Features vs Targets

① What is a feature?

inputs/Features

Target

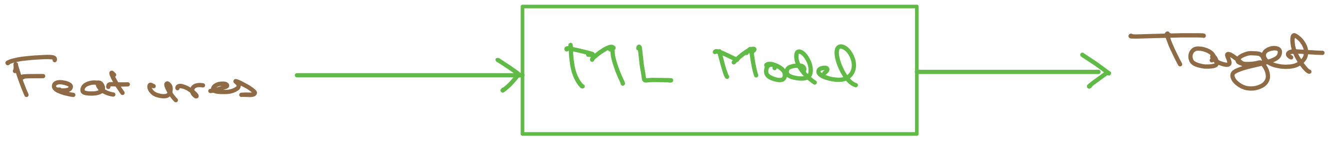
Customer Dataset

1 cust

index	education	Gender	Income	fitness . . . . - - - - - - - . . usage	Product
0	10th	M	20,000	----------	P1
1	12th	F	30,000	----------	P1
2	Masters	M	60,000	----------	P3
,,	,,	,,	,,	----------	,,
,,	,,	,,	,,	----------	,,
,,	,,	,,	,,		,,
,,	,,	,,	,,	----------	,,
n					

Goal: Given 'details' of a person build ML 'Model' to Recommend Best Product.

Features	Targets
Education Gender Income ,, ,,	Product



Feature Engineering

Processing features to improve ML Models

⇒ create New Feature

⇒ Apply transformation on Existing Features

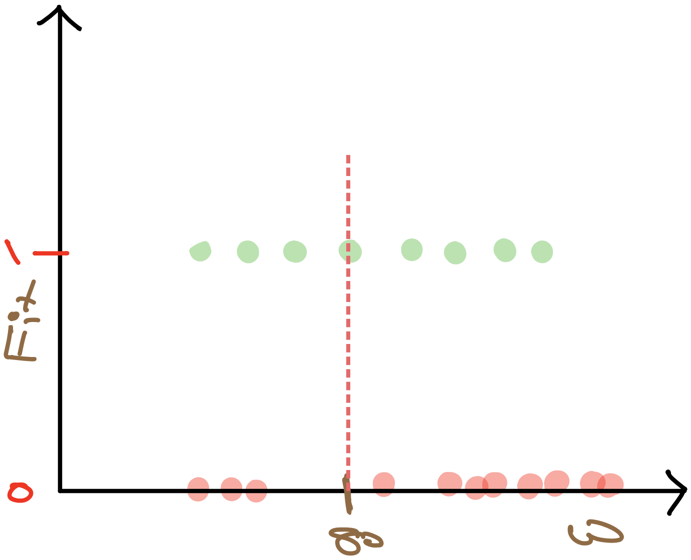
Ex: Given weight and Height of a person predict if they are Fit or Not Fit

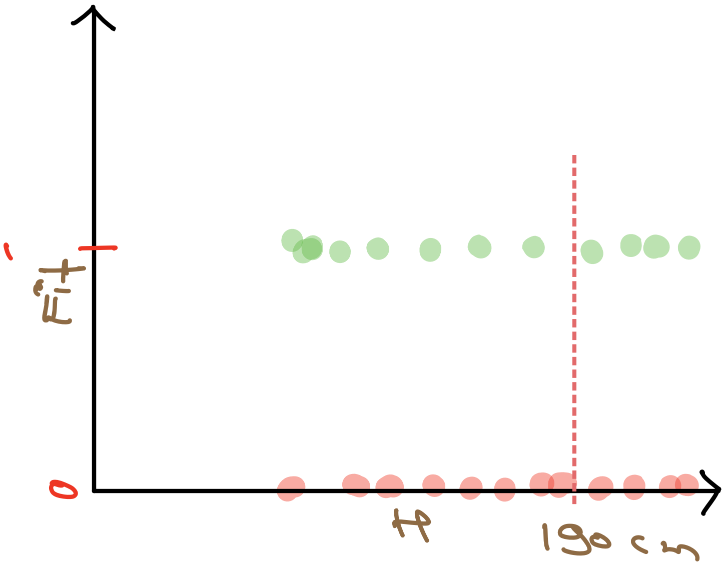
H	W	Fit
170	70	Yes
160	65	Yes
165	80	No
150	90	No
"	"	"
"	"	"

