Data Pre-Processing

- Handling missing value - Measure of central tendency — KNN Imputation & Iterative-MICE
- Handling Char variable - Encoding concept — Uni-Variate Analysis
- Handling Outlier - Transformation approach
- Scaling the data - Normalization / Standardization
- Handling Imbalanced dataset - Imblearn
  - applicable only with Classification Problem

* Imbalanced Vs Banlanced dataset

2 class <
- (+)ve = Yes
- (-)ve = NO

$n = 1000$

n_1 = (+)ve
n_2 = (-)ve

$n = n_1 + n_2$

Case I :- if $n_1 \approx n_2$ == Balance dataset

$n_1 = 500$, $n_2 = 500$

$n_1 = n_2$

$n_1 = 580$, $n_2 = 420$

58% 42%.

$n_1 \approx n_2$ $n_1 \neq n_2$

Case II :- If $n_1 << n_2$ or $n_1 >> n_2$ → Imbalanced dataset

$n_1 = 900$, $n_2 = 100$

n_1 → Majority class → 90%

n_2 → minority class → 10%

Severely imbalanced

Minority * 2 ≤ Majority

$200 < 900$

$n_1 = 200$, $n_2 = 800$

$n_1 = 700$, $n_2 = 300$

0 class - Real train = 284315 = n_1 = 99.7%

1 class - Fraud = 488 = n_2 = 0.01%

Split the data into training & testing

70%. — Building — Purpose — both datasets

30%. — Predicting —

$n_1 = 50$ 1.
$n_2 = 50$ 1

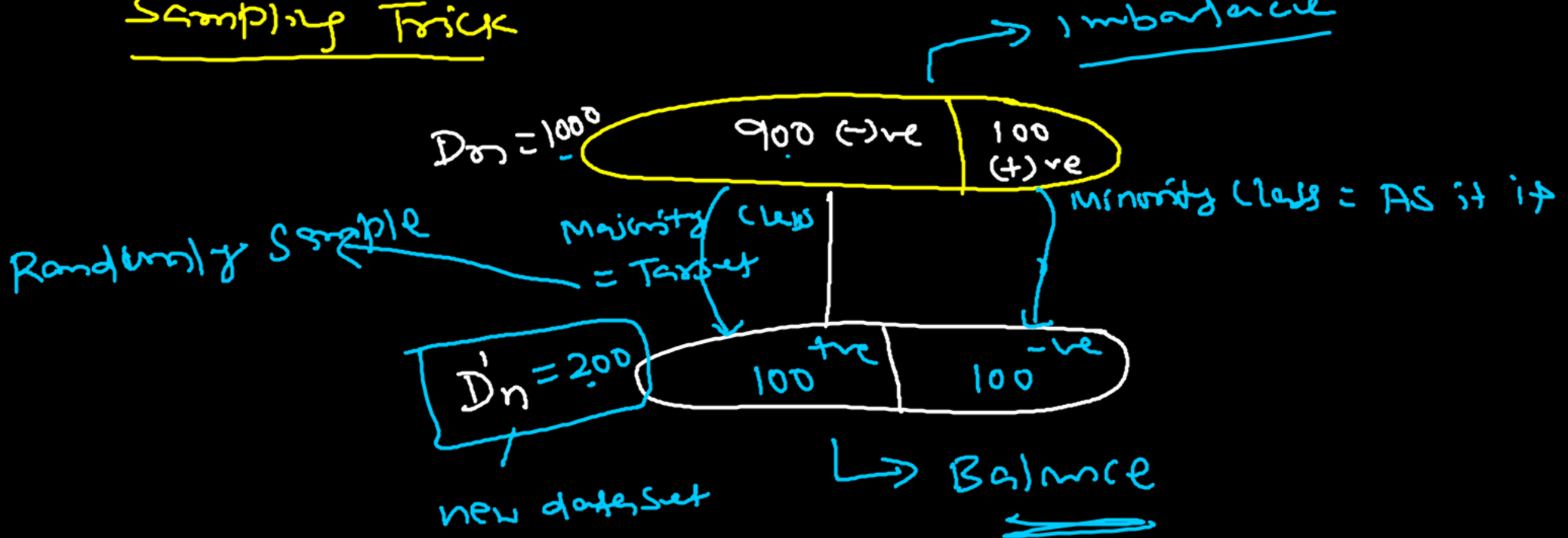
* How to work around imbalance dataset issue

1. Undersampling method → Nearmiss
2. oversampling method.
3. Artificial/synthetic technique (SMOTE)
4. Class weight - Technique.

