Performance Metrics in Machine Learning Classification Model

By Sunil Ghimire - "**Be Unique, Be Identifiable, Be You**"

Today I am going to talk about 5 of the most widely used Evaluation Metrics of the Classification Model. Before going into the details of performance metrics, let's answer a few points:

WHY DO WE NEED EVALUATION METRICS?

Being Humans we want to know the efficiency or the performance of any machine or software we come across. For example, if we consider a car we want to know the Mileage, or if we there is a certain algorithm we want to know about the Time and Space Complexity, similarly there must be some or the other way we can measure the efficiency or performance of our Machine Learning Models as well.

That being said, let's look at some of the metrics for our Classification Models. Here, there are separate metrics for Regression and Classification models. As Regression gives us continuous values as output and Classification gives us discrete values as output, we will focus on Classification Metrics.

ACCURACY

The most commonly and widely used metric, for any model, is **accuracy**, it basically does what It says, calculates what is the prediction accuracy of our model. The formulation is given below:

$$\text{Accuracy} = \frac{\text{\# of correct Prediction}}{\text{Total \# of points}} \times 100$$

As we can see, it basically tells us among all the points how many of them are correctly predicted.

Advantages:

1. Easy to use Metric.
2. Highly Interpretable.
3. If data points are balanced it gives proper effectiveness of the model.

Disadvantages:

1. Not recommended for Imbalanced data, as results can be misleading. Let me give you an example. Let's say we have 100 data points among which 95 points are negative and 5 points are positive. If I have a dumb model, which only predicts negative results then at the end of training I will have a model that will only predict negative. But still, be 95% accurate based on the above formula. Hence not recommended for imbalanced data.
2. We don't understand where our model is making mistakes.

CONFUSION METRICS

As the name suggests it is a 2×2 matrix that has Actual and Predicted as Rows and Columns respectively. It determines the number of Correct and Incorrect Predictions, we didn't bother about incorrect prediction in the Accuracy method, and we only consider the correct ones, so the Confusion Matrix helps us understand both aspects.

Let's have a look at the diagram to have a better understanding of it:

Confusion Matrix		Predicted	
		Positive	Negative
Actual	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

WHAT DOES THESE NOTATION MEANS?

Imagine I have a binary classification problem with classes as positive and negative labels, now, If my actual point is Positive and my Model predicted point is also positive then I get a True Positive, here "True" means correctly classified, and "Positive" is the predicted class by the model, Similarly If I have actual class as Negative and I predicted it as Positive, i.e. an incorrect predicted, then I get False Positive, "False" means Incorrect prediction, and "Positive" is the predicted class by the model.

We always want diagonal elements to have high values. As they are correct predictions, i.e. TP & TN.

Advantages:

1. It specifies a model is confused between which class labels.
2. You get the types of errors made by the model, especially Type I or Type II.
3. Better than accuracy as it shows the incorrect predictions as well, you understand in-depth the errors made by the model, and rectify the areas where it is going incorrect.

DIsadvantages

1. Not very much well suited for Multi-class.

PRECISION & RECALL

Precision is the measure which states, among all the predicted positive class, how many are actually positive, formula is given below:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

Recall is the measure which states, among all the Positive classes how many are actually predicted correctly, formula is given below:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

We often seek for getting high precision and recall. If both are high means our model is sensible. Here, we also take into consideration, the incorrect points, hence we are aware of where our model is making mistakes, and Minority class is also taken into consideration.

Advantages

1. It tells us about the efficiency of the model
2. Also shows us how much of the data is biased towards one class.
3. Helps us understand whether our model is performing well in an imbalanced dataset for the minority class.

Disadvantages:

1. Recall deals with true positives and false negatives and precision deals with true positives and false positives. It doesn't deal with all the cells of the confusion matrix. True negatives are never taken into account.
2. Hence, precision and recall should only be used in situations, where the correct identification of the negative class does not play a role.
3. Focuses only on Positive class.
4. Best suited for Binary Classification.

F1-SCORE

F1 score is the harmonic mean of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall). The F1 score is also known as the Sorensen–Dice coefficient or Dice similarity coefficient (DSC).

It leverages both the advantages of Precision and Recall. An Ideal model will have precision and recall as 1 hence F1 score will also be 1.

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

Advantages and Disadvantages

1. It is as same as Precision and Recall.

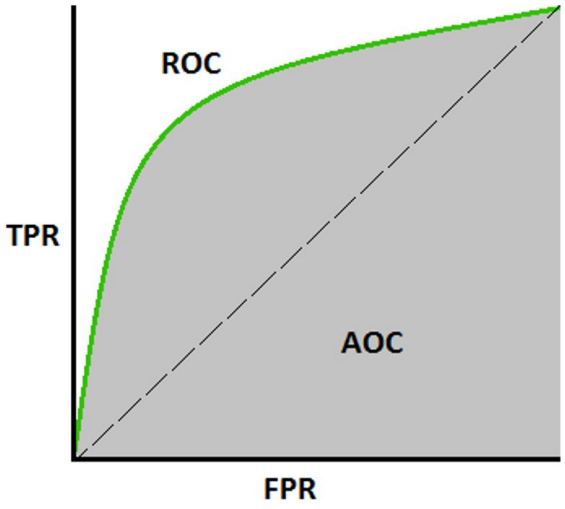
AU-ROC

AU-ROC is the Area Under the Receiver Operating Curve, which is a graph showing the performance of a model, for all the values considered as a threshold. As AU-ROC is a graph it has its own X-axis and Y-axis, whereas X-axis is FPR and Y-axis is TPR

a. **TPR = True Positive / (True Positive + False Negative)**
b. **FPR = False Positive / (False Positive + True Negative)**

ROC curve plots are basically TPR vs. FPR calculated at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives i.e. basically correct predictions.

All the values are sorted and plotted in a graph, and the area under the ROC curve is the actual performance of the model at different thresholds.



Advantages:

1. A simple graphical representation of the diagnostic accuracy of a test: the closer the apex of the curve toward the upper left corner, the greater the discriminatory ability of the test.
2. Also, allows a more complex (and more exact) measure of the accuracy of a test, which is the AUC.
3. The AUC in turn can be used as a simple numeric rating of diagnostic test accuracy, which simplifies comparison between diagnostic tests.

Disadvantages:

1. Actual decision thresholds are usually not displayed in the plot.
2. As the sample size decreases, the plot becomes more jagged.
3. Not easily interpretable from a business perspective.

So there you have it, some of the widely used performance metrics for Classification Models.

"Happy Math, Happy AI"

☺ Thanks for your time ☺

What do you think of this "**Performance Metrics in Machine Learning Classification Model**"? (Appreciation, Suggestions, and Questions are highly appreciated).

[50 Questions on Statistics & Machine Learning – Can you answer?]