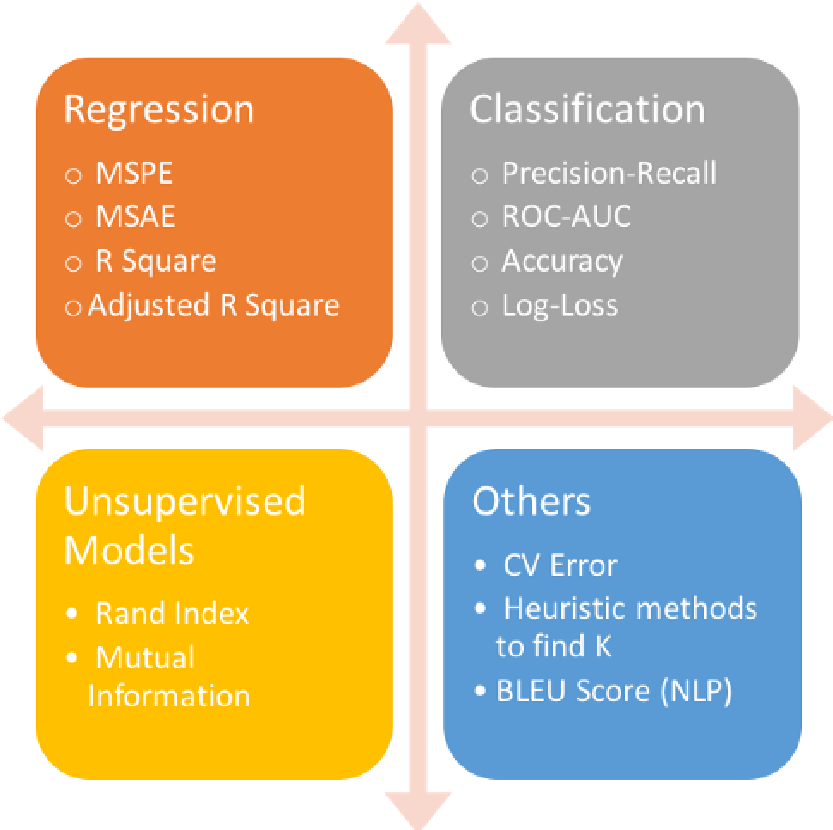
Performance Metrics for Classification



Basic Vocabulary related to Metrics

Predicted: **Outcome** of the **model** on the **validation set**.

Actual: **Values seen** in the **training set**.

Positive (P): **Observation** is **positive**.

Negative (N): **Observation** is **not positive**.

True Positive (TP): **Observation** is **positive**, and is **predicted correctly**.

False Negative (FN): **Observation** is **positive**, but **predicted wrongly**.

True Negative (TN): **Observation** is **negative**, and **predicted correctly.**

False Positive (FP): **Observation is negative**, but **predicted wrongly**.

1. Confusion Matrix

Also known as an **Error Matrix**, the** Confusion Matrix is a two-dimensional matrix that allows visualization of the algorithm's performance**.

While this **isn't an actual metric to use for evaluation**, it's an **important starting point**.

Predictions are highlighted and divided by class (true/false), before being **compared with the actual values**.

The **matrix's size is compatible with the amount of classes** in the label column.

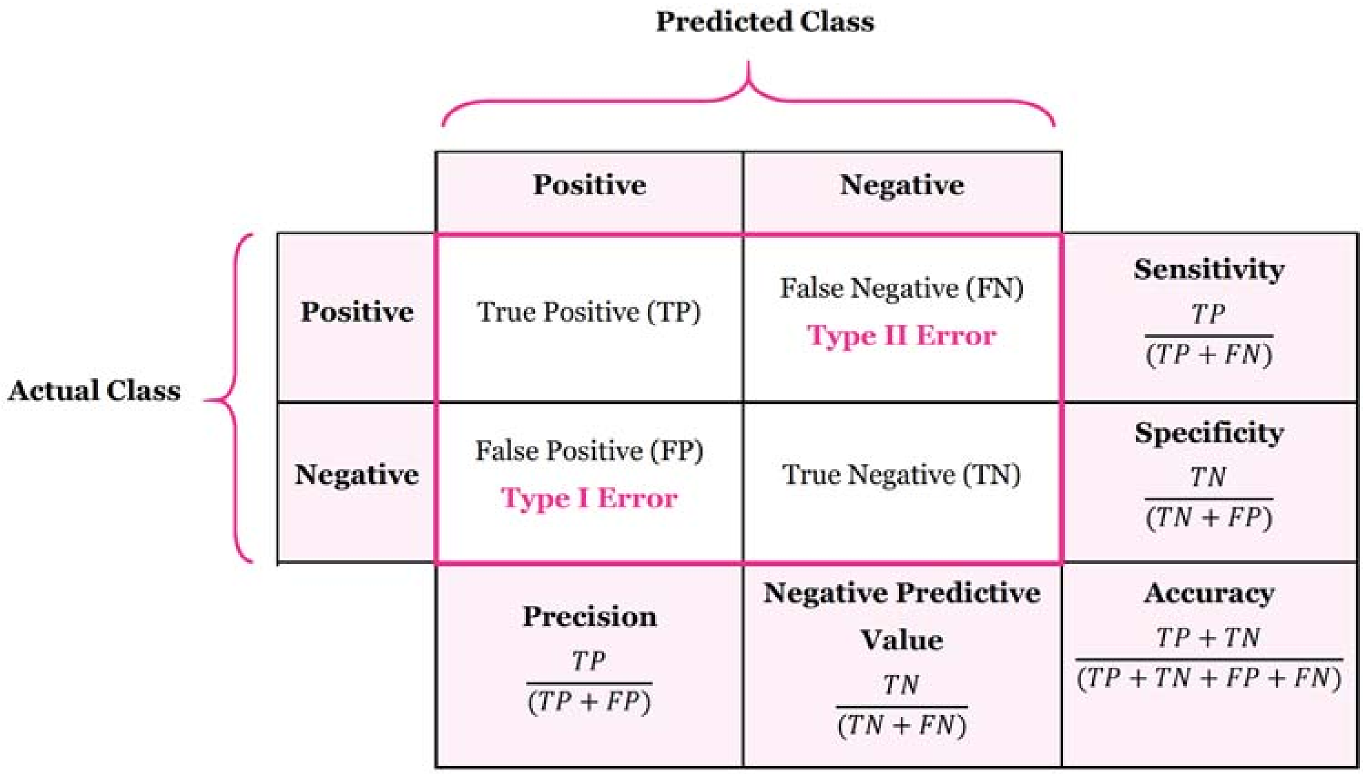
In a **binary classification, the matrix will be 2X2**.

If there are **3 classes**, the matrix **will be 3X3**, and so on.

This **matrix essentially helps you determine if the classification model is optimized.**

It **shows what errors are being made and helps to determine their exact type**.

Besides machine learning, the Confusion Matrix is also used in the fields of statistics, data mining, and artificial intelligence.

