
UNIT-12 PERFORMANCE EVALUATION MEASURES

- 12.0 Introduction
- 12.1 Expected Learning Outcomes
- 12.2 Confusion matrix,
- 12.3 Accuracy, Precision and Recall
- 12.4 Sensitivity and Specificity
- 12.5 F1-Score,
- 12.6 ROC Curve,
- 12.7 PR Curve
- 12.8 Summary
- 12.9 Answers
- 12.10 References/Further Readings

12.0 INTRODUCTION

Performance evaluation is essential in predictive and classification models because building a model alone does not guarantee that it is reliable, accurate, or useful in real-life situations. A model may produce outputs, but without proper evaluation, we cannot know whether those outputs are correct, consistent, and meaningful. Performance evaluation helps measure how well a model is learning from data and how effectively it makes predictions on new, unseen cases. It also allows us to compare different models and choose the one that performs best for a given problem.

In predictive modelling, evaluation tells us how close the predicted values are to the actual values. In classification models, it helps us understand whether the model is correctly identifying categories or classes. This becomes especially important in practical applications such as medical diagnosis, fraud detection, email spam filtering, or loan approval, where wrong predictions may lead to serious consequences. Therefore, performance evaluation is not just a technical step, but a necessary process to ensure trust, accuracy, and decision-making quality.

For example, consider a medical test model designed to classify whether a patient has a disease or not. Suppose the model is applied to 100 patients and predicts 90 patients as disease-free and 10 as diseased. At first glance, the model may appear accurate. However, if in reality 15 patients actually had the disease and the model correctly identified only 8 of them, then it failed to detect 7 diseased patients. In healthcare, such errors are critical because missed cases may delay treatment. This example shows that simply looking at the number of correct predictions is not enough; proper performance evaluation is needed to fully understand the strengths and weaknesses of a model.

Thus, performance evaluation helps answer important questions such as: Is the model dependable? Is it identifying the right cases? Can it be trusted in real-world use? Hence, it forms a vital part of every predictive and classification modelling process.

In the previous units, different categories of data analysis models were introduced, each designed to solve a particular type of problem. **Unit 8 – Regression Models** focused on predicting continuous numerical values, while **Unit 9 – Time Series Models** dealt with data observed over time and the identification of trends, seasonality, and future patterns. Moving further, **Unit 10 – Supervised Learning Models for Data Analysis** explained how labelled data can be used for prediction and classification, and **Unit 11 – Unsupervised Learning Models for Data Analysis** discussed techniques for discovering hidden structures, patterns, and groupings in unlabelled data. Together, these units provide a broad understanding of how different analytical models are built and applied to real-world problems.

However, building a model is only one part of the analytical process. It is equally important to assess how well the model performs and whether its predictions are reliable enough for practical use. This is the focus of **Unit 12 – Performance Evaluation Measures**. Performance evaluation helps compare models, identify their strengths and weaknesses, and determine their suitability for a given task. In predictive analytics, especially in classification problems, evaluation measures provide a systematic way to judge whether a model is making correct decisions or not.

This unit introduces the fundamental concepts and tools used to evaluate model performance, with particular emphasis on classification-based measures. Among the most important of these are the **confusion matrix**, **accuracy**, **precision**, **recall**, **sensitivity**, **specificity**, and **F1-score**, which help quantify different aspects of prediction quality. In addition, graphical evaluation techniques such as the **ROC curve** and **PR curve** provide deeper insight into a model's behaviour across different decision thresholds. These measures will be discussed in detail later in this unit, forming the basis for understanding how to interpret and compare analytical models effectively.

12.1 EXPECTED LEARNING OUTCOMES

After studying this unit, you would be able to:

- understand the need and importance of performance evaluation in predictive and classification models.
- introduce the **confusion matrix** as a basic tool for comparing actual and predicted outcomes.
- explain key evaluation measures such as **accuracy**, **precision**, **recall**, **sensitivity**, **specificity**, and **F1-score**.
- develop an understanding of how different performance measures reflect different aspects of model effectiveness.
- explain the role of **ROC curve** and **PR curve** in evaluating classification performance at different decision thresholds.

- enable learners to compare, interpret, and validate models more effectively for practical data analysis applications.

12.2 CONFUSION MATRIX

A confusion matrix is a simple and important table used to evaluate the performance of a classification model. It compares the actual outcomes with the predicted outcomes given by the model. In this way, it helps us clearly understand how many predictions are correct and how many are incorrect. It is called a “confusion” matrix because it shows where the model gets confused between classes. The confusion matrix is especially useful because it provides the foundation for almost all classification performance measures. Instead of giving only one overall number, it presents a complete picture of the model’s prediction behaviour.

Basic Structure of a Confusion Matrix: For a **binary classification problem**, the confusion matrix is generally represented as follows:

Actual / Predicted	<u>Predicted Positive</u>	<u>Predicted Negative</u>
<u>Actual Positive</u>	<i>True Positive (TP)</i>	<i>False Negative (FN)</i>
<u>Actual Negative</u>	<i>False Positive (FP)</i>	<i>True Negative (TN)</i>

A confusion matrix has four main components:

1. *True Positive (TP)*: Cases where the actual class is positive and the model also predicts positive. Example: A diseased patient is correctly predicted as diseased.
2. *True Negative (TN)*: Cases where the actual class is negative and the model also predicts negative. Example: A healthy patient is correctly predicted as healthy.
3. *False Positive (FP)*: Cases where the actual class is negative but the model predicts positive. Example: A healthy patient is wrongly predicted as diseased.
4. *False Negative (FN)*: Cases where the actual class is positive but the model predicts negative. Example: A diseased patient is wrongly predicted as healthy.

The confusion matrix is directly related to **Type-I Error** and **Type-II Error** in statistics:

<u>Error Type</u>	<u>Confusion Matrix Component</u>	<u>Meaning</u>
Type-I Error	False Positive (FP)	Rejecting something that is actually true in negative class terms; predicting positive when actual is negative

<u>Error Type</u>	<u>Confusion Matrix Component</u>	<u>Meaning</u>
Type-II Error	False Negative (FN)	Failing to detect something that is actually positive; predicting negative when actual is positive

Important Note:

- *Type-I Error relates to False Positive (FP):* This happens when the model gives a false alarm. Example: A healthy person is predicted to have a disease.
- *Type-II Error relates to False Negative (FN):* This happens when the model misses a real case. Example: A diseased person is predicted to be healthy.
- In many real-life situations, Type-II error may be more dangerous than Type-I error, especially in medical diagnosis, fraud detection, or security systems.

The Performance measures that can be calculated from the confusion matrix, they are as follows:

1. Accuracy
2. Precision,
3. Recall
4. Sensitivity,
5. Specificity,
6. F1-Score,
7. False Positive Rate (FPR)
8. False Negative Rate (FNR)

A brief introduction to these performance measures is given below:

CONFUSION MATRIX

Actual Class	Predicted Class	
	<i>Positive</i>	<i>Negative</i>
<i>Actual Positive</i>	TP	FN
<i>Actual Negative</i>	FP	TN

1. **Accuracy:** Measures the overall correctness of the model.

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

2. **Precision:** Measures how many predicted positive cases are actually positive.

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

3. **Recall / Sensitivity / True Positive Rate:** Measures how many actual positive cases are correctly identified.

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

4. **Specificity / True Negative Rate:** Measures how many actual negative cases are correctly identified.

$$Specificity = \frac{TN}{TN + FP}$$

5. **F1-Score:** Harmonic mean of precision and recall.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

6. **False Positive Rate (FPR):**

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

7. **False Negative Rate (FNR)**

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

These measures help in understanding different aspects of classification performance.

Example 1: Suppose a hospital develops a machine learning model to predict whether a patient has diabetes or not. Out of **100 patients** tested:

- 40 patients actually have diabetes
- 60 patients are actually healthy

Solution: The model gives the following results:

- 35 diabetic patients are correctly predicted as diabetic i.e. **TP = 35**
- 5 diabetic patients are wrongly predicted as healthy i.e. **FN = 5**
- 10 healthy patients are wrongly predicted as diabetic i.e. **FP = 10**
- 50 healthy patients are correctly predicted as healthy i.e. **TN = 50**

So, the confusion matrix becomes:

<u>CONFUSION MATRIX</u>		
		Predicted Class
		<i>Diabetic</i> <i>Healthy</i>
<i>Actual</i>	<i>Diabetic</i>	TP=35 FN=5
	<i>Healthy</i>	FP=10 TN=50

From this confusion matrix, it can be interpreted that:

- The model correctly identified **35 diabetic patients**
- The model missed **5 diabetic patients**
- The model wrongly alarmed **10 healthy patients**
- The model correctly classified **50 healthy patients**

Now some of the performance measures that can be calculated for the above example are:

Accuracy

$$Accuracy = \frac{35 + 50}{100} = \frac{85}{100} = 0.85 = 85\%$$

Precision

$$Precision = \frac{35}{35 + 10} = \frac{35}{45} = 77.78\%$$

Recall / Sensitivity

$$Recall = \frac{35}{35 + 5} = \frac{35}{40} = 87.5\%$$

Specificity

$$Specificity = \frac{50}{50 + 10} = \frac{50}{60} = 83.33\%$$

This example shows that the confusion matrix gives a much deeper understanding than only saying the model is **85% accurate**.

Thus, we can say that the **confusion matrix** is a basic yet powerful tool for comparing **actual outcomes** with **predicted outcomes** in classification problems. It helps identify correct predictions and different types of errors, especially **Type-I error (False Positive)** and **Type-II error (False Negative)**. It also forms the basis for calculating key performance measures such as **accuracy, precision, recall, sensitivity, specificity, and F1-score**. Therefore, learning the confusion matrix is the first and most important step in understanding classification model evaluation.

Now, we will discuss the performance evaluation metrics in a bit detail in the subsequent sections of this unit. Let's begin with Accuracy Precision & Recall.

☛ Check Your Progress 1

Q1. What is a Confusion Matrix? Why is it important in classification models?

.....

Q2. Explain the four components of a Confusion Matrix.

.....

Q3. What is the difference between Type-I and Type-II errors in the confusion matrix?

.....

Q4. How does a confusion matrix provide better insight than accuracy alone?.

.....

