Performance Metrics - Classification

The ultimate purpose of preparing a model in data science is that it should be able to perform reliably over unseen data. This requires multiple testing of the model before applying it to unseen data. To understand the performance of a model, we need some numeric metrics that can tell about the performance of the model. Such metrics are called **performance metrics**.

In the case of classification problems, there is an abundance of performance metrics that have found suitable use in different applications. A few examples of these metrics are accuracy, recall and precision.

Performance of a classification model

There are many metrics used in classification to check model performance. One of the important elements to do performance analysis of classification problems is the confusion matrix.

Let's first define what a confusion matrix is, and then we will move forward to discuss some of the important metrics used in classification. A confusion matrix is a cross-tabular representation of the **frequency of actual vs predicted classes** in a classification problem. It helps in the comparison of the actual results vs predicted results. Refer to the image given below:

		Predicted Values	
		Positive(1)	**Negative(0)**
Actual Values	**Positive(1)**	TP	FN
	Negative(0)	FP	TN

It consists of the following numerical entities that represent important aspects of the model's performance. Using these entities, we will further derive performance metrics for classification problems.

TP = *True Positive,* i.e., the actual value is 1 and the model also predicted 1

TN = *True Negative, i.e.,* the actual value is 0 and the model also predicted 0

FN = *False Negative, i.e.,* the actual value is 1 and the model predicted 0

FP = *False Positive, i.e.,* the actual value is 0 and the model predicted 1

Before getting into performance metrics, it is important to understand the concept of **balanced and imbalanced data**. If the classes present in the target variable of the data are **equal (or approximately equal) in count,** then it is called **balanced data,** otherwise, it is called **imbalanced data**. This affects the type of performance metrics suitable for a specific problem.

Let us now understand the performance metrics used for classification problems:

Accuracy: Accuracy is one of the important performance metrics for classification problems. It is the **fraction of total observations** that are **predicted correctly** by the model. It is suitable for cases where the data is **balanced among the existing classes**. Mathematically, it can be given as follows:

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

Precision: It is the fraction of **total positive predictions** that are **actually correct**. In other words, out of all the values predicted as 1, how many are correctly predicted.

Mathematically, it can be given as follows:

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

In real-life problems, there are cases where precision is more suitable in comparison to other metrics. It is preferred in cases where we cannot afford to make false-positive errors. The higher **the number of false-positive cases,** the **lower the precision, and vice versa**.

For example, in the case of **email spam detection**, precision is more important because **missing out on detecting/classifying a spam email might be okay but no important email should go into the spam folder (false-positive)**.

Recall: Recall is again an important performance metric derived from the confusion matrix. It is the **fraction of total actually positive cases** that are **predicted correctly.** In other words, out of all the actual positive cases, how many were correctly classified.

Mathematically it is given as follows:

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

It is preferred to use recall where we cannot afford to make false-negative errors. The higher **the number of false-negative errors, the lower the recall, and vice versa**.

For example, in the case of cancer detection, recall is more important because classifying a **healthy person as having cancer (false-positive) is tolerable** because in further tests that patient would be identified to not have cancer, but it is certainly **not okay to classify a cancer patient as healthy (false-negative)** which can be fatal for the patient if it went untreated.

F1-score: Sometimes precision and recall both seem to be equally important, or any one of them may not be preferred over the other. In such a case, we need to combine both precision and recall. We can do so by creating another performance metric called the **F1-score**. It is the **harmonic mean of the precision and recall** of the model. It is majorly preferred in case of an **imbalanced class distribution** of the data.

Mathematically, it can be given as follows:

$$F1\ score = \frac{(2 * precision * recall)}{(precision\ +\ recall)}$$

< Previous

Next >