1 Introduction

Assessing the performance of a machine learning model is an essential step in a predictive modeling pipeline. Once a model is ready, it has to be evaluated to establish its correctness. Building a model is easy, but creating a useful model is difficult. Evaluation metrics explain the performance of a model. An important aspects of evaluation metrics is their capability to discriminate among model results.

We consider different kinds of metrics to evaluate our models. The choice of metric completely depends on the type of model and the implementation plan of the model. Now, we will go through some of these metrics which will help us in evaluating our model accuracy.

2 Confusion Matrix

The Confusion matrix is one of the most intuitive and easiest metrics used for finding the correctness and accuracy of the model. It is used for Classification problem where the output can be of two or more types of classes.

Before diving into what the confusion matrix is all about, Let's say we are solving a classification problem where we are predicting whether a person is having cancer or not.

Let's give a label to our target variable: "1" When a person is having cancer "0" When a person is not having cancer.

The confusion matrix is a table with two dimensions ("Actual" and "Predicted"), and sets of "classes" in both dimensions. Our Actual classifications are columns and Predicted ones are Rows.

		Actual	
		Positives(1)	Negatives(0)
Predicted	Positives(1)	TP	FP
	Negatives(0)	FN	TN

Terms associated with Confusion matrix:

True Positives (TP): When the actual class of the data point was 1(True) and the predicted is also 1(True).

True Negatives (TN): When the actual class of the data point was 0(False) and the predicted is also 0(False).

False Positives (FP): When the actual class of the data point was 0(False) and the predicted is 1(True). False is because the model has predicted incorrectly and positive because the class predicted was a positive one.

False Negatives (FN): When the actual class of the data point was 1(True) and the predicted is 0(False). False is because the model has predicted incorrectly and negative because the class predicted was a negative one.

3 Accuracy

Accuracy in classification problems is the number of correct predictions made by the model over all kinds predictions made.

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

One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when we have symmetric datasets where values of false positive and false negatives are almost same. Therefore, we have to look at other parameters to evaluate the performance of our model.

4 Precision

Precision is a measure that tells us what proportion of patients that we diagnosed as having cancer, actually had cancer. The predicted positives (People predicted as cancerous are TP and FP) and the people actually having a cancer are TP.

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

5 Recall or Sensitivity

Recall is a measure that tells us what proportion of patients actually had cancer was diagnosed by the algorithm as having cancer. The actual positives (People having cancer are TP and FN) and the people diagnosed by the model having a cancer are TP.

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

6 Specificity

Specificity is a measure that tells us what proportion of patients did NOT have cancer, were predicted by the model as non-cancerous. The actual negatives (People actually NOT having cancer are FP and TN) and the people diagnosed by us not having cancer are TN (FP is included because the Person did NOT actually have cancer even though the model predicted otherwise).

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

Ex: In our cancer example with 100 people, 5 people actually have cancer. Let's say that the model predicts every case as cancer.

So our denominator (False positives and True Negatives) is 95 and the numerator, person not having cancer and the model predicting his case as no cancer is 0 (Since we predicted every case as cancer).

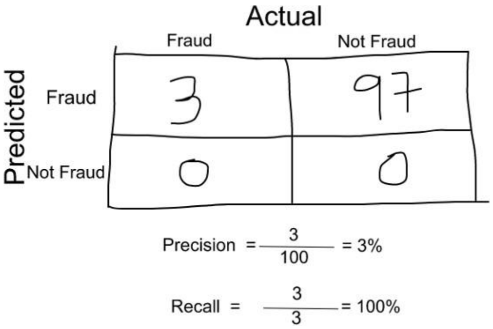
7 F1 Score

When we make a model for solving a classification problem we really don’t want to carry both Precision and Recall every time. So it’s best if we can get a single score which represents both Precision(P) and Recall(R).

One way to do that is simply taking their arithmetic mean. i.e (P + R) / 2 where P is Precision and R is Recall. But that’s pretty bad in some situations.

Suppose we have 100 credit card transactions, of which 97 are legal and 3 are fraud and let’s say we came up with a model which predicts everything as fraud. (Horrendous right!?)

Precision and Recall for the example is shown in the figure below.



Predicted \ Actual	Fraud	Not Fraud
Fraud	3	97
Not Fraud	0	0

$$\text{Precision} = \frac{3}{100} = 3\%$$

$$\text{Recall} = \frac{3}{3} = 100\%$$

Now, if we simply take arithmetic mean of both, then it comes out to be nearly 51%. We shouldn’t be giving such a moderate score to a terrible model since it’s just predicting every transaction as fraud. So, we need something more balanced than the arithmetic mean and that is harmonic mean.

The Harmonic mean is given by 2xy/x + y

Harmonic mean is a kind of an average when x and y are equal. But when x and y are different, then it’s closer to the smaller number as compared to the larger number.

For our previous example, F1 Score = Harmonic Mean(Precision, Recall)

F1 Score = 2 * Precision * Recall / (Precision + Recall) = 2*3*100/103 = 5

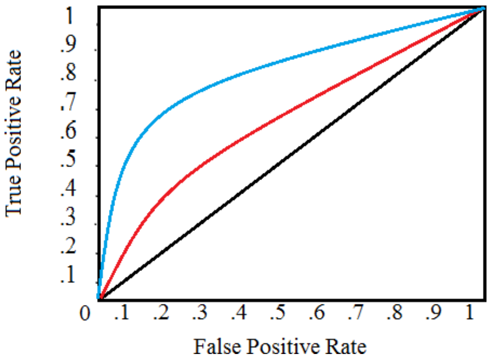
So if one number is really small between precision and recall, the F1 Score raises a flag and is more closer to the smaller number than the bigger one, giving the model an appropriate score rather than just an arithmetic mean.

8 Receiver Operating Characteristic (ROC) curve

ROC curve shows how the recall vs precision relationship changes as we vary the threshold for identifying a positive in our model. The threshold represents the value above which a data point is considered in the positive class. If we have a model for identifying a disease, our model might output a score for each patient between 0 and 1 and we can set a threshold in this range for labeling a patient as having the disease (a positive label). By altering the threshold, we can try to achieve the right precision vs recall balance.

An ROC curve plots the true positive rate on the y-axis versus the false positive rate on the x-axis. The true positive rate (TPR) is the recall and the false positive rate (FPR) is the probability of a false alarm. Both of these can be calculated from the confusion matrix:

True positive rate $= \frac{true\ positives}{true\ positives\ +\ false\ negatives}$ False positive rate $= \frac{false\ \ positives}{false\ positives\ +\ true\ negatives}$



The black diagonal line indicates a random classifier and the red and blue curves show two different classification models. For a given model, we can only stay on one curve, but we can move along the curve by adjusting our threshold for classifying a positive case. Generally, as we decrease the threshold, we move to the right and upwards along the curve. With a threshold of 1.0, we would be in the lower left of the graph because we identify no data points as positives leading to no true positives and no false positives (TPR = FPR = 0).

As we decrease the threshold, we identify more data points as positive, leading to more true positives, but also more false positives (the TPR and FPR increase). Eventually, at a threshold of 0.0 we identify all data points as positive and find ourselves in the upper right corner of the ROC curve (TPR = FPR = 1.0).

we can quantify a model’s ROC curve by calculating the total Area Under the Curve (AUC), a metric which falls between 0 and 1 with a higher number indicating better classification performance. In the graph above, the AUC for the blue curve will be greater than that for the red curve, meaning the blue model is better at achieving a blend of precision and recall. A random classifier (the black line) achieves an AUC of 0.5.