# Interview Prep Session 1

## [Classical Machine Learning]

# Warm Up Questions

Q: If you are training a LR to predict Farhite using Celcius as independent var. what will be the value of the intercept? (Quiz)

→ Automated | Easy

Soln → We know →

$$F = \frac{9}{5} \times C + 32$$

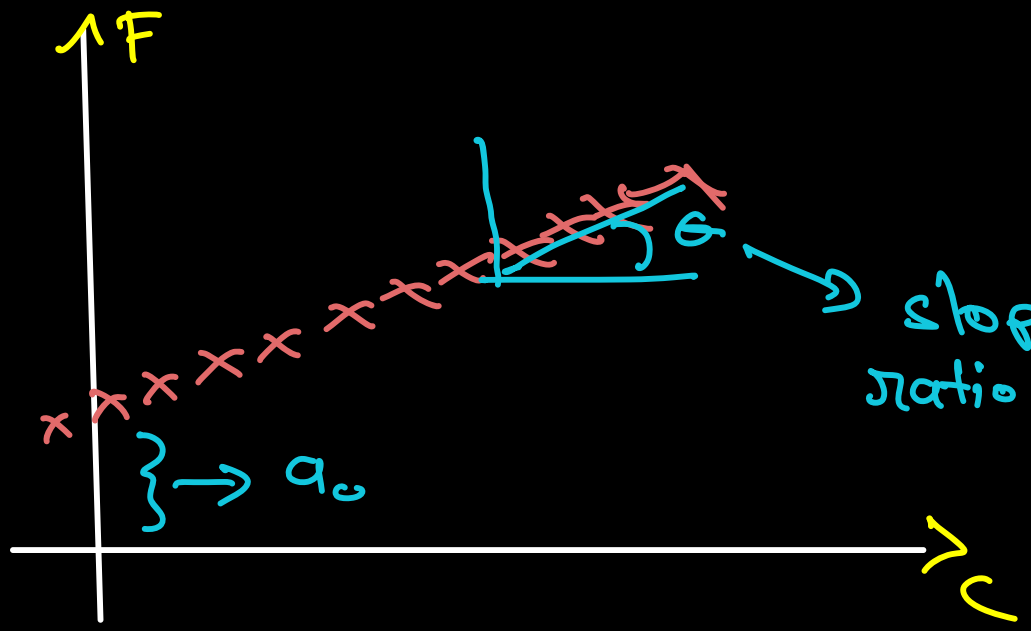
Train a LR, st.

$X = [-, -, -, -, -]$ ] Readings in C°

$y = [\ldots]$ Readings in $F$

We need $f(x)$, s.t

$$X \to f \to \hat{y} \sim y = \frac{9}{5} \cdot x + 32$$

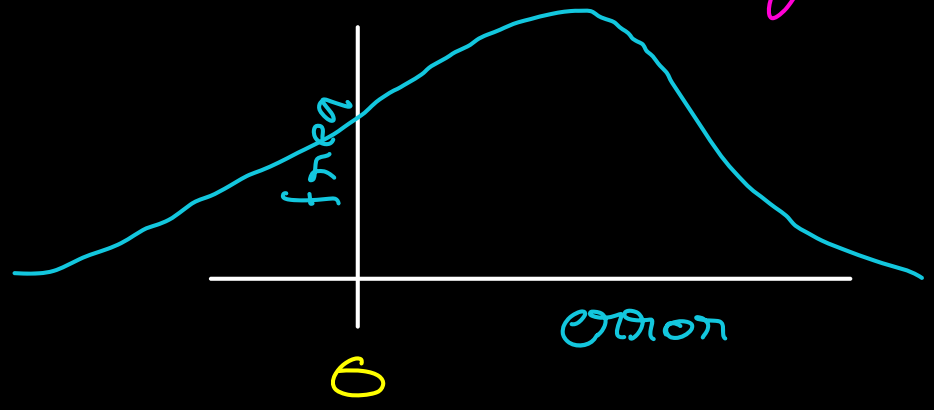



$$\therefore \hat{y} = \underbrace{a_1}_{\frac{9}{5}} C + \underbrace{a_0}_{32}$$ → intercept

Hence answer = **32.**

Q: After training a LR model, the residual error has the following distribution

Should this model go into production? (Quiz)




$-\infty$ $+\infty$

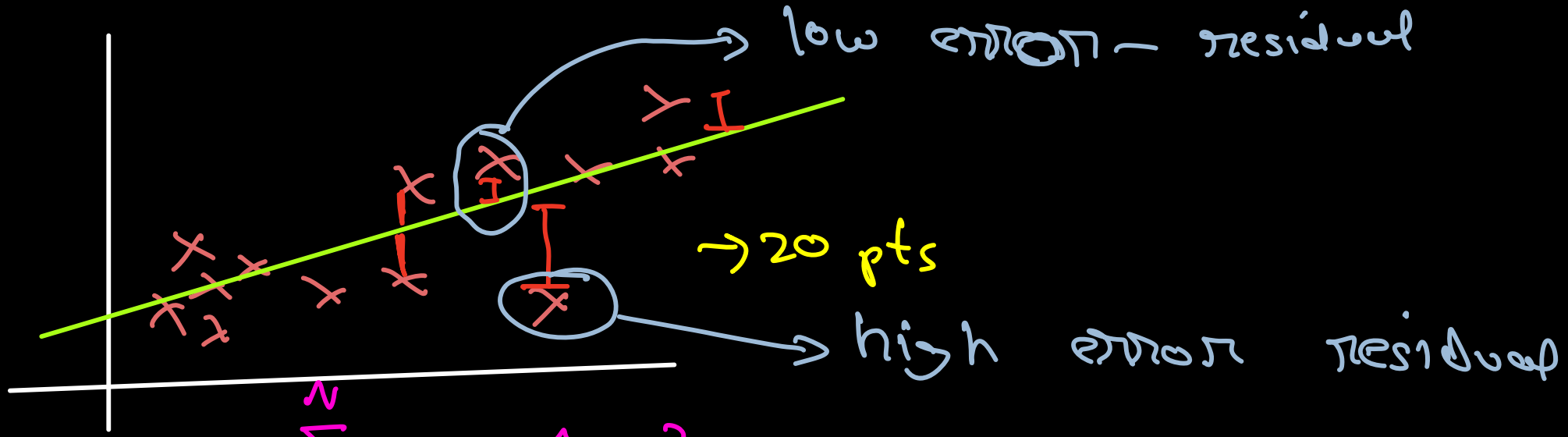
→ Automated | Easy

Sol$^n$ → No

→ The error dist is not normal, assuption violated

→ The mean of residual is non-zero, +ve, indicating high tendency to over predict

# Concept: Regression Residuals




$$MSE = \sum_{i=0}^{n} (y - \hat{y})^2$$

$(y - \hat{y})$ → residual

$$R = \underbrace{y - \hat{y}}$$

$$MSE = \sum_{i=0}^{n} r_i^2$$

$$r_i = y_i - \hat{y}_i$$

# Importance of Residuals

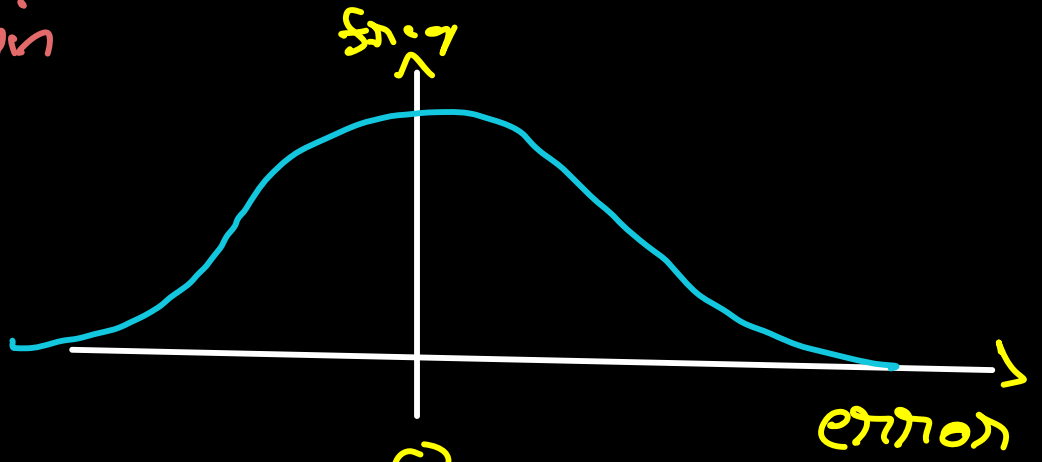
→ We just saw → all error metrics are computed on residuals

↳ **Analysis**

$X, y \xrightarrow[\text{train}]{} f(x) \rightarrow \hat{y}$ ; $y - \hat{y} = r$

n pts.

↳ histogram

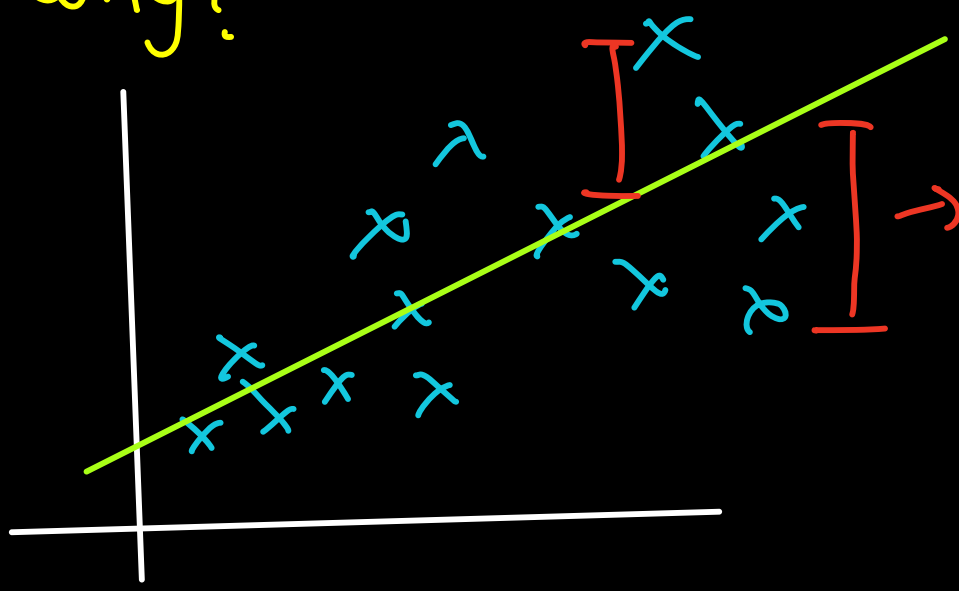



→ Ideally for LR → dist should be normal

→ Mean should be 0

→ Variance : lower the better

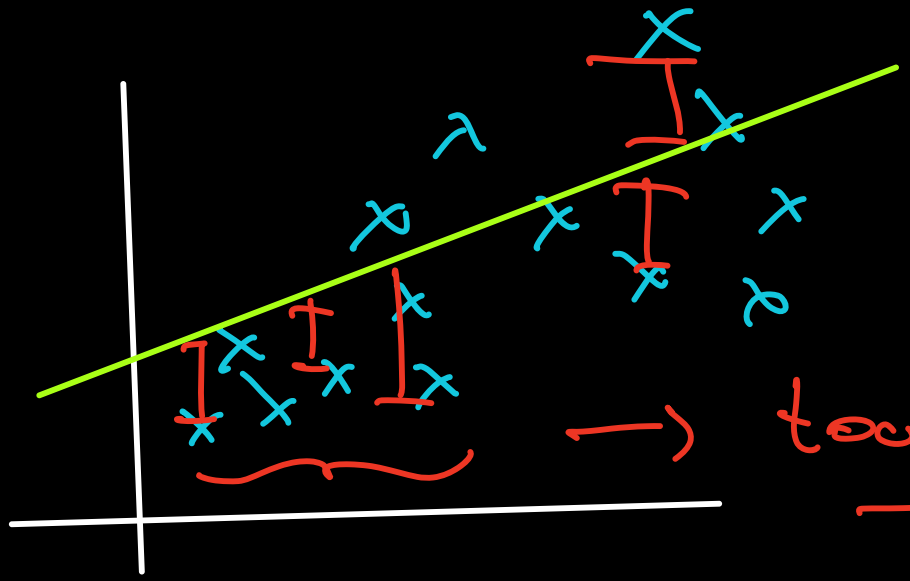
Why?



