Understanding different classification metrics - Accuracy, Precision, Recall, and F1-score

A machine learning model tends to make some mistakes by incorrectly classifying data points, resulting in a difference between the actual and predicted class of the data point. This difference in actual vs predicted values gives rise to different combinations such as -

Suppose we have a binary classification problem in which we have to predict two classes: 1 and 0.

- True Positive (TP): The values which belonged to class 1 and were predicted 1.
- False Positive (FP): The values which belonged to class 0 and were predicted 1.
- False Negative (FN): The values which belonged to class 1 and were predicted 0.
- True Negative (TN): The values which belonged to class 0 and were predicted 0.

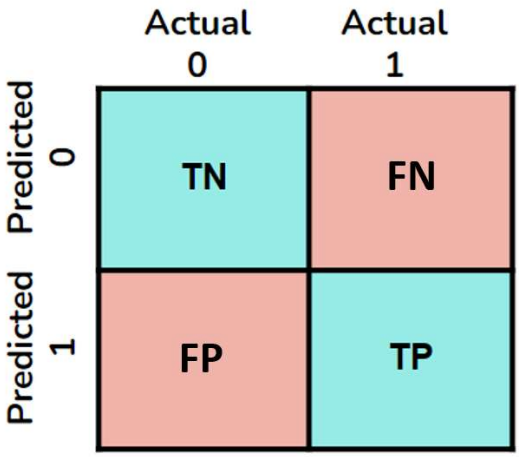
Understanding it with the help of an example

Let us consider a task to classify whether the credit card transaction is fraudulent (positive or '1') or not (negative or '0'). If this classification task is carried out by a machine learning algorithm, the output can be mapped to one of the following categories -

1. The transaction is fraudulent and classified as ‘fraudulent’. This is called True Positive (TP)
2. The transaction is genuine and classified as ‘not fraud’. This is called True Negative (TN)
3. The transaction is fraudulent but classified as ‘not fraudulent’. This is called False Negative (FN)
4. The transaction is genuine but classified as ‘fraudulent’. This is called False Positive (FP)

Clearly, we want True Positives and True Negatives to be predicted. However, no machine learning algorithm is completely perfect and we end up with False Positive and False-negative due to misclassifications.

This confusion in classifying the data can be easily shown by a matrix, called the Confusion Matrix -



From a confusion matrix, we can obtain different measures like Accuracy, Precision, Recall, and F1 scores.

How do we calculate Accuracy from the confusion matrix for a classification task?

Accuracy represents the number of correctly classified data instances (TN+TP) over the total number of data instances (TN+TP+FN+FP) which is as follows -

$$Accuracy = \frac{TN+TP}{TN+FN+TP+FP}$$

Accuracy is a very good measure **if negative and positive classes have the same number of data instances**, which means that the data is balanced. In reality, we can hardly find balanced data for classification tasks.

Let's see with an example why accuracy may not be the best measure for data with imbalanced classes -

Suppose, there are 90 people who are healthy (negative) and 10 people who have some disease (positive).

Now let's say our machine learning model perfectly classified the 90 people as healthy but it also classified the 10 unhealthy people as healthy. So, in this scenario,

True Positive cases = 90

True Negative cases = 0

False Positive cases = 10

False Negative cases = 0

and, Accuracy = (90+0) / (90+0+10+0) = 0.9 i.e, 90%

which implies that our machine learning model is working perfectly fine.

However, we can imagine the severity of the instances where our machine learning model totally misclassified unhealthy people as healthy ones. If we make any decision solely based on accuracy, 10 people will receive no medication as they are classified as 'Healthy' by the ML model.

What would be the best measure in this scenario?

In the above example, we are interested only in spotting the real positives (i.e the people who have some disease) in the dataset as often as possible, and identifying the real negatives is less important to us.

Recall can be used as a measure in such cases, where Overlooked Cases (False Negatives) are more costly and the focus is on finding the positive cases.

The recall is calculated as follows -

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

So, when should we use Precision?

Precision is a good evaluation metric to use when the cost of a false positive is very high and the cost of a false negative is low. Let's understand this with the help of an example.

Suppose we are creating a machine learning model that predicts whether a stock will be a good investment or not. If we were to invest our whole net worth in this single stock, we'd best hope our model is correct.

Precision is the appropriate statistic to employ in this case because it determines the number of stocks that are truely profitable and were also predicted as a good investment.

Precision is defined as follows -

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

In this case, we can afford to miss a few good stock investments here and there, as long as our money is going to an appreciating stock that our model accurately forecasted. Thus, Recall is not an important measure in this case.

What about the F1 score?

F1 score is the combination of Precision and Recall. If we want our model to be correct and not miss any correct predictions, then we want to maximize both Precision and Recall scores. There comes F1 score.

F1 score is defined as follows -

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Based on the above fact, we can infer that the F1 Score is the one to use above all others. However, depending on the scenario, it may be appropriate to prioritize Precision or Recall scores over the other.

Each metric has its own advantages and disadvantages, and we must decide which one to use to evaluate our model based on the problem description.