1

Evaluation Metrics for Classification Problems

The most common metrics used for classification are—

→ Accuracy →Precision (P) → Recall (R)

→ F_β score → Area under the ROC (Receiver Operating Characteristic) (AUC)

→ Log loss → Precision at k (P@k) → Average Precision at k (AP@K)

→ Mean Average Precision at k (MAP@k)

It is really important to use the appropriate metric because if it is not used properly our model will not give good results. It mostly depends on the distribution of target variables in the dataset.

* When we have an equal number of positive and negative samples in a binary classification metric, we generally use accuracy, precision, recall and F-1 score.

When our dataset is imbalanced / skewed i.e., number of samples in one class out number the number of samples in another class by a lot, In these type of cases we are **not** advised to use **accuracy** as the evaluation metric as it is not representative of the data, so we might get high accuracy, but our model will not perform that well when it comes to real-world samples.

For such imbalanced datasets we use Precision, Recall and F-beta score as the evaluation metric.

② Classification Metric in case of imbalanced datasets

Type-I error

For binary classification problem.

		Actual	
		1	0
Predicted	1	TP	FP
	0	FN	TN

Type-II error

* TP (True Positive): When our actual and predicted labels match exactly with each other in terms of +ve labels, we treat them as True Positive. [Eg. 1 – 1]

TN (True Negative): When our actual and predicted labels match exactly with each other in terms of -ve labels, we treat them as True Negatives. [Eg 0 – 0].

* In simple words, if our model accurately predicts positive class, it is true positive and if your model accurately predicts negative class, it is a true negative.

FP (False Positive): When our model does not accurately predict the positive class, we say it as false positive.
Eg: [1 – 0].

FN (False Negative): When our model does not actually predict the negative class, we say it as false negative.
Eg: [0 – 1].

Goal: In classification problem, our main goal is to reduce the no. of False Positives and False Negatives.

Precision →

$$\boxed{Precision = \frac{TP}{TP + FP}}$$

Recall →

$$\boxed{Recall = \frac{TP}{TP + FN}}$$

F-1 score

$$\boxed{F1\text{-}score = \frac{2 \times P \times R}{(P + R)}} \quad \text{or} \quad \boxed{F1\text{-}score = \frac{2TP}{2TP + FP + FN}}$$

F-1 score is a metric that combines both Precision and Recall. It is defined as a simple weighted average (harmonic mean) of precision and recall.

** When dealing with datasets that have skewed targets, we should look at F1 score (or precision and recall) instead of accuracy.

Imp. Ques. When to choose Precision over Recall and vice-versa?

Ans → Suppose, we have a situation where we want to predict the fraudulent transactions in a dataset. So, in this case we will surely want to decrease the no. of false positives (FP) i.e., we want to classify genuine transactions more accurately because if we classify a genuine transaction as fraud, we may lose some important data. So, in such cases our false positives should be reduced. So, in such cases, we give more preference to **Precision** as evaluation metric.

Similarly, in case of Spam classification we want to correctly classify each email, i.e, to **reduce the no of false positives**. So, we use **Precision** as the evaluation metric.

Suppose another scenario where we have to classify two patients as being COVID +ve or COVID -ve based on their reports and symptoms, then in such cases we want to reduce the False negatives in our classification report because, if we classify a having COVID +ve as covid-ve then in such scenario, that person can loose his/her life which is very costly. So, **we have more focus on reducing the False Negatives**, in such cases we use **Recall** as the evaluation metric.

F-Beta Score -

$$F_{\beta eta} = \frac{(1+\beta^2)\ P \times R}{\beta^2 * P + R}$$

When $\beta = 1$, It is called $\boxed{F1\text{-score} = \frac{2\ P \times R}{P + R}}$

We use FP and FN are equally important then, we use β value equal to 1.

- When FN is more impactful than FP, we increase the β-value, to $\beta = 2$, $\beta = 3$, $\beta = 4$, ----
- When FP is more impactful than FN, we decrease the β-value, to $\beta = 0.1$, $\beta = 0.2$, ----

* **True Positive Rate (TPR)** ⇒ It is same as recall.

$$\boxed{TPR = \frac{TP}{TP + FN}}$$

It is also known as sensitivity.