12.3 ACCURACY, PRECISION AND RECALL

In classification problems, a model is not judged only by its ability to generate predictions, but also by how correctly and meaningfully those predictions represent the actual class labels. This is why **accuracy, precision, and recall** are among the most important performance evaluation measures in data analysis. These measures are derived from the confusion matrix and are widely used to understand the quality, reliability, and usefulness of a classification model. Each metric conveys a different aspect of model performance, and together they help in developing a more complete understanding of both the model and the dataset on which it is applied.

We learned about these performance evaluation parameters in the above section; here we are going to discuss them in a bit detail. We begin with Accuracy first and later we will take up other performance evaluation parameters viz. Precision and Recall.

Accuracy is the proportion of total predictions that are correct out of all predictions made by the model. It is one of the simplest and most commonly used evaluation measures because it gives an overall idea of how often the model is right. Mathematically, accuracy is defined as:

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

where TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives.

Accuracy mainly conveys the **overall correctness** of the model. If a model has high accuracy, it means that it is able to classify a large number of observations correctly. This makes accuracy a useful starting point when evaluating a classifier. It is particularly meaningful when the dataset is balanced, that is, when the number of instances in each class is approximately equal, and when the cost of false positives and false negatives is similar.

However, the usefulness of accuracy becomes limited in the case of **imbalanced datasets**. For example, if 95 out of 100 cases belong to one class, a model may achieve 95% accuracy simply by always predicting the majority class, even though it completely fails to identify the minority class. Therefore, while analysing a model or dataset, accuracy must not be interpreted in isolation. It gives a broad picture of performance, but it does not explain the kind of mistakes the model is making.

For instance, suppose a model is used to classify 100 email messages, and it correctly classifies 80 of them. Then the accuracy is:

$$Accuracy = \frac{80}{100} = 0.80 = 80\%$$

This means the model is correct 80% of the time. Although this seems good, it still does not tell us whether the model is correctly identifying the important class, such as spam emails, and whether it is making too many false alarms.

Precision is the proportion of predicted positive cases that are actually positive. It measures the correctness of positive predictions made by the model. Precision is calculated as:

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

Precision conveys the **reliability of positive predictions**. In other words, when the model predicts that an observation belongs to the positive class, precision tells us how often that prediction is truly correct. A high precision means that the model makes fewer false positive errors. Therefore, precision becomes especially important in situations where false positives are costly or undesirable.

For example, in a fraud detection system, low precision would mean that many normal transactions are incorrectly flagged as fraudulent. In email spam filtering, low precision would mean that many genuine emails are wrongly classified as spam. In medical diagnosis, low precision may cause healthy individuals to be incorrectly identified as diseased, creating unnecessary stress, cost, and further testing.

Precision is useful in understanding whether the model is over-predicting the positive class and whether its positive predictions can be trusted. If precision is low, it may suggest that the model is confusing negative cases with positive ones. This may occur due to class overlap, noisy features, poor model design, or an inappropriate decision threshold. Thus, precision helps in analysing the **quality of positive decisions** made by the model.

Consider a disease prediction model that predicts 50 patients as diseased, but among them only 40 actually have the disease and 10 are healthy. Then the precision is:

$$Precision = \frac{40}{40 + 10} = \frac{40}{50} = 0.80 = 80\%$$

This means that when the model predicts disease, it is correct 80% of the time. Therefore, precision focuses not on all predictions, but specifically on how accurate the positive predictions are.

Recall is the proportion of actual positive cases that are correctly identified by the model. It is also called **sensitivity** or the **true positive rate**. The formula for recall is:

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

Recall conveys the **ability of the model to capture actual positive cases**. It tells us how many of the truly positive observations were successfully identified by the classifier. A high recall means that the model is missing very few positive cases, and therefore false negatives are low. Recall is particularly important in situations where missing a positive case may have serious consequences.

For example, in medical diagnosis, low recall means that diseased patients may go undetected. In fraud detection, low recall means that many fraudulent activities remain unnoticed. In security systems, low recall may lead to actual threats being missed. In such

cases, recall becomes more important than precision, because identifying as many real positive cases as possible is the primary concern.

Recall is useful for understanding whether the classifier is too conservative in assigning the positive label. If recall is low, it may indicate that the model is failing to capture many true positives. This may happen because the threshold for predicting the positive class is too strict, the positive class is underrepresented in the data, or the features are not sufficiently informative. Thus, recall provides insight into the **completeness of positive detection**.

For example, suppose there are 60 patients who actually have a disease, but the model correctly identifies only 48 of them while missing 12. Then the recall is:

$$Recall = \frac{48}{48 + 12} = \frac{48}{60} = 0.80 = 80\%$$

This means the model is able to detect 80% of the actual diseased cases. Hence, recall is concerned with how well the model finds all relevant positive instances.

Combined Understanding of Accuracy, Precision and Recall provides a much deeper understanding of model behaviour than any one of them alone. Accuracy gives the overall correctness of the model, precision shows how reliable the positive predictions are, and recall indicates how effectively the model identifies the actual positive cases. But it is to be noted that:

- A model may have high accuracy but low recall if it mostly predicts the majority class correctly while missing many positives.
- Similarly, a model may have high recall but low precision if it predicts many cases as positive in order to capture most of the real positives, thereby increasing false positives.
- A model may also have high precision but low recall if it predicts positive only when it is very certain, thereby missing many actual positive cases.

Therefore, these measures must always be interpreted together and in the context of the specific application. Their importance depends on the type of problem and on whether false positives or false negatives are more serious in that setting.

Example 2: Consider a binary classification model tested on 100 observations with the following confusion matrix values: true positives = 30, true negatives = 50, false positives = 10, and false negatives = 10.

The accuracy of the model is:

$$Accuracy = \frac{30 + 50}{30 + 50 + 10 + 10} = \frac{80}{100} = 80\%$$

The precision of the model is:

$$Precision = \frac{30}{30 + 10} = \frac{30}{40} = 75\%$$

The recall of the model is:

$$Recall = \frac{30}{30 + 10} = \frac{30}{40} = 75\%$$

This example shows that although the model is 80% accurate overall, its positive predictions are correct only 75% of the time, and it is able to identify 75% of the actual positives. This gives a more balanced understanding of performance than accuracy alone.

Now, it's time to Learn about the **Role of Accuracy, precision, and recall in Understanding the Dataset**: Accuracy, precision, and recall are not only useful for evaluating the model, but also for understanding the dataset itself.

For example:

- if accuracy is very high but recall is low, it may indicate that the dataset is imbalanced and the model is biased towards the majority class.
- If precision is low, it may suggest that there is overlap between the classes, noisy observations, or insufficiently informative features.
- If recall is low, it may indicate that the positive class is difficult to detect or that it is under-represented in the data.

Thus, these measures help reveal the structure and challenges of the dataset in addition to describing the model's behaviour.

Another important point is that precision and recall are sensitive to the classification threshold used by the model. A change in threshold may increase recall while reducing precision, or increase precision while reducing recall. This threshold-based behaviour is one of the reasons why graphical evaluation tools such as ROC and PR curves are important.

Later in this unit, we will learn about ROC curve and PR curves in detail

Now, we extend our discussion to understand the **Relation of Performance evaluation metrics Viz. Accuracy, precision, and recall with the Receiver Operating Characteristic (ROC) Curve and Precision–Recall (PR) curve**

- **Relation with ROC curve:** The ROC curve, or Receiver Operating Characteristic curve, is a graphical method used to evaluate the performance of a classification model across different threshold values. It plots the True Positive Rate (TPR) on the vertical axis against the False Positive Rate (FPR) on the horizontal axis.

It is important to note that, sincerecall is the same as the true positive rate, recall is directly connected to the ROCcurve:

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

The false positive rate is defined as:

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

Thus, each point on the ROC curve corresponds to a particular decision threshold and represents a trade-off between detecting actual positives and incorrectly classifying negatives as positives.

Recall is directly represented in the ROC curve, whereas accuracy and precision are not directly plotted. However, since all of these measures come from the same confusion matrix values, they are indirectly related. At each threshold, a different confusion matrix is formed, and from that confusion matrix, accuracy, precision, recall, and false positive rate can all be computed.

ROC analysis is useful for understanding how well the model separates the positive and negative classes overall. A better model achieves higher recall at lower false positive rates.

- **Relation with PR Curve:** The PR curve, or Precision–Recall curve, is another graphical evaluation tool that plots precision on the vertical axis and recall on the horizontal axis for different threshold values. Thus, both precision and recall are directly involved in the construction of the PR curve.

The PR curve shows the trade-off between precision and recall at different threshold values. Usually:

- when recall increases, precision may decrease,
- when precision increases, recall may decrease.

This is because making the model more sensitive to capture more positives often also increases false positives.

Unlike the ROC curve, which uses recall and false positive rate, the PR curve focuses entirely on the trade-off between how many positive cases are captured and how reliable those positive predictions are. This makes the PR curve especially useful when dealing with **imbalanced datasets**, where the positive class is rare and where accuracy may be misleading.

As the threshold changes, precision and recall usually move in opposite directions. Increasing recall often means identifying more actual positives, but it may also increase false positives, thereby reducing precision. On the other hand, increasing precision may require stricter conditions for predicting positive, which may reduce recall. The PR curve provides a visual understanding of this balance.

Accuracy is not directly represented in the PR curve, but it still depends on the same confusion matrix from which precision and recall are computed. Therefore, while accuracy provides a general summary, the PR curve offers a deeper understanding of positive class performance.

Finally, we need to address the Conceptual Relationship Among the Metrics and Curves: The confusion matrix forms the basis for all these measures. From the values of TP, TN, FP, and FN, we calculate accuracy, precision, and recall. The ROC curve uses recall and false positive rate derived from the confusion matrix at different thresholds, while the PR curve uses precision and recall derived from the confusion matrix at different thresholds. In this sense, ROC and PR curves are visual extensions of the same underlying evaluation framework.

Thus, recall plays a direct role in both ROC and PR curve analysis, precision plays a direct role in PR curve analysis, and accuracy remains an important overall measure, though it is not directly plotted in either curve.

Together, these metrics and curves help analysts understand not only whether a model is performing well, but also **how** and **why** it is performing in a particular way.