Features = H, W

Target = Fit

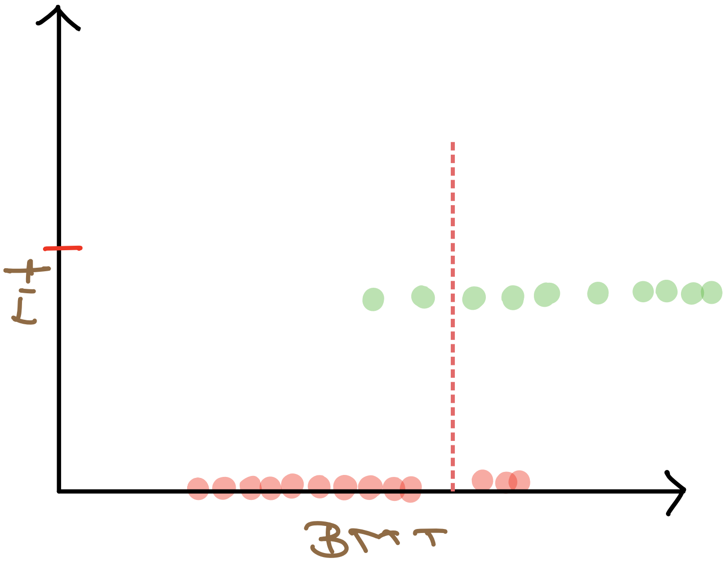
Can we plot the features against Targets





Q Is there anything that we can do to seperate Fit and Non-Fit people cleanly?

H	W	BMI	Fit
170	70	N_1	Yes
160	65	N_2	Yes
165	80	N_3	No
150	90	,,	No
,,	,,		,,
,,	,,	,,	,,



