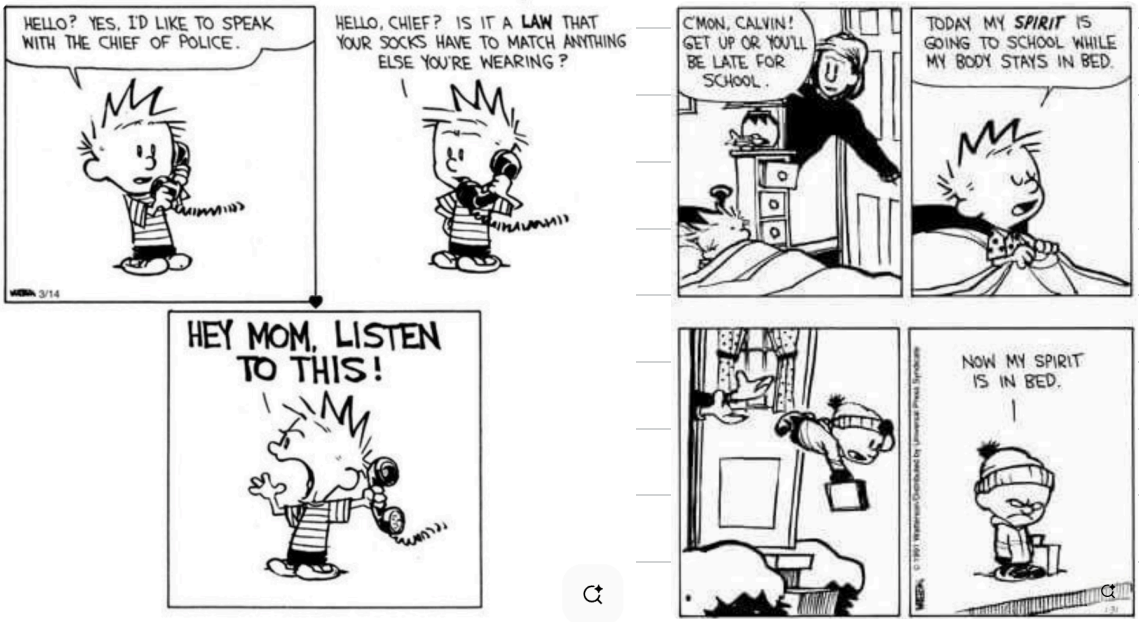
Session-7

BOOSTING - 2

Sep 08, 2025



AGENDA

1. Questions
2. GBDT - Pseudo Code
3. Explanation

GBDT- Algorithm

Input: training set $\{(x_i, y_i)\}_{i=1}^{n}$, a differentiable loss function $L(y, F(x))$, number of iterations M.

Algorithm:

1. Initialize model with a constant value:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma).$$

2. For m = 1 to M:
   1. Compute so-called *pseudo-residuals*:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, n.$$

   2. Fit a base learner (or weak learner, e.g. tree) closed under scaling $h_m(x)$ to pseudo-residuals, i.e. train it using the training set $\{(x_i, r_{im})\}_{i=1}^{n}$.
   3. Compute multiplier γ_m by solving the following one-dimensional optimization problem:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right).$$

   4. Update the model:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x).$$

3. Output $F_M(x)$.

Input: training set $\{(x_i, y_i)\}_{i=1}^{n}$, a differentiable loss function $L(y, F(x))$, number of iterations M.

STEP-1

1. Initialize model with a constant value:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma).$$

$$L = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

$$\text{Argmin}_{\gamma} \sum_{i=1}^{n} (y_i - \gamma)^2$$

derivative = 0

$$\sum_{i=1}^{n} 2(y_i - \gamma)(-1) = 0$$

$$\sum_{i=1}^{n} (y_i - \gamma) = 0$$

$$(y_1 - \gamma) + (y_2 - \gamma) + (y_3 - \gamma) + \cdots + (y_n - \gamma) = 0$$

$$\therefore \sum_{i=1}^{n} y_i - n \times \gamma = 0$$

$$\Rightarrow \gamma = \sum_{i=1}^{n} y_i \times \frac{1}{n}$$

$$\underset{x}{\text{Argmin}} \; (5 - x)^2 \; ; \quad x = 5$$

$$= 0$$

2. For m = 1 to M:
    1. Compute so-called *pseudo-residuals*:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, n.$$

    2. Fit a base learner (or weak learner, e.g. tree) closed under scaling $h_m(x)$ t to pseudo-residuals, i.e. train it using the training set $\{(x_i, r_{im})\}_{i=1}^{n}$.

