Imbalance Data

02 April 2022 07:33

2 class classification

Total = 1000 pts.

DF → n_1 +ve 100 : 1

n_2 -ve 900 : 9

n_1 0 100

n_2 1 900

Balanced ..

$n_1 = 580$ 58% ✓

$n_2 = 420$ 42% ✓

1) Majority class n_2

2) Minority class n_1

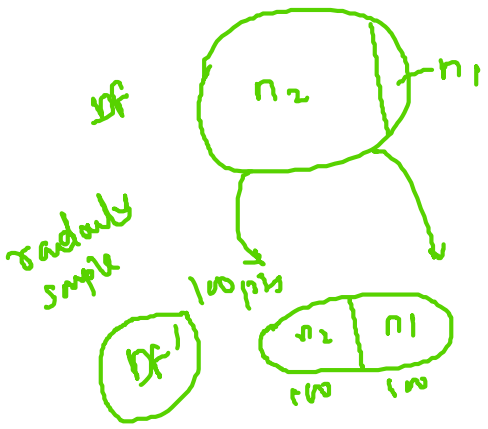
Majority class dominates minority class

1) Undersampling

2) oversampling

* Undersampling DF → n_1 : 100 +ve

$n_2 = 900$ -ve



$n_1 = 100$

$n_2 = 100$ } Balanced data.

original df = 1000

df = 1000 $n_1 = 100$, $n_2 = 100$ } smaller.

100 → 900 → 900

250 → 750 → 500

Oversampling

DF n_1 100 +ve

n_2 900 -ve

DF1 n_1 100 x 9 = 900 +ve } 1800

n_2 900 -ve

n_1 = 100 → 500 = 450

n_2 = 900

Balanced → 70 30 / → 65 35 / → Preferred

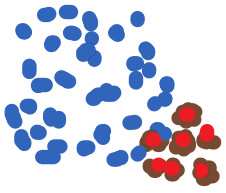
3 : 2 : 1

33 % 66 %

Fraud :- 100 10 /

90 % .

oversampling



● → +ve

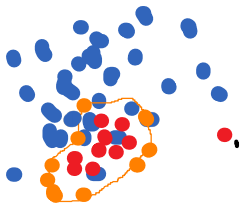
● → -ve

● → oversampled / replicated duplicate.

repeatation is simplest -

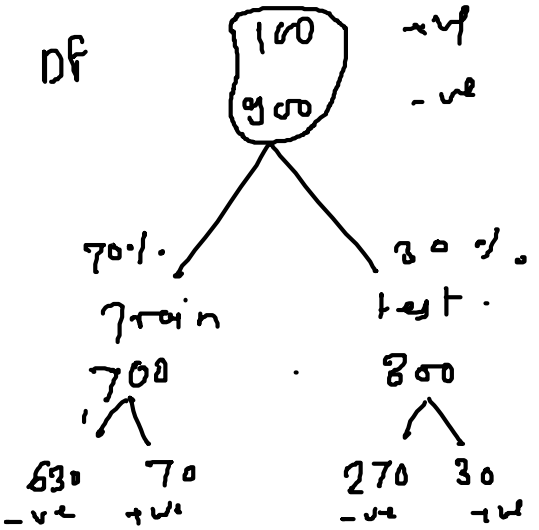
Artificial points | synthetic SMOTE

hi



→ extrapolating

DF



g.ve : 1 +ve.

→

accuracy : $\frac{270}{300} = 90\%$ → dumb

200

EX medical :- 10% cancer

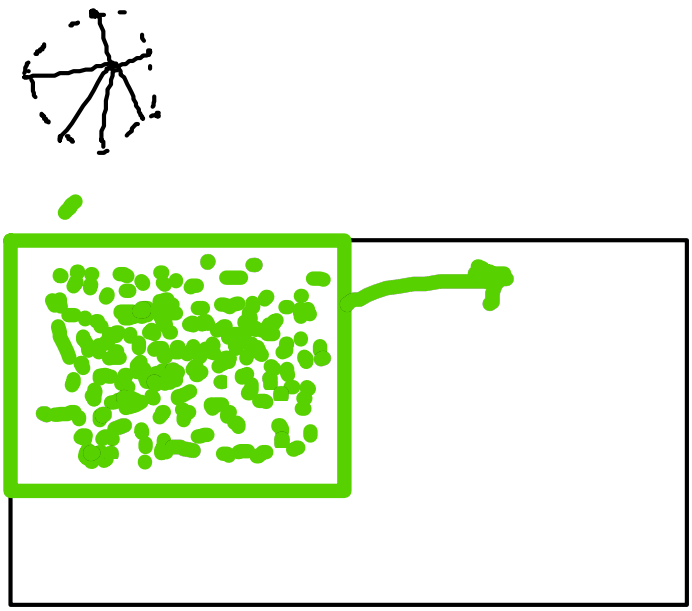
90% no-cancer

ecommerce :- 90% of customer do not buy

10% buy

SMOTE :- synthetic minority oversampling Technique

ADASYN :- Adaptive synthetic sampling.



hi