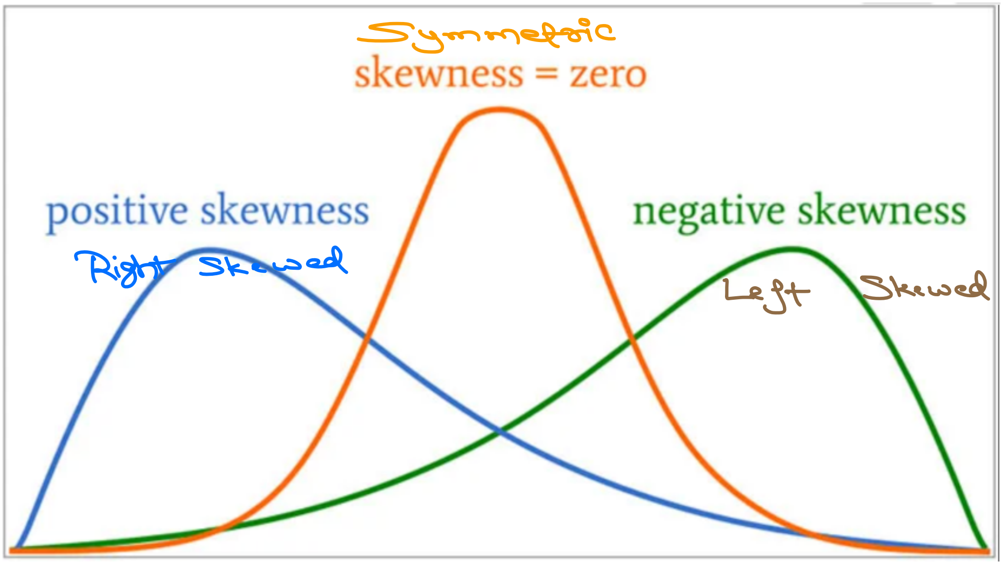
Conclusion

Feature Engineering can help build Better ML Models Efficiently

Lets Dive Deeper with Loan Status Case study in Colab Notebook

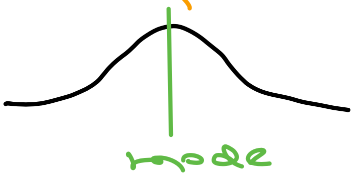
Skewness

Skewness is a measure of Asymmetry in the distribution of Data



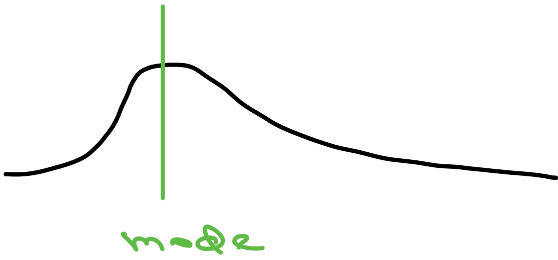
* No Skewness (Symmetric)

- Perfectly Balanced with even spread on both sides
- Mean ≃ mode ≃ Median



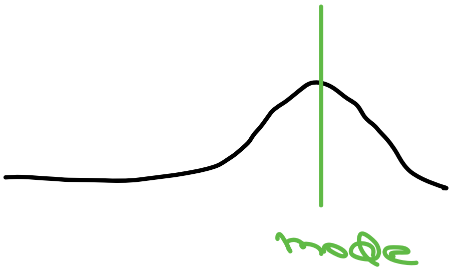
* Positive Skew (Right Skewed)

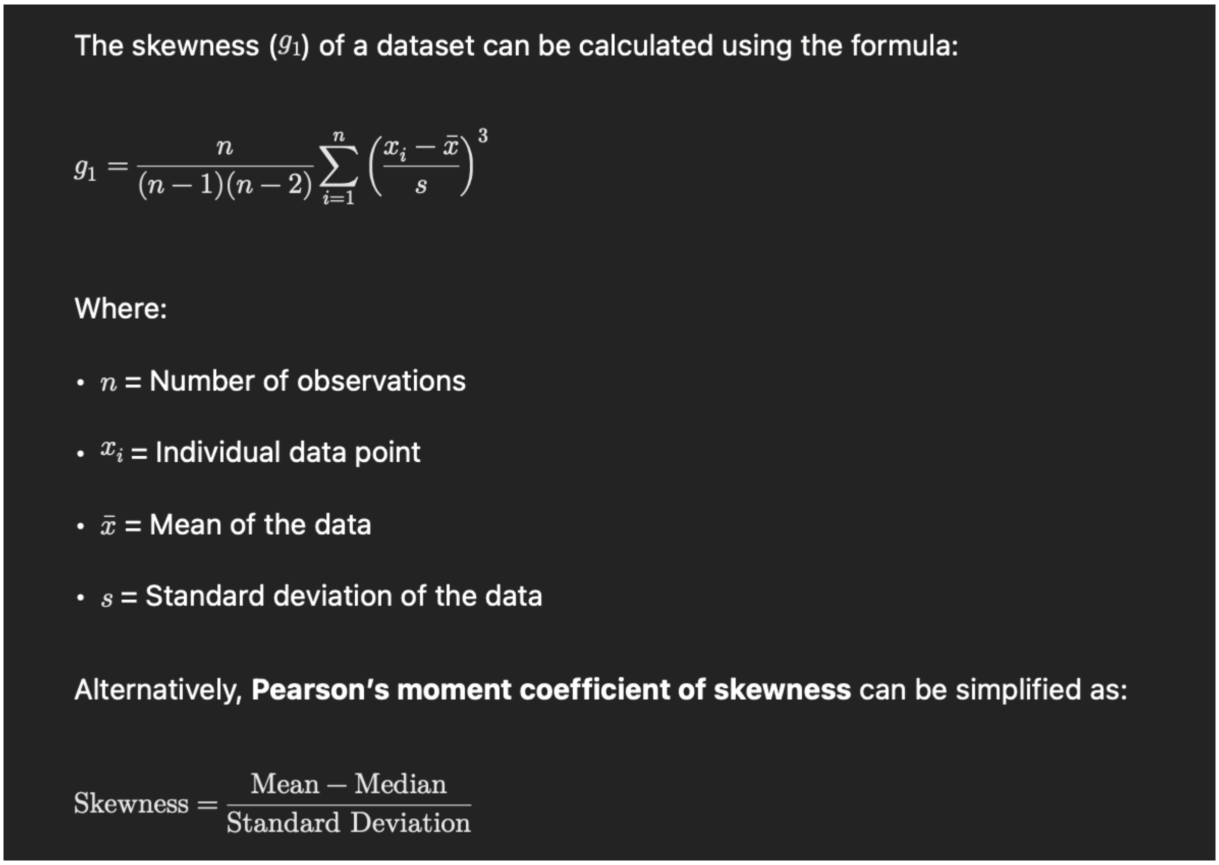
- Long Tail on Right Side
- Bulk data on Left



