Performance measurement of models

1. What is Accuracy? (on test data)

$$\text{Accuracy} = \frac{\text{no. of correctly classified points}}{\text{Total no. of points in } D_{test}}$$

Problem: 1. imbalanced data

$D_{test} \rightarrow$ [90% −ve | 10% +ve]

even a dumb model $\rightarrow$ [$x_q \longrightarrow$ Model $\longrightarrow$ −ve]

If it return −ve for all the values

Accuracy : 90% or 0.9 with dumb model

Accuracy is not a good measure for imbalanced data.

2

x	y	M_1	M_2	$\hat{Y}_1$	$\hat{Y}_2$
x_1	1	0.9	0.6	1	1
x_2	1	0.8	0.65	1	1
x_3	0	0.1	0.45	0	0
x_4	0	0.15	0.4	0	0

model M_1 & $M_2 \rightarrow$ returns probability score

$\therefore x_q \longrightarrow prob(y_q = 1)$ · - - $\rightarrow (0 \le P \le 1)$

* predicted class labels are exactly same as M_1 & M_2
* But looking at probability score, M_1 is better than M_2.

Accuracy cannot use probability score.

2. Explain about Confusion matrix, TPR, FPR, FNR, TNR?

A.] Confusion Matrix :-

Actual Values

Predicted Values	Positive (1)	Negative (0)
Positive (1)	TP ↑	FP ↓
Negative (0)	FN ↓	TN ↑

Principal diagonal element

* it can be similarily extended for multiclass classification
* principal diagonal element should be large.

$\boxed{T \mid P}$ → P: what is the predicted label; T: are you correct (model)

* confusion matrix cannot process probability score.

Performance measurement of models

1. True positive rate:

$$TPR = \frac{TP}{P} \uparrow$$

2. True negative rate

$$TNR = \frac{TN}{N} \uparrow$$

3. false positive rate:

$$FPR = \frac{FP}{N} \downarrow$$

4. false negative rate

$$FNR = \frac{FN}{P} \downarrow$$

* it works for imbalanced data

Problem:- Out of TPR / TNR / FPR / FNR → which one is more important

→ it is very domain specific

ex: cancer or not :-

↑ TPR → very high

FPR → high (that is fine)

↓ FNR → very low

3. What do you understand about Precision recall, F1-score? How would you use it?

	0	1
0	TP	FP
1	FN	TN

1. Precision :- $\frac{TP}{TP + FP}$ ← Actual true positives ← All positives predicted by model

* only cares about positive rate

Out of all pts the model predicted to be true, what percentage of them are actually true.

2.] Recall = True positive rate = $\frac{TP}{P}$:- out of all actually true point how many predicted to be true.

* only cares about positive rate.

3.] F1 Score: $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

F1 Score ↑ = Pr ↑ & Re ↑

$0 \leq f_1 \text{ score} \leq 1$

Performance measurement of models

4. What is the ROC Curve and what is AUC (a.k.a. AUROC)?

ROC (Reciever operating characteristic curve)
AUC (Area under curve)
} used for binary classification.

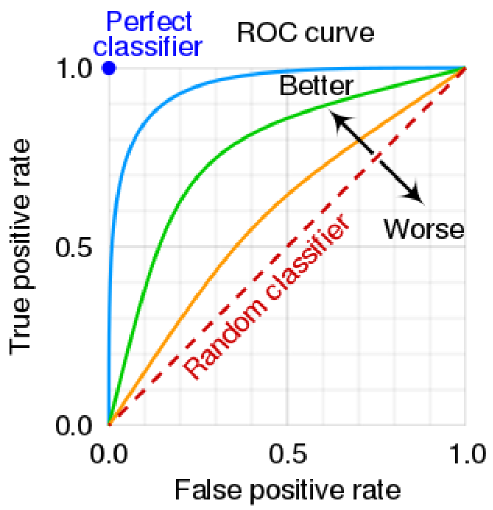
* During WW2 was developed by electronics and radio engineers to see how well missiles are working.

