predicted	1	0
1	$\boxed{TP}$	FP $\rightarrow$ I
0	FN $\rightarrow$ II	TN

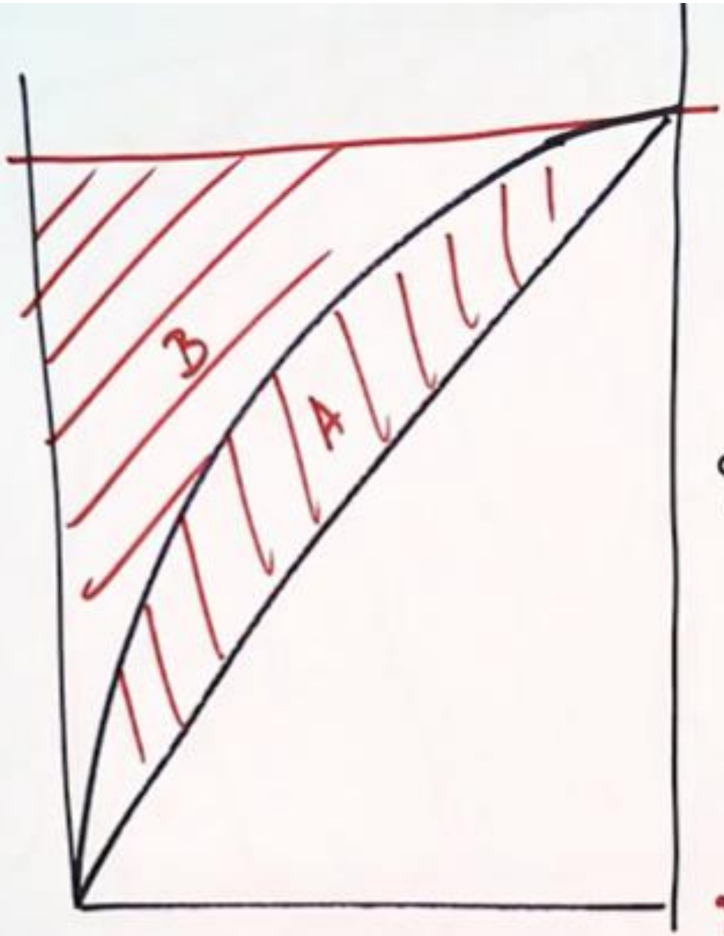
- Classification Error Rate: sum of Type 1 (FP) and Type 2 (FN) Errors (in percentage). Accuracy is 1-(error rate).

- Sensitivity (also called Recall or True Positive Rate): proportion of Total Positives that were correctly identified $\rightarrow \frac{TP}{TP + FN}$

- Specificity (also called True Negative Rate): proportion of Total Negatives that were correctly identified $\rightarrow \frac{TN}{TN + FP}$

P	A	$\xrightarrow{0.5}$
0.9	1	1
0.51	0	1
0.3	1	0
0.8	1	1
0.3	0	0


TPR → TP / TP+FN
FPR → FP / FP+TN
model 3
model 2
model 1
AUC
TPR ↑
FPR ↑
0.1
1
0.9
0
TPR ↓
FPR ↓