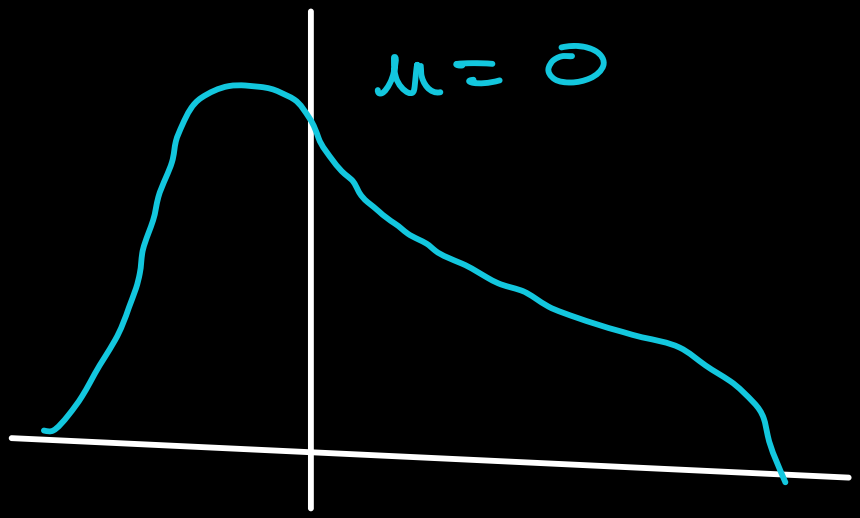
→ error!!

But overall
-ve error
~ +ve error

→ We want symmetric errors.



→ too much -ve error } → Over prediction




→ We want to predict "error rate" in test set.

→ In this model
  → if error $> 0$, error is low
  if " $< 0$, error is high
  ⇒ assymetry !!

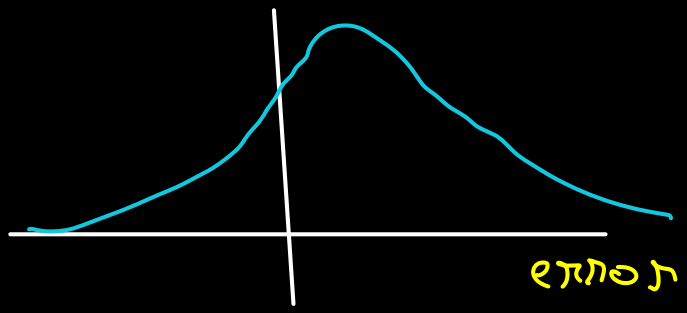
To make decisions we need confidence

→ my prediction is $53.5 \pm 3$ (95%).

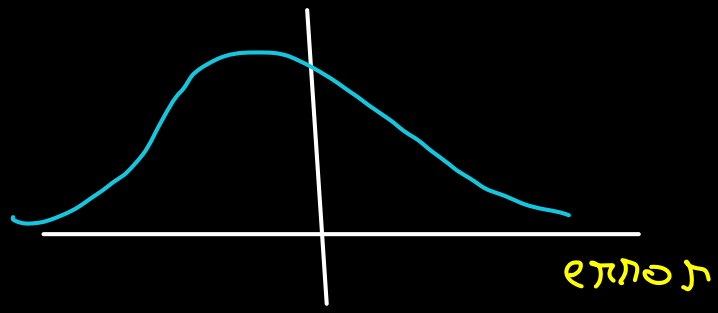
Q If a retail company has 50% profit margin for each sale, and 80% salvage price for unsold goods, while predicting Order Quantity, which of the following models seem best? (Quiz)

→ Automated / Live | Medium

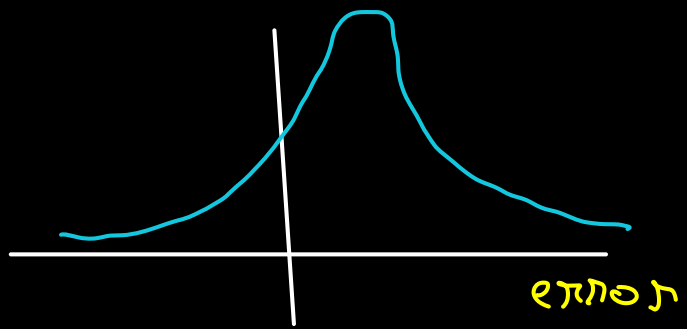
a)



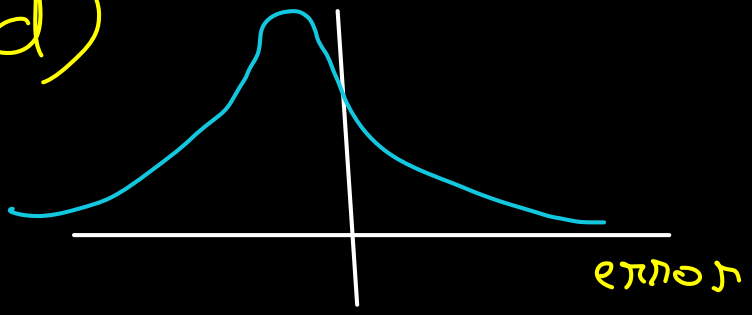

b)




c)




d)



Salvage price = Discounted price @ unsold goods are sold (Company looses money)

The company wants max profits

if over-produce → need to sell at salvage price
↳ Loss = 20%

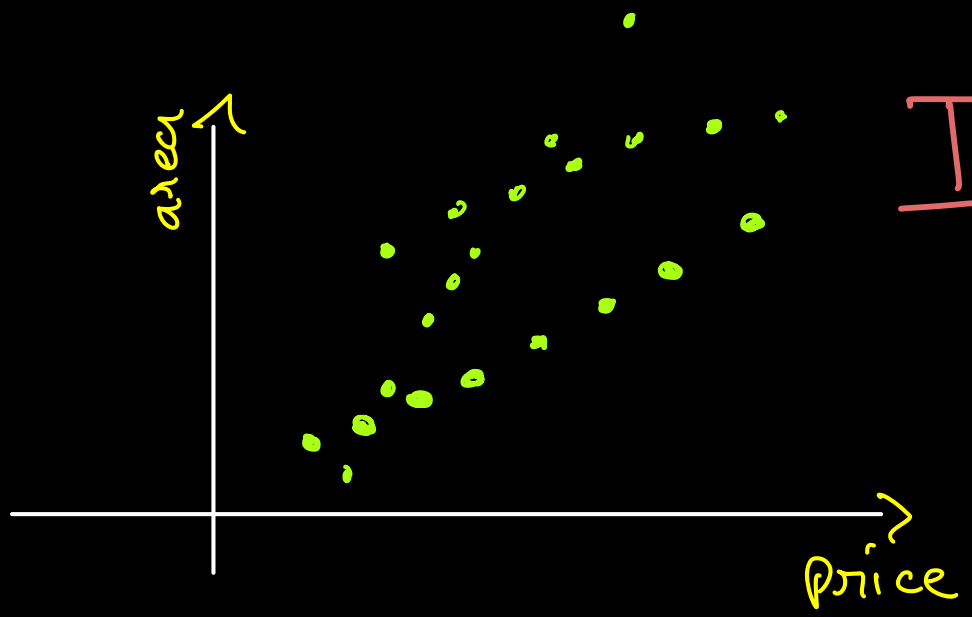
if under-produce → Loosing potential profit
↳ Loss = 50%

Here, better to overpredict → $y - \hat{y} \geq 0$
→ Next, choose one with low variance (more accurate)

Hence Ans → **d**

Q You have been given house prices for a certain city. The only variables available are latitude, longitude and square footage. One scatter plot is shown below: How would you approach to build a price prediction mode?

→ Live | Medium (open ended)



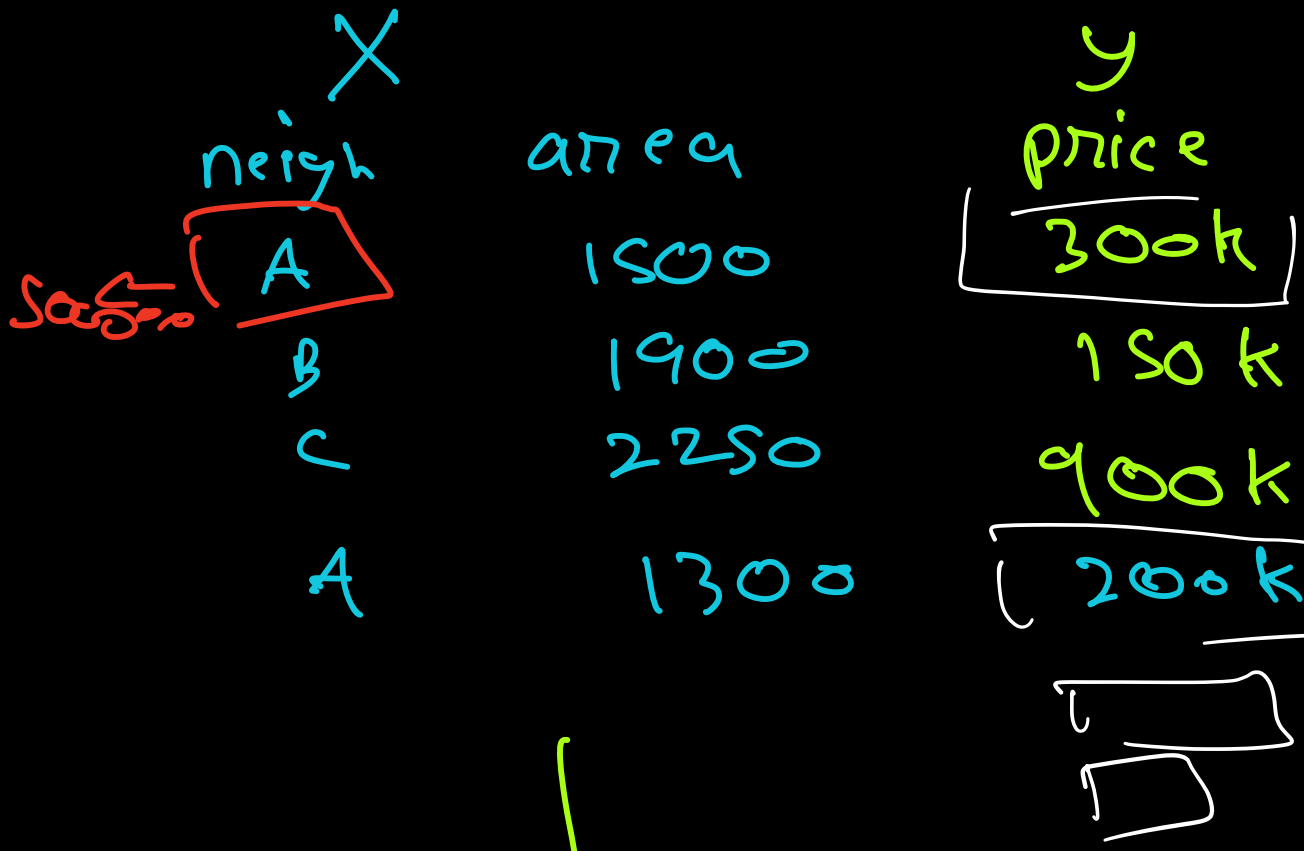

I → ISO

→ Soln → train a simple regression using any kind of model SVM, DT, LR..

incorrect → bad performance.

