

# Feature Engineering

(Also see 111)  
(Also see 107)

- is the process of creating new more informative features from existing raw data to improve a model's performance. (involves Domain knowledge + creativity)

Ex: instead of using raw Date values you might create features like Day of week / holiday. This can better explain something like sales (less or more on weekends?) helps model focus on relevant parts of data another ex is BMI

Raw Data	Engineered Features
Date	Day   Is holiday
Jan 1, 2024	Sunday   Yes
Feb 3, 2024	Monday   No
Mar 5, 2024	Thursday   No

## Feature Scaling (Normalization, standardization)

- is the process of transforming numeric features to a similar scale, to prevent features with larger ranges from dominating the learning process

Ex: here the values 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

id	norm age	norm salary
1	0.9095	0.8387
2	0	0.0
3	0.1428	0.1935

min max norm.  
(0-1 range)

Common methods:

**\* Standardization**  

$$X_{std} = \frac{X - \text{mean}(X)}{\text{standard deviation}(X)}$$

$$\Rightarrow \text{transforms mean} = 0 \text{ and std} = 1$$
 (same scale)

**\* Normalization**  

$$X_{norm} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

$$\Rightarrow \text{transforms range (0-1)}$$
 (same range)

- Why?

- in short many ML algo's are sensitive to scale of features by scaling we avoid bias in learning (larger features are more important) and help some algo's like GD be faster (smaller vals)

- When to use what?

Scenario	norm	std
Features have different scales	✓	✓
Data not gaussian	✓	✗
Data gaussian (normal dist)	✗	✓
Distance based model (KNN, SVM, GD)	✓	✓
NN	✓	✓ (sometimes)