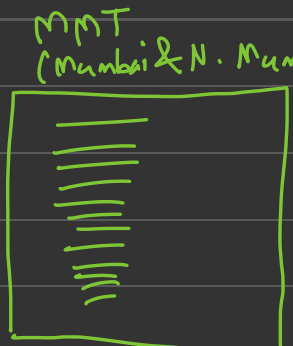
03 Sept 2022

Term 4 - ML April 2022

- Data is the new oil.
- ML is a powerful concept & framework for making the best out of it.

- Age of Information or Age of data.

eg: Google knows everything about my location.



Information
[Darshan + 49 more ppl]

A ⟷ B
8am 4pm

Bus Service	Car
✗	✓ convenient & easy
AC	AC
✓	Guaranteed Seat
P2P	Home – Destn
15 min	Time Conv.
✗ 2950/- pm	10 000 – 120000 rs / month
✗ No Driving Sleep, Meetings, Netflix	Driving – Tired & fatigue Home – fresh

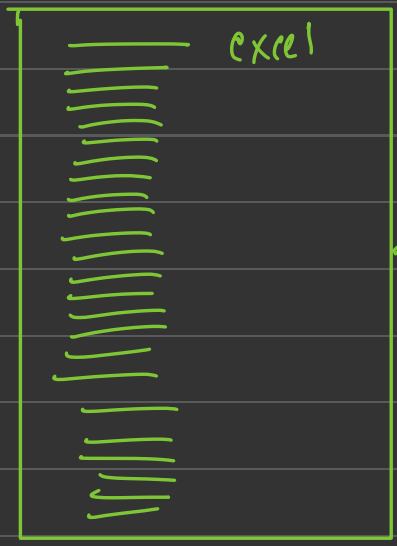
Who is most intelligent lifeform on the planet at the moment?
Human.

– m/c : To mk data driven decisions at a scale.

your friend (10yrs)

Will you lend ₹ 10000 rs?

10yrs (100 interaction)



M/c

80 yes ✓ ∴ Money Lended

20 No

10 interaction Human

7 yes ✓ ∴ Money lended.