Fit a base learner on r_{im} of $F_{m-1}(x)$

    3. Compute multiplier γ_m by solving the following one-dimensional optimization problem:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right).$$

For each base learner, we will have different level of contribution, so model will assign diff weight parameter for each base-learner.

x	y	$\Pi_0(x)$	ε_0	$\Pi_1(x)$	$\hat{y}_1(x)$
1	3	0	3	2.5	$0 + \gamma_1 \times 2.5$
2	0	0	0	0.1	$0 + \gamma_1 \times 0.1$
3	−3	6	−3	−2.5	$0 + \gamma_1 \times (-2.5)$

(x, ε_0)

$\hat{y} = F_1(x)$

$$\hat{y} = \Pi_0(x) + \gamma_1 \times \Pi_1(x)$$

$$L = \frac{1}{3} \sum_{i=1}^{3} (y_i - \hat{y})^2$$

$$= \frac{1}{3} \left[(3 - 2.5 \times \gamma_1)^2 + (0 - \gamma_1 \times 0.1)^2 + (-3 - (-2.5\gamma_1))^2 \right]$$

$$= \frac{1}{3} \left[\underbrace{(3 - 2.5 \times \gamma_1)^2}_{A} + \underbrace{(0 - \gamma_1 \times 0.1)^2}_{B} + \underbrace{(-3 + 2.5\gamma_1)^2}_{C} \right]$$

$$(a-b)^2 = a^2 + b^2 - 2ab$$

$$A = 9 + 6.25\gamma^2 - 15\gamma$$

$$C = 9 + 6.25\gamma^2 - 15\gamma$$

$$B = 0.01\gamma^2$$

$$A + B + C = 12.51\gamma^2 - 30\gamma + 18$$

$$\therefore \quad 2 \times 12.51\gamma - 30 = 0$$

$$\therefore \quad \gamma = \frac{30}{2 \times 12.51} = 1.199$$

4. Update the model:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x).$$

$$F_3(x) = h_0(x) + h_1(x) + h_2(x) + h_3(x)$$

$$= F_2(x) + h_3(x) \times \gamma_3$$

After learning h1 (x) - We need to find γ

STEP 2 . 3

3. Compute multiplier γ_m by solving the following one-dimensional optimization problem:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right).$$

In simple terms, this equation is saying ;

- To find γ which minimizes given **loss**

$$L(y_i\ ,\ F_{m-1}(x_i) + \gamma h_m(x_i))$$

Actual Value — *ŷ (Predicted Value)*

To find

$$\hat{y} = F_{m-1}(x^i) + \gamma h_m(x^i)$$

$$For\ m = 1,\quad \hat{y} = F_{1-1}(x^i) + \gamma_1 h_1(x^i)$$

$$= F_0(x^i) + \gamma_1 h_1(x^i)$$

Stage 0 model (mean model) — Calculated is step 2.2

Only variable here is γ

(Rest everything is constant - already calculated)

To find γ which minimizes given **loss**

- **Take derivative of loss w.r.t γ & equate it to 0**

STEP 2 . 4

After finding $h_1(x)$ *and* γ_1

- We need to make final prediction

(i.e combine previous model with current model predictions)

4. Update the model:
$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$.

$$For\ m = 1$$

$$= F_0(x) + \gamma_1 h_1(x)$$

calculated in Step - 1

calculated in Step - 2.3

calculated in Step - 2.2

We keep doing sub-steps for M iteration i.e finding $h_1(x)\ , h_2(x)\ , h_3(x) \ldots h_m(x)$ and $\gamma_1\ , \gamma_2\ , \gamma_3 \cdots \gamma_m$