* Negative Skew (Left skewed)

- Long Tail on Left Side
- Bulk data on Right





The skewness (g_1) of a dataset can be calculated using the formula:

$$g_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3$$

Where:

- n = Number of observations
- x_i = Individual data point
- $\bar{x}$ = Mean of the data
- s = Standard deviation of the data

Alternatively, **Pearson's moment coefficient of skewness** can be simplified as:

$$\text{Skewness} = \frac{\text{Mean} - \text{Median}}{\text{Standard Deviation}}$$

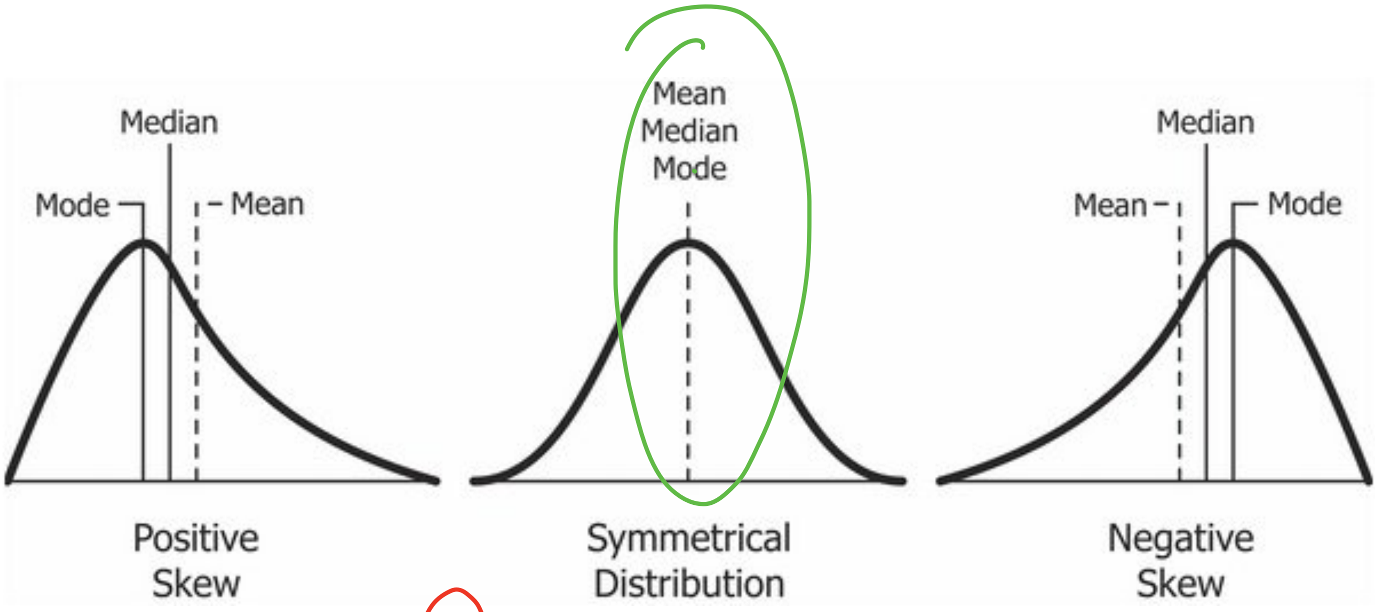
Interpretation of Skewness :