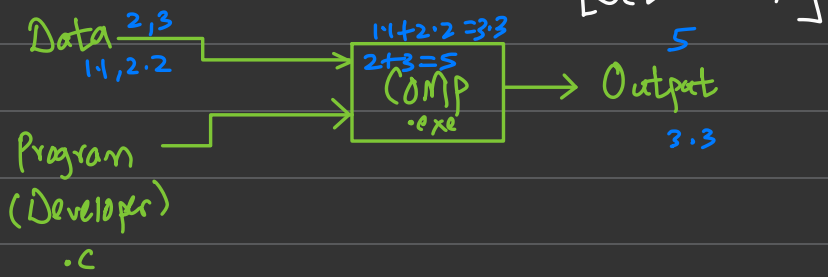
3 No

Traditional Programming Paradigm

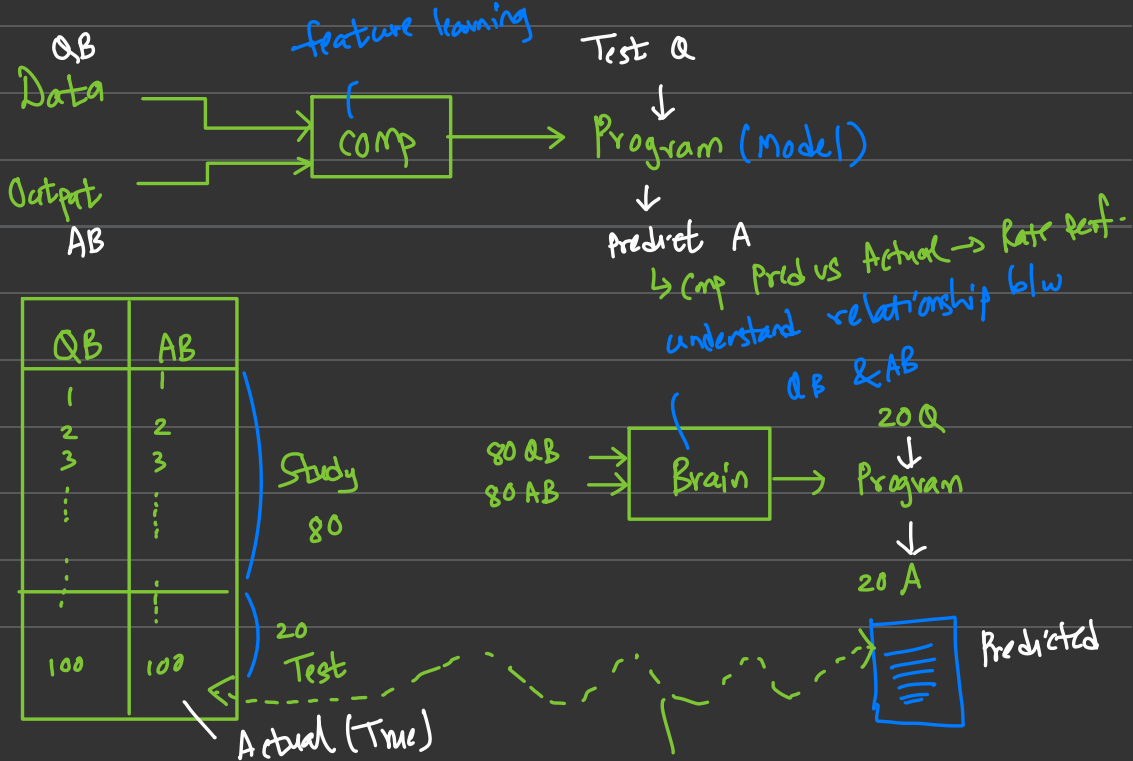
HDFC → Website → Infosys

Task: Add two nos

[Deterministic]



ML Programming Paradigm



Compare Actual & predicted answers & rate the performance of a student.

04 Sept 2022

y — Dependent Variable (1)

X — Independent Variables (1 or more)

① Price of car ~ Brand + EnginePower + Mileage + Transmission Type + Color + Variant + FuelType + BodyType + ABS + EBD + ADAS + HP + Torque

② Price of Mobile ~ Brand + RAM + Storage + Battery + ScreenSize + Camera + OS + 5G/4G + OLED/AMOLED + Processor

→ Regression

③ Apple vs Orange ~ Color + Taste + Shape + Texture + NutritionValue + CalorieValue + Price

④ Iris (Setosa vs Versicolor vs Virginica) ~ SepalLength + SepalWidth + PetalLength + PetalWidth

→ Classification

https://miro.medium.com/max/1000/1*IFC_U5j_Y8IXF4Ga87KNVg.png

https://miro.medium.com/max/638/0*2c7voFri9cIXGrc4

ML Problems

Regression

y is Continuous

Classification

y is Categorical

Binary (2)