Finally, we use these to make final model $F_M(x)$

3. Output $F_M(x)$.

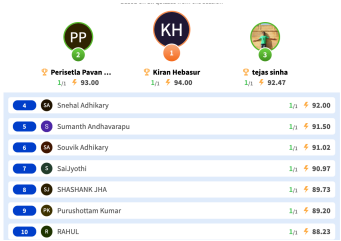
$$F_m(x) = h_0(x) + \gamma_1 h_1(x) + \gamma_2 h_2(x) + \ldots + \gamma_m h_m(x)$$

Which of the following statement(s) are True?

0 users have participated

A	F_0(x) model will be a mean model	0%
B	For additive combining, we do weighted addition of consecutive models instead of simple addition	0%
C	We perform weighted addition in order to control the influence of the base learners.	0%
✓ D	All of the above	0%

End Quiz Now

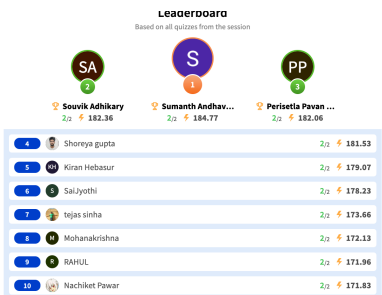


As we increase the value of M (number of base learners), the model will:

0 users have participated

A	Underfit	0%
✓ B	Overfit	0%
C	No change	0%
D	None of the above	0%

End Quiz Now



Hyperparameter : # of base learners (M)

M ↑ ⟶ Overfit

As M increases , model will overfit .

WHY ?

- Because as base learners increase more likely training error will tend to 0

M ↓ ⟶ Underfit

As M decreases , model will underfit.

WHY ?

- Say M = 1 , Stage 0 & Stage 1 model.
- Prediction will be close to mean model - Underfit

Hyperparameter

Depth

- As depth increases

→ Model will overfit

Why ?

- Increase in depth -> Variance Increase

→ Model will overfit quickly

What will happen if we increase max_depth of base learner in boosting?

0 users have participated

A	Overfit	0%
B	Underfit	0%
C	No change	0%
D	None of the above	0%

End Quiz Now

Leaderboard

Based on all quizzes from the session

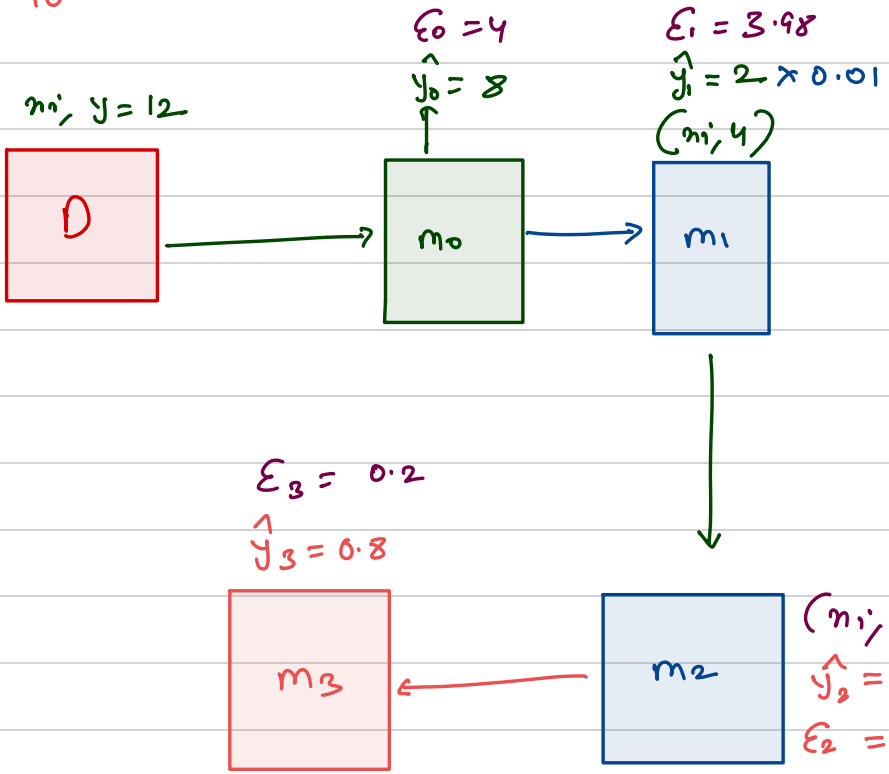
2	Souvik Adhikary	3/3 276.27
1	Shoreya gupta	3/3 276.95
3	Sumanth Andhav...	3/3 274.01
4	Kiran Hebasur	3/3 270.75
5	tejas sinha	3/3 267.62
6	Perisetla Pavan Kalyan	3/3 266.37
7	Purushottam Kumar	3/3 265.51
8	Nachiket Pawar	3/3 264.54
9	Mohanakrishna	3/3 260.84
10	SHASHANK JHA	3/3 258.68