~~AC~~

AUC

$$\text{gini coeff} = \frac{A}{A+B}$$

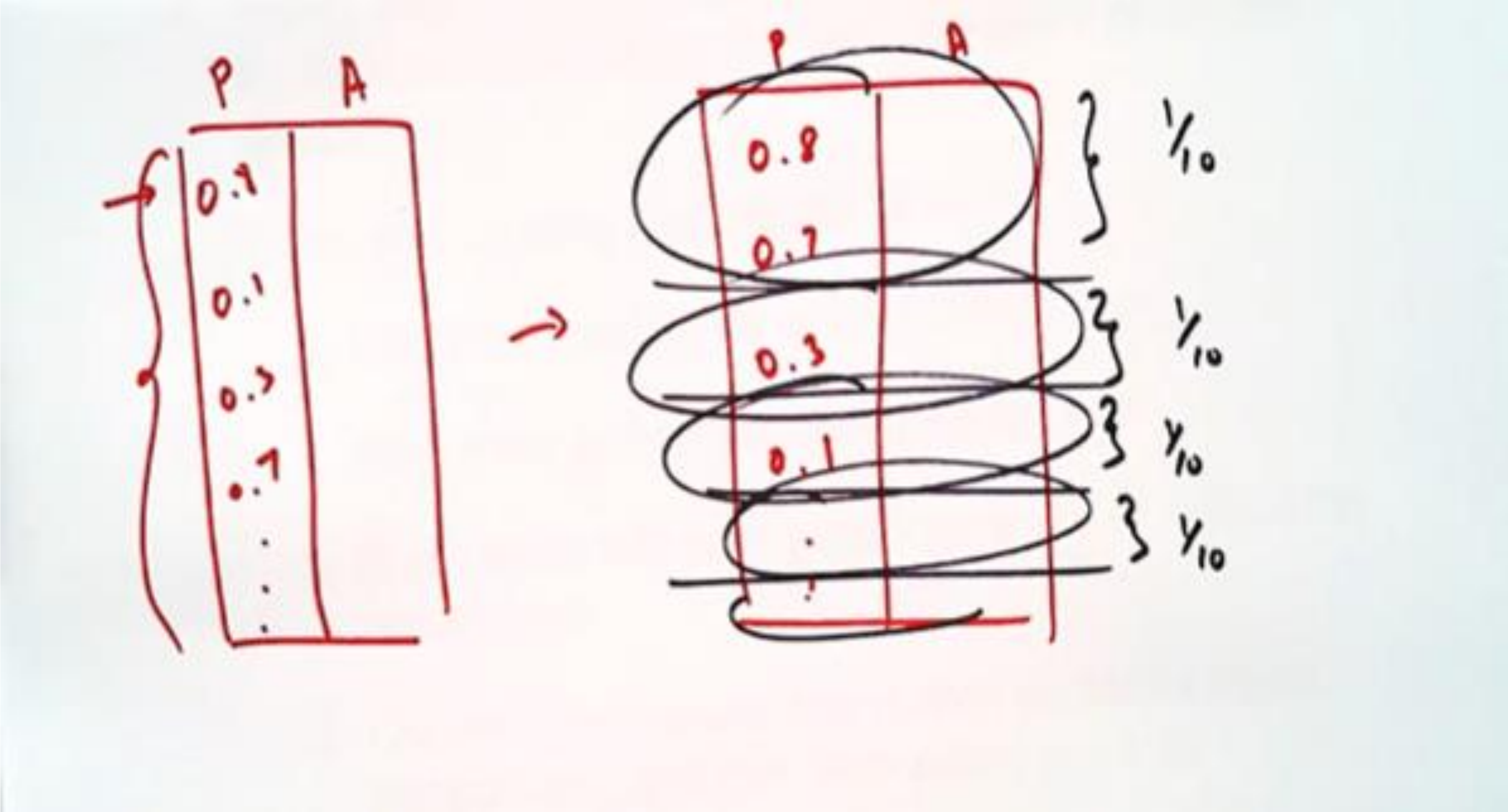
$$A = AUC - 0.5$$

$$A + B = 0.5$$

$$\text{gini coeff} = \frac{AUC - 0.5}{0.5}$$

$$\boxed{Gini = 2AUC - 1}$$

Gain & Lift Charts



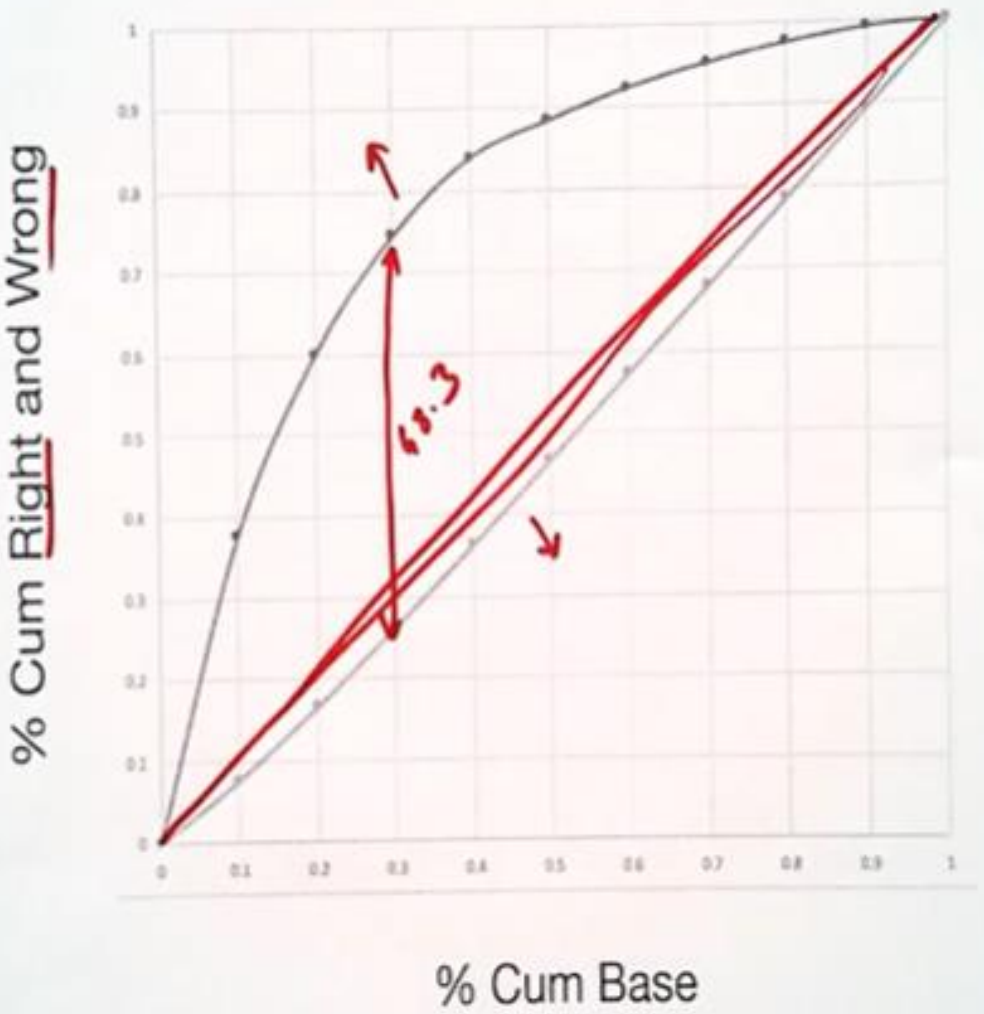
Rank Ordered Table Example

greatlearning
Learning for Life

Rgt wron Ri wro
787

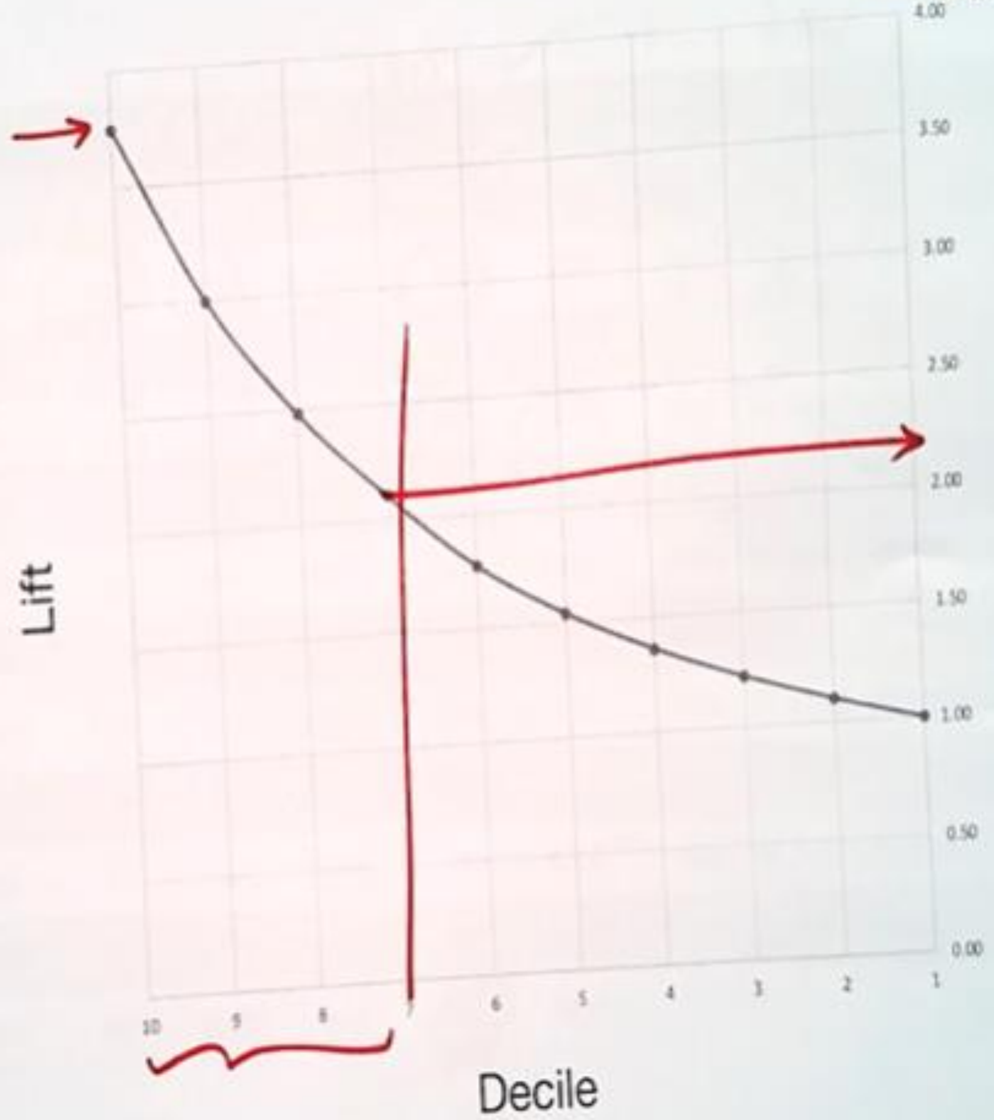
Decile	Base Cnt	#Right	# Wrong	%Right	Cum. Resp.	Cum. Non-Resp.	%Cum. Base	%Cum. Resp.	%Cum. Non-Resp.	KS	Lift
A	B	C	D=B-C	E=C/B	F=CumSum(C)	G=CumSum(D)	H=CumSum(B)/Total	I=F/Total	J=G/Total	I-J	I/H
