DT - 1

Imbalanced data

Agenda

* Imbalanced Dataset → Under Sampling, Over Sampling, SMOTE
* DT Intro
* Geometric Intuition
* Algorithm (mathematical)
* PURE NODE → Entropy → Information Gain

Imbalanced dataset

Binary Classification. → 1000

One class → dominated.

950 → majority class

50 → minority class

50% - 50% → Balanced

60% - 40% → Slightly balanced

70% - 30% → Slightly Imbalanced

80% - 20% → Imbalanced

90% - 10%
95% - 5%
} → Extremely Imbalanced

Fraudulent Transaction

HDFC

10000 → 1% → 100 Fraud.

99% → 9900 Normal

Data

- 1% Fraud
- 99% Normal

(1) Classification - Modul

Dumbest Modul → every trans is → Normal

Accuracy ⇒ 99%

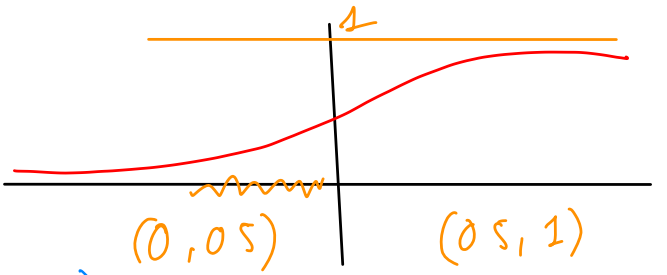
(2) Modul

1% - F

99% - NF

↳ Tends to become biased towards Majority Class.

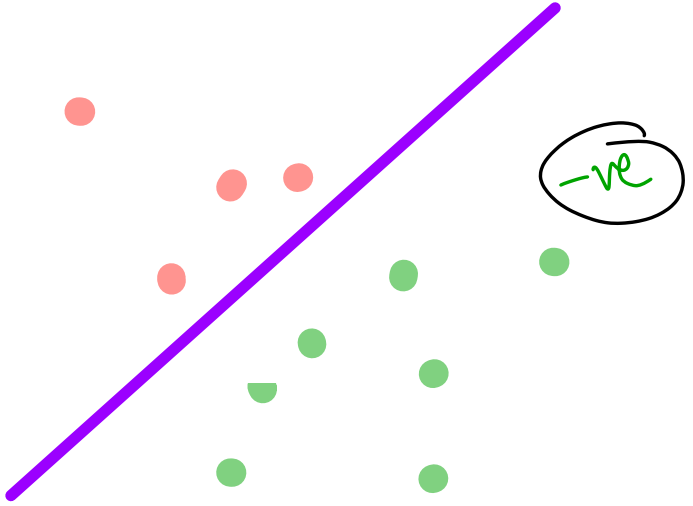
Impact of Imb. data on Log. Reg.