eg: Apple vs Orange

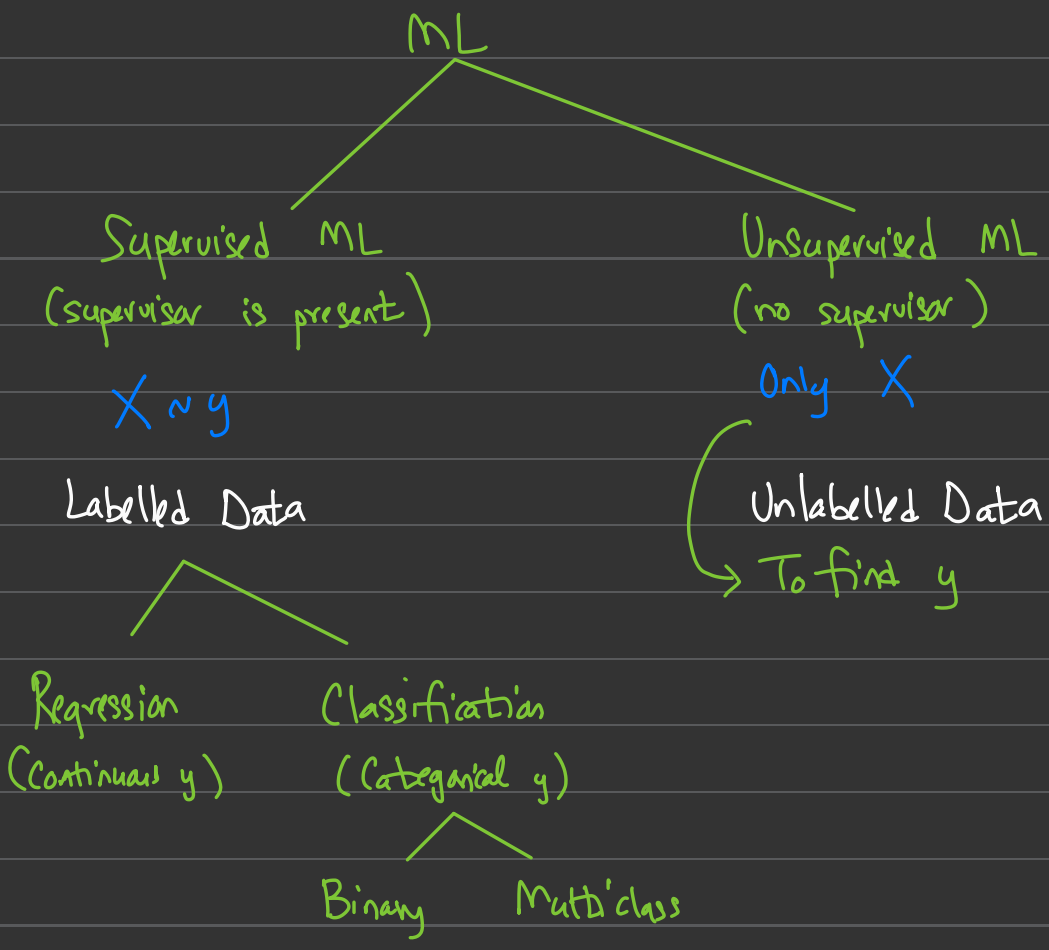
eg: Male vs Female

eg: Spam vs Ham

Multi-class (more than 2)

eg: Iris Dataset

eg: Apple vs Orange vs Mango vs Grapes



https://www.aplustopper.com/wp-content/uploads/2017/05/Correlation-1.jpg

Take Home Task: Difference b/w pearsons, Spearman Rank & Kendal's Correlation Coefficient + Applications

https://www.mathsisfun.com/data/correlation.html

Inbox (28,604) - ingledarshan | Slope (Gradient) of a Straight L | You are screen sharing | Stop Share

mathsisfun.com/geometry/slope.html

Graph Index

To calculate the Slope:

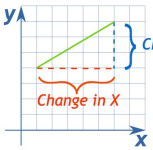
Divide the **change in height** by the **change in horizontal distance**

Equation of a Straight Line

Slope
m = 0.5

If we move 1 unit on X-axis, we climb 0.5 units on Y-axis

Slope = Change in Y / Change in X = $\frac{y_2 - y_1}{x_2 - x_1}$

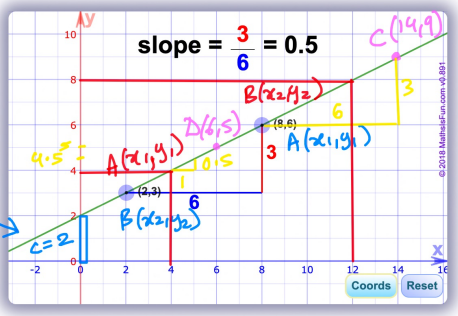


$m = \frac{6-3}{8-2} = \frac{3}{6} = 0.5$

$m = \frac{3-6}{2-8} = \frac{-3}{-6} = 0.5$

y-intercept
c = 2

Have a play (drag the points):



$m = \frac{9-5}{14-6} = \frac{4}{8} = 0.5$

$m = \frac{5-9}{6-14} = \frac{-4}{-8} = 0.5$

Eqn. of Line

$y = mx + c$

$y = 0.5x + 2$

Given x, we can find y.

If $x = 4$, $y = 0.5(4) + 2 = 4$

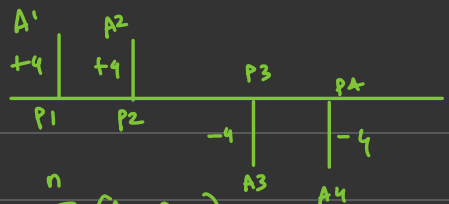
Examples:

If $x = 12$, $y = 0.5(12) + 2 = 8$

10th Sept 2022

Linear Regression
- Simple LR $(X \sim y)$
- Multiple LR $(X_1 + X_2 + \cdots + X_n \sim y)$

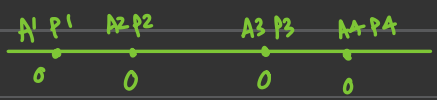
Pb. Stmnt: Predict price of house depending on its sqft area.