Following table tabulates the Main focus and the best to use cases for the Performance evaluation metrics and the Curves

<u>Measure/Curve</u>	<u>Main Focus</u>	<u>Best Used When</u>
Accuracy	Overall correctness	Classes are balanced
Precision	Reliability of positive predictions	False positives are costly
Recall	Ability to detect positives	False negatives are costly
ROC Curve	Trade-off between recall and false positive rate	General classification comparison
PR Curve	Trade-off between precision and recall	Imbalanced datasets

This section concludes with the understanding that the Accuracy, precision, and recall are fundamental performance evaluation measures used in classification problems. Accuracy conveys the overall correctness of the model, precision conveys the reliability of positive predictions, and recall conveys the ability of the model to identify actual positive cases. These metrics are highly valuable in understanding both the performance of a model and the characteristics of the dataset, such as class imbalance, overlap, and threshold sensitivity.

They are also closely related to graphical evaluation methods. Recall directly contributes to the ROC curve as the true positive rate, while both precision and recall together form the PR curve. Therefore, a proper understanding of accuracy, precision, and recall provides the conceptual foundation for interpreting ROC and PR curves and for performing meaningful evaluation of predictive classification models.

☛ Check Your Progress 2

Q1. What is Accuracy? When is it most useful in model evaluation?

.....
.....

Q2. Define Precision and explain its importance.

.....
.....

Q3. What is Recall? Why is it important in critical applications?

.....
.....

Q4. Why should Accuracy, Precision, and Recall be interpreted together?

.....
.....

12.4 SENSITIVITY AND SPECIFICITY

The discussion of Accuracy, Precision, and Recall made in the last section, naturally leads to two more important performance evaluation measures: Sensitivity and Specificity. These metrics are especially useful when we want to understand not only how many predictions are correct, but also we want to understand that what kind of errors the model is making.

In many real-life classification problems, such as disease detection, fraud identification, spam filtering, or defect detection in manufacturing, a model must be evaluated from both perspectives: how well it identifies the positive cases and how well it rejects the negative cases. This is where Sensitivity and Specificity become essential.

As discussed earlier, a confusion matrix forms the foundation for these metrics. For a binary classification problem, the confusion matrix contains four outcomes: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Accuracy gives the overall correctness of the model, Precision tells us how many predicted positives are truly positive, and Recall tells us how many actual positives are correctly detected. Sensitivity is closely related to Recall, while Specificity complements it by focusing on the negative class.

In this section we are going to discuss about Sensitivity and Specificity in detail. Let's begin with Sensitivity, and later Specificity.

Sensitivity is the ability of a model to correctly identify actual positive cases. It answers the question: Out of all real positive instances, how many did the model correctly detect? Because of this meaning, Sensitivity is also called the True Positive Rate (TPR) or simply Recall in many machine learning contexts.

The formula for Sensitivity is:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Here, $TP + FN$ represents all actual positive cases. So, Sensitivity measures the proportion of actual positives that are successfully recognized by the model.

A high Sensitivity means the model misses very few positive cases. This is very important in situations where, failing to detect a positive case is costly or dangerous.

For example, in medical diagnosis for a serious disease, if a sick patient is predicted as healthy, it becomes a False Negative, which may delay treatment and create severe consequences. Therefore, screening tests are often designed to have very high Sensitivity.

Whereas, **Specificity** is the ability of a model to correctly identify actual negative cases. It answers the question like: Out of all real negative instances, how many did the model correctly classify as negative? It is also called the True Negative Rate (TNR).

The formula for Specificity is:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Here, $TN + FP$ represents all actual negative cases. So, Specificity measures the proportion of actual negatives that are correctly rejected by the model.

A high Specificity means the model generates very few False Positives. This is important in situations where incorrectly labelling a negative case as positive causes unnecessary cost, anxiety, or intervention.

For instance, if a fraud detection system wrongly flags too many normal banking transactions as fraudulent, customers may face inconvenience and the bank may waste resources in manual verification. In such cases, Specificity becomes highly valuable.

Relationship of Sensitivity and Specificity with Accuracy, Precision, & Recall: Sensitivity and Specificity should be understood as part of the same family of performance measures, the usual purpose is listed below:

- Accuracy shows overall correctness.
- Precision focuses on the reliability of positive predictions.
- Recall measures how many actual positives are detected.
- Sensitivity is effectively the same as Recall for the positive class.
- Specificity measures how well the model identifies actual negatives.

Thus, if Recall tells us how well the model captures the positive class, Specificity tells us how well the model protects the negative class from being incorrectly classified. Together, Sensitivity and Specificity provide a more balanced understanding of classifier behaviour, especially when the dataset is imbalanced.

A model can have high Accuracy but poor Sensitivity if the positive class is rare. For example, if only 5 out of 100 patients have a disease, a model that predicts “healthy” for

everyone will still achieve 95% Accuracy, but its Sensitivity will be 0, because it fails to detect any actual patient. This shows why Sensitivity and Specificity are often more informative than Accuracy alone.

Important Note: Recall and Sensitivity are not different in terms of calculation or meaning. Both represent the model's ability to correctly identify actual positive cases. In other words, both answer the same question: out of all the real positive instances present in the dataset, how many did the model correctly predict as positive.

Mathematically, both Recall and Sensitivity are calculated using the same formula:

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

where **TP** stands for True Positives and **FN** stands for False Negatives. This means both metrics focus only on the actual positive class and measure how effectively the model captures it.

The main difference lies only in the terminology and context of usage. The term **Recall** is more commonly used in machine learning, data science, and information retrieval, while **Sensitivity** is more commonly used in medical diagnosis, healthcare testing, and statistics. For example, in a disease detection problem, a doctor may say that the test has high sensitivity, whereas in a machine learning classification problem, a data scientist may describe the same result as high recall.

Thus, Recall and Sensitivity are essentially the same metric with different names used in different domains. There is no numerical difference between them; only the field of application changes the preferred term.

Example 3: Let us consider a medical screening test for detecting a disease among 100 people. Suppose the confusion matrix is as follows:

- True Positive (TP) = 32
- False Negative (FN) = 8
- True Negative (TN) = 50
- False Positive (FP) = 10

Solution: This means:

- 40 people actually have the disease (32+8)
- 60 people are actually healthy (50+10)

Now let us compute the metrics, Sensitivity

$$\text{Sensitivity} = \frac{TP}{TP + FN} = \frac{32}{32 + 8} = \frac{32}{40} = 0.80$$

So, the Sensitivity is 0.80, or 80%.

The results can be interpreted that Out of all actual diseased patients, the model correctly identifies 80% of them. However, it misses 20% of the truly diseased patients. In a real-life medical setting, this means that the screening test is fairly good at detecting disease, but some patients may still go undiagnosed.

Whereas the computation of Specificity reveals that:

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{50}{50 + 10} = \frac{50}{60} \approx 0.8333$$

So, the Specificity is about 83.33%.

The results can be Interpretated that Out of all healthy people, the model correctly identifies 83.33% as healthy. The remaining 16.67% are wrongly classified as diseased. In a medical context, these are people who may experience unnecessary fear, additional tests, and extra cost even though they are actually healthy.

This numerical example helps us understand how both measures work together.

- Sensitivity = 80% means the test is reasonably effective at finding sick patients.
- Specificity = 83.33% means the test is also reasonably effective at avoiding false alarms among healthy people.

Now suppose this is a cancer screening test. In such a case, doctors may prefer to maximize Sensitivity, because missing a cancer patient is far more dangerous than sending a healthy person for an extra confirmatory test. On the other hand, if this were a fraud detection system, too many false alarms could create customer dissatisfaction. In that case, Specificity would also need to be carefully controlled.

So, the “best” metric depends on the application domain and the cost of the two error types.

Example 4: Consider a fraud detection model is used by a bank and following are the outcomes:

- TP = 18 fraudulent transactions correctly detected
- FN = 2 fraudulent transactions missed
- TN = 160 genuine transactions correctly approved
- FP = 20 genuine transactions wrongly flagged as fraud

Solution: The computation of Sensitivity reveals that:

$$\text{Sensitivity} = \frac{TP}{TP + FN} = \frac{18}{18 + 2} = \frac{18}{20} = 0.90$$

So, Sensitivity = 90%.

The results can be Interpretated that the model catches 90% of fraudulent transactions. This is good, because very few fraud cases are missed.

Whereas, the computation of Specificity reveals that:

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{160}{160 + 20} = \frac{160}{180} \approx 0.8889$$

So, Specificity = 88.89%.

The results can be Interpretated that the model correctly approves almost 89% of genuine transactions, but about 11% of normal customer transactions are wrongly blocked.

In banking, this may be frustrating for customers, so even though high Sensitivity is desirable, the bank must also improve Specificity to reduce inconvenience.

This real-life example shows that Sensitivity protects against missed positives, while Specificity protects against false alarms.

Trade-off between Sensitivity and Specificity can be understood as follows:

In many classification systems, Sensitivity and Specificity have an inverse relationship. When the classification threshold is adjusted to make the model more sensitive, it may classify more borderline cases as positive. This increases Sensitivity but may also increase False Positives, thereby reducing Specificity. Conversely, if the threshold is made stricter, Specificity may improve, but Sensitivity may decrease because more actual positives are missed.

For example, in a medical test:

- A low threshold may detect almost all diseased patients, giving very high Sensitivity, but it may also wrongly label many healthy people as diseased, lowering Specificity.
- A high threshold may reduce false alarms and improve Specificity, but it may fail to detect some real patients, lowering Sensitivity.

Hence, model evaluation often requires a balance between the two rather than maximizing only one.

Relationship of Sensitivity and Specificity with ROC Curve: Sensitivity and Specificity are directly connected to the ROC (Receiver Operating Characteristic) Curve. In fact, the ROC curve is built using these two measures across different classification thresholds.

The ROC curve plots:

- True Positive Rate (TPR) on the Y-axis, which is the same as Sensitivity
- False Positive Rate (FPR) on the X-axis,

Where,

$$FPR = \frac{FP}{FP + TN} = 1 - \text{Specificity}$$

Thus:

- Sensitivity = TPR
- 1 - Specificity = FPR

This means every point on the ROC curve represents a pair of values: one for Sensitivity and one for False Positive Rate. A classifier with better performance has points closer to the top-left corner of the ROC space, because that indicates high Sensitivity and low False Positive Rate, or equivalently, high Specificity.

