**Evaluation Metrics for your Machine Learning Classification Models**

The most important part in any Machine Learning Model is to know how good or accurate your model is. Okay, so I am a budding Data Scientist and I start building models. But, how do I know the model that I built is good enough. You need to have certain parameters that will define the quality of model. Isn’t it? Evaluating the quality of the model is very important in improving the model until it performs the best.

So, when it comes to Classification models, the evaluation metrics compare both expected and predicted output to come up with the probability for the class labels. Let’s just understand what is Classification problem. These are the one where you can clearly see the target variable is divided into classes. If it can be divided to two class, then it is called Binary Classification Problem and if you can divide it to more classes then it is called Multi Class Classification Problem.

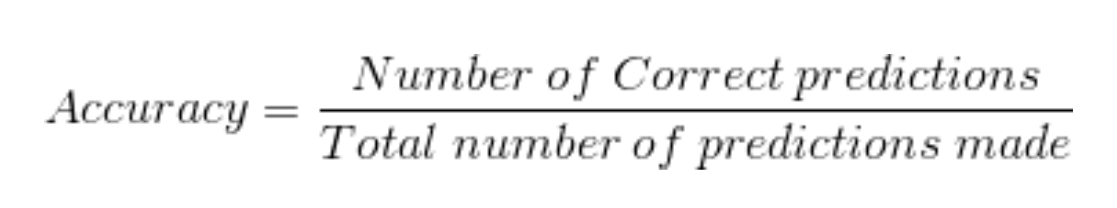
So, moving ahead with Evaluation Metrics for Classification Models. Below listed are the metrics used and we will discuss one by one.

1. Accuracy (Not in Case of Imbalanced Classes).
2. Confusion Matrix.
3. Precision.
4. Recall.
5. F1 Score.
6. AUC/ROC.

Let us understand further.

**Accuracy:**

Okay, let us get this straight into our minds. By Accuracy, what we mean is Classification Accuracy. So, it can be defined as the ration of ‘Number of Correct Prediction’ to Total prediction/Total number of input samples.



Let’s just say we had 5 input out of which we predicted 4 to be correct. Then,

Accuracy = 4/5 = 0.8 = 80%.

Accuracy is one of the simplest Metrics to be used. But, is it the best metrics? Well, answer is a big NO. Let’s find out why with an example.

Let’s assume that we building a model to predict whether the transaction is fraudulent or not. Well we built a model with accuracy 99%. Why is accuracy such high, well its because of the class imbalance. Most of the transaction would not be fraudulent. So, if you fit a model that predicts the transaction to be not fraudulent, the accuracy remains 99% owing to class imbalance. Because of the class imbalance the accuracy shoots up and is not the correct metrics to be used.

**Confusion Matrix:**

A Confusion Matrix is a table that is often used to describe performance of classification model on the given set of data for which true values are known. It is a table with four different combinations of predicted and actual values. This matrix relatively is very simple to understand but the terminologies related to this can confuse many. Let’s understand this with an example.

|  |  |  |
| --- | --- | --- |
|  | Predicted Positive (1) | Predicted Negative (0) |
| Actual Positive (1) | TP | FN |
| Actual Negative (0) | FP | TN |

Let us define the most basic terms associated with Confusion Matrix.

**True Positive (TP):** This is the number of times our model predicted YES and Actual output was YES.

**True Negative (TN):** The number of times out model predicted NO and Actual output was NO.

**False Positive (FP):** This is the number of times our model predicted YES and output was NO. (Type I Error)

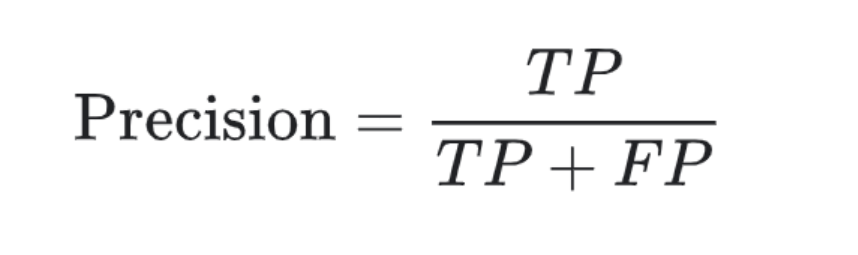
**False Negative (FN):** The number of times our model predicted NO and the output was YES. (Type II Error)

There is a list of other metrices that we can define from Confusion Matrix.

**Precision:**

Precision is correctly predicted as positive compared to total predicted as positive. Confusing? Not much. Let’s see. Precision answers the question, ’What proportion of positive instance was actually correct?’ It is the ratio of Number of true positive to total number of predicted positive.