2) Use lat - long info to get neighbours

→ Google → Assign 3 neigh..



| N | A | P |
|---|---|---|
| 500k | 1700 | |
| 350k | 225 | |
| 1000k | 130 | |
| | | |

LR

$\hat{y} = a_0 + a_1 N + a_2 A$

lets assume $a_1 = 1$

For $N = A$ (1st neighbour)

$\hat{y} = a_0 + 500K + a_2 A$

For $N = B$ (2nd neighbour)

$\hat{y} = a_0 + 350K + a_2 A$

By design -> $A > B$

-> 150 K

# Use clustering (ML based on Manual)

→ Different LR for each Neighbourhood.

$x = (A, 1000)$ → $N = A$

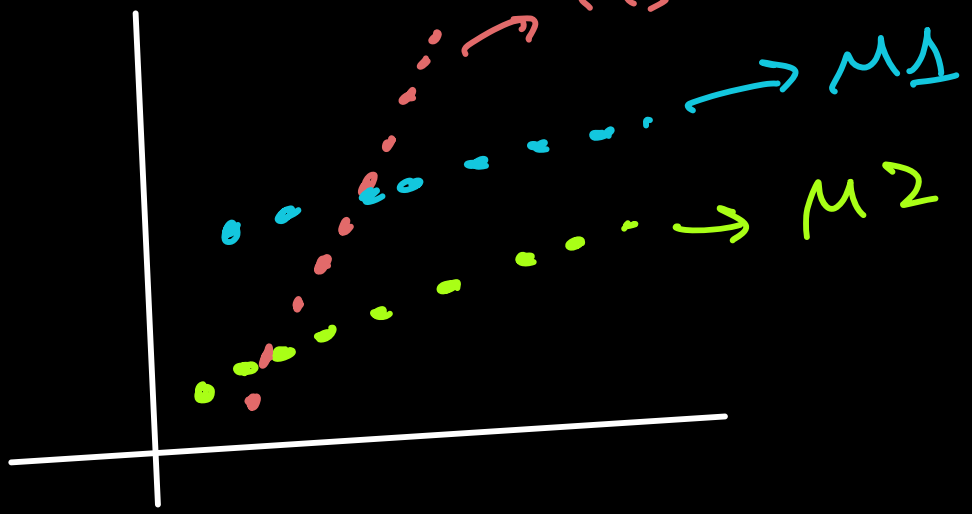
(input now)

hence → select M1

$M_1 = a_0^1 + a_1^1 A = \hat{y}$

if

$N = B$, $M_2 = a_0^2 + a_1^2 A = \hat{y}$




Each model is highly accurate

Q: For a KNN classifier, as we increase the value of K, does the model move towards

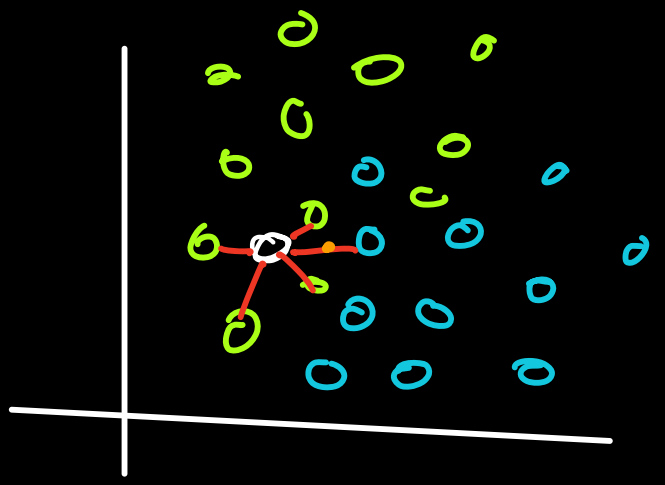
(Quiz)

→ overfitting

→ underfitting

→ Automated + Live | Easy

Sol^n.



→ check dist from K nearest neighbours

K = defined by DS

↳ hyper parameter

let K = 5

4 + 1

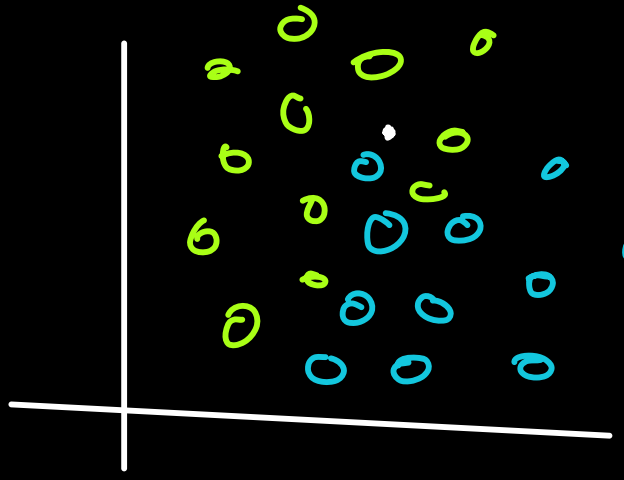
↳ voting → green

As we increase K, what happens?

Let K = 100M

↳ will always take majority class

→ underfit

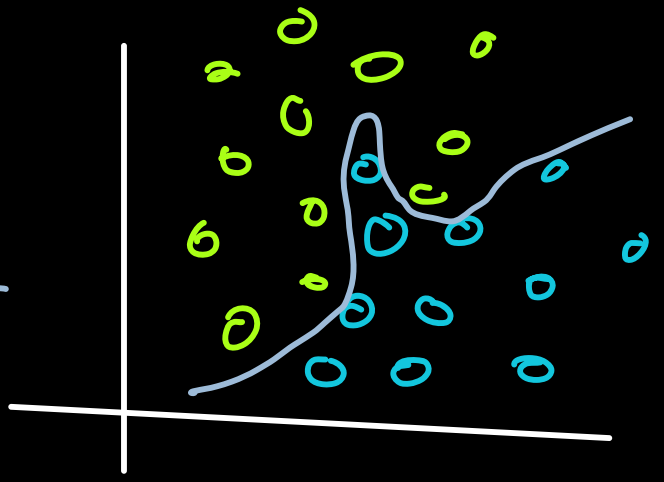


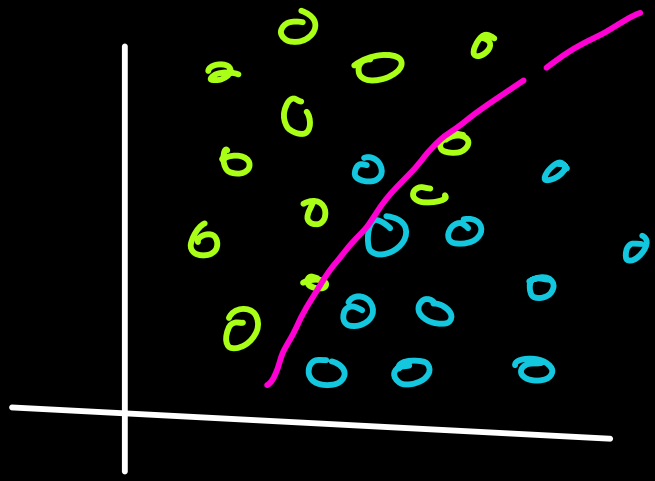
K=3, class = green

K=1, class = blue

→ K=1, DB

K=1





K=4

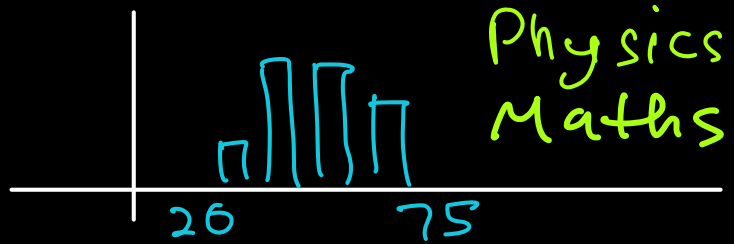
Q: Your colleague has trained a Tree based regression model, to predict the marks of students for a very hard test. The max marks for all exams are 100.

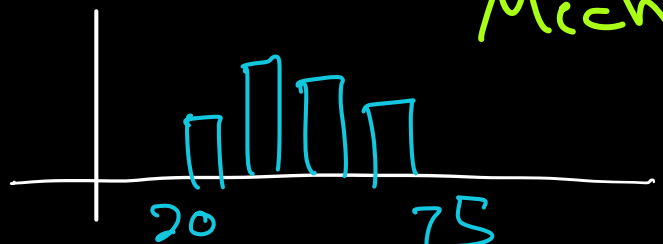
→ Live | Medium (open ended)

| Physics | Maths | Mechanics → target |
|---|---|---|
| 65 | 68 | 74 |
| 72 | 67 | 70 |
| 35 | 40 | 43 |

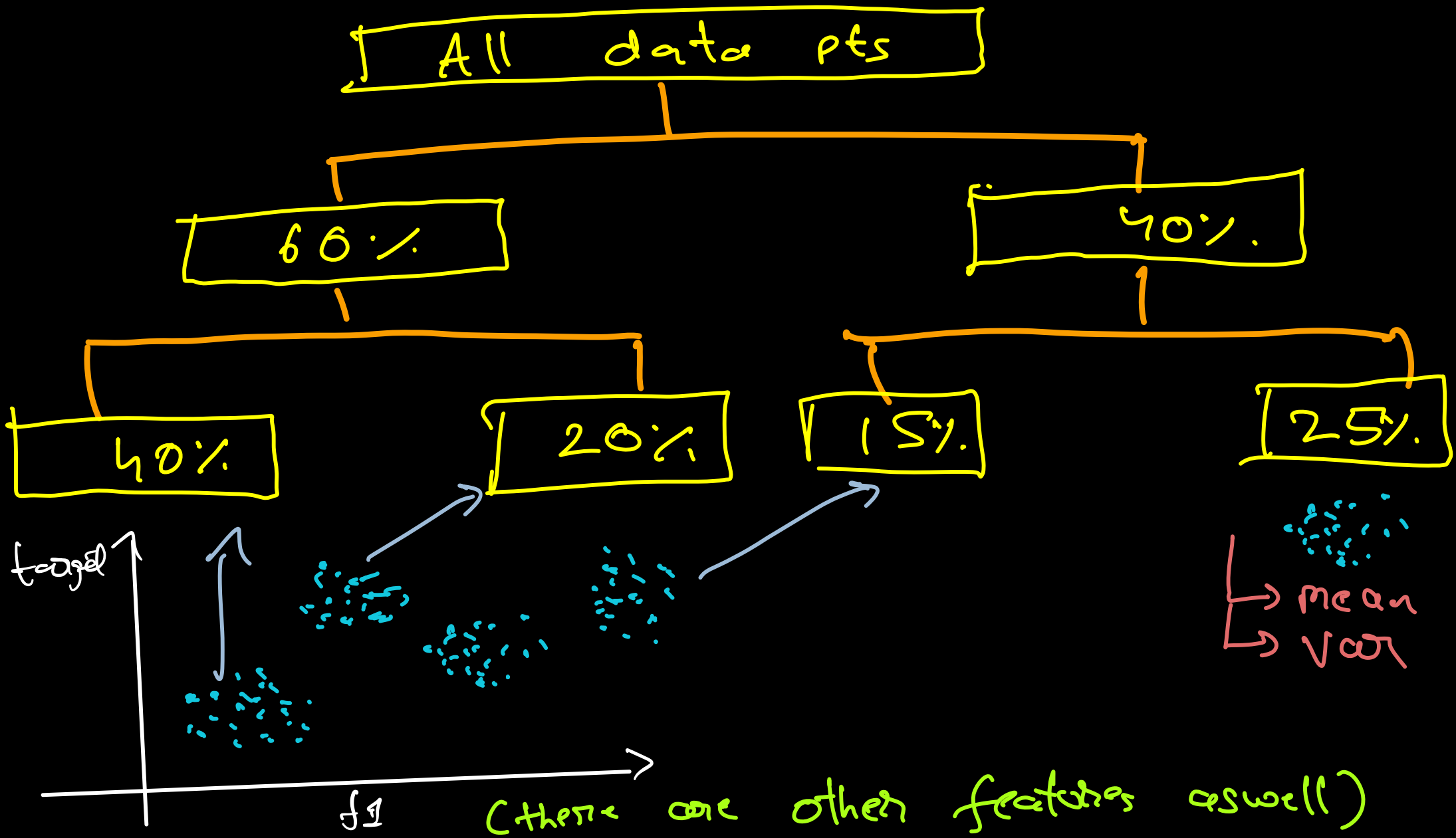
The model MAPE is 1%. Can we put this into production?