Thus, an ROC curve can be interpreted in terms of Sensitivity and Specificity, as follows:

- A model near the top-left corner has high Sensitivity and high Specificity.
- A model near the diagonal line performs like random guessing.

- The Area Under the ROC Curve (AUC-ROC) summarizes the model's ability to distinguish between positive and negative classes across all thresholds.

Important Note: So, ROC analysis is essentially a threshold-wise study of the balance between Sensitivity and Specificity.

Relationship of Sensitivity and Specificity with PR Curve: The Precision-Recall (PR) Curve focuses on:

- Precision on the Y-axis
- Recall (Sensitivity) on the X-axis

Thus, Sensitivity is directly part of the PR curve because Recall and Sensitivity are equivalent in binary classification.

However, Specificity does not appear directly in the PR curve. The PR curve is more concerned with how accurately positive predictions are made and how many actual positives are captured. Therefore:

- Sensitivity has a direct relationship with PR curves
- Specificity has an indirect relationship, because if False Positives increase, Precision decreases, and Specificity also decreases

So although Specificity is not explicitly plotted in the PR curve, it still influences the curve through its effect on False Positives.

Practically speaking, the key points from application point of view are as follows:

- In ROC curves, Sensitivity and Specificity are the central concepts.
- In PR curves, Sensitivity remains important, but the counterpart is Precision instead of Specificity.
- For imbalanced datasets, PR curves are often more informative than ROC curves, especially when the positive class is rare. But if we want to understand how well the model distinguishes both positives and negatives, Sensitivity and Specificity through ROC analysis are very useful.

This section concludes with the understanding that Sensitivity and Specificity are powerful performance evaluation measures that deepen our understanding beyond Accuracy, Precision, and Recall.

Sensitivity tells us how effectively the model identifies actual positive cases, while Specificity tells us how effectively it identifies actual negative cases. Together, they reveal the model's behaviour from both sides of the classification task.

In real-life applications, these measures must be interpreted in the context of error costs:

- In healthcare, high Sensitivity is often critical to avoid missing patients.
- In fraud detection, both Sensitivity and Specificity matter because the system must catch fraud while not disturbing genuine users.

- In industrial quality control, Sensitivity helps detect defective products, while Specificity prevents good products from being wrongly discarded.

The connection of Sensitivity and Specificity with ROC curves is very strong, since ROC plots Sensitivity against $(1 - \text{Specificity})$. Their connection with PR curves is also important, especially because Sensitivity appears directly as Recall. Therefore, Sensitivity and Specificity do not stand alone; rather, they form an essential continuation of the broader framework of classification performance evaluation.

Now we are going to discuss the next important performance evaluation metrics i.e. F-1 Score in the next section.

☛ Check Your Progress 3

Q1. What is Sensitivity? Why is it important in classification problems?

.....

Q2. Define Specificity and explain its significance.

.....

Q3. Explain the relationship between Sensitivity and Recall.

.....

Q4. Why is there a trade-off between Sensitivity and Specificity?

.....

12.5 F1-SCORE

After understanding Accuracy, Precision, Recall (Sensitivity), and Specificity, the next important performance evaluation metric is the F1-Score. This metric becomes especially useful when we do not want to judge a model only by overall correctness, but rather by how well it balances the detection of positive cases and the reliability of positive predictions.

In many real-world classification problems, a model may achieve high Accuracy simply because one class dominates the dataset, yet still perform poorly on the class that actually matters.

Similarly, a model may have high Recall but low Precision, or high Precision but low Recall. In such situations, F1-Score provides a single balanced measure that combines both Precision and Recall into one value.

This makes F1-Score a natural continuation of the previous discussion, where:

- Accuracy tells us the overall correctness of the classifier.

- Recall or Sensitivity tells us how many actual positive cases are correctly detected.
- Specificity tells us how well actual negative cases are correctly rejected.
- Precision tells us how trustworthy the positive predictions are.

Based on these performance evaluation metrics the F1-Score is derived to answer a very practical question like: how good is the model at identifying positive cases while also keeping its positive predictions reliable? And this is a valid question in general.

Mathematical the Meaning of F1-Score states that the F1-Score is the harmonic mean of Precision and Recall. It is used because the harmonic mean gives a balanced score only when both values are reasonably high. If one of them is very low, the F1-Score also becomes low. This makes it a better summary measure than simply taking the arithmetic average of Precision and Recall.

The formula for F1-Score is:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Since Recall is the same as Sensitivity in binary classification, the formula can also be written as:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$$

Where,

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

and

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

Thus, F1-Score depends on False Positives as well as False Negatives, because Precision is affected by FP and Recall is affected by FN. This is what makes F1-Score especially important in applications where both kinds of errors matter.

There might be a question in your mind that ***why harmonic mean is used to compute F1-Score?*** The Question is quite valid and the answer is as follows:

The F1-Score uses the harmonic mean, not the arithmetic mean, because the harmonic mean penalizes imbalance between Precision and Recall.

For example, if Precision = 1.0 and Recall = 0.1, Then arithmetic average would be:

$$\frac{1.0 + 0.1}{2} = 0.55$$

This seems moderate.

But F1-Score would be:

$$F1 = 2 \times \frac{1.0 \times 0.1}{1.0 + 0.1} = \frac{0.2}{1.1} \approx 0.1818$$

This value is much lower and more realistic, because the model is clearly poor if it performs well on only one metric and badly on the other. So, the harmonic mean ensures that the F1-Score becomes high only when both Precision and Recall are high.

Now, there might be another question in your mind that when we already have so many metrics for performance evaluation and again F1-score is also derived from those metrics only, so what is the need of F1-Score or Why do we need F1-Score?

The answer is as follows:

Suppose a model predicts many cases as positive. It may detect most actual positives, giving a high Recall, but it may also produce many false alarms, lowering Precision. On the other hand, a model may be very strict and predict positive only when it is highly certain. In that case, Precision may be high, but many true positive cases may be missed, causing low Recall. If we look at only one of these metrics, we may get a misleading impression of model performance.

F1-Score solves this problem by combining both into one number. A high F1-Score means the model has found a good balance between:

- capturing actual positives
- avoiding incorrect positive predictions

That is why F1-Score is widely used in medical diagnosis, spam detection, information retrieval, fraud detection, and defect detection, especially when the positive class is important and class imbalance exists.

Now it's time to understand the Relationship of F1-Score with Accuracy, Precision, Recall, Sensitivity, and Specificity. F1-Score is closely related to the previous evaluation measures:

- Accuracy gives overall correctness, but may hide poor positive-class performance in imbalanced data.
- Precision tells us how many predicted positives are actually correct.
- Recall/Sensitivity tells us how many actual positives are successfully captured.
- Specificity focuses on the negative class.

F1-Score does not directly include Specificity, because it focuses mainly on the quality of positive-class prediction. However, it strongly depends on Precision and Recall, which themselves are affected by the confusion matrix values.

So, if the study objective is to measure how well the classifier handles the positive class, then F1-Score is often more meaningful than Accuracy. This is especially true when false negatives and false positives are both important.

The Interpretation scale for F1-Score, describes that the value of F1-Score lies between 0 and 1, the meaning of these values of F1-Score are as follows:

- $F1 = 1$ means perfect Precision and perfect Recall
- $F1 = 0$ means the model completely fails on either Precision or Recall
- A higher F1-Score indicates a better balance between Precision and Recall

Example 5: *Case of Disease Screening:* Let us continue with an example for medical diagnosis so that the discussion remains connected with Sensitivity and Specificity. Compute F1-Score for the following case and interpret the results.

Suppose a disease screening model gives the following confusion matrix for 100 people:

- True Positive (TP) = 32
- False Positive (FP) = 10
- False Negative (FN) = 8
- True Negative (TN) = 50

Solution:

Step 1: Calculate Precision

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{32}{32 + 10} = \frac{32}{42} \approx 0.7619$$

So, Precision = 76.19%

From the results of Precision, it can be interpreted that Out of all people predicted as diseased, about 76.19% are actually diseased. The remaining predicted positives are false alarms.

Step 2: Calculate Recall / Sensitivity

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{32}{32 + 8} = \frac{32}{40} = 0.80$$

So, Recall or Sensitivity = 80%

From the results of Recall it can be interpreted that Out of all actual diseased people, the model correctly identifies 80%.

Step 3: Calculate F1-Score

$$\begin{aligned} F1 &= 2 \times \frac{0.7619 \times 0.80}{0.7619 + 0.80} \\ F1 &= 2 \times \frac{0.6095}{1.5619} \\ F1 &\approx 0.7805 \end{aligned}$$

So, the F1-Score is approximately 0.78 or 78.05%

From the results of F1-Score it can be interpreted that this value tells us that the disease screening model has a fairly good balance between identifying sick patients and ensuring that predicted positive results are trustworthy. In medical screening, this is important because:

- If Recall is high but Precision is low, too many healthy people may be incorrectly told they are sick.
- If Precision is high but Recall is low, many real patients may go undetected.

An F1-Score of 78.05% indicates that the test is reasonably balanced, but there is still room for improvement. In a real hospital setting, whether this is acceptable depends on the seriousness of the disease. For life-threatening conditions, doctors may tolerate more false positives in order to increase Recall. For confirmatory diagnostic tests, they may want both Precision and Recall to be high, which would increase the F1-Score.

Example 6:Case of Spam Email Detection: Compute F1-Score for the following case and interpret the results.

Consider a spam filter with the following confusion matrix:

- TP = 90 spam emails correctly identified
- FP = 30 normal emails wrongly classified as spam
- FN = 10 spam emails missed
- TN = 170 normal emails correctly classified

Solution:

Compute Precision

$$\text{Precision} = \frac{90}{90 + 30} = \frac{90}{120} = 0.75$$

$$\text{Precision} = 75\%$$

Compute Recall

$$\text{Recall} = \frac{90}{90 + 10} = \frac{90}{100} = 0.90$$

$$\text{Recall} = 90\%$$

Computer F1-Score

$$F1 = 2 \times \frac{0.75 \times 0.90}{0.75 + 0.90}$$

$$F1 = 2 \times \frac{0.675}{1.65}$$

$$F1 \approx 0.8182$$

So, F1-Score = 81.82%

From the results of Precision, Recall and F1-Score it can be interpreted that This spam filter catches most spam emails, because Recall is high at 90%. However, its Precision is only 75%, meaning some genuine emails are wrongly sent to spam. The F1-Score of 81.82%

shows that overall, the model performs well in balancing spam detection with prediction reliability.