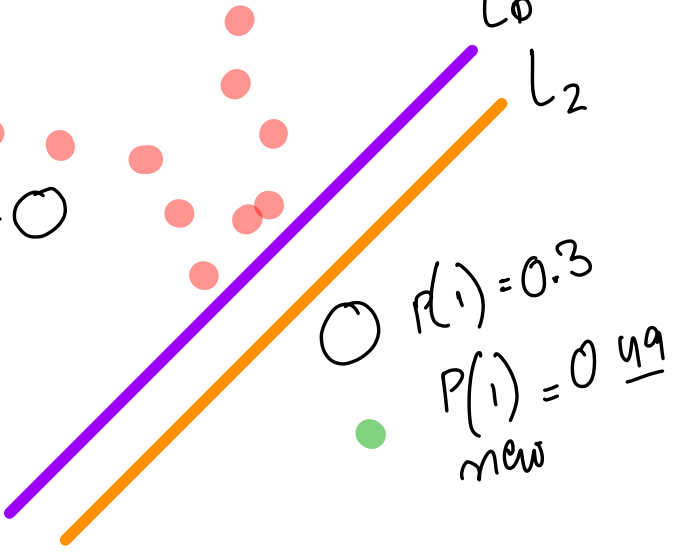


$$\text{log loss} = \sum \text{log loss}_i$$

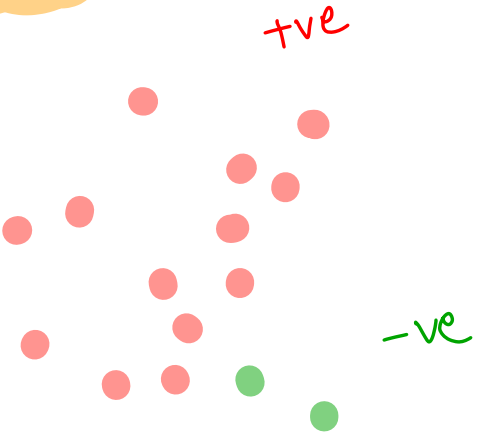
$$y \log(\hat{y}) + (1-y)\log(1-\hat{y})$$

tve





Impact of Imbalanced data on KNN



$k = 1, 3$

low

Impact of Imb. data for low values of k is comparatively low

$k = 100$

high

Impact of Imb data for high values of k is HIGH.

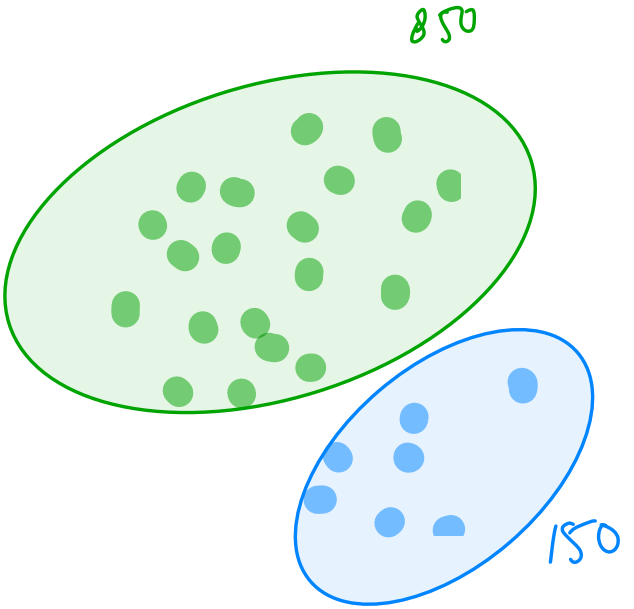
Handle Imb. dataset.

① Class weights

1000 emails (Imbalanced)

850 Normal (majority)

150 Spam (minority)



$$\frac{850\ NS}{150\ S} \approx 5.66 \approx 6$$

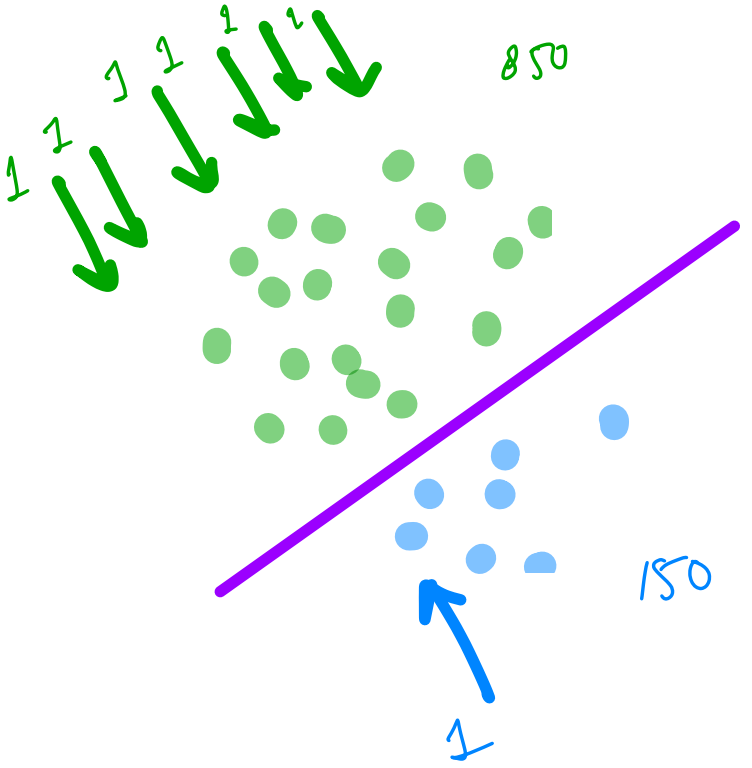
For every 1 Spam email, I have 6 NS emails.