① 

Mean error = $\frac{(4)+(4)+(-4)+(-4)}{4}$

$= 0$

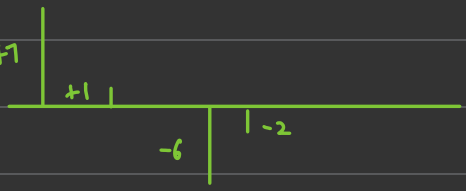
$ME = \frac{\sum_{i=1}^{n}(A_i - P_i)}{n}$



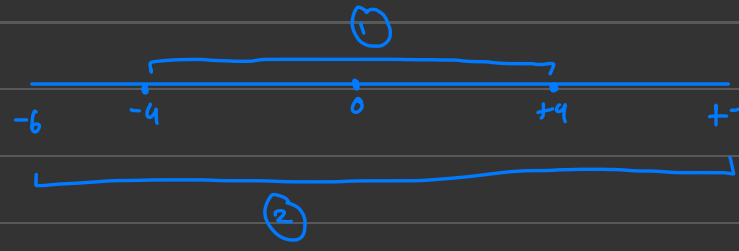
$ME = \frac{0+0+0+0}{4} = 0$

Mean Absolute error : $\frac{\sum_{i=1}^{n}|A_i - P_i|}{n}$

① $MAE = \frac{|+4|+|+4|+|-4|+|-4|}{4} = \frac{16}{4} = 4.$

② 

$MAE = \frac{|+7|+|+1|+|-6|+|-2|}{4} = 4$



Spread of errors

② > ①

Mean Squared Error = $\frac{\sum_{i=1}^{n}(A_i - P_i)^2}{n}$

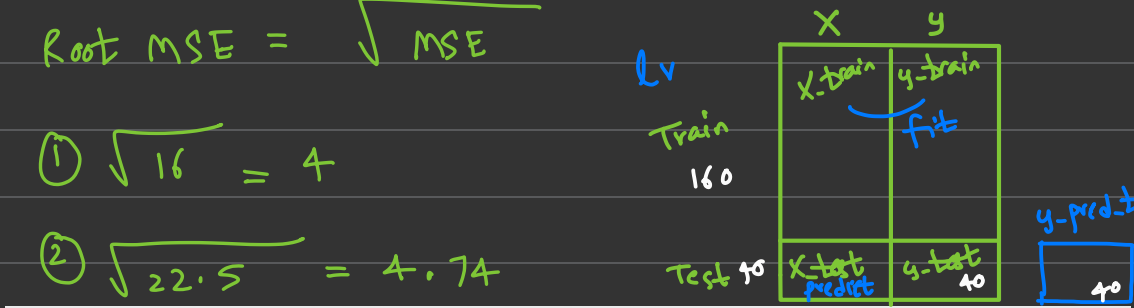
① $MSE = \frac{(+4)^2 + (+4)^2 + (-4)^2 + (-4)^2}{4} = 16$

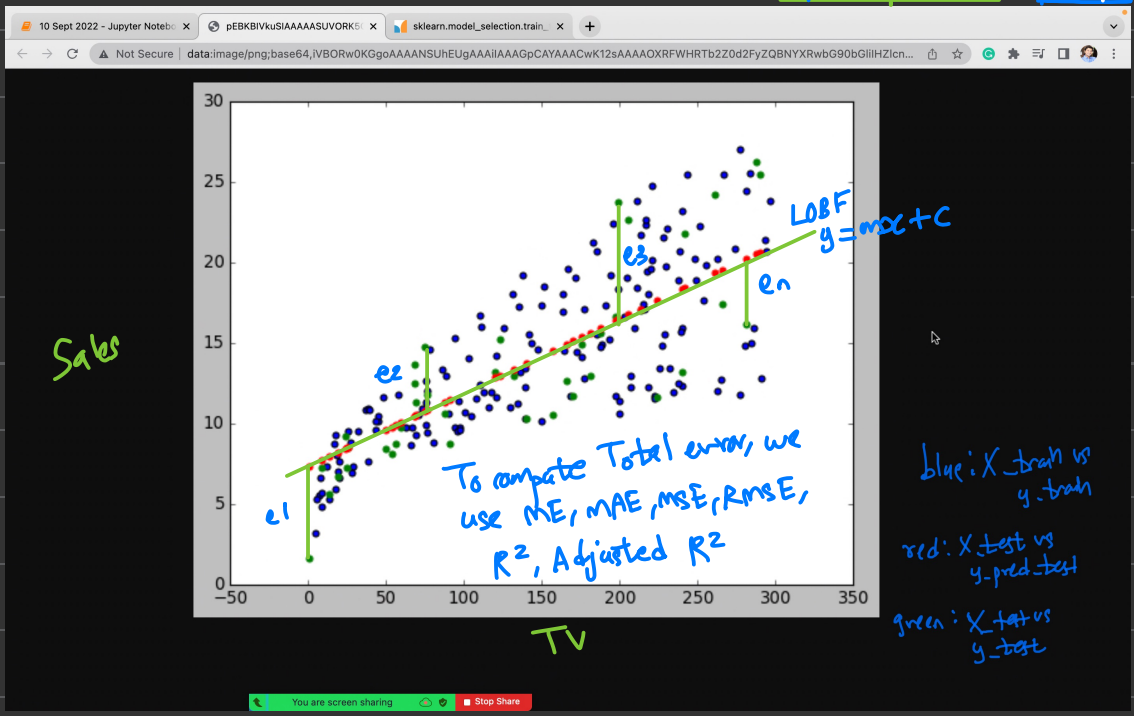
② $MSE = \frac{(+7)^2 + (+1)^2 + (-6)^2 + (-2)^2}{4} = \frac{90}{4} = 22.5$