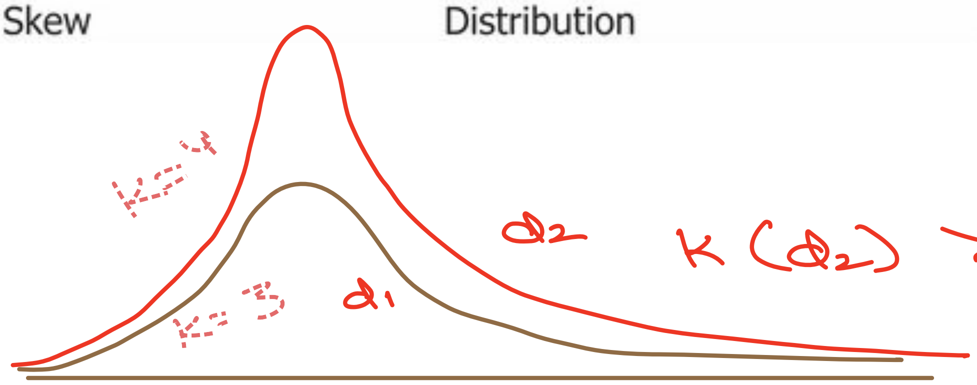
Skewness > 0   Right skewed (Positive)

Skewness < 0   Left Skewed (Negative)

Skewness = 0   No skew (Symmetric)