w_i (+ve) = 1 | y_i = 1 (+ve)

w_i (-ve) = 6 | y_i = 0 (-ve)

$$\sum_{i=1}^{n} \text{log loss} \cdot (w_i)$$

$$\{0 \cdot 6 \;,\; 1 \cdot 1\}$$



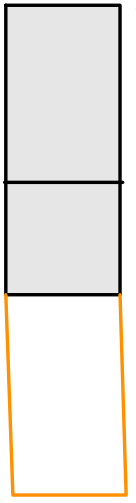
(2) Oversampling

Advantage:

* No data loss
* Preferred over under sampling

Disadvantage

* Possibility of overfitting increases.



850 + 850 = 1700

↑ Modelling

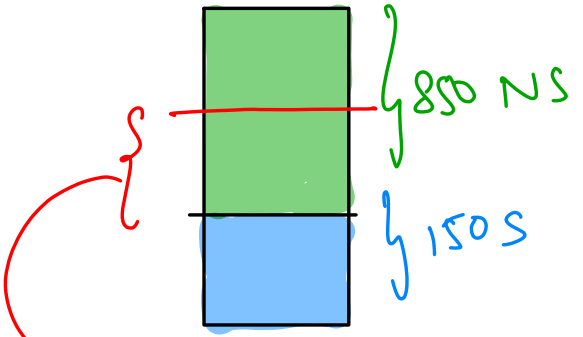
Imbalanced ↓ Balanced.

→ From where we will get this ?

Random Sampling with Repetition

Duplicating minority class Randomly.

③ Under Sampling



850 + 150 = 1000

150 + 150 ⇒ 300

Randomly select equal no. of data point equal to minority class.

Advantages:

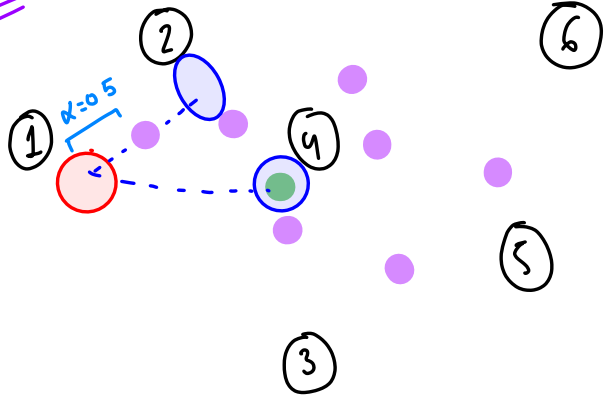
* Time & space effective.

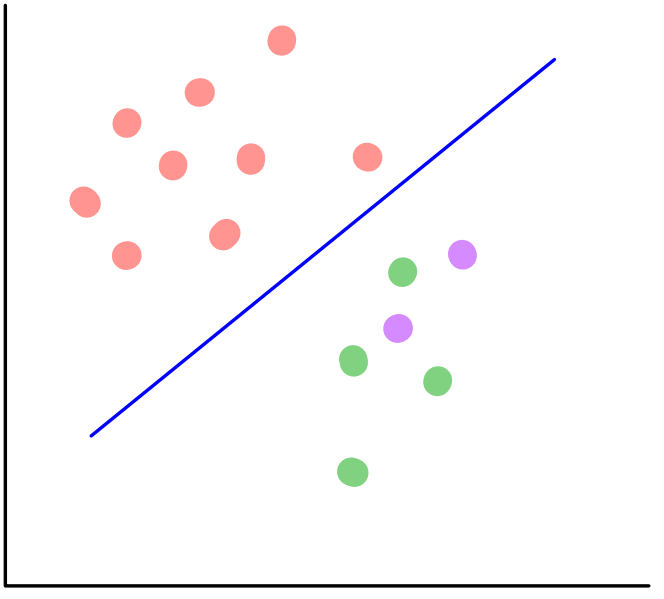
Disadvantages:

* Loss of data {expensive}
* Lose out on potentially important data points.
* Sample chosen might be biased.