x	Y	$\hat{Y}$	$\hat{Y}_{\tau_1=0.95}$	$\hat{Y}_{\tau_2=0.92}$	----	$\hat{Y}_{\tau_n=0.12}$
x_1	1	0.95	1	1	--	1
x_2	1	0.92	0	1	--	1
x_3	0	0.80	0	0	--	1
x_4	1	0.76	0	0	--	1
⋮	⋮		⋮	⋮	--	⋮
x_n	Y_n	0.12	0	0	0	0
			↓ FPR_1 TPR_1	↓ FPR_2 TPR_2		↓ FPR_n TPR_n

1.] Take data and sort it in descending order

2.] Thresholding (τ)
→ for each τ we will get FPR & TPR

3.] plot it on graph

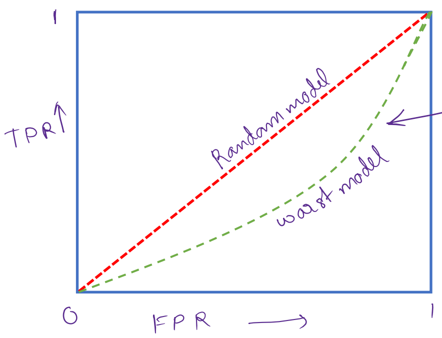


<u>Cases</u> :-

1. imbalanced data :- AUC can be high for dumb/simple model
(disadvantage)

2.] AUC is not dependent on $\hat{Y}$ score.
it only depends on order

3.] AUC (random model) = 0.5
If your model is random $x_q \longrightarrow 1$ or 0
AUC (random model) = 0.5



4.) suppose you train a model (M)
AUC (M) = 0.2
∴ worse than random model

* This tells us that there is something wrong

solutions <u>swapping</u>

$\hat{Y} = 0 \longrightarrow 1$
$\hat{Y} = 1 \longrightarrow 0$
} convert or exchange the model class label

AUC: 0 to 0.5 → worse model
0.5 → random model
0.5 to 1 → good model.

After swapping AUC = 1 − 0.2 = 0.8

Performance measurement of models

5. What is Log-loss and how it helps to improve performance?

Log Loss: it uses probability score (can lie between 0 to ∞)
$\uparrow$ better

ex: Binary classification

x	y	$\hat{y} = P$		Log(P)
x_1	1	0.9	$\rightarrow \log(0.9) \rightarrow$	0.0457
x_2	1	0.6	$\rightarrow \log(0.6) \rightarrow$	0.22
x_3	0	0.1	$\rightarrow \log(0.9) \rightarrow$	0.0457
x_4	0	0.4	$\rightarrow \log(0.6) \rightarrow$	0.22

log loss

$$\text{Log Loss} = \underbrace{-\frac{1}{n}\sum_{i=1}^{n}}_{\text{Average}} \left\{ \underbrace{\left(\log(P_i) * y_i\right)}_{\text{if } y=1} + \underbrace{(1-y_i) * \log(1-P_i)}_{\text{if } y=0} \right\}$$

Case 1: x_1 where $y_1 = 1$ $\rightarrow \{(0.0457 \times 1) + \underbrace{(1-1)}_{0} \times \log(1-0.0457)\}$

This whole term became 0

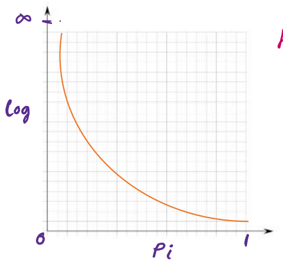
Case 2: x_3 where $y_3 = 0$ $\{(0.0457 \times 0) + (1-0) \times \log(1-0.0457)\}$

This whole term becames 0

* for 1 (+ve) if p_i value is high (0.9) then log loss is very low (0.0457). similarily for 0 (-ve).