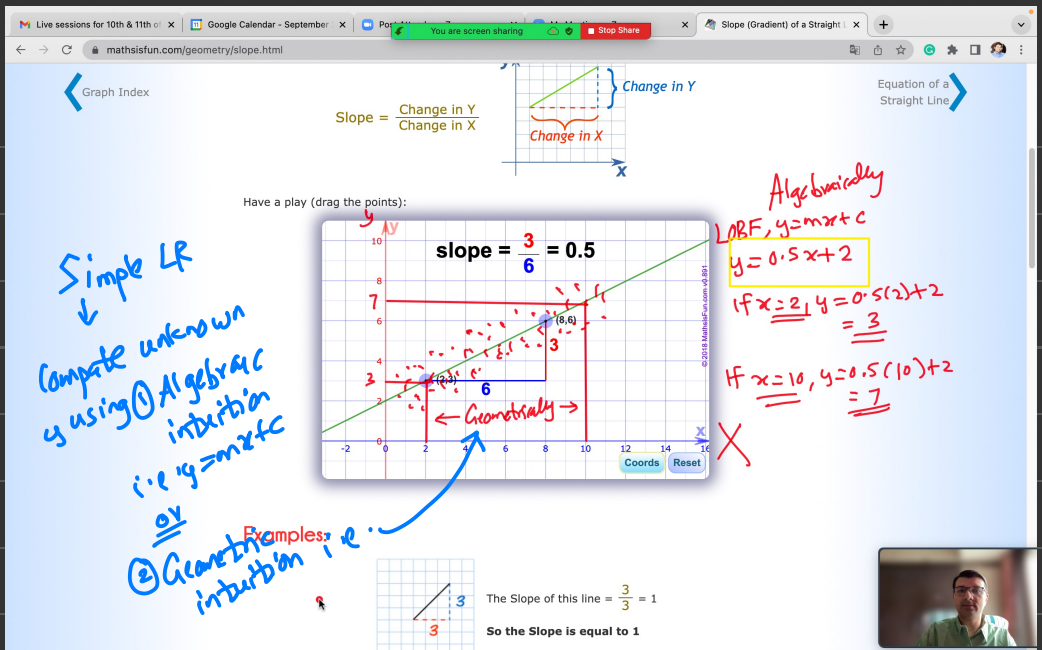
$Root\ MSE = \sqrt{MSE}$

① $\sqrt{16} = 4$

② $\sqrt{22.5} = 4.74$





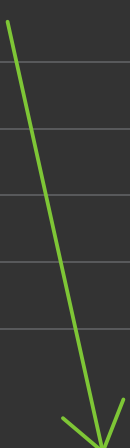


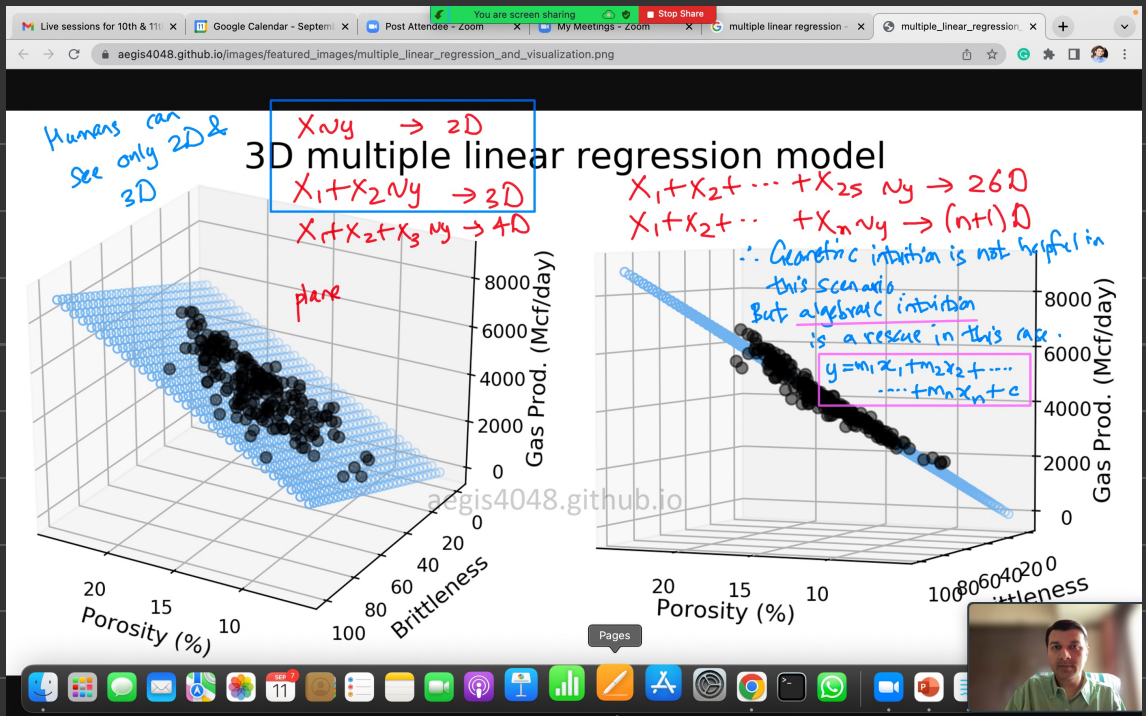
Multiple LR:

$X_1 + X_2 \sim y$

Algebraic intuition: $y = m_1X_1 + m_2X_2 + c$

Geometric intuition:





Code for Simple LR:

1. read dataset
2. X vs y Separation
3. TTS (70:30)
   X_train, X_test, y_train, y_test
4. import LR
5. lr = LR() # LR object
6. lr.fit(X_train, y_train)
7. y_pred_test = lr.predict(X_test)
   y_pred_train = lr.predict(X_train)
8. Check error metrics like MAE, MSE, MAPE, RMSE, R^2, AR^2.

Model 1 ⟶ MAE = 1.42

How can I create more models?

Repeat all 8 Steps as before with a different random_state like (0, 1, 2, ...)

or Change TTS ratio (75:25, 80:20, ...)

Hyperparameter Tuning