(4) SMOTE [Synthetic Minority Oversampling Technique]

KNN

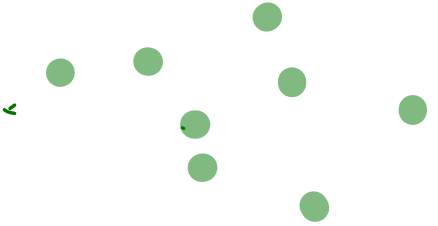




(1) K = 2

(2) Pick any value b/w (0,1) = α

(3) New point $x_{New} = x_1 + \alpha\, d(1,2)$



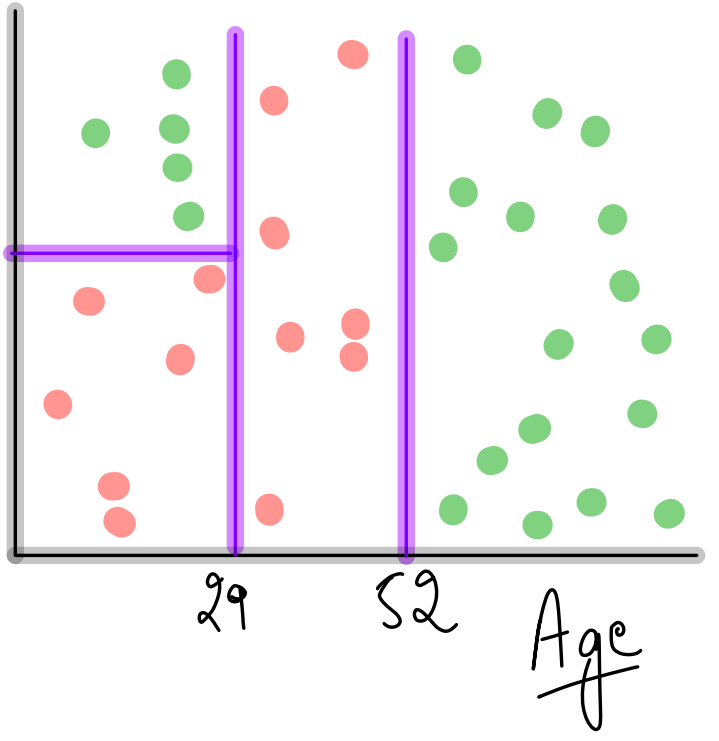
Decision Tree

+ve → Churn

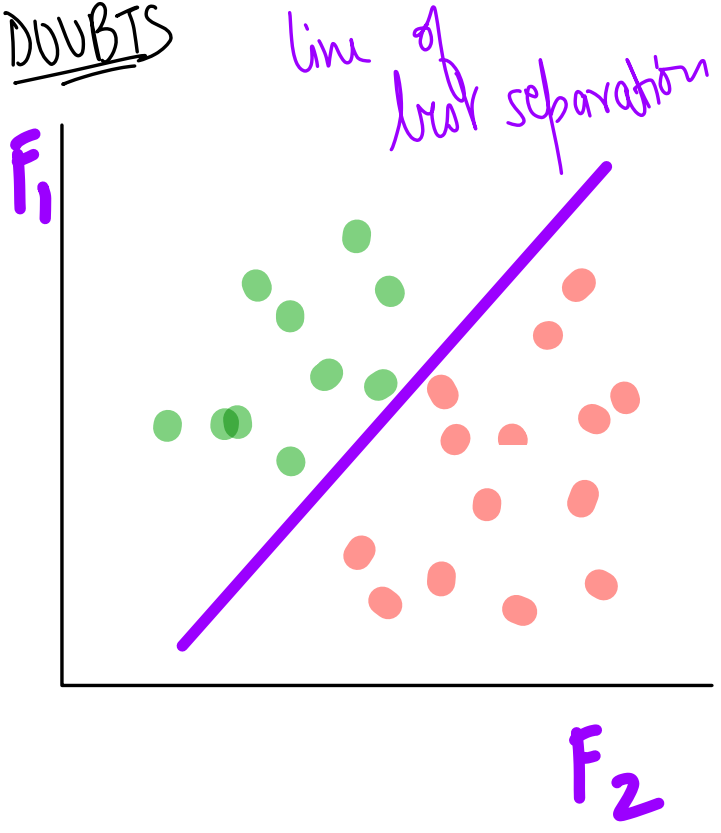
-ve → Stay

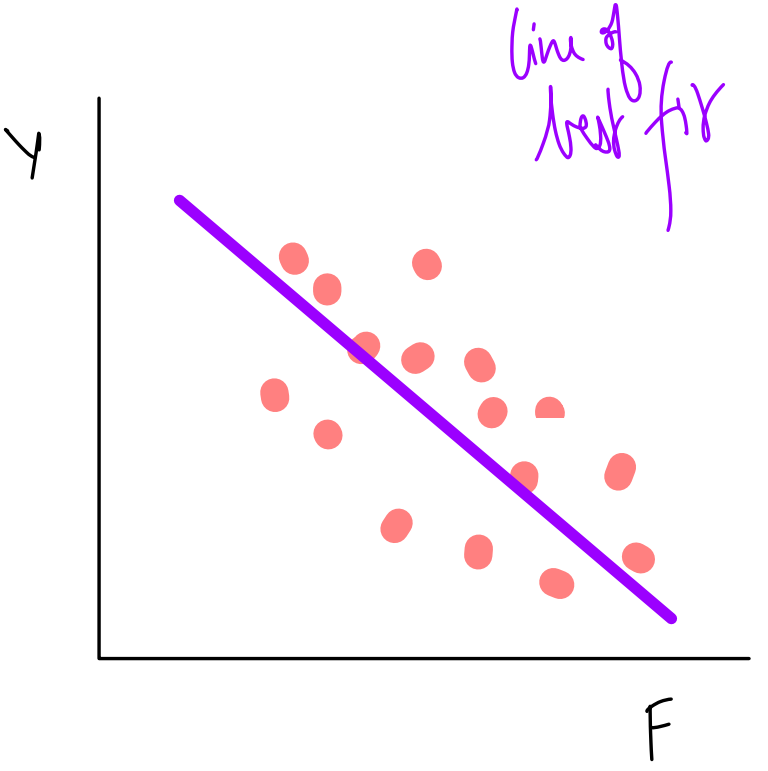
classification