* for 1 (+ve) if p_i value is low (0.4) then log loss is very high (0.22). similarily for 0 (-ve).

penalizing for small deviation in probability score



Log Loss = Average negative log of probability of correct class label

$0 \leq P \leq 1$

Ø Log Loss for multi class label

$x_q \rightarrow P_1, P_2, P_3 \ldots P_c$ (c classes)

$$\text{Log Loss} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c} y_{ij}\log(p_{ij})$$

Disadvantage: it is hard to interprete because log loss can be 0 – ∞

Performance measurement of models

6. Explain about R-Squared/ Coefficient of determination?

R^2 / coeff. of determination :- $\quad Y_i \in \mathbb{R}$ <- - - (real continous value)

$x_i, Y_i, \hat{Y}_i$ (predicted) ; where i : 1 to n

① error e_i : $(y_i - \hat{y}_i)$

y_i: Actual label; $\hat{y}_i$: Predicted label

② $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$ ($\bar{Y}$ = average value y_i's in test data)

③ Total sum of squares :- $SS_{total} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$

in regression ⟶ simple model ⟶ x_q ⟶ mean (y_i) = $\bar{Y}$

↓

Average model

↳ x_q ⟶ mean (y_i) i.e $\bar{Y}$ $\quad \therefore \quad \bar{Y} = \hat{Y}_i$

sum of squares of error using simple mean model

$$SS_{total} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

4. sum of squares of Residuals :- $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
(SS_{res})

$= \sum_{i=1}^{n} (e_i)^2$ ⟵ sum of all errors squared

5). $R^2 = \left(1 - \frac{SS_{res}}{SS_{total}}\right)$

$$R^2 = \left[1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}\right]$$

case 1: $SS_{res} = 0$; i.e $e_i = 0$ ⟶ $R^2 = 1$ (best value)

Case 2: $SS_{res} < SS_{total}$; $R^2 = 0$ to 1 (0 (worse) — to 1 (better))

case 3: $SS_{res} = SS_{total}$; $R^2 = 1 - 1 = 0$ model is same as simple mean model

case 4: $SS_{res} > SS_{total}$; $R^2 = -ve$ model is worse than a simple model.

Performance measurement of models

7. Explain about Median absolute deviation (MAD)? Importance of MAD?

Problem with R^2 = R^2 is not very robust to outlier.

ex: $SS_{res} = \sum_{i=1}^{n} e_i^2$ { If one error(es) is very large the R^2 will be impacted heavily.

Median Absolute Deviation:-

$x_i, \hat{y}_i, e_i$

$|e_i| = 0 \longrightarrow$ great

$|e_i| = \longrightarrow$ large $\longrightarrow$ not so good

e_i: random variable

$\text{median}(e_i)$ = central value of errors ← mean

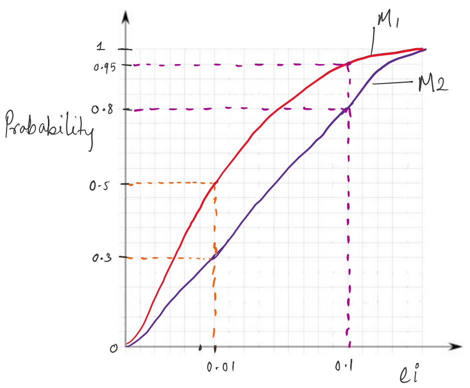
$$MAD = \underbrace{\text{median}}_{\text{median}}\underbrace{(|e_i - \text{median}(e_i)|)}_{\text{Absolute deviation}} \leftarrow \text{std dev}$$

impacted by outliers — Mean or medians of e_i's — robust to outliers

std dev or MAD of e_i's

8. Define Distribution of errors?

we can use CDF & PDF to compare 2 models (regression) with each other.



* M_2 cdf is below M_1

M_1: 95% of errors are below 0.1 ← better

M_2: 80% of errors are below 0.1

Similarly

M_1: 50% of errors are below 0.01 ← better

M_2: 30% of errors are below 0.01