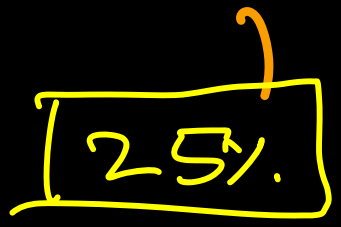
Distributions:



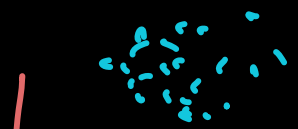
# Let's understand how a DT works



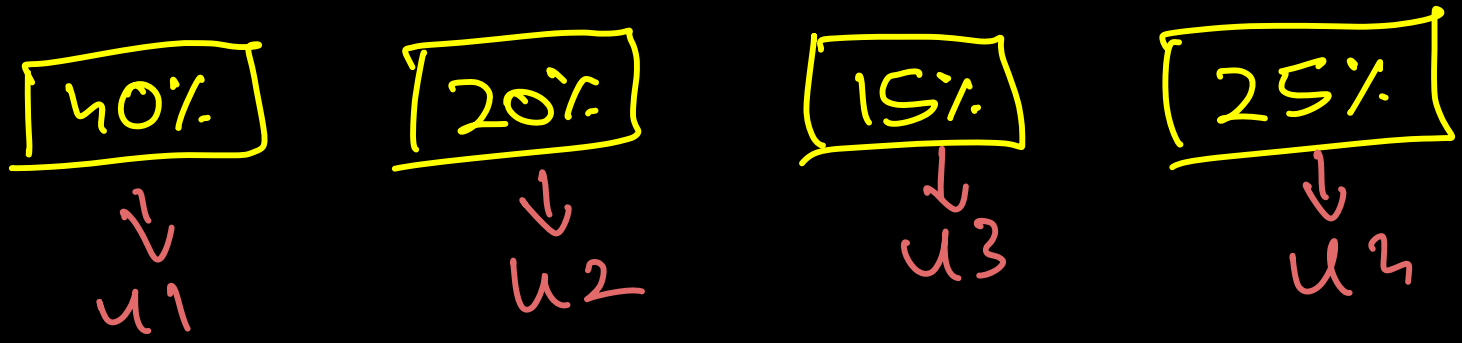



output → mean :→ more reliable

if ↓ Low



→ mean
→ var → Variance




Q: What will be the o/p range??

→ [u1, u2, u3, u4]

=> Yes, there will be only 4 unique values in regresson o/p in this case

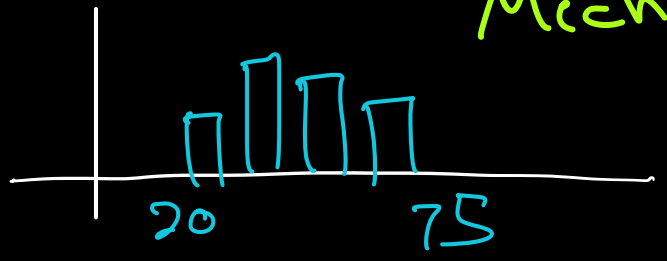
Had you guys realized this?

For regressions, we use RFR → so we get many (but finite) unique values.

because it averages many trees.

Tree based Regressor can never o/p anything outside of training data 'target' range.

Target has a range of only 20–75 in training data

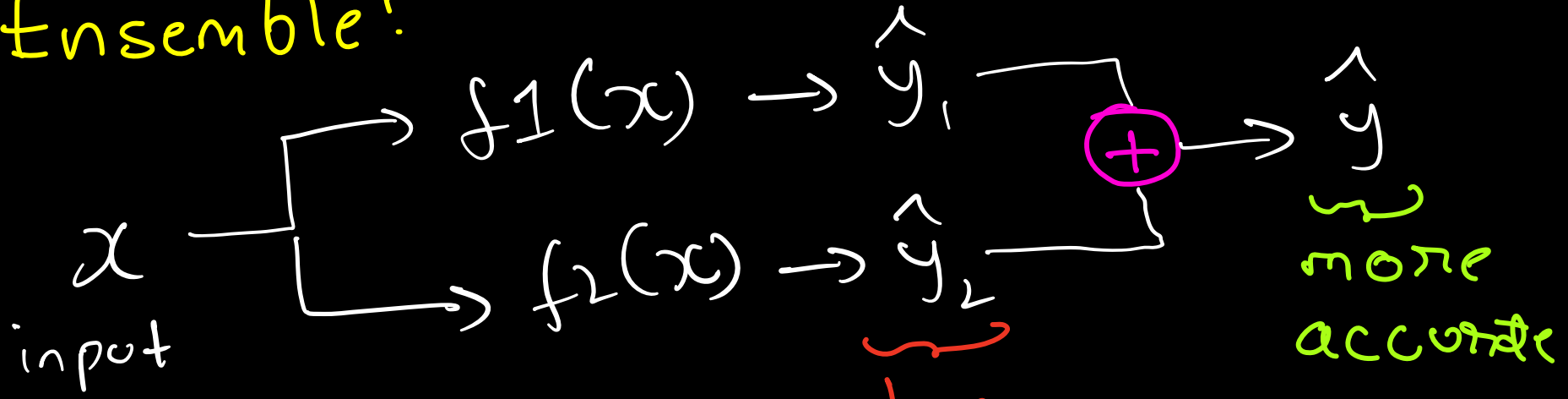



Mechanics

Linear regressor would be better

Q: How do you create ensemble model from KNN and DT ? (open ended)

Ensemble:



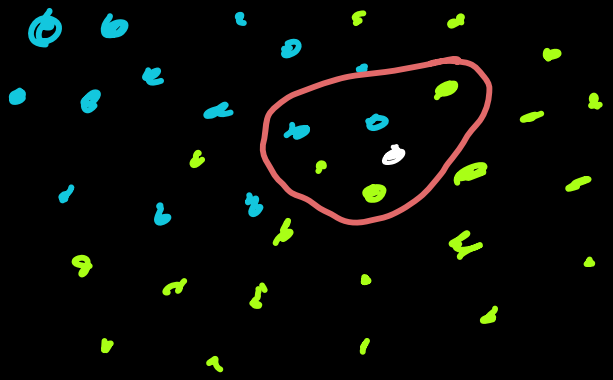

$f_1$ = KNN

$f_2$ = DT

$\oplus$ = ?

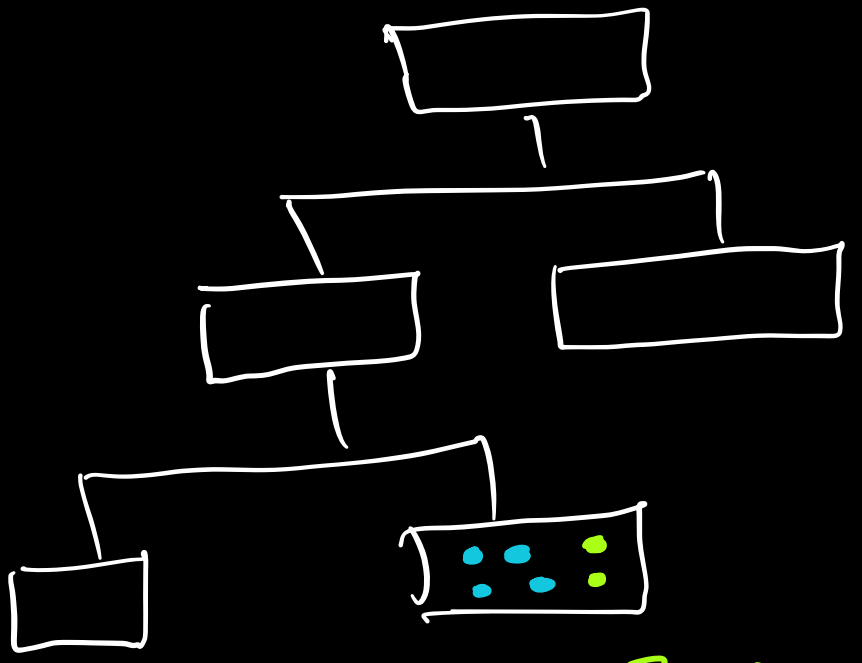
Soln

Since KNN is purely a classification model, we need to create ensemble for



2 blue 3 green

output → green



4 blue 3 green

output → Blue

Voting ? = Tie

↓

keep odd number.

Better Idea → Use Probabilities

KNN Prob :

$\frac{2}{5}$ $\frac{3}{5}$

DT Prob :

$\frac{4}{6}$ $\frac{2}{6}$

Adding up:

blue $(2/5 + 4/6)/2 = 16/30 = 8/15$

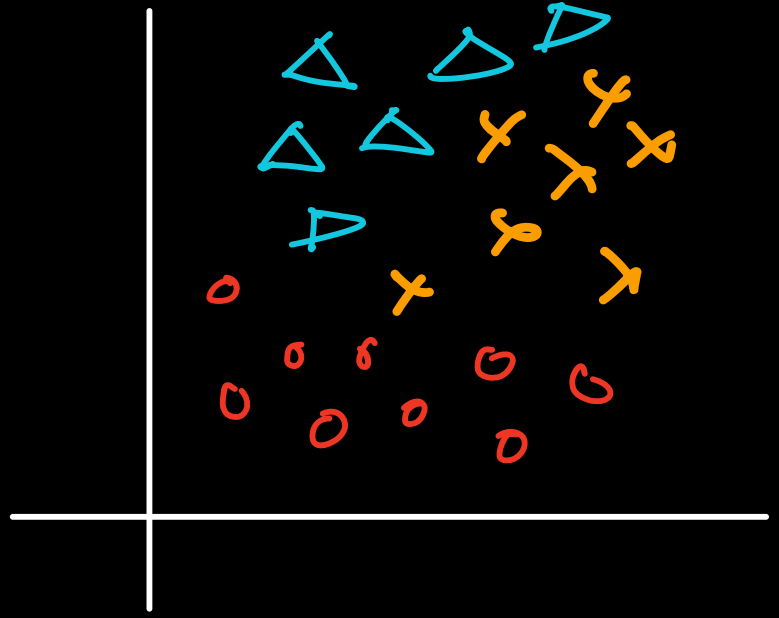
green $(3/5 + 2/6)/2 = 14/30 = 7/15$

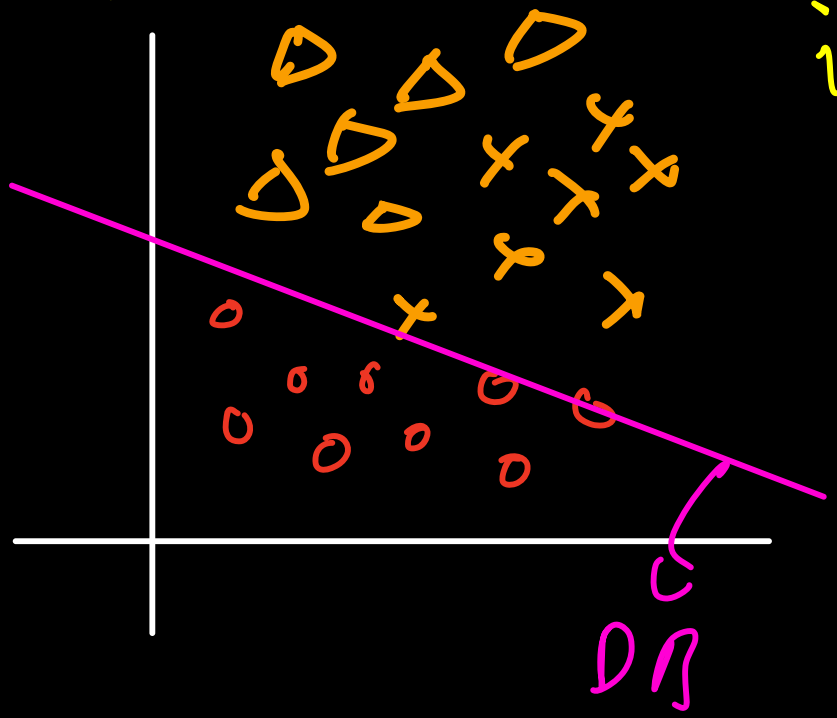
↳ output → Blue

Q: How can you use Log Reg for a multiclass problem statement ?