This means that the filter is effective, but users may still occasionally lose important emails in the spam folder. If the business goal is aggressive spam blocking, this may be acceptable. If the goal is to avoid misclassifying important client emails, Precision should be improved, which would also improve the F1-Score.

Example 7: Case of Fraud Detection with Imbalanced Data. Compute Accuracy, and F1-Score for the following case and interpret the results. Consider a bank fraud detection system with the following confusion matrix:

- $TP = 18$
- $FP = 20$
- $FN = 2$
- $TN = 160$

Solution: Compute Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{18 + 160}{200} = \frac{178}{200} = 89\%$$

$$\text{Accuracy} = 89\%$$

This seems high Accuracy.

Compute Precision

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{18}{18 + 20} = \frac{18}{38} \approx 0.4737$$

$$\text{Precision} = 47.37\%$$

Compute Recall

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{18}{18 + 2} = \frac{18}{20} = 0.90$$

$$\text{Recall} = 90\%$$

Compute F1-Score

$$F1 = 2 \times \frac{0.4737 \times 0.90}{0.4737 + 0.90}$$

$$F1 = 2 \times \frac{0.4263}{1.3737}$$

$$F1 \approx 0.6207$$

So, F1-Score = 62.07%

This is a very important example. From the results of Accuracy, Precision, Recall and F1-Score it can be interpreted that Accuracy is 89%, which looks excellent. Recall is also very

high at 90%, meaning the model catches most fraudulent transactions. But Precision is poor at only 47.37%, which means that more than half of the transactions flagged as fraud are actually genuine.

The F1-Score of 62.07% reveals the true picture much better than Accuracy. It shows that although fraud detection is good, the model creates too many false alarms. In a real banking system, this would lead to customer frustration, transaction blocking, and increased manual verification workload. Thus, the F1-Score provides a more realistic evaluation than Accuracy alone.

Now it's time to discuss the **Relationship of F1-Score with ROC Curve**: The relationship of F1-Score with the ROC curve is indirect, because the ROC curve plots:

- True Positive Rate (TPR) = Recall = Sensitivity on the Y-axis
- False Positive Rate (FPR) = $1 - \text{Specificity}$ on the X-axis

Whereas, the F1-Score depends on Recall and Precision, and ROC depends on Recall and FPR, the two are related but not identical.

Key point to be noted is that a model with a good ROC performance may not always have the best F1-Score, especially in imbalanced datasets. This happens because ROC considers the trade-off between Sensitivity and False Positive Rate, but does not directly account for Precision. Since Precision depends heavily on the number of False Positives and the prevalence of the positive class, F1-Score may give a different impression.

Practically this means that:

- ROC is useful for studying classifier performance across thresholds.
- F1-Score is useful when you want one threshold-specific measure that balances Precision and Recall.
- In imbalanced datasets, F1-Score is often more informative than ROC-AUC for the positive class.

Thus, F1-Score is not directly plotted on the ROC curve, but it is still influenced by threshold changes that affect Precision and Recall.

Relationship of F1-Score with PR Curve: The relationship between F1-Score and the Precision-Recall (PR) Curve is much stronger, because the PR curve plots:

- Recall on the X-axis
- Precision on the Y-axis

Since, F1-Score is calculated directly from Precision and Recall, every point on the PR curve corresponds to a possible F1-Score at a given threshold.

The important insight from PR curve is that the F1-Score can be seen as a summary of a point on the PR curve. A better point on the PR curve, where both Precision and Recall are high, usually gives a higher F1-Score. In fact, F1-Score is often used to choose the best

classification threshold from the PR curve by selecting the threshold where the balance between Precision and Recall is strongest.

Practically this means that:

- PR curves help visualize trade-offs between Precision and Recall across thresholds.
- F1-Score helps summarize this trade-off at one chosen threshold.
- For imbalanced datasets, the PR curve and F1-Score together are often more meaningful than ROC analysis.

So, among all curve-based performance tools, F1-Score is most closely tied to the PR curve.

Finally, we learned that apart from other metrics the F1-Score is also an important performance evaluation metrics, but we should know when to use it i.e. When to use F1-Score? So, it is to be noted that F1-Score is especially appropriate when:

- the dataset is imbalanced
- the positive class is important
- both False Positives and False Negatives matter
- we need a single measure combining Precision and Recall

The analysis of F1-score is widely required in the case like disease diagnosis, fraud detection, spam classification, document retrieval, defect detection in manufacturing, intrusion detection in cybersecurity

We learned that F1-Score is very useful but there are some **Limitations of F1-Score** and they are listed below:

- It ignores True Negatives, so it does not reflect Specificity.
- It focuses mainly on the positive class.
- It may not be enough when negative-class performance is equally important.
- In such cases, it should be used along with Accuracy, Specificity, and ROC analysis.

So, F1-Score should not replace all other metrics. It should be used as part of a broader evaluation framework.

This section concludes with the understanding that F1-Score is an important performance evaluation metric that naturally extends the earlier discussion on Accuracy, Precision, Recall, Sensitivity, and Specificity. While Accuracy gives overall correctness and Recall/Sensitivity and Specificity explain class-wise detection ability, F1-Score provides a balanced single measure of how well the classifier handles the positive class by combining Precision and Recall.

Its main strength lies in situations where class imbalance exists and where both false alarms and missed positive cases are important. Through real-life examples such as medical diagnosis, spam filtering, and fraud detection, it becomes clear that F1-Score often gives a more realistic view of performance than Accuracy alone.

Its connection with the ROC curve is indirect, because ROC focuses on Recall and False Positive Rate. Its connection with the PR curve is direct, because F1-Score is derived from Precision and Recall, the two axes of the PR plot itself. Therefore, F1-Score is one of the most practical and meaningful metrics for evaluating classification models, especially when the positive class is of central interest.

☛ Check Your Progress 4

Q1. What is F1-Score? Why is it used in model evaluation?

.....
.....

Q2. Write the formula of F1-Score and explain its components.

.....
.....

Q3. Why is F1-Score preferred over Accuracy in some cases?

.....
.....

Q4. What does a high F1-Score indicate about a model?

.....
.....

12.6 ROC CURVE

After understanding Accuracy, Precision, Recall (Sensitivity), Specificity, and F1-Score, the next important performance evaluation tool is the ROC Curve, that is, the Receiver Operating Characteristic Curve. The ROC curve is not just a single numerical metric like Accuracy or F1-Score; rather, it is a graphical performance evaluation tool that helps us study how a classification model behaves at different decision thresholds. This makes it a natural continuation of the previous discussion. Accuracy, Precision, Recall, Sensitivity, Specificity, and F1-Score evaluate model performance at a chosen threshold, whereas the ROC curve shows how the classifier performs across all possible thresholds.

This is especially important because many classification models do not directly output final class labels such as “positive” or “negative.” Instead, they often output a score or probability. A threshold is then applied to convert that score into a predicted class. For example, if a disease prediction model gives a patient a probability of 0.72 of having a disease, a threshold such as 0.50 may classify the patient as diseased. But if the threshold is changed to 0.80, the same patient may be classified as healthy. Thus, model performance changes with the threshold, and the ROC curve helps us visualize this behaviour.

Basically, ROC Curve is a graph that plots:

- **True Positive Rate (TPR)** on the Y-axis
- **False Positive Rate (FPR)** on the X-axis

We already learned that the **True Positive Rate** is the same as **Recall** or **Sensitivity**:

$$TPR = \text{Sensitivity} = \text{Recall} = \frac{TP}{TP + FN}$$

And, the **False Positive Rate** is defined as:

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

Since Specificity is:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

we can also write:

$$FPR = 1 - \text{Specificity}$$

Thus, the ROC curve directly connects with the previously discussed metrics i.e. Sensitivity and Specificity where:

- **Sensitivity / Recall** appears on the Y-axis
- **Specificity** appears indirectly through $(1 - \text{Specificity})$ on the X-axis

So, the ROC curve is fundamentally a graphical representation of the trade-off between **Sensitivity** and **Specificity** across different threshold values.

Now, there might be a question in your mind that when we already have so many metrics for performance evaluation, why do we need ROC Curve for performance evaluation?

The answer is as follows:

A classifier rarely performs equally well at all thresholds. If we choose a very low threshold, the classifier may label many cases as positive. This increases Sensitivity because fewer actual positives are missed, but it may also increase False Positives, thereby reducing Specificity. On the other hand, if we choose a very high threshold, the classifier becomes stricter, reducing False Positives and improving Specificity, but it may miss more actual positive cases, lowering Sensitivity.

Therefore, instead of checking only one threshold, the ROC curve helps us evaluate the model over the entire range of threshold values. This gives a broader understanding of how well the model can separate positive and negative classes.

Now, let's understand the **Construction of ROC Curve**: To construct an ROC curve, we take the predicted probabilities or scores from the classifier and apply different thresholds, such as 0.9, 0.8, 0.7, 0.6, and so on. For each threshold, we calculate:

- True Positive Rate (Sensitivity / Recall)

- False Positive Rate (1–Specificity)

These pairs of values are then plotted as points on a graph. Joining these points gives the ROC curve.

It is Important to note for ROC points that:

- **(0, 0)** means the classifier predicts everything as negative
- **(1, 1)** means the classifier predicts everything as positive
- The ideal classifier is near **(0, 1)**, where False Positive Rate is 0 and True Positive Rate is 1
- A random classifier tends to lie near the diagonal line from **(0,0)** to **(1,1)**

Thus, the closer the ROC curve is to the **top-left corner**, the better the classifier is at distinguishing between the two classes.

Example 8: Let us consider a **disease detection model** for 10 patients. Suppose the actual class and predicted probability of disease are:

<u>Patient</u>	<u>Actual Class</u>	<u>Predicted Probability</u>
P1	1	0.95
P2	1	0.90
P3	1	0.80
P4	1	0.60
P5	0	0.70
P6	0	0.55
P7	0	0.40
P8	0	0.30
P9	0	0.20
P10	0	0.10

Solution: Here Actual Class can be coded as, **1 = diseased, 0 = healthy**. Referring to the table above it can be understood that there are 4 actual positive cases and 6 actual negative cases. Now, let us compute TPR and FPR at different thresholds.