Mean vs Median vs Mode





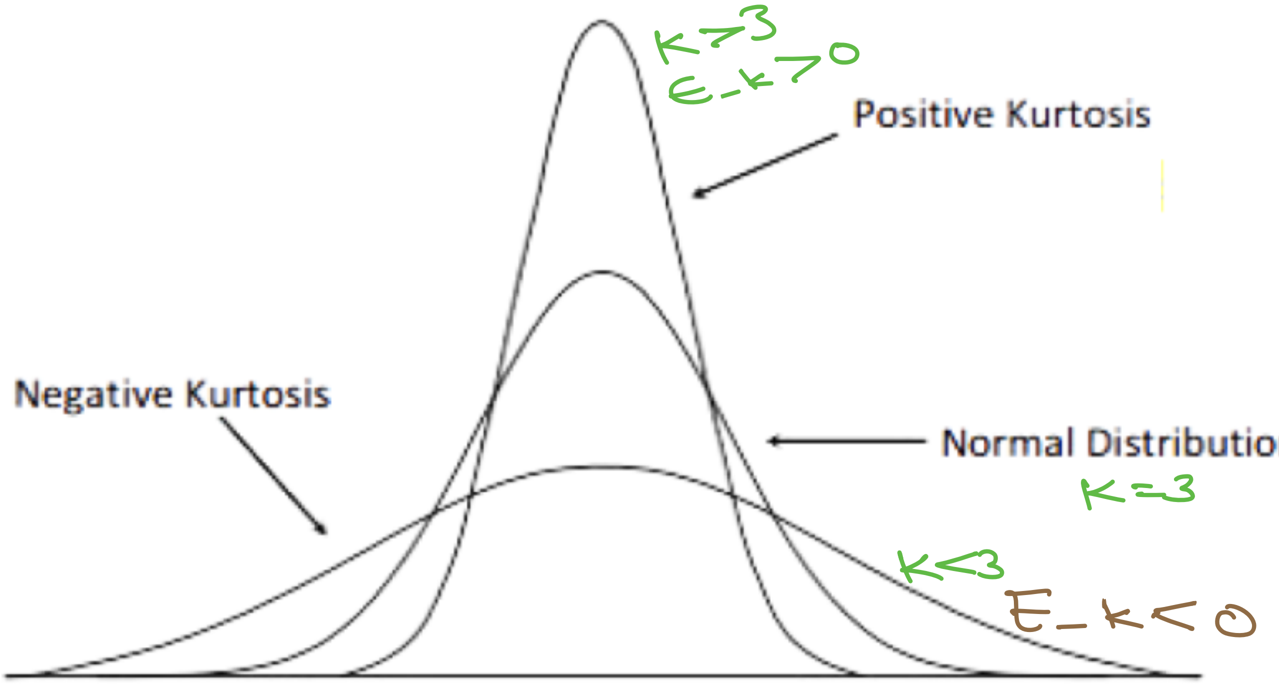
$K(d_2) > K(d_1)$

Kurtosis

Kurtosis measure 'Sharpness of Peak' of Data Distribution

High kurtosis	Low kurtosis
⇒ High Peak	⇒ Low Peak
⇒ Heavy tails	⇒ Light tails
⇒ More Outliers	⇒ Less Outliers

Excess Kurtosis ⇒ Measures kurtosis wrt Normal Distribution (kurt = 3)



$\text{Excess\_K} = \text{Kurt}_{\text{dist}} - 3$

① Leptokurtic $(EK > 0)$

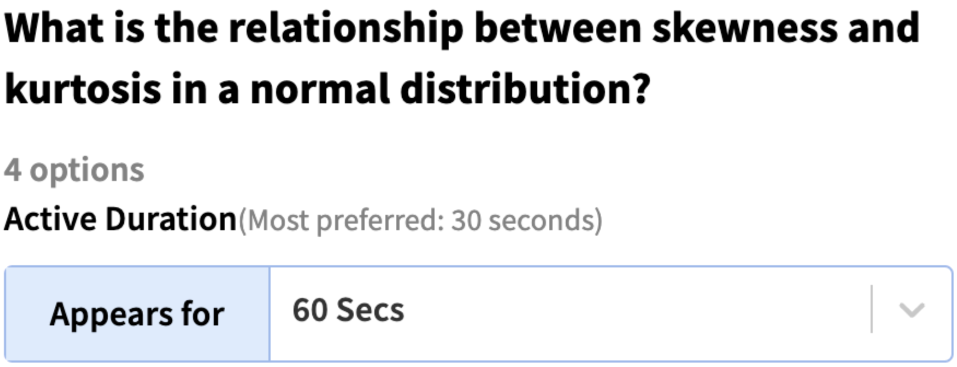
$d \Rightarrow 3.1$
$E\text{-}k \Rightarrow 0.1$

② Mesokurtic $(E\text{-}k \simeq 0)$

Approx Normal

③ Platykurtic $(E_{-k} < 0)$

A dist with negative E-K



What is the relationship between skewness and Excess kurtosis in a normal distribution?

4 options

Active Duration (Most preferred: 30 seconds)

$\text{Skew} = 0$
$E\text{-}k = 0$

$K = 3$
$E\text{-}k = K - 3$
$3 - 3 = 0$

New Feature

Feature: **Able to Pay EMI**

Consider Following two Loan applicants

① P1 earns 20L and applies for 50L Loan

② P2 earns 30L and applies for 5CR Loan

Who is more Likely to Get Loan?