Regularization by Shrinkage

$$F_n(x) = h_0(x) + \eta \times \gamma_1 \times h_1(x) + \eta \times \gamma_2 \times h_2(x) + \eta \times \gamma_3 \times h_3(x) + \eta \times \gamma_4 \times h_4(x) + \vdots + \eta \times \gamma_n \times h_n(x)$$

weightage for each base learner

$\eta = 10^{-2}$



Regularization by Shrinkage

Final Model equation is : $F_m(x) = h_0(x) + \sum_{m=1}^{M} \gamma_m . h_m(x)$

To regularize, we add an regularization term i.e learning rate

$$F_m(x) = h_0(x) + \nu \sum_{m=1}^{M} \gamma_m . h_m(x)$$

Learning Rate

Range : $0 \leq \nu \leq 1$

Notes :

- Adding learning rate is reducing the impact of Mth .

  Hence, reducing overfit.

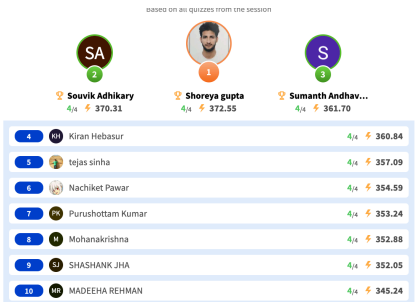
This learning-rate was applied post-hoc. Need to check in sklearn.

What happens if we have learning rate as a very small value ?

0 users have participated

A	It will take less iterations (M) for model to converge	0%
B	It will take more iterations (M) for model to converge	0%
C	No impact. Model will converge as it was before introduction of learning rate	0%
D	None of the above	0%