Threshold > 0.90: No instance predicted positive. TP = 0, FP = 0, FN = 3, TN = 6;

Therefore, $TPR = \frac{0}{3} = 0$ and $FPR = \frac{0}{6} = 0$.

Hence, ROC Point = (0, 0)

Threshold = 0.90: The Predicted positive are patient: P1, P2 with TP=2; FN=2; FP=0; TN=6

$$\text{Therefore } TPR = \frac{2}{4} = 0.5 \text{ and } FPR = \frac{0}{6} = 0$$

Hence, Point on ROC curve: **(0, 0.5)**

Threshold = 0.70: The Predicted positive patients are: P1, P2, P3, P5 with
TP=3; FN=1; FP=1 & TN = 5.

$$\text{Therefore } TPR = \frac{3}{4} = 0.75 \text{ and } FPR = \frac{1}{6} \approx 0.167.$$

Hence, Point on ROC curve: **(0.167, 0.75)**

Threshold = 0.50: The Predicted positive patients are: P1, P2, P3, P4, P5, P6 with
TP = 4; FN = 0; FP = 2; TN = 4

$$\text{Therefore, } TPR = \frac{4}{4} = 1 \text{ and } FPR = \frac{2}{6} \approx 0.333$$

Hence, Point on ROC curve: **(0.333, 1)**

Threshold = 0.30: The Predicted positive patients are: P1, P2, P3, P4, P5, P6, P7, P8 with
TP = 4; FN = 0; FP = 4; TN = 2

$$\text{Therefore, } TPR = \frac{4}{4} = 1 \text{ and } FPR = \frac{4}{6} \approx 0.667$$

Hence, Point on ROC curve: **(0.667, 1)**

Threshold = 0.10: The Predicted positive patients are: all patients, with
TP = 4; FN = 0; FP = 6; TN = 0.

$$\text{Therefore, } TPR = \frac{4}{4} = 1 \text{ and } FPR = \frac{6}{6} = 1.$$

Hence, Point on ROC curve: **(1, 1)**

It can be interpreted from the above results, that as the threshold decreases the **Sensitivity / Recall increases, the False Positive Rate also increases but the Specificity decreases**. This shows the core idea of the ROC curve, that there is a trade-off between detecting more actual positives and avoiding false alarms.

Suppose this model is used for **cancer screening then the ROC curve can be interpreted as:**

- At a **high threshold** like 0.90, the test is very strict. It identifies only the most obvious positive cases. This leads to **no false alarms**, but it misses some actual patients. In real life, this means fewer healthy people are disturbed, but some cancer cases remain undetected.
- At a **moderate threshold** like 0.50, the model detects all cancer patients, which is excellent in terms of Sensitivity. However, some healthy people are wrongly identified as diseased, leading to additional medical tests.
- At a **very low threshold** like 0.10, the model classifies everyone as diseased. This ensures no patient is missed, but it creates too many false positives, which is impractical.

Thus, the ROC curve helps doctors or analysts choose a threshold depending on the seriousness of false negatives and false positives.

After understanding the Construction of ROC Curve and interpretation of the results at various thresholds, it is required the utility and concept of **Area Under the ROC Curve (AUC-ROC)**.

The ROC curve is often summarized by a single value called the **Area Under the Curve (AUC)**. The AUC gives the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. So, the higher the AUC, the better the model is at separating the two classes.

The **Area Under the ROC Curve** is known as **(AUC-ROC)**, where the meaning of AUC values can be understood as follows:

- AUC = 1.0 means perfect classification
- AUC = 0.5 means the classifier performs like random guessing
- AUC < 0.5 means worse than random, suggesting very poor discrimination

General interpretation scale for the values of AUC is as follows:

- 0.90 to 1.00 = excellent
- 0.80 to 0.90 = very good
- 0.70 to 0.80 = acceptable
- 0.60 to 0.70 = weak
- 0.50 to 0.60 = poor

These ranges are not absolute, but they provide a practical understanding.

Example 9: Calculate the **Area Under the ROC Curve** is known as **(AUC-ROC)**, for the Example 8 given above. The ROC points are determined for respective thresholds, now List ROC points in order. So, the calculated ROC points are **(0, 0) ; (0, 0.5) ; (0.167, 0.75) ; (0.333, 1) ; (0.667, 1) ; (1, 1)**

Solution: Now, Compute AUC using trapezoidal rule: The Area Under the ROC Curve is computed as the sum of trapezoid areas:

$$\text{Area} = \frac{(y_1 + y_2)}{2} (x_2 - x_1)$$

For the given ROC points: **(0, 0) ; (0, 0.5) ; (0.167, 0.75) ; (0.333, 1) ; (0.667, 1) ; (1, 1)**

the **Area Under the ROC Curve (AUC)** is computed using the trapezoidal formula:

$$\text{Area} = \frac{(y_1 + y_2)}{2} (x_2 - x_1)$$

We now compute the area **segment wise**.

Segment 1: From (0, 0) to (0, 0.5) : Here, $x_1 = 0$; $y_1 = 0$ and $x_2 = 0$, $y_2 = 0.5$. Therefore,

$$\text{Area}_1 = \frac{(0 + 0.5)}{2} (0 - 0) = \text{Area}_1 = \frac{0.5}{2} \times 0 = 0$$

Segment 2: From (0, 0.5) to (0.167, 0.75): Here, $x_1 = 0$; $y_1 = 0.5$ and $x_2 = 0.167$, $y_2 = 0.75$. Therefore, $\text{Area}_2 = \frac{(0.5+0.75)}{2} (0.167 - 0) = \frac{1.25}{2} \times 0.167 = 0.625 \times 0.167 = 0.104375$

Segment 3: From (0.167, 0.75) to (0.333, 1): Here, $x_1 = 0.167$; $y_1 = 0.75$ and $x_2 = 0.333$, $y_2 = 1$.
Therefore, $\text{Area}_3 = \frac{(0.75+1)}{2} (0.333 - 0.167) = \frac{1.75}{2} \times 0.166 = 0.875 \times 0.166 = 0.14525$

Segment 4: From (0.333, 1) to (0.667, 1): Here, $x_1 = 0.333$; $y_1 = 1$ and $x_2 = 0.667$, $y_2 = 1$. Therefore, $\text{Area}_4 = \frac{(1+1)}{2} (0.667 - 0.333) = \frac{2}{2} \times 0.334 = 1 \times 0.334 = 0.334$

Segment 5: From (0.667, 1) to (1, 1): Here, $x_1 = 0.667$; $y_1 = 1$ and $x_2 = 1$, $y_2 = 1$.
Therefore, $\text{Area}_5 = \frac{(1+1)}{2} (1 - 0.667) = \frac{2}{2} \times 0.333 = 1 \times 0.333 = 0.333$

Thus, the Total Area Under the Curve (AUC) can be obtained by adding all segment-wise areas, Giving: $\text{AUC} = 0 + 0.104375 + 0.14525 + 0.334 + 0.333 = 0.916625$
So, the final result is: $\text{AUC} \approx 0.917$

Based on the General interpretation scale for the values of AUC given above it can be interpreted that model performs excellent classification because $\text{AUC} \approx 0.917$ which lies between the range 0.90 to 1.00 that means excellent classification.

So, we learned how to determine AUC. Now, It is to be noted that *the ROC curve should not be seen as separate from Accuracy, Precision, Recall, Sensitivity, Specificity, or F1-Score. Rather, it extends them.* Let's understand the *Link of ROC curve with Accuracy, Precision, Recall / Sensitivity, Specificity and F1-Score*

- **ROC Link with Accuracy:** Accuracy gives a single summary of correct predictions at one threshold. However, if the threshold changes, Accuracy may also change. ROC analysis goes beyond a single threshold and evaluates the classifier more comprehensively.
- **ROC Link with Recall / Sensitivity:** Recall or Sensitivity is directly the Y-axis of the ROC curve. So, every ROC point tells us how well the model detects the positive class at that threshold.
- **ROC Link with Specificity:** Specificity is directly connected to the X-axis, because:

$$FPR = 1 - \text{Specificity}$$

Hence, moving right on the ROC graph means losing Specificity.

- **ROC Link with Precision:** Precision is not directly shown on the ROC curve. This is an important limitation. A classifier may have a good ROC curve and still have poor Precision if the positive class is rare.
- **ROC Link with F1-Score:** F1-Score combines Precision and Recall at a particular threshold. ROC, on the other hand, studies Recall and False Positive Rate across multiple thresholds. So, while F1-Score is a threshold-specific summary, ROC is a threshold-varying performance curve.

To better *understand the relevance of ROC curve let's understand a Real-life Example: Fraud Detection Case* - Consider a bank fraud detection system. The model assigns a fraud probability to each transaction.

- If the threshold is set very low, almost every suspicious transaction is flagged. This increases Sensitivity, so fraud cases are rarely missed. But it also increases False Positives, meaning many genuine customer transactions are blocked.
- If the threshold is set very high, fewer genuine customers are disturbed, so Specificity improves. However, some fraud cases may go undetected.

The ROC curve allows the bank to choose a threshold depending on business priorities. If preventing fraud losses is the top priority, a threshold with high Sensitivity may be selected. If customer convenience is equally important, a more balanced point on the ROC curve may be preferred.

This example illustrates the practical importance of the ROC curves in Real-life.

Now we are in the position to understand the ***Relationship between ROC Curve and PR Curve***: The **ROC curve** and the **PR curve** are both graphical tools for evaluating classification models, but they focus on different aspects.

- The **ROC Curve plots** X-axis as False Positive Rate and Y-axis as True Positive Rate / Recall / Sensitivity. Whereas,
- The **PR Curve plots** X-axis as Recall / Sensitivity and Y-axis as Precision

Thus, the ROC curve emphasizes the trade-off between **Sensitivity and Specificity**, while the PR curve emphasizes the trade-off between **Precision and Recall**.

Further we need to understand that, **When ROC is more useful?** the ROC curves are useful under following use cases, i.e. when:

- both positive and negative classes are important
- the dataset is reasonably balanced

- we want to understand overall class separation ability

AND

The PR curve is more useful under following use cases, i.e. when:

- the dataset is **imbalanced**
- the positive class is rare
- we care strongly about the quality of positive predictions.