* **False Positive Rate (FPR)** ⇒

$$\boxed{FPR = \frac{FP}{TN + FP}}$$

* **True Negative Rate (TNR) or Specificity**

$= 1 - FPR$

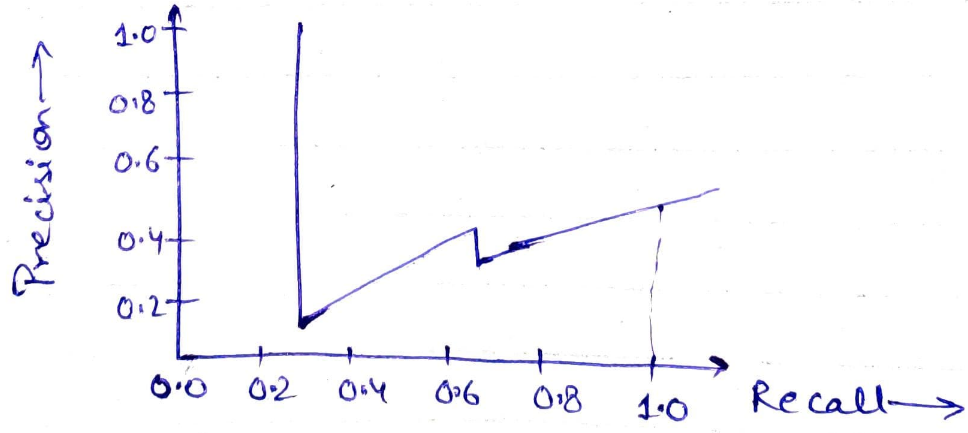
In classification problems, most models predict a probability and based on that value we have a certain **threshold** over which our classes are decided.

So, it is obvious that when we have different value of threshold, we will certainly have changed value of prediction of classes and hence Precision and Recall will also differ as per changing threshold. Thus, for different values of threshold, we can plot a graph between changing Precision and Recall which is known as **Precision-Recall curve.**

So, we need to choose an appropriate value of threshold based on the domain knowledge where we can get both good precision and recall.

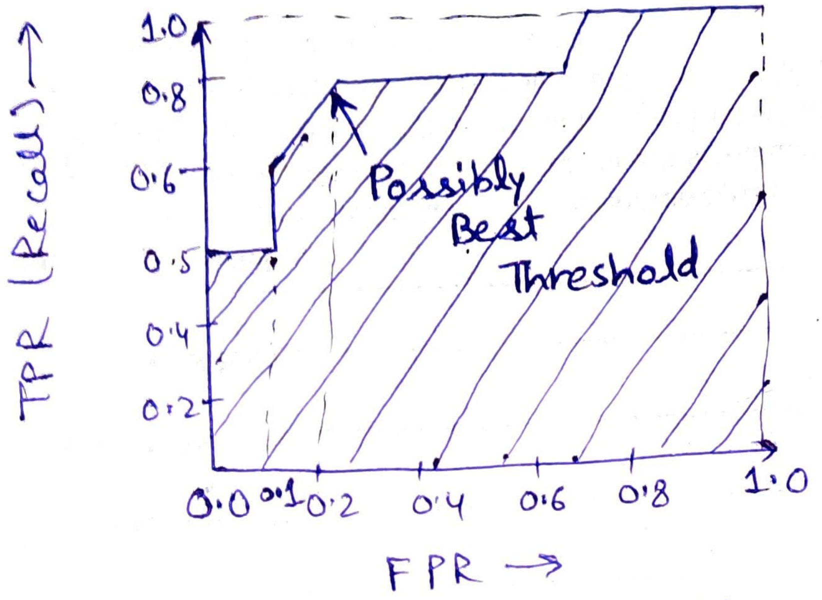
→ If the threshold will be too high, then we have smaller no. of TP and high no. of FN which decreases our recall, however precision will be increased.

→ If the threshold will be very low, FP will increase a lot and thus precision will become less.



A typical Precision-Recall curve

As we saw in the case of Precision and Recall, different threshold value can lead to different values of P&R, similarly different values of P&R will lead to different TPR and FPR and for different TPR & FPR we can plot a TPR & FPR curve.



This curve is also known as Receiver Operating Characteristic (ROC). Now, the area under the curve or simply AUC score is another metric for binary classification problems.

→ AUC=1 implies that we have a perfect model, which is not possible all the time, because some error is bound to happen.

→ AUC=0 implies that our model is dumb (i.e., it is performing very badly).

→ AUC=0.5, means that model is producing results randomly.

* We should look for those models which gives AUC score close to 1.

* Most of the time the top-left value on ROC curve should give us a quite good threshold

Threshold should be chosen such that we do not have a lot of TP & FP, a trade-off between these is observed.

In case of binary classification, we define log loss as:

$$\boxed{\text{Log Loss} = -1.0 * (\text{target}) * \log(\text{prediction}) + (1-\text{target}) * \log(1-\text{prediction})}$$

When the target is 0 or 1 and prediction is probability of a sample belonging to class 1.

For multiple samples in the dataset, the log-loss overall samples is a mere average of all individual log losses.

* One thing to remember is that log loss penalizes quite high for an incorrect or a far-off prediction, i.e. log loss punishes us for being very sure and very wrong.

→ Log loss penalizes a lot more than other metrics.

For Multi-label Classification problem

In multi-label classification, each sample can have one or more classes attatched to it.

Evaluation metrics for this type of classification problem are-

→ Precision at k (P@k)

→ Average Precision at k (AP@k)

→ Mean Average Precision at k (MAP@k)

→ Log Loss