→ Live | Easy (Quiz)

Soln One -vs- Rest framework.





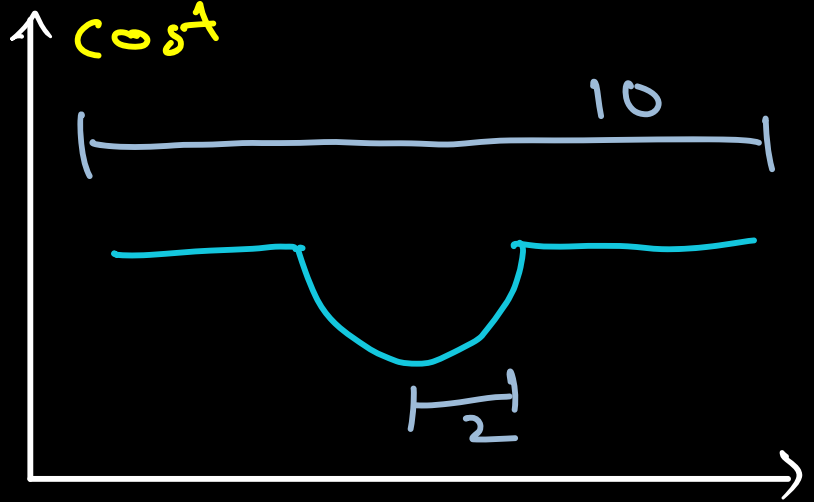

if M1 = 1
↳ Red
elif M2 = 1
↳ Blue
else
↳ Orange

Make 2 LogRegs → Red vs Rest
blue vs Rest

Q: What is the probability of finding the optimum value using simple G.D $(\theta_1, \theta_2)$ if the cost function looks like this?

↳ 2D view w.r.t $\theta_1$

↳ Same view w.r.t $\theta_2$




→ Automated | Medium (Quiz)

So (~→ If you init on flat surface, G.D will be stuck

$10 \times 10 = 100$

$\pi (2)^2 = 4\pi$ → prob = $\frac{4\pi}{100}$

Q: Your teammate tried adding a new feature to the training data and has reported that $R^2$ has dropped.

What is your suggestion? (Quiz)

→ Live | Medium

→ New feature is not important, should remove it.

→ Colleague has made a mistake, $R^2$ cannot reduce in this case

→ Feature could be correlated with another feature. → Deep dive

→ Do hyper-param tuning first, then take a decission

Soln

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad ; \bar{y} = \text{mean}$$

⇒ How much better is the prediction compared to just the mean

Now, let's recall LR.

$$\min_{\beta} SSE = \min_{\beta} \sum_{i=1}^{n} (y - \hat{y})^2$$

$$= \min_{\beta} \sum_{i=1}^{n} (y - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots \beta_n x_{in})$$

↳ G.D will make $\vec{\beta}$ such that SSE is min → S1

Now when we add one more Var.

$$= \min_{\beta} \sum_{i=1}^{\hat{}} (y - \beta_0 - \beta_1 x_{i1} \dots \beta_n x_{in} - \underbrace{\beta_{n+1} x_{in+1}}_{\text{This is extra}})$$

G.D returns → S2

$\therefore$ $\boxed{S2 \leq S1}$ Because, if $\beta_{n+1} = 0$, $S1 = S2$

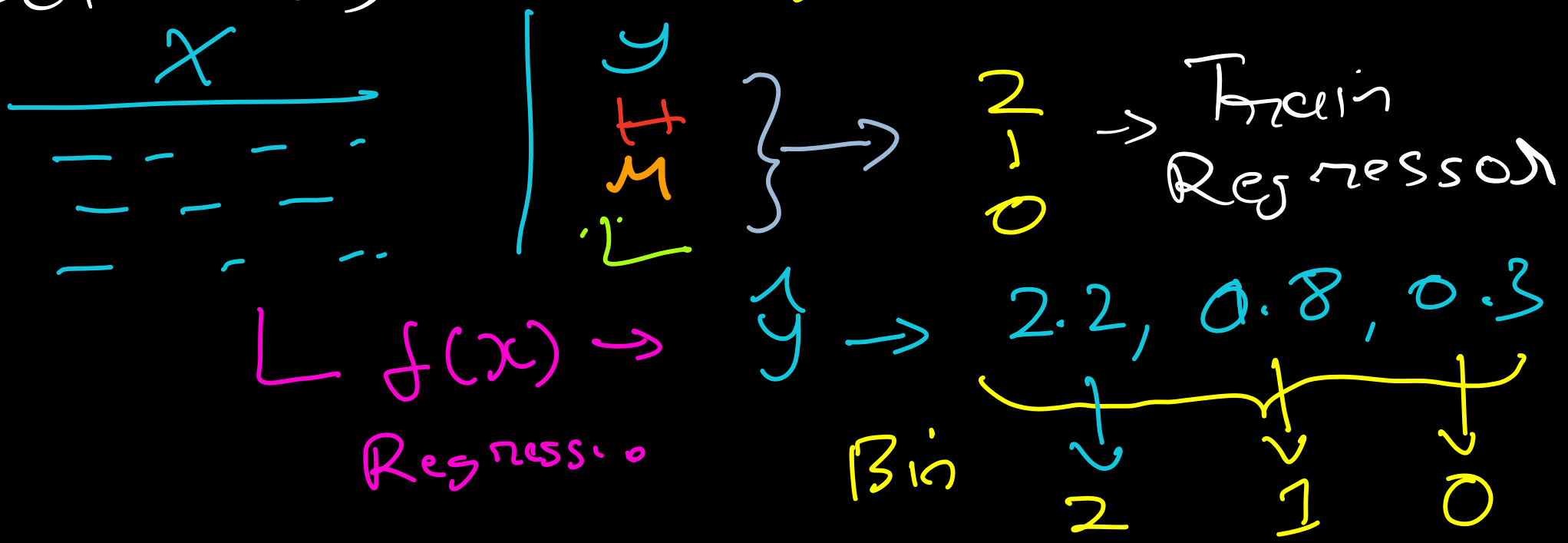
OR, $\beta_{n+1} \neq 0$; in a way that $S2 < S1$

Q: How would you approach a cancer severity classification problem? H M L ?

(open ended)

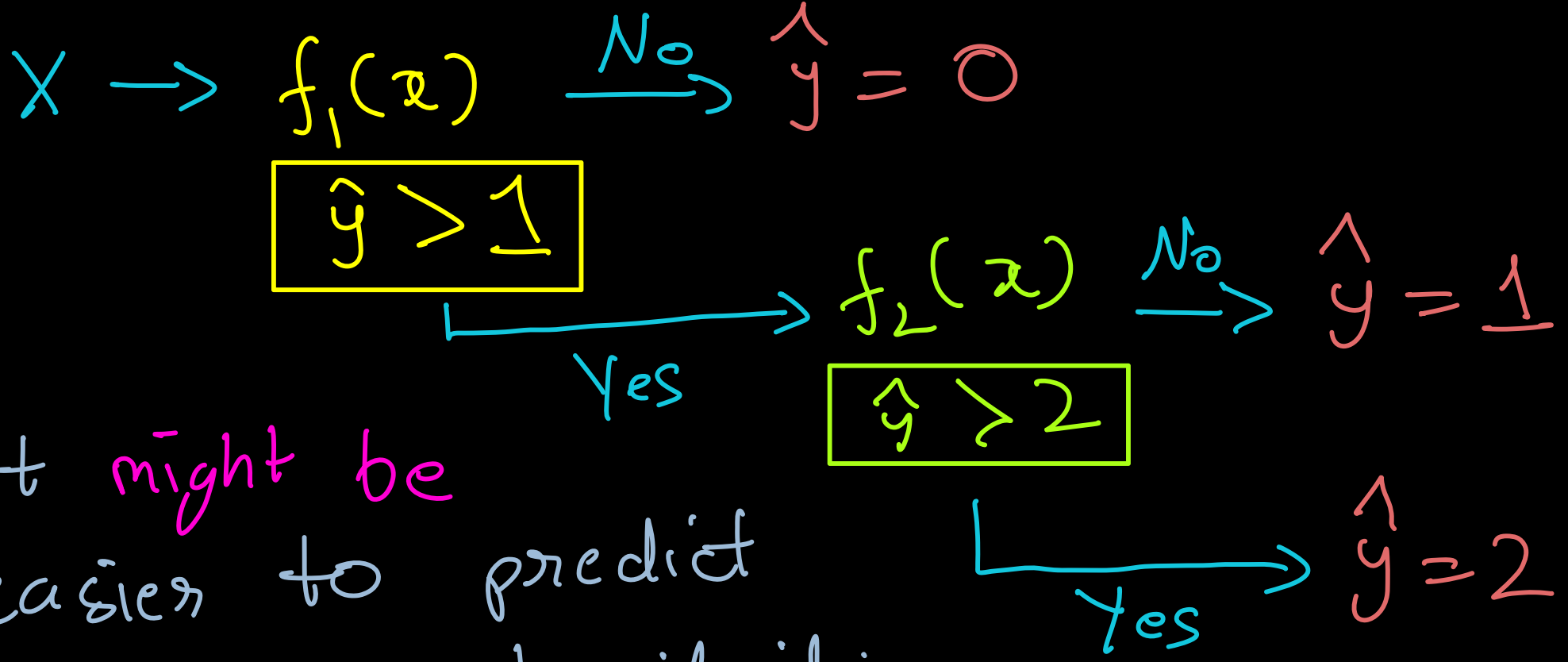
→ Live | Medium

Soln → 1) Use Regression and bin




Binning may have issues at the boundaries

# Sol$^{n}$ - 2) Classification for Ordinal Classes




It might be easier to predict $\hat{y} > 1$ as a classification compared to predicting exact value of $\hat{y}$.

Q: Can you desing a Gradient Descent algorithm, to optimize a cost function fith 4 parrams $\alpha, \beta, \gamma, \theta$, when $\beta$ is a boolean variable and can only be 0 or 1?