In highly imbalanced problems, ROC curves can sometimes look overly optimistic because even a moderate number of false positives may correspond to a low False Positive Rate when negatives are very large. In such cases, the PR curve gives a clearer picture because Precision directly penalizes false positives.

Relationship of ROC Curve with PR Curve through previous metrics becomes easier to understand in continuity with the earlier sections, where we observed that:

- Recall / Sensitivity appears in both ROC and PR curves
- Specificity is represented only in ROC through $(1 - \text{Specificity})$
- Precision is represented only in the PR curve
- F1-Score is closely related to the PR curve because it is the harmonic mean of Precision and Recall
- ROC is therefore more connected with Sensitivity and Specificity
- PR is more connected with Precision, Recall, and F1-Score

So, if the discussion of Sensitivity and Specificity leads to ROC, then the discussion of Precision, Recall, and F1-Score leads to the PR curve.

The Advantages of ROC Curve are as follows:

- it evaluates performance across all thresholds
- it shows the trade-off between Sensitivity and Specificity
- it helps compare multiple classifiers visually
- AUC provides a single summary of discrimination ability
- it is easy to interpret in diagnostic and binary classification problems

Limitations of ROC Curve are as follows:

- it does not directly include Precision
- it may appear optimistic in highly imbalanced datasets
- it does not directly tell the best threshold unless application needs are considered
- two models with similar AUC may behave differently at a threshold important for the problem

Therefore, ROC should be interpreted along with other metrics such as Precision, Recall, Specificity, F1-Score, and sometimes the PR curve.

This section on the ROC curve concludes with the remark that ROC curve is an important performance evaluation tool that extends the earlier ideas of Accuracy, Precision, Recall,

Sensitivity, Specificity, and F1-Score. While those metrics evaluate a model at a fixed threshold, the ROC curve studies model behaviour across all thresholds. It plots Sensitivity (i.e. Recall) against False Positive Rate *i.e.* (1–Specificity), thereby visually capturing the balance between identifying actual positives and avoiding false alarms.

Through threshold-based numerical examples and real-life interpretation in disease screening and fraud detection, it becomes clear that the ROC curve is highly useful when we need to choose a suitable threshold and compare classifiers. Its summary measure, AUC-ROC, indicates the model’s overall ability to separate positive and negative classes.

Its relationship with the PR curve is also important. ROC is closely tied to Sensitivity and Specificity, while PR is closely tied to Precision, Recall, and F1-Score. Therefore, ROC and PR curves should be seen as complementary tools rather than competing ones. Together with the previously discussed metrics, the ROC curve provides a complete and meaningful framework for evaluating classification models.

☛ Check Your Progress 5

Q1. What is an ROC Curve? Explain its purpose in model evaluation.

.....
.....

Q2. What are the axes of the ROC Curve?

.....
.....

Q3. What does the Area Under the ROC Curve (AUC) indicate?

.....
.....

Q4. How is ROC Curve useful in model comparison?

.....
.....

12.7 PR CURVE

After understanding Accuracy, Precision, Recall, Sensitivity, Specificity, F1-Score, and the ROC Curve, the next important graphical performance evaluation tool is the PR Curve, that is, the Precision-Recall Curve. The PR curve is especially useful in classification problems where the positive class is more important than the negative class, or where the dataset is imbalanced, meaning one class occurs much less frequently than the other. Thus, the PR curve naturally continues the earlier discussion. While the ROC curve emphasizes the relationship between Sensitivity and Specificity, the PR curve shifts the focus toward Precision and Recall, and therefore connects more directly with F1-Score and the practical usefulness of positive predictions.

In many real-life applications, it is not enough for a model to simply separate positive and negative classes overall. We often want to know two things very clearly: first, how many actual positive cases the model is able to find, and second, how many of the predicted positive cases are truly correct. These two concerns are captured by Recall and Precision, respectively. The PR curve studies the trade-off between these two measures across different threshold values.

We learned from our earlier discussions that the PR Curve is a graph in which, Recall is plotted on the X-axis and Precision is plotted on the Y-axis. Where, Recall is given by:

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

and Precision is given by:

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

Since Recall is the same as Sensitivity, the PR curve also indirectly continues the earlier section on Sensitivity and Specificity. However, unlike the ROC curve, the PR curve does not use Specificity or False Positive Rate directly. Instead, it focuses on how Recall changes along with Precision when the classification threshold is varied.

This means the PR curve tells us that as the model tries to identify more positive cases, does it maintain the reliability of its positive predictions, or does Precision start falling because too many false positives are introduced?

Even if so many performance evaluation parameters are studied, we need to understand the relevance/need of the PR curves i.e. **Why do we need PR curves?**

The answer is as follows:

A single value of Precision, Recall, or F1-Score tells us model performance only at one chosen threshold. But many classification models produce probabilities or confidence scores. If we change the threshold, both Precision and Recall will also change.

- At a high threshold, the model predicts positive only when it is very confident. This usually leads to high Precision but lower Recall, because some actual positives are missed.
- At a low threshold, the model predicts more cases as positive. This usually increases Recall, but Precision may decrease because more false positives are included.

Therefore, instead of looking at only one threshold, the PR curve helps us understand this trade-off across many thresholds. This makes it especially useful when the aim is to select a threshold that balances the ability to find positives with the reliability of positive predictions.

In Continuum with previous performance measures, the PR curve should be understood in continuity with the previously discussed metrics, and we need to understand the Link of PR curves with Accuracy, Precision, Recall, F1-Score, Specificity etc.

- **Link of PR curve with Accuracy:** Accuracy measures overall correctness, but it may be misleading in imbalanced datasets. For example, if 95 out of 100 cases are negative, a model that predicts all cases as negative will still achieve 95% Accuracy, even though it completely fails to identify the positive class. The PR curve avoids this problem by focusing only on the positive class performance through Precision and Recall.
- **Link of PR curve with Precision and Recall:** The PR curve is directly built from these two measures. Every point on the PR curve corresponds to a pair of Precision and Recall values at a certain threshold.
- **Link of PR curve with Sensitivity:** Recall and Sensitivity are mathematically the same:

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

Thus, the Recall axis of the PR curve is also the Sensitivity axis.

- **Link of PR curve with Specificity:** Specificity is not directly represented in the PR curve. This is an important difference from the ROC curve. While ROC emphasizes the balance between Sensitivity and Specificity, the PR curve emphasizes the balance between Precision and Recall.
- **Link of PR curve with F1-Score:** The connection between the PR curve and F1-Score is very strong. Since F1-Score is the harmonic mean of Precision and Recall,

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

every point on the PR curve can be associated with an F1-Score. A point with both high Precision and high Recall generally gives a high F1-Score. Therefore, the PR curve is often used to identify the threshold where the classifier achieves the best balance between Precision and Recall.

- **Link of PR curve with ROC Curve:** The ROC curve plots Recall/Sensitivity against False Positive Rate, whereas the PR curve plots Recall against Precision.

Thus:

1. ROC is more connected with Sensitivity and Specificity
2. PR is more connected with Precision, Recall, and F1-Score

In balanced datasets, ROC and PR curves may both provide useful insight. But in highly imbalanced datasets, the PR curve is often more informative because Precision directly reflects the impact of false positives.

In the above section we learned about the Construction of ROC curve and interpreted the results, here we are going to extend our discussion by learning about the Construction of PR Curve. To draw a PR curve, we take the predicted probabilities or scores produced by the classifier and apply different thresholds. For each threshold, we calculate Recall and Precision. Finally, the The resulting pairs are plotted on a graph, and the joined points form the PR curve.

A good PR curve stays near the top-right region, because:

- high Recall means most actual positives are being found
- high Precision means most predicted positives are actually correct

Important Note: If the curve drops sharply downward as Recall increases, it means the model is finding more positives only by sacrificing Precision, that is, by generating many false positives.

Let us consider the same disease-detection style example to maintain continuity.

Example 10: Suppose a classifier gives the following predicted probabilities:

Patient	Actual Class	Predicted Probability
P1	1	0.95
P2	1	0.90
P3	1	0.80
P4	1	0.60
P5	0	0.70
P6	0	0.55
P7	0	0.40
P8	0	0.30
P9	0	0.20
P10	0	0.10

From the above tabular data, it can be observed that:

- Total actual positives = 4
- Total actual negatives = 6

Compute Precision and Recall at different thresholds.

Solution:

Threshold = 0.90: Predicted positive Patients are: P1, P2. With TP = 2; FP = 0; FN = 2

Therefore, Recall = $\frac{2}{4} = 0.5$ and Precision = $\frac{2}{2} = 1.0$. Thus, PR point = (0.5, 1.0)

Threshold = 0.70: Predicted positive Patients are: P1, P2, P3, P5. With TP = 3; FP = 1; FN = 1

Therefore, Recall = $\frac{3}{4} = 0.75$ and Precision = $\frac{3}{4} = 0.75$. Thus, PR point = (0.75, 0.75)

Threshold = 0.50: Predicted positive Patients are: P1, P2, P3, P4, P5, P6 with TP=4, FP=2, FN=0

Therefore, Recall = $\frac{4}{4} = 1$ and Precision = $\frac{4}{6} \approx 0.667$. Thus, PR point = (1.0, 0.667)

Threshold = 0.30: Predicted positive Patients are: P1, P2, P3, P4, P5, P6, P7, P8 with

TP = 4; FP = 4; FN = 0. Therefore, Recall = $\frac{4}{4} = 1$ and Precision = $\frac{4}{8} = 0.5$. Thus,

PR point = (1.0, 0.5)

Threshold = 0.10: Predicted positive Patients are: all patients. With TP = 4; FP = 6; FN = 0.

Therefore, Recall = $\frac{4}{4} = 1$ and Precision = $\frac{4}{10} = 0.4$. Thus, PR point = (1.0, 0.4)

From these results, we can see the central behaviour of the PR curve, and it is observed that:

- At a high threshold, Precision is very high because only the most confident positive cases are selected. But Recall is lower because some actual positives are missed.
- As the threshold decreases, Recall improves because more actual positives are captured.
- However, Precision starts to fall because the model also begins to include more false positives.