End Quiz Now

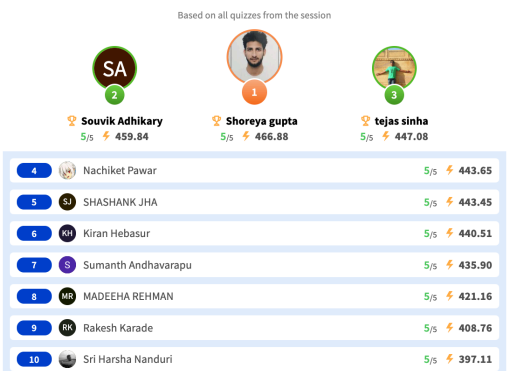


What does it mean when model is overfit ? Model will have ___

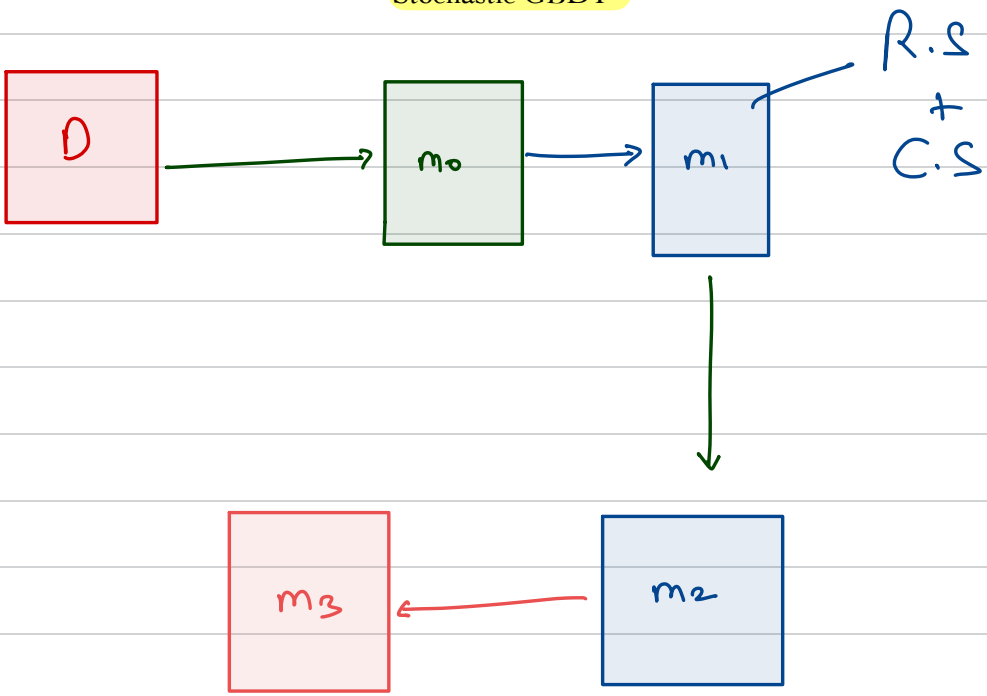
0 users have participated

A	Low variance	0%
B	High bias	0%
✓ C	High variance	0%
D	None of the above.	0%

End Quiz Now



Stochastic GBDT



	f_1	f_2	f_3	f_4	f_5	f_6	y
x_1							
x_2							
x_3							
x_4							
x_5							
x_6							
x_7							
x_8							

$m_2 \rightarrow x_2, x_3 ; f_2, f_5$

$m_3 \rightarrow x_7, x_8 ; F_1, F_6$

This is HW.

Stochastic Gradient Boosting

PROBLEM : GBDT overfit a lot

(High Variance)

How to reduce variance ? - **Randomization** → **Row sampling + column sampling**

- Can use the some concept of randomization (as used in RF) to reduce variance

- This variation of GBDT is called **' Stochastic Gradient Boosting'**

GBDT → **Pseudo Residual + Additive Combing**

Stochastic GDBT → **GBDT + Randomization**

Row Sampling + Column Sampling

- skLearn provides ability of row samping
  - Using **subsample** hyperparameter
  - And column sampling using **max_features** hyperparameter

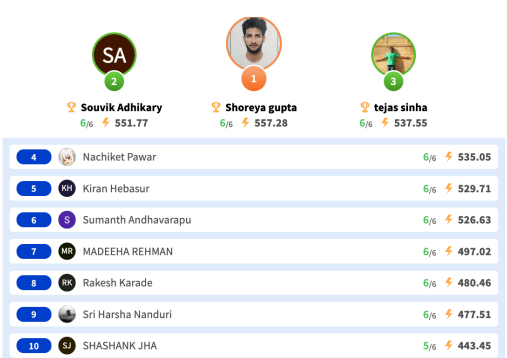
Time Left: 28s

Which of the following techniques can help in regularizing GBDT

0 users have participated

A	Using Learning rate	0%
B	Using Sampling (Stochastic GBDT)	0%
✓ C	All of the above	0%

End Quiz Now

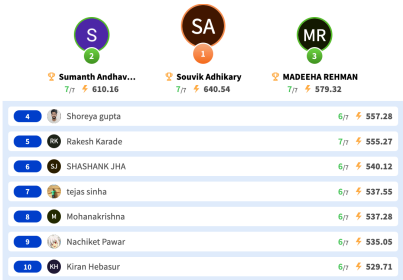


What do you think: Do outliers impact GBDT?