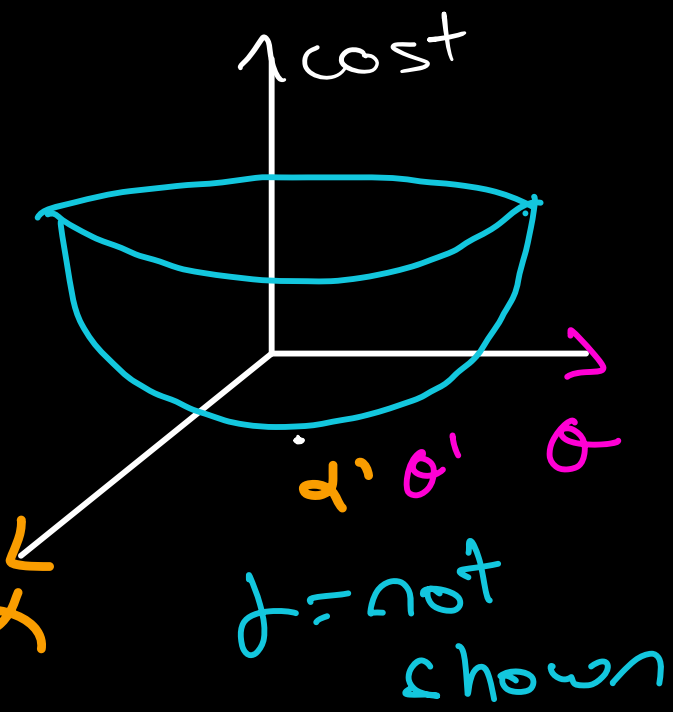
→ Live | Hard (open ended)

Soln

Problem → G.D needs cost func$^n$ to be differentiable w.r.t $\alpha, \beta, \gamma, \theta$

But $\beta$ is not differentiable

$\frac{\partial}{\partial \beta}$ = not defined



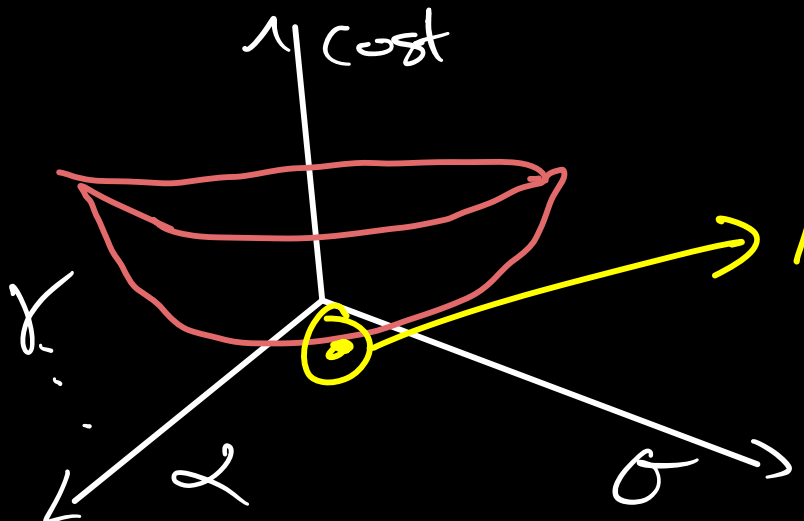
# Q: Any Ideas?

1) Assume $\beta$ to be continuous from $0 \rightarrow 1$. Use GD where $\beta$ is constrained

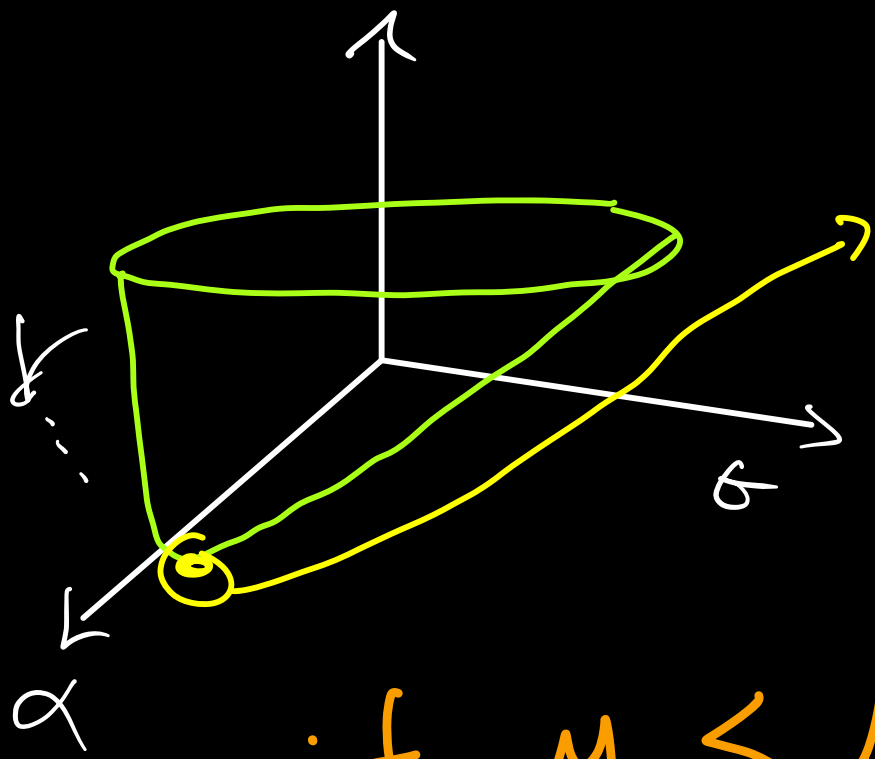
↳ IF, GD outputs $\beta = 0.09$
↳ $\beta = 0$

if, $\beta = 0.65 \rightarrow \beta = 1$ (suboptimal)

2) Fix $\beta = 0$, Solve GD with $\alpha, \gamma, \theta$.




$M_1 \rightarrow$ $\alpha = \alpha_1$, $\beta = \boxed{0}$, $\gamma = \gamma_1$, $\theta = \theta_1$ } set 1




$M_2 \rightarrow$

$\alpha = \alpha_2$

$\beta = \boxed{1}$

$\gamma = \sigma_2$

$\theta = \theta_2$

} set 2

if $M_1 < M_2$

$out = \{\alpha_1, 0, \gamma_1, \theta_1\}$

else

$out = \{\alpha_2, 1, \gamma_2\ \theta_2\}$

Q: Based on product attributes, how would you model, a product assortment problem at Amazon? Please ack any further data points that you require.

(open ended)
(long disc)

→ Live | Hard

Soln: First, the interviewer wants to know how much you can imagine about product data. [long discussion]

→ Applications:

→ Search query → Filters

→ UI segmentation

→ **Clustering**: But, these clusters need to have names, so they can be filtered / UI.

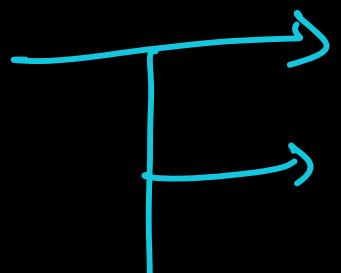
→ **Classification**:

- → Using business logic
- → Using Unsupervised learning + human.

**Product Hierarchy**

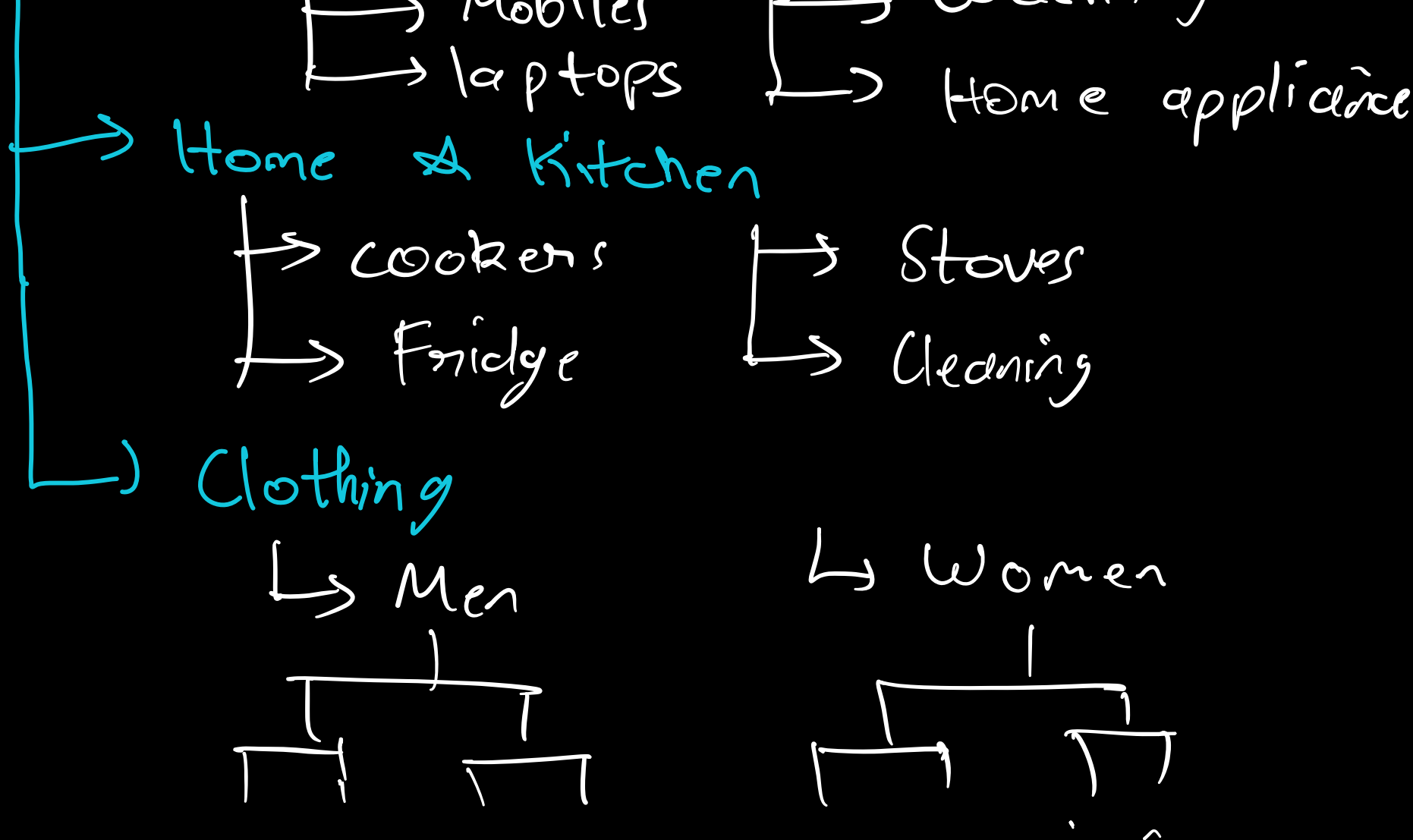
Do not miss the point about P. H.

Eg:



→ Electronics

→ Washing Mch

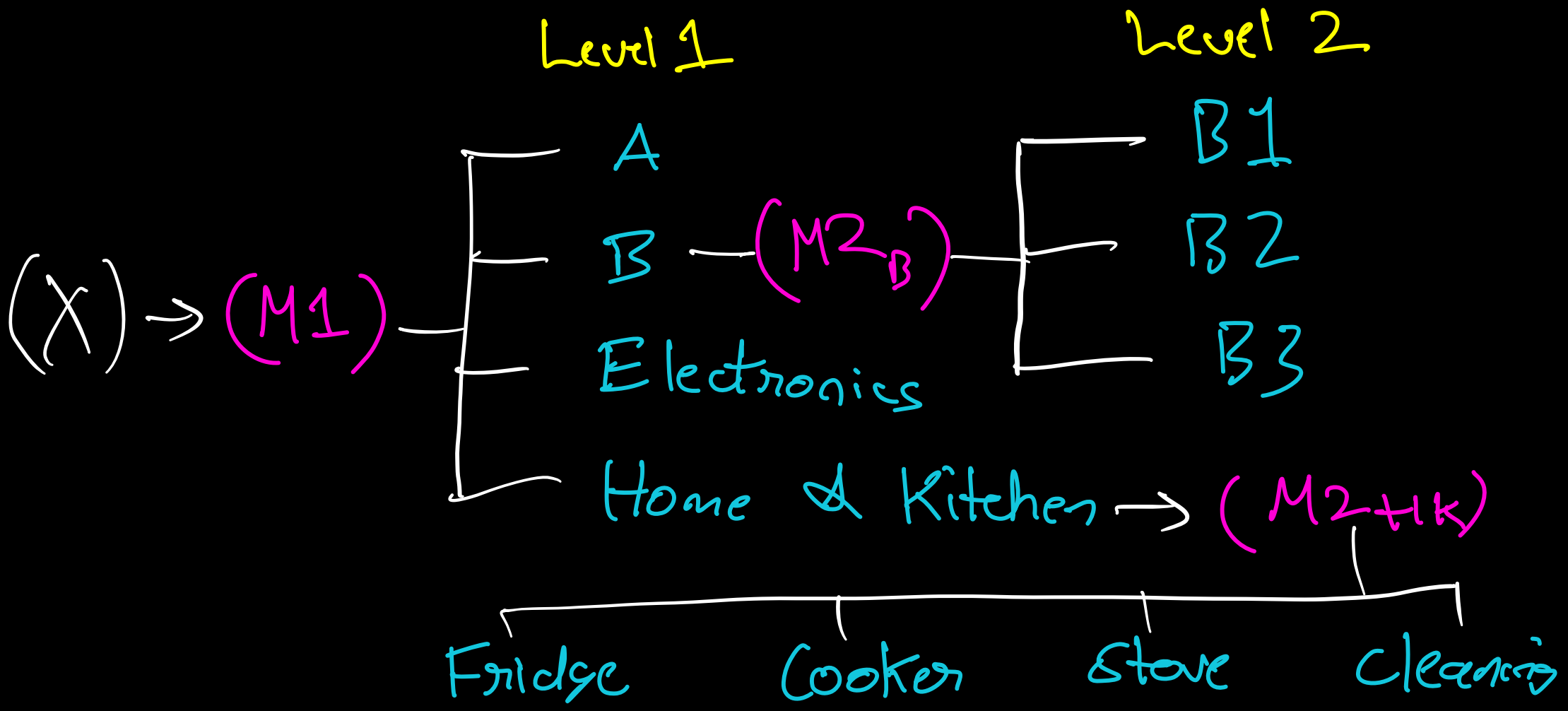



Q: So how do you solve a hierarchical, multi class, multi label problem?