This is the exact trade-off shown by the PR curve.

Real-life Example: Disease Detection Case - Suppose this model is used to detect a serious disease, then it can be observed that:

- At a high threshold, the model is very strict. If it predicts disease, it is usually correct, so Precision is high. But it may miss some patients, so Recall is lower.
- At a lower threshold, the model identifies more patients who truly have the disease, so Recall becomes high. But some healthy individuals are also wrongly labeled as diseased, lowering Precision.

In practical terms, the doctor or analyst must decide what is more important:

- avoiding missed disease cases, which means prioritizing Recall

OR

- avoiding unnecessary panic and extra tests for healthy people, which means prioritizing Precision

The PR curve helps in choosing this balance.

Now, let's understand the **Mapping of PR Curve with F1-Score** i.e. *How to connect the above PR points with F1-Score?*

At threshold = 0.90: Precision = 1.0, Recall = 0.5. So, $F1 = 2 \times \frac{1.0 \times 0.5}{1.0 + 0.5} = \frac{1}{1.5} \approx 0.667$

At threshold = 0.70: Precision = 0.75, Recall = 0.75. So, $F1 = 2 \times \frac{0.75 \times 0.75}{0.75 + 0.75} = \frac{1.125}{1.5} = 0.75$

At threshold = 0.50: Precision = 0.667, Recall = 1.0. So, $F1 = 2 \times \frac{0.667 \times 1.0}{0.667 + 1.0} \approx 0.80$

It can be interpreted from the results that the PR point with a stronger balance of Precision and Recall gives a better F1-Score. So, the PR curve visually supports threshold selection, while the F1-Score gives a numerical summary at that chosen threshold.

Real-life Example: Spam Detection Case - Consider an email spam filter.

- If the spam filter uses a high threshold, then only emails that look very strongly suspicious are marked as spam. This gives high Precision, meaning most emails classified as spam are truly spam. But some spam mails may still remain in the inbox, causing lower Recall.
- If the threshold is lowered, more spam emails are captured, increasing Recall. However, some genuine emails may also get moved to spam, lowering Precision.

In such applications, the PR curve is highly useful because users care deeply about the quality of positive predictions, that is, whether emails marked as spam are really spam. Thus, PR analysis often becomes more meaningful than ROC analysis in such contexts.

Now, we need to address the Question, *Why PR Curve is more informative in imbalanced datasets?*

The answer is as follows:

The PR curve is particularly important in imbalanced datasets. Consider a case of fraud detection problem with 10,000 transactions, of which only 100 are fraudulent. A classifier might have a good ROC curve because even several hundred false positives may still produce a relatively low False Positive Rate when divided by the huge number of negatives. But from a business perspective, those false positives can be highly problematic, because many genuine transactions may be blocked.

The PR curve reveals this more clearly because Precision directly decreases when false positives rise. Thus, in rare-event detection problems such as fraud detection, disease detection, cyberattack detection, and fault detection, the PR curve often gives a more realistic view of classifier usefulness.

Shape and interpretation of PR Curve, reveals that A good PR curve remains high across a large range of Recall values, with following observations:

- A curve near the top-right region indicates strong performance
- A curve that falls sharply indicates that Recall is being increased at the cost of much lower Precision
- A higher curve is generally better than a lower one
- Unlike ROC, there is no universal diagonal baseline with exactly the same visual meaning.
- The baseline for a PR curve depends on the proportion of positive cases in the dataset. If the positive class is rare, the baseline Precision is low. This again shows why PR curves are sensitive to class imbalance.

In context of the Average Precision and area under PR curve it is important to understand that just as the ROC curve can be summarized using AUC-ROC, the PR curve can be summarized by the area under the PR curve, often called Average Precision (AP) in practical machine learning. A higher area means the classifier maintains strong Precision across a wide range of Recall values.

However, in academic explanation, it is often enough to state that the PR curve provides a threshold-wise graphical view of Precision and Recall, and that higher area indicates better positive-class performance.

Advantages of PR Curve are as follows:

- it directly focuses on the positive class
- it is more informative for imbalanced datasets
- it connects naturally with Precision, Recall, and F1-Score
- it helps choose thresholds when false positives matter significantly
- it gives a realistic picture in rare-event detection problems

Limitations of PR Curve are as follows:

- it does not directly show Specificity
- it focuses mainly on positive-class performance
- it may be less intuitive than ROC for balanced datasets
- its baseline changes with class distribution, so comparisons across datasets must be done carefully

Therefore, like ROC, it should not be used in isolation. It should be interpreted along with Accuracy, Precision, Recall, Specificity, F1-Score, and application context.

This section concludes with the understanding that the PR curve is an important performance evaluation tool that naturally extends the earlier discussion of Accuracy, Precision, Recall, Sensitivity, Specificity, F1-Score, and the ROC curve. While the ROC curve emphasizes the

trade-off between Sensitivity and Specificity, the PR curve emphasizes the trade-off between Precision and Recall. This makes it especially valuable when the positive class is rare or when the correctness of positive predictions is of primary importance.

Through numerical examples and real-life interpretation in disease screening, spam filtering, and fraud detection, it becomes clear that the PR curve helps us understand how well a model identifies true positives without generating too many false alarms. It is also closely connected with F1-Score, because every point on the PR curve represents a particular balance between Precision and Recall. Therefore, the PR curve is one of the most meaningful graphical tools for evaluating classification models, especially in imbalanced and high-stakes real-life applications.

☛ Check Your Progress 6

Q1. What is a Precision–Recall (PR) Curve? Explain its purpose.

.....
.....

Q2. What are the axes of the PR Curve?

.....
.....

Q3. Why is the PR Curve preferred over ROC Curve for imbalanced datasets?

.....
.....

Q4. Explain the trade-off between Precision and Recall in PR Curve.

.....
.....

12.8 SUMMARY

It is learned that the Unit 12 on Performance Evaluation Measures is focused on assessing how well a predictive or classification model performs. It begins with the confusion matrix, which is the foundation of evaluation and summarizes predictions into four categories: true positives, true negatives, false positives, and false negatives. Based on this matrix, key metrics such as accuracy, precision, and recall are derived. Accuracy measures overall correctness, precision focuses on how many predicted positives are actually correct, and recall evaluates how well the model identifies all actual positives. These metrics help in understanding different aspects of model performance rather than relying on a single measure.

The unit further introduces sensitivity and specificity, which are particularly important in domains like healthcare and fraud detection. Sensitivity (same as recall) measures the ability to correctly identify positive cases, while specificity measures how well the model identifies negative cases. To balance precision and recall, the F1-score is used, which provides a

harmonic mean of both and is especially useful when dealing with imbalanced datasets. These measures ensure that the evaluation captures both correctness and reliability of predictions.

Finally, the unit covers graphical evaluation techniques such as the ROC (Receiver Operating Characteristic) curve and the Precision-Recall (PR) curve. The ROC curve illustrates the trade-off between true positive rate and false positive rate, helping in selecting optimal thresholds, while the PR curve focuses on precision versus recall, making it more informative for imbalanced datasets. Together, these evaluation measures provide a comprehensive framework for analyzing and comparing model performance, enabling data scientists to select the most suitable model for real-world applications.

12.9 ANSWERS

☛ Check Your Progress 1

Q1. What is a Confusion Matrix? Why is it important in classification models?

Solution: Refer to Section 12.2

Q2. Explain the four components of a Confusion Matrix.

Solution: Refer to Section 12.2

Q3. What is the difference between Type-I and Type-II errors in the confusion matrix?

Solution: Refer to Section 12.2

Q4. How does a confusion matrix provide better insight than accuracy alone?.

Solution: Refer to Section 12.2

☛ Check Your Progress 2

Q1. What is Accuracy? When is it most useful in model evaluation?

Solution: Refer to Section 12.3

Q2. Define Precision and explain its importance.

Solution: Refer to Section 12.3

Q3. What is Recall? Why is it important in critical applications?

Solution: Refer to Section 12.3

Q4. Why should Accuracy, Precision, and Recall be interpreted together?

Solution: Refer to Section 12.3

☛ Check Your Progress 3

Q1. What is Sensitivity? Why is it important in classification problems?

Solution: Refer to Section 12.4

Q2. Define Specificity and explain its significance.

Solution: Refer to Section 12.4

Q3. Explain the relationship between Sensitivity and Recall.

Solution: Refer to Section 12.4

Q4. Why is there a trade-off between Sensitivity and Specificity?

Solution: Refer to Section 12.4

☛ Check Your Progress 4

Q1. What is F1-Score? Why is it used in model evaluation?

Solution: Refer to Section 12.5

Q2. Write the formula of F1-Score and explain its components.

Solution: Refer to Section 12.5

Q3. Why is F1-Score preferred over Accuracy in some cases?

Solution: Refer to Section 12.5

Q4. What does a high F1-Score indicate about a model?

Solution: Refer to Section 12.5

☛ Check Your Progress 5

Q1. What is an ROC Curve? Explain its purpose in model evaluation.

Solution: Refer to Section 12.6

Q2. What are the axes of the ROC Curve?

Solution: Refer to Section 12.6

Q3. What does the Area Under the ROC Curve (AUC) indicate?

Solution: Refer to Section 12.6

Q4. How is ROC Curve useful in model comparison?

Solution: Refer to Section 12.6

☛ Check Your Progress 6

Q1. What is a Precision–Recall (PR) Curve? Explain its purpose.

Solution: Refer to Section 12.7

Q2. What are the axes of the PR Curve?

Solution: Refer to Section 12.7

Q3. Why is the PR Curve preferred over ROC Curve for imbalanced datasets?

Solution: Refer to Section 12.7

Q4. Explain the trade-off between Precision and Recall in PR Curve.

Solution: Refer to Section 12.7

12.10 REFERENCES/FURTHER READINGS

1. *Introduction to Information Retrieval* by Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, published by Cambridge University Press (1st Edition, 2008, ISBN: 978-0521865715).
2. *Pattern Recognition and Machine Learning* by Christopher M. Bishop, published by Springer (1st Edition, 2006, ISBN: 978-0387310732).

3. *Machine Learning: A Probabilistic Perspective* by Kevin P. Murphy, published by MIT Press (1st Edition, 2012, ISBN: 978-0262018029).