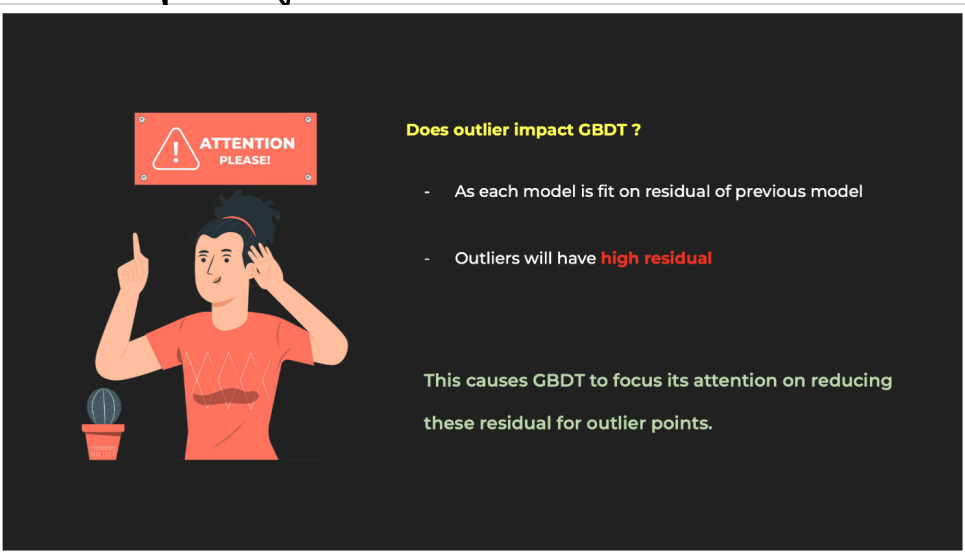
11 users have participated

✓ A	Yes. as it is based on residuals	73%
B	No. as it is based on residuals	27%

End Quiz Now

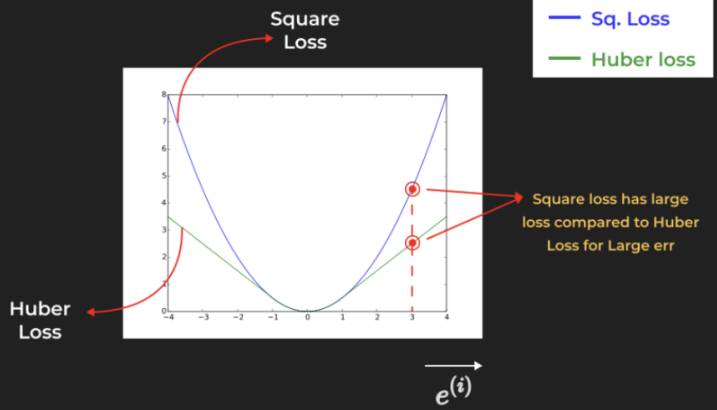


n	y	To
1	5	70.33
2	6	70.33
3	200	70.83



Instead of using squared loss ,

- We can use **huber loss**



Notice :

- For smaller residual values.

  Both squared loss & hyper loss have same values
- As value of residual increases

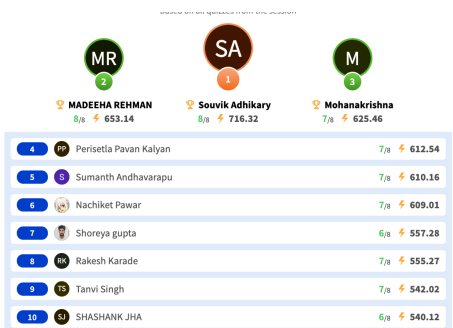
  Huber loss doesn't explode like squared loss

How can we deal with outliers in GBDT? Statement 1: Remove the outliers Statement 2: Use loss function which isn't impacted by outliers

0 users have participated

A	Statement 1	0%
B	Statement 2	0%
C ✓	Statement 1 & 2	0%
D	None of the above	0%

End Quiz Now



MSE vs MAE

$(y_i - \hat{y}_i)^2$ $\quad$ $|y_i - \hat{y}_i|$

2 → 4

4 → 16

(0.01 − 0.03)

↓ w/o fraction → ur MAE