Variation Q: How would you approach

a classification problem with classes as follows? A1 A2 A3, B1 -- B5, C1 -- C3, etc?

⇒ Muliple Levels of Models



M1 -> Classification | $M2_B$ -> ORDINAL Classificatn

$M2_{1HE}$ -> U

Follow up Question: Wont this make too many models if there are 3-4 layers

-> Yes, decide based on, How many products each model is classifying. Eg: 100 K per model in last layer.

-> Layer depth can also vary: Eg:

3 layers for electronics, 5 for clothing

FUQ: What are the accuracy trade offs)

Concerns / priorities ?

SJT → L1 needs more accuracy, since

→ It segregates more data pts.

→ If User goes in wrong category he is misguided from the start.

FUQ: What changes would you make if one product can belong to multiple cat?

→ Each model is a binary model to identify own class (one-vs-rest). Input needs to be passed through the entire tree.

Q: You are working on a binary classf^ model. When you changed threshold from 0.5 to 0.55, the confusion matrix flipped. What can you say abt this model? (discussion)

→ Live | Hard

| | |
|---|---|
| 100 | 200 |
| 50 | 900 |

T = 0.5

| | |
|---|---|
| 102 | 56 |
| 197 | 895 |

T = 0.55

Soln → The model is **not confident**

Use of threshold:

$X \rightarrow f(x) \rightarrow$ probability $\rightarrow$ thresh $\rightarrow$ label

($f(x)$: trained model; probability: (0 − 1); thresh: $> T$)

Ideally, when the model returns class '0' it should be 100% confident. Realistically a good model will be >90% confident for most data points.

Hence, changing threshold from 50% → 55% should not alter the predictions.

This confusion matrix implies, most predictions are around 50 - 60% confidence.

**Unstable Model**

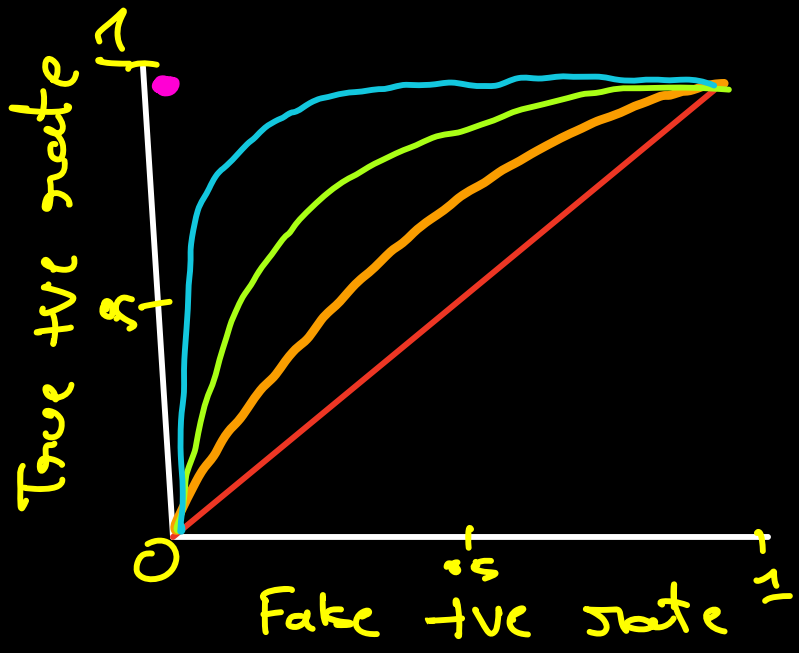
→ Slight changes in I/P will cause different predictions

FUQ: How can you best represent this?

→ **ROC Curve.**

- • = perfect
- • = Great
- • = Better
- • = Bad
- • = Random

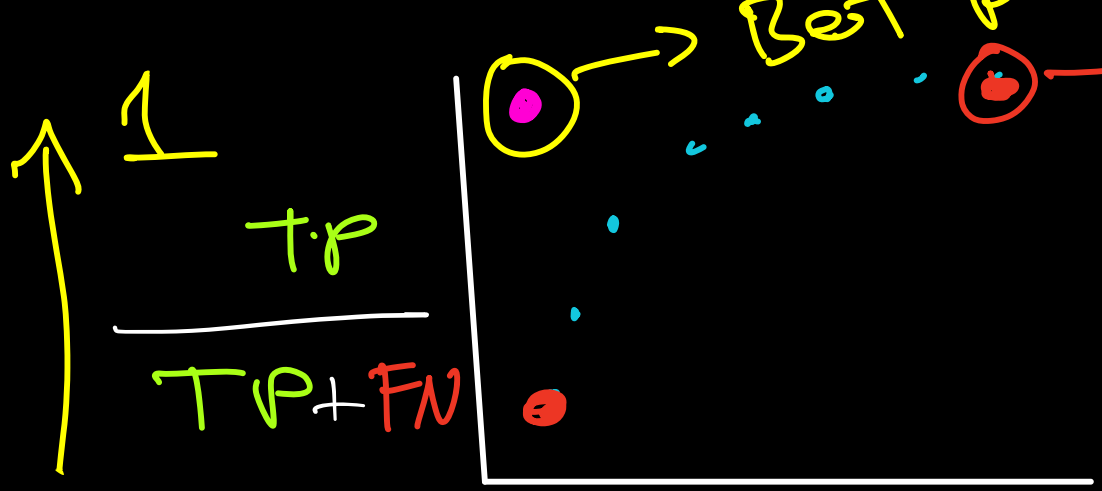
AUC ~ 0.6 (guess)



| TP | FP |
|---|---|
| FN | TN |

Green should be more

Red should be less




For a good model, TPR = 1 and FPR = 0.

At highest T, TPR = 1, FPR = 1

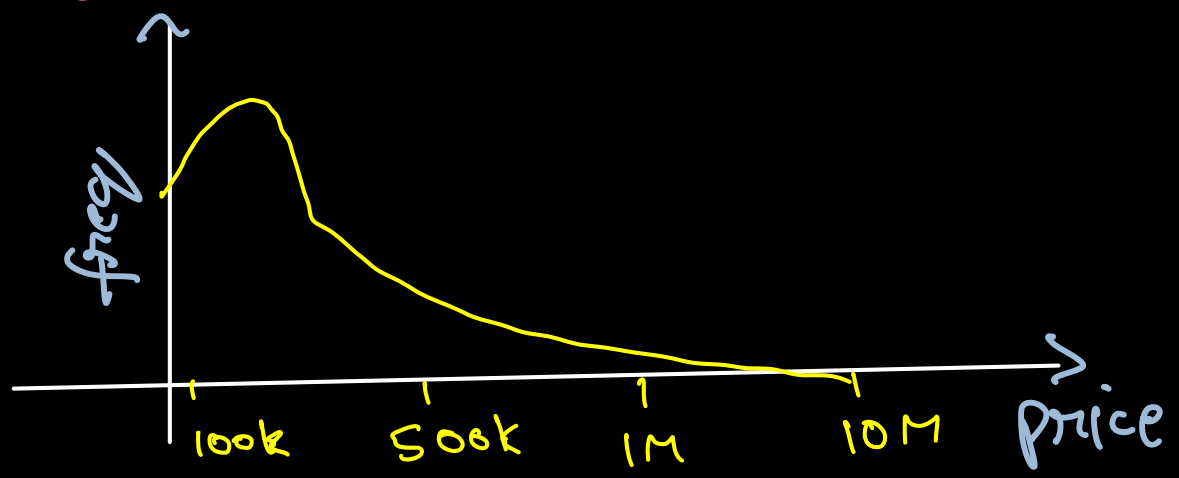
T1 → (TPR1, FPR1)

T2 → (TP2, FPR2)

Best

⋮

Q: Predicting House prices: The dist of prices are shown below: (Assume LR)




What metrics would you use to evaluate your predictions?

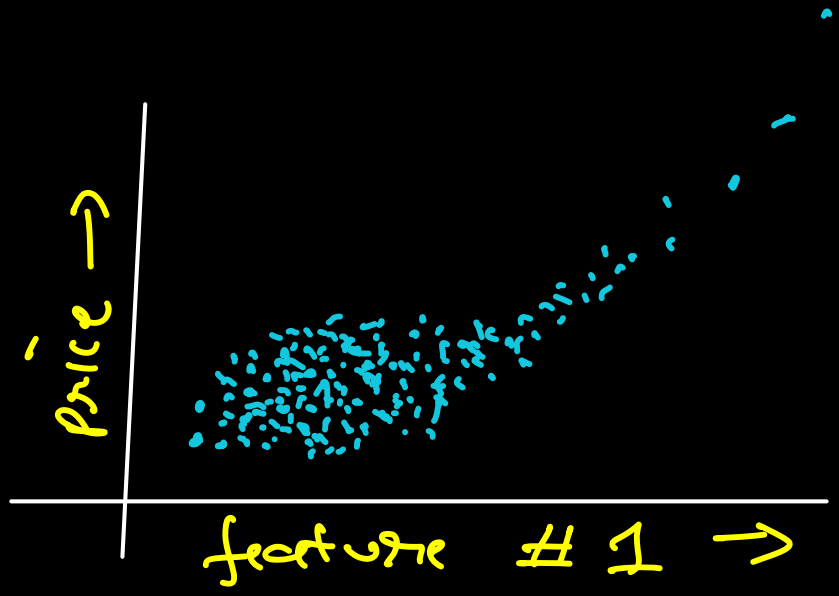
a) MAE b) MAPE
c) RMSE d) MSLE

FUQ: Any other suggestions while modelling?