Axis Parallel decision Boundaries

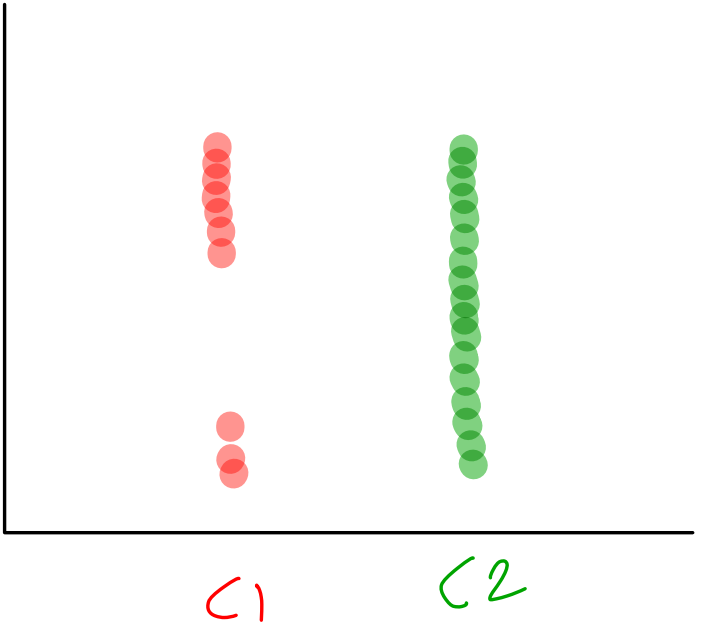


DOUBTS





Cont



odds 1 +ve 0 -ve

4 : 1

$$P(y=1|x) = \frac{4}{5}$$

$$P(y=0|x) = \frac{1}{5}$$

$$\frac{P(\text{Success})}{P(\text{Failure})} = \frac{P}{1-P}$$

$$\log\left(\frac{P}{1-P}\right) = w^T x + w_0$$

$$\log\left(\frac{P}{1-P}\right) = w_1 x_1 + w_2 x_2 \cdots w_d x_d + w_0$$