EVALUATION METRICS IN MACHINE LEARNING

Evaluation metrics are essential in machine learning to measure how well a model performs on a given dataset. They provide a standardized way to assess the effectiveness of models, helping data scientists decide whether a model is ready for deployment or needs further improvement. Without appropriate evaluation metrics, selecting the best model for a specific task would be challenging. For instance, metrics like accuracy, precision, or mean squared error allow us to quantify performance, enabling better decision-making.

Evaluation metrics can vary depending on the type of problem: classification or regression. This guide explains the key evaluation metrics for both types and how to choose the right ones for your machine learning projects.

Classification Metrics

Classification metrics evaluate models that predict discrete outcomes, such as determining whether an email is spam or not. These metrics help assess how well the model distinguishes between different classes. Below are some commonly used metrics for classification tasks:

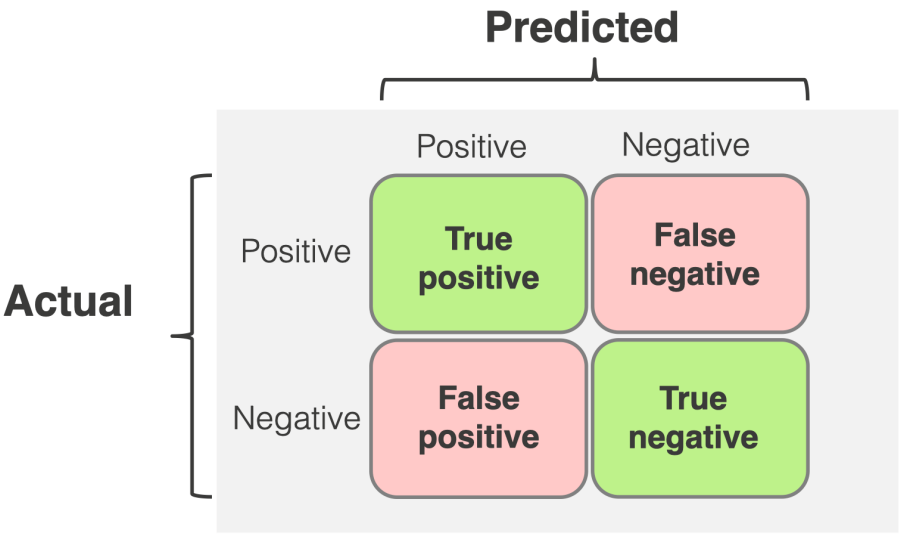
1. CONFUSION MATRIX

A confusion matrix provides a comprehensive overview of a model's predictions, breaking them down into:

- **True Positive (TP):** The model correctly predicted a positive outcome i.e the actual outcome was positive.

- **True Negative (TN):** The model correctly predicted a negative outcome i.e the actual outcome was negative.
- **False Positive (FP):** The model incorrectly predicted a positive outcome i.e the actual outcome was negative. It is also known as a Type I error.
- **False Negative (FN):** The model incorrectly predicted a negative outcome i.e the actual outcome was positive. It is also known as a Type II error.

Example of Confusion Matrix:



Use Case: The confusion matrix is a versatile tool for calculating other metrics like Precision, Recall, and Specificity.

2. ACCURACY

Accuracy is one of the simplest metrics. It measures the percentage of correctly predicted labels out of the total predictions.

Formula:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Predictions}}$$

Example: If a model correctly classifies 90 out of 100 test samples, the accuracy is 90%.

Limitations: Accuracy can be misleading when working with imbalanced datasets. For example, in a dataset where 95% of the samples belong to one class, a model predicting only the majority class will have high accuracy but poor overall performance.

Solution: Use additional metrics like **Precision**, **Recall**, or the **F1 Score** for imbalanced datasets.

3. RECALL OR SENSITIVITY

Recall measures the model's ability to correctly identify all actual positive instances. It is especially important in scenarios where missing true positives has severe consequences, such as detecting diseases or security breaches.

Formula:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

Example: If a model detects 90 out of 100 actual spam emails, its recall is

Use Case: In fraud detection, a high recall ensures that as many fraudulent cases as possible are flagged, even if it means some false positives.

7. SPECIFICITY

Specificity measures the proportion of correctly identified negatives out of all actual negatives. It complements Recall in scenarios where minimizing false positives is also critical.

Formula:

$$\text{Specificity} = \frac{\text{True Negatives (TN)}}{\text{True Negatives (TN)} + \text{False Positives (FP)}}$$

Use Case: In medical screening, high specificity ensures that healthy individuals are not incorrectly diagnosed with a disease.