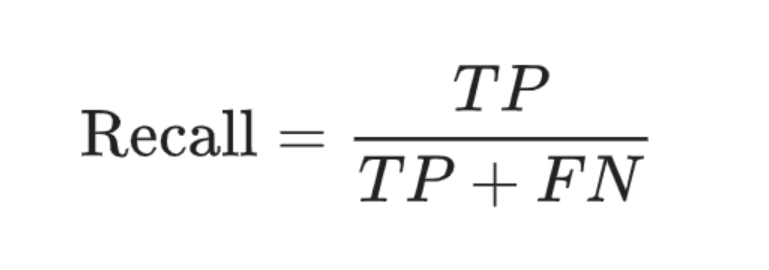
The formula for Precision goes as below:



**Recall/Sensitivity:**

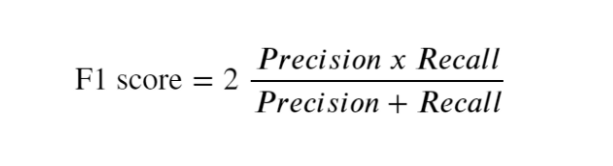
Recall/Sensitivity is correctly predicted as positive to total number of positive. Be very careful here. In case of Precision we have denominator as total predicted as positive i.e. TP+FP while here we have total number of positive i.e. TP+TN. This means denominator takes total number of outputs we predicted correctly be it positive or negative class. The question that we answer here is, ‘What proportion of positive instance was actually classified correctly?’

The formula for Recall goes as below:



**F1 Score:**

F1 Score is dependent on precision and recall. It is a function of both. In short F1 determines how many instances your model identified correctly without missing a significant number of instances.

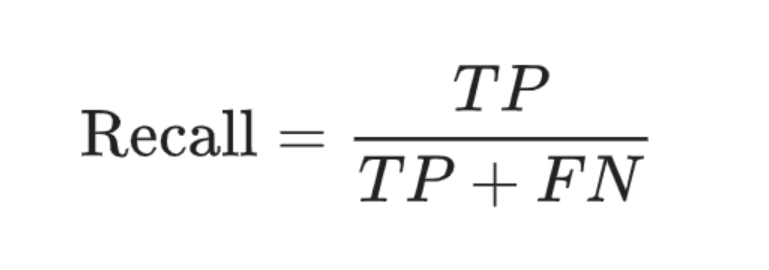


Our goal here will be to maximise F1 score. The more the score, the better the performance of the model. So, F1 score is a better metric when there is imbalance in the dataset. But if the dataset is balanced and small, we can use Accuracy.

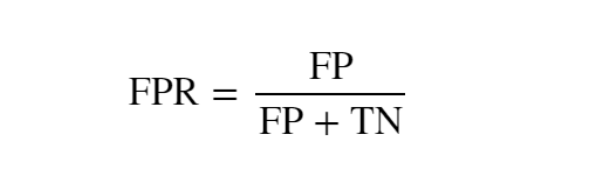
**AUC (Area Under Curve) ROC (Receiver Operating Characteristics) Curve:**

ROC Curve is a commonly used method to visualise the performance of a binary classifier. We already discussed what a binary classifier is. To Recap, it is a classifier with two possible output classes. ROC curve shows the performance of a classification model at all threshold values. It plots two parameters.

1. True Positive rate/Recall (TPR):

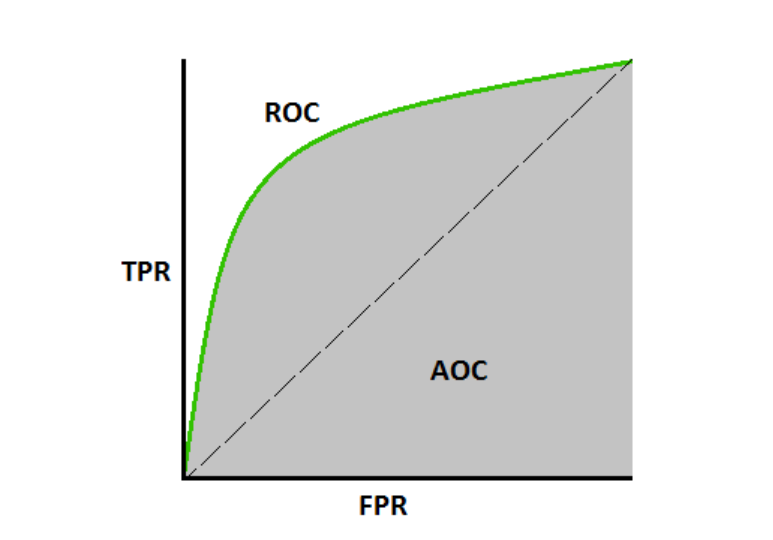


1. False Positive rate (FPR):



AUC stands for **Area under the ROC Curve**. AUC provides aggregate measure of performance across all possible classification thresholds.

Both TPR and FPR are within the range [0,1] and the curve determine the positioning of TPR vs FPR at different points in the range [0,1]. And, the larger the AUC, better your model is.



NOTE: There is no metric that will work for all the problem every time. It depends on the problem statement, kind of data and industry.

**References:**

<https://en.wikipedia.org/wiki/F1_score>

<https://en.wikipedia.org/wiki/Receiver_operating_characteristic>

<https://en.wikipedia.org/wiki/Precision_and_recall>

<https://en.wikipedia.org/wiki/Confusion_matrix>