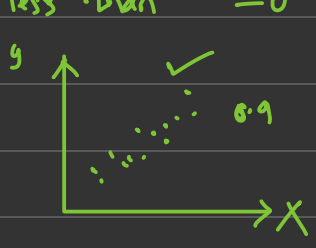
18/09/22

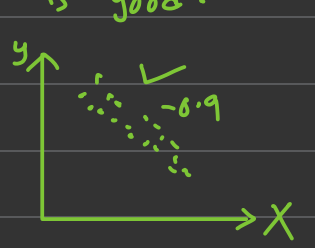
We want:

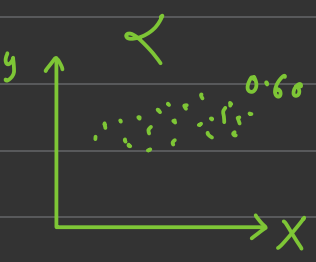
① I.V.N D.V. to hv highest correlation.

X vy [correlation value = +1 or −1 or near then]

Any value greater than 0.70 or less than −0.70 is good.







② I.V. ∿ IV should be least correlated else it gives rise to Multicollinearity.

eg: Mumbai to Pune (y)

X_1 ① My car
X_2 ② My wife car
X_3 ③ My father's car

∴ To remove MC, we can choose any one of the IVs & drop the rest.

Take home task:

We hv created 2 Models : Model 1: TV+RadNP ∿ Sales

Model 2: Tv+Rad ∿ Sales

Where in Model 2 was found out to be better.

Now, by creating following models:

① TTS: 60:40, 75:25
② Scaling: SS, MMS
③ Random state: 0, 1, 100

12 Models & choose & comment on the best one.

—— X ——