10	1000	295	705	29.50%	295	705	10%	37.48%	7.65%	29.83%	3.75
9	1000	176	824	17.60%	471	1529	20%	59.85%	16.60%	43.25%	2.99
8	1000	115	885	11.50%	586	2414	30%	74.46%	26.20%	48.26%	2.48
7	1000	75	925	7.50%	661	3339	40%	83.99%	36.24%	47.75%	2.10
6	1000	35	965	3.50%	696	4304	50%	88.44%	46.72%	41.72%	1.77
5	1000	30	970	3.00%	726	5274	60%	92.25%	57.25%	35.00%	1.54
4	1000	23	977	2.30%	749	6251	70%	95.17%	67.85%	27.32%	1.36
3	1000	18	982	1.80%	767	7233	80%	97.46%	78.51%	18.95%	1.22
2	1000	13	987	1.30%	780	8220	90%	99.11%	89.22%	9.89%	1.10
1	1000	7	993	0.70%	787	9213	100%	100.00%	100.00%	0.00%	1.00
Total	10000	787	9213	7.87%	787	9213					

K-S Chart



Lift Chart

greatlearning
Learning for Life



Performance Measures

greatlearning
Learning for Life

- Confusion Matrix
- ROC Curves, Gini Coefficient
- Gain and Lift Chart
- Kolomogorov-Smirnov (K-S) chart
- Concordance-Discordance ratio
- Root Mean Square Error, Mean Absolute Error

greatlearning
Learning for Life

Name	Right?	Prob
A	0	0.056
B	0	0.134
C	0	0.156
D	1	0.512
E	0	0.235
F	0	0.25
G	1	0.25
H	1	0.2
I	0	0.135
J	0	0.089

A P

D, A → 0.512, 0.056 → ✓ C

D, B → 0.512, 0.134 → ✓ C

D, C

D, E

D, F

D, I

D, J

G, A

G, F → 0.25 0.25 → O T

H, E → 0.2, 0.235 → × D

Concordance Ratio = $\frac{18}{21}$