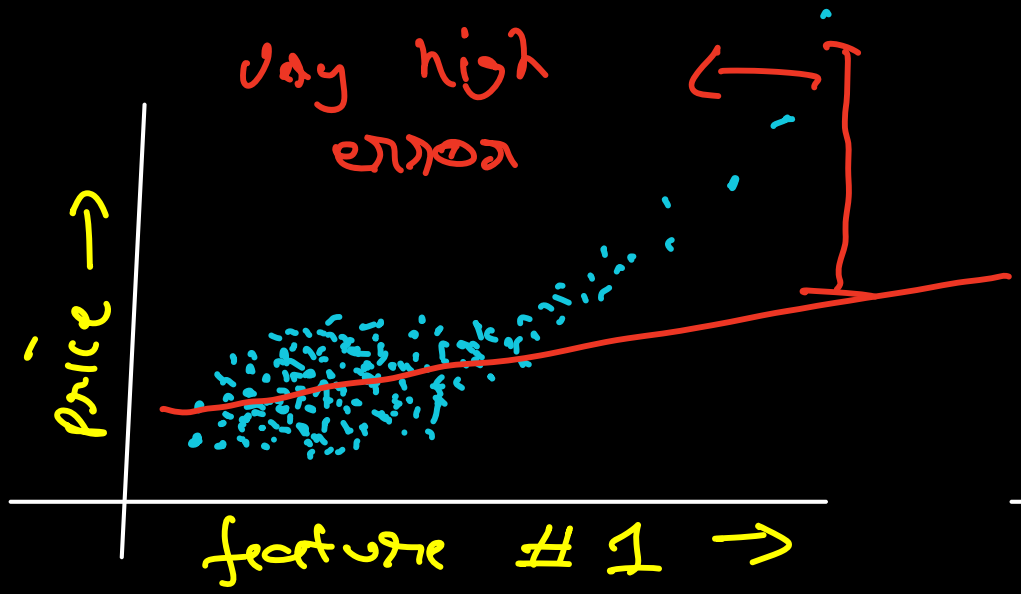
$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|(y_i - \hat{y}_i)|}{y_i}$$

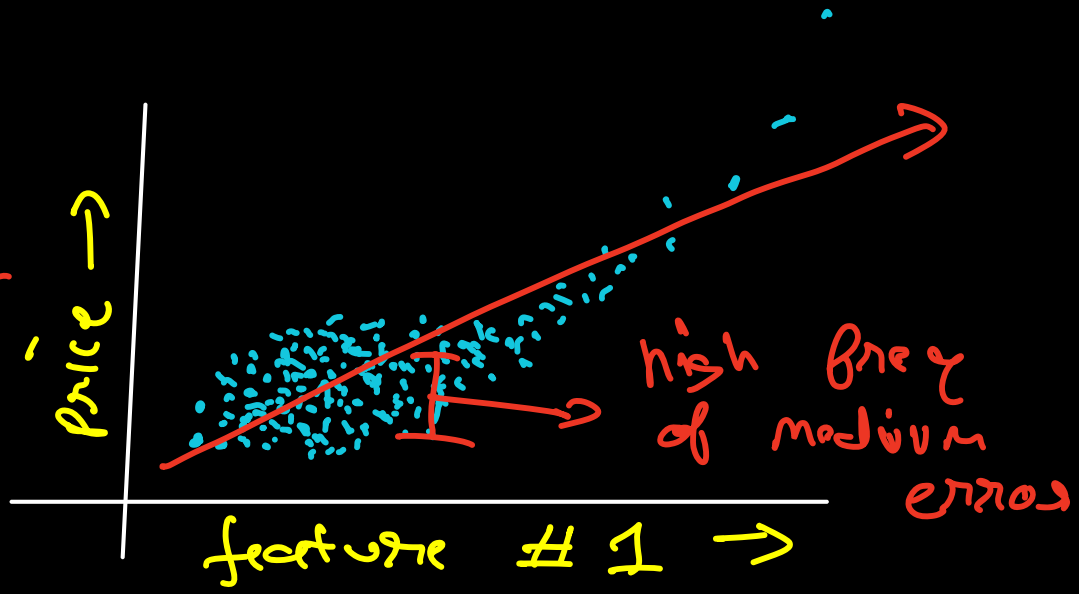
$$MSLE = \frac{1}{N}\sum_{i=1}^{N}\left[\ln(y_i) - \ln(\hat{y}_i)\right]^2$$

[Note: Many times interviewer will give you a new metric / formula and see if you can make sense]




→ Notice that density of pts is higher closer to the y-axis.




Trade off.

MAE → X: because it's not at the same scale.

RMSE → X: As we see we will need to trade-off some where:

Eg:

| | $y$ | $\hat{y}$ | $e$ | |
|---|---|---|---|---|
| case 1 | 100k | 150k | 50k | 4.75 M |
| | 5M | 300k | 4.7M | |
| case 2 | 100k | 200k | 100k | 1.1 M |
| | 5M | 4M | 1M | |

which one of them is more wrong?

case-2 ✓

Will total error $\frac{1}{N}\sum(\dots)$ make sense?

Percentage Error:

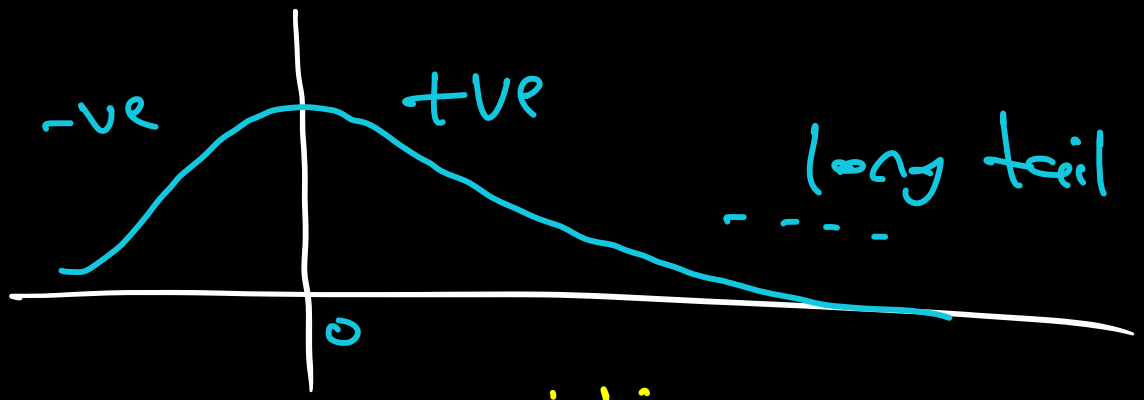
$$\% e = \frac{(y - \hat{y})}{y} = \frac{100}{100} = 100\%$$ error (2x value)

$$= \frac{1M}{5M} = 20\%$$ error.

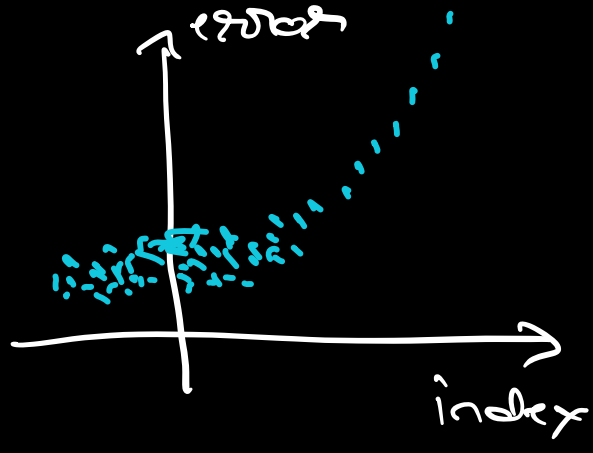
So for such a varying scale 100K – 10M percentage error, i.e MAPE makes more sense.

Now, moving Further:

Let's look at the dist of error for case -1




distribution




scatter

Notice the non-linear scatter plot. And let's look at MSLE:

$$\frac{1}{N}\sum_{i=2}^{N}\left[\ln(y) - \ln(\hat{y})\right]^2 = \frac{1}{N}\sum_{i=1}^{N}\left[\ln(y/\hat{y})\right]^2$$

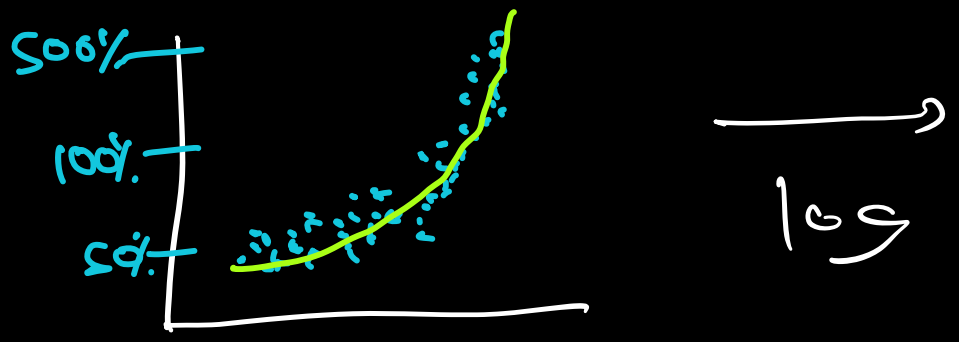
$\because$ $\boxed{\ln(a/b) = \ln(a) - \ln(b)}$

$y/\hat{y}$ → percentage

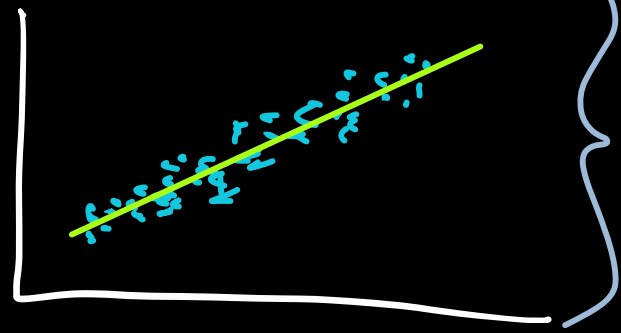
best pred $\leftarrow$ $y/\hat{y} = 1$

$\ln(1) = 0$ ⟹ No error.

Let's say % error is rising fast



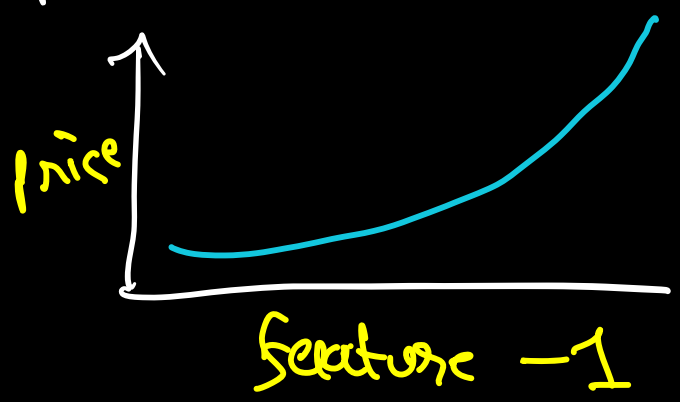

log →



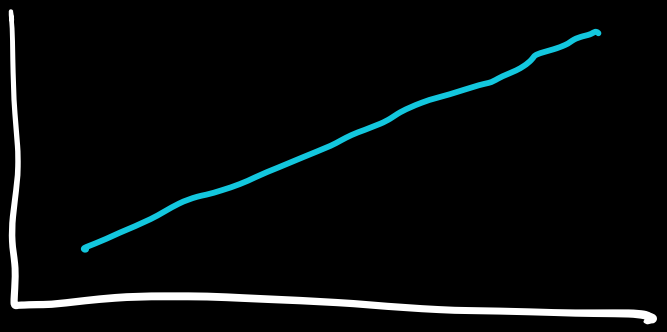
Reduces error for high error case.

Best way:

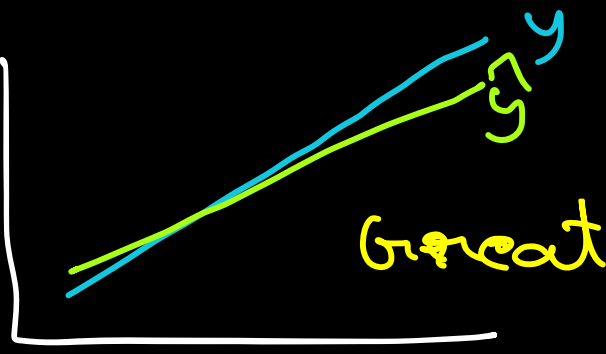
Step -1: Take log of price




→



↓ X_train




← LR( )

↙ Measure RMSE

→ X → LR(x) → $\hat{y}$ → $e^{\hat{y}}$ → # pred

Anti log