* Undersampling method – Target (Majority class)

D_n < $n_1 = 100$ (+)ve — Minority class
$n_2 = 900$ (−)ve — Majority class

Sampling Trick



Problem with undersampling

$D_n = 1000$ → undersampling → 200 → what about 800 datapoint

↓

20% - Model builders 80% → Throwaway

$|D'_n| << |D_n|$

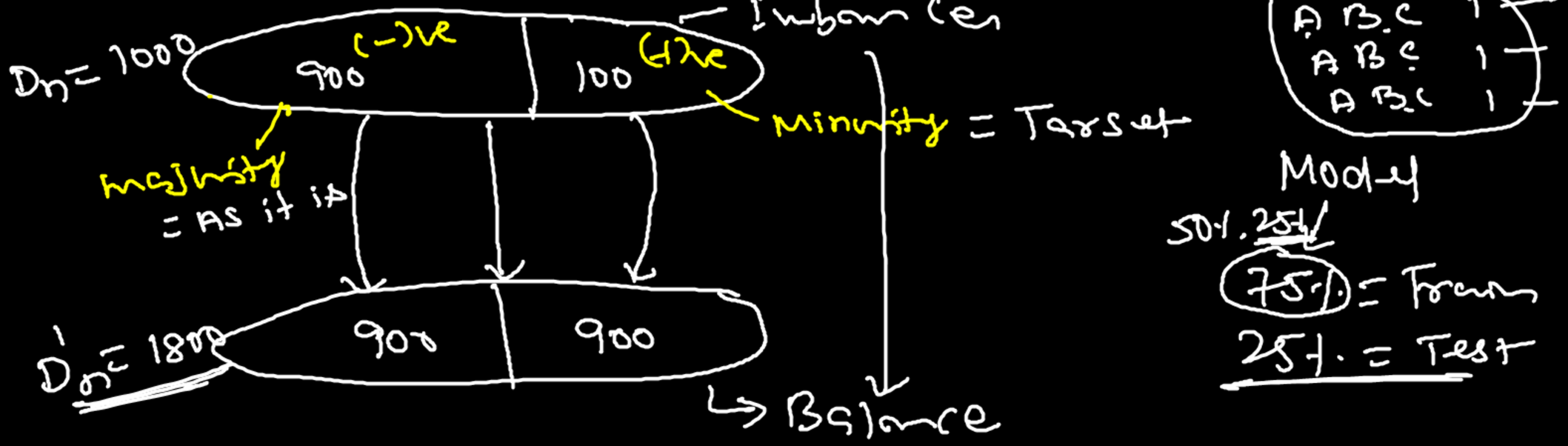
↳ Throwing away data

→ Should be avoided

→ Not good idea to remove 80% data

* Oversampling method = Target (Minority) 100 ① - very very less ① - Nirav m

D_n < n_1 = 100 = 10.1. , n_2 = 900 = 90.1. > Imbance – how to handle it ?



Model

50:1.25+/

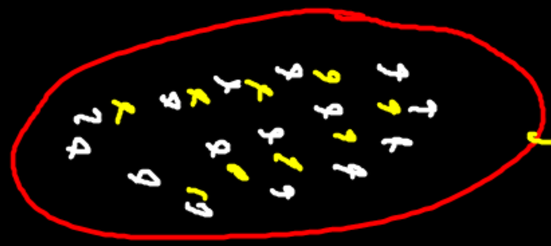
75.1. = Train

25+. = Test

Oversampling

→ Placing more points from the minority class of the dataset

Advanced Techniques → Not use very frequently



→ extrapolation — creating synthetic approach

SMOTE

Class weight 100 +ve 900 -ve $Y = mx + c$

slope/co-efficient

$W_+ = 9 \rightarrow$ More weight to the minority class

$W_- = 1 \rightarrow$ less " " " majority "

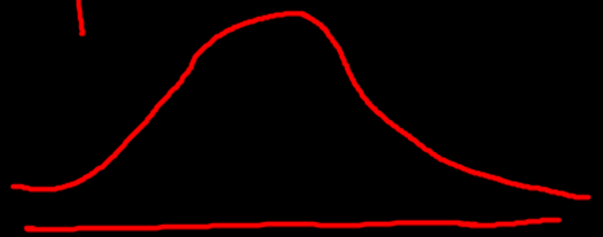
AI

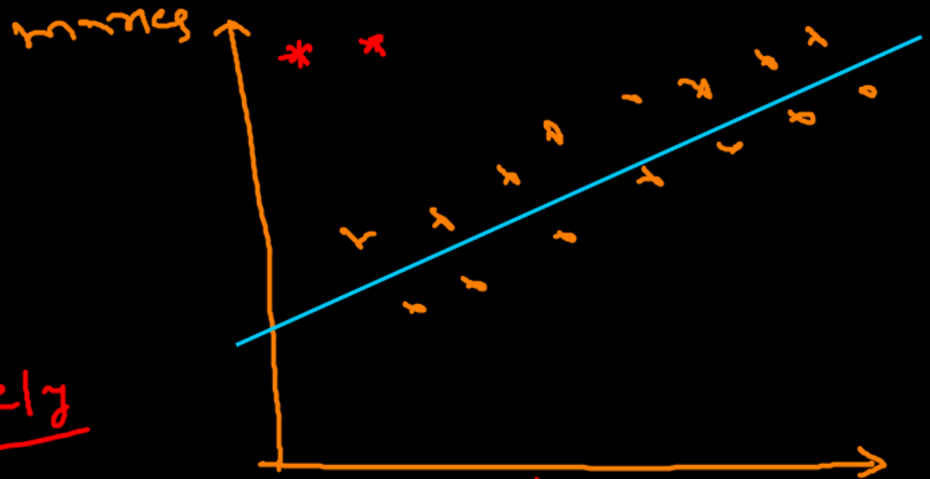
Outlier Treatment

How to handle outlier

- Trimming
- Capping
- remove - rarely
- Discritization

thin
Fast





Milk	
50	50
60	60
80	80
500	80

(+)ve outlier = Q3 + 1.5 * IQR - iski

(-)ve outlier = Q1 - 1.5 * IQR — iski

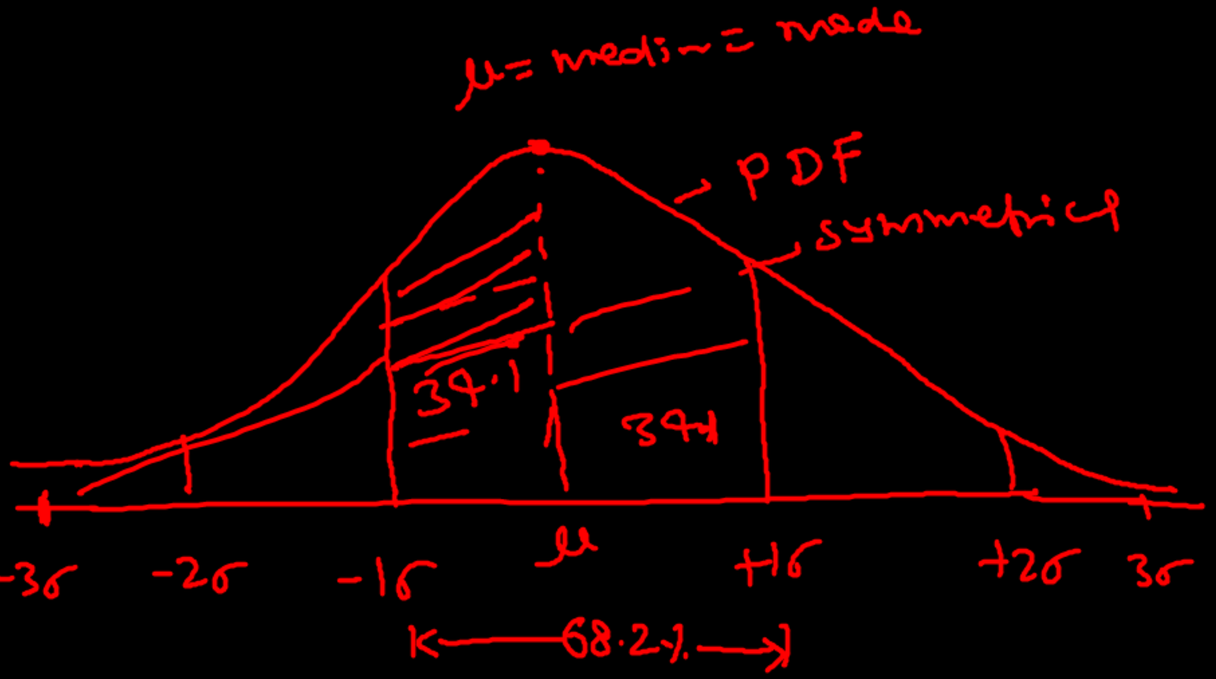
Normal Distribution

Empirical Rule

(i) $\text{Prob}[\sigma - \mu \le x \le \sigma + \mu] \approx 68.2\%$

(ii) $\text{Prob}[2\sigma - \mu \le x \le 2\sigma + \mu] \approx 95.4\%$

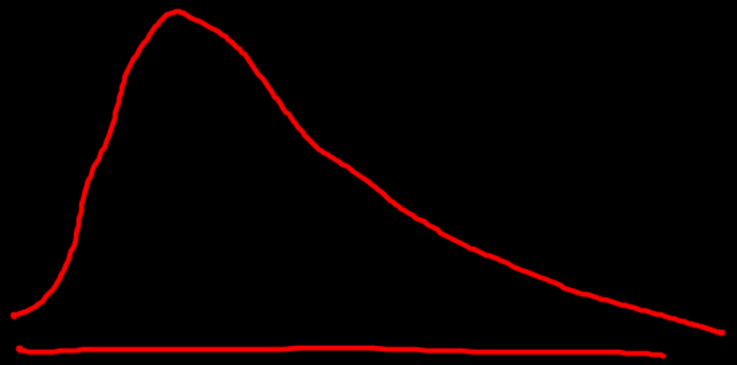
(iii) $\text{Prob}[3\sigma - \mu \le x \le 3\sigma + \mu] \approx 99.7\%$



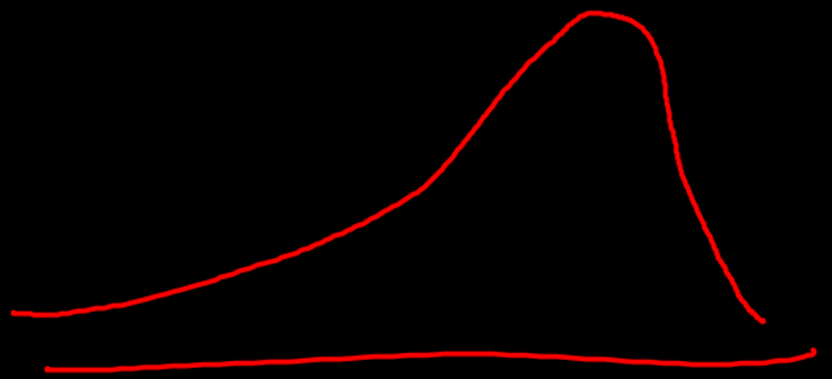
$(\mu + 3\sigma) >$
$(\mu - 3\sigma) <$
] outlier

↳ Z-Score method

(2) Skewed dist - Box Plot = IQR



Right tail Skew



Left Sk