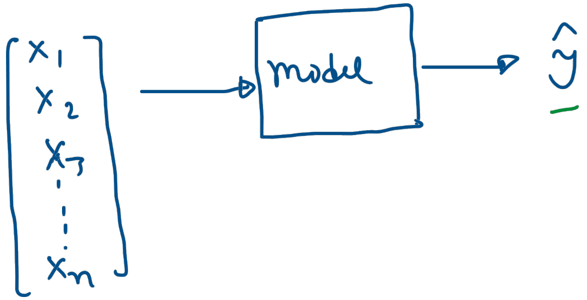
Feature Engineering

Feature ? → Problem and data specific

The input variables of your model.



Churn Prediction (Bank)

i/p variables : Salary
Credit Score
Bank balance
# products
Credit Card User or not
favourite color

NLP : Sentiment analysis
(+ve / -ve)

The direction and the cinematography of the movie is really good. → +ve

Feature Engineering

1) **Feature pre-processing** → preprocessing the i/p variables so that it can be used in the model.

2) **Feature Extraction** → i/p variables ⟶ features.

$$\phi(x) \longrightarrow x^*$$

$$\underline{f}^* : \underset{i/p}{\underline{\mathbb{R}^{n\times d}}} \longrightarrow \underset{o/p}{\underline{\mathbb{R}}}$$

$\underline{d}$ → dimension

$\underline{d' < d}$ ↳ dimensionality reduction

$$\underline{\phi : \mathbb{R}^{n\times d} \longrightarrow \mathbb{R}^{n\times d'}}$$

$$\hat{y} = f^*(x)$$

$$\hat{y}' = \hat{f}(\phi(x))$$

→ model
→ visualization
→ improve the performance

3) **Feature Selection** :

Feature Preprocessing

$x_1 \rightarrow x_1^*$

$x_2 \rightarrow x_2^*$

$\vdots$

$x_n \rightarrow x_n^*$

Feature Extration

$x_1 x_2$, x_1/x_2 , $(1 + x_1 x_2)^3$

Units Sold (x_1) price per unit (x_2) amount $(x_1 \cdot x_2)$

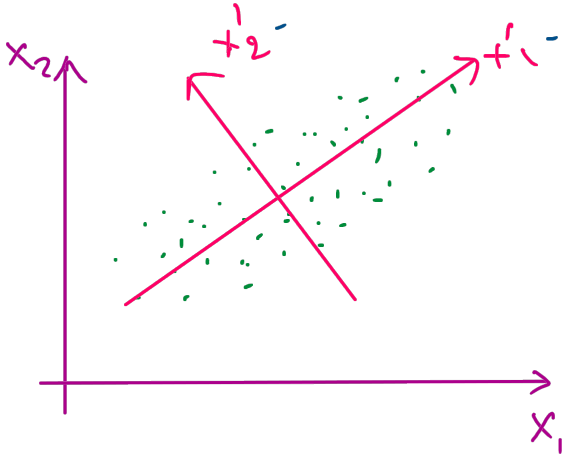
RFM : Recency Frequency Monetary

Recency	Frequency	Monetary
1	1	1
$\vdots$	$\vdots$	$\vdots$
5	5	5

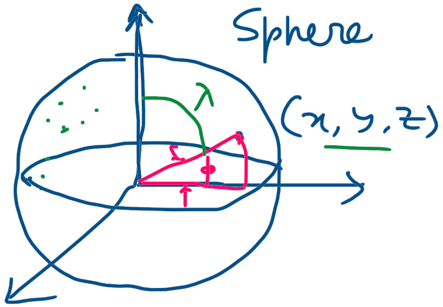
111 → 555

GDP   # population   GDP / person

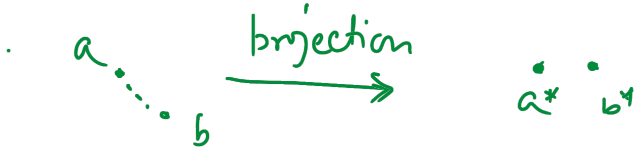
Raw features → feature Extration → preprocessing → Feature Selection

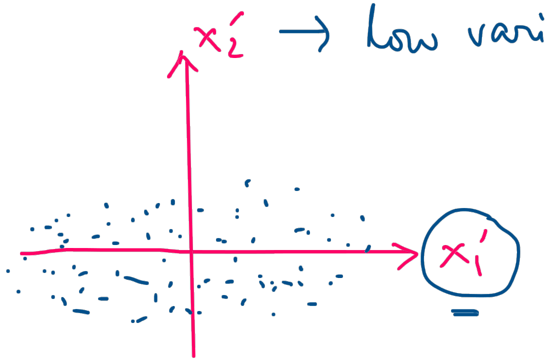


$\underline{(x_1, x_2)} \xrightarrow{PCA} (\underline{x_1'}, \underline{x_2'})$



$\overset{2}{(\phi, \underline{\lambda})}$





x_2' → low variance along x_2'

high variance along x_1'

Feature Selection

→ Filter Technique : Filter out less important features using some criteria

→ Wrapper Technique : (set of features) ⟶ (subset) → $\boxed{\text{model}}$

(RFE)

Filter Methods :

1) Variance threshold

$f_1 \longrightarrow$ variance = 0

1

1

1

1

⋮

1

f_2 (80% = 1, 20% = 0)

1

1

0

1

1

0

⋮

1

$p(1-p)$

$= 0.8 \times 0.2$

$= \underline{0.16}$

f_3 $\boxed{0.8}$

0.3

0.5

1.2

⋮

2) Univariate F-test

	feature (x)	Class(y)
	2	A
3	3	A
	4	A
	5	B
6	6	B
	7	B
1.5	8	C
0.5	9	C
9.5 0.5	10	C
1:5	11	C

Step-1 : Compute Group mean. of x

$$\bar{X}_A = \frac{2+3+4}{3} = 3$$

$$\bar{X}_B = 6 \quad , \quad \bar{x}_C = 9.5$$

Step-2 : Overall mean

$$\bar{x} = \frac{65}{10} = 6.5$$

Step-3 : Between Group Variance

$$SSB = 3(\bar{x}_A - \bar{x})^2 + 3(\bar{x}_B - \bar{x})^2 + 4(\bar{x}_C - \bar{x})^2$$

$$= 3 \cdot 3.5^2 + 3 \cdot (0.5)^2 + 4 \cdot 3^2$$

$$= 73.5$$

$$MSB = \frac{SSB}{df} = \frac{73.5}{2} = 36.75$$

Step-4 : Within Group variance :

$$SSW = \sum_j \sum_i (x_{ij} - \bar{x}_j)^2 = 9 \checkmark$$

$$MSW = \frac{SSW}{dof} = \frac{SSW}{\underset{\#\text{data}}{\underset{\downarrow}{N}} - \underset{\#\text{class}}{\underset{\downarrow}{k}}} = \frac{SSW}{10-3} = \frac{9}{7} = 1.28$$

Step-5 : F - statistics :

$$F = \frac{MSB}{MSW} = \frac{36.75}{1.28} \simeq \underline{28.7}$$