Feature Engineering

- is the process of creating new more informative features from existing raw data to improve a models preformance. (involes Domain knolage + creativity)

Ex: instad of using raw Date values you might create fetures like Day of week / holidy. This can better explain something like sales (less or more on weekends?) helps model focaos on relevent parts of data another ex is BMI

Raw Data

Date
Jan 1, 2024
Feb 3, 2024
Mar 5, 2024

→

Engineered Features

Day	Is holiday
Sunday	Yes
monday	NO
Thursday	NO

Feature Scaling (Normalization / standardization)

- is the process of Transforming numeric fetures to a similar scale, to prevent fetures with larger ranges from Dominating the learning process

Ex: here the nums for Salary are much larger than those for age and thus dominate model fitting

id	age	salary
1	44	73000
2	27	47000
3	30	53000

Normalization →

Ex: min = 27, max = 48 for age col

- min max norm. (0-1 range)

id	norm age	norm salary
1	0.8095	0.8387
2	0	0.0
3	0.1428	0.1935

range (0-1)

Common methods:

✱ Standerdization

$$x_{stand} = \frac{x - \text{mean}(x)}{\text{Standerd deviation}(x)}$$

⇒ transforms mean = 0, std = 1

(same scale)

✱ normilization

$$x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)}$$ [0, 1]

⇒ transforms range (0-1) or (-1-1)

(same range)

- Why?

- in short many ML algos Are sensitive to scale of fetures by scaling We avoid bais in learning (larger Feature ≠ more important) and help some algos like GD be faster (smaller vals)

- When to use what?

Senario	norm	stand
Features have different scales	✓	✓
Data not gaussian	✓	✗
Data gaussian (normm dist)	✗	✓
Distance based model (KNN, SVM, GD)	✓	✓
NN	✓	✓ (sometimes)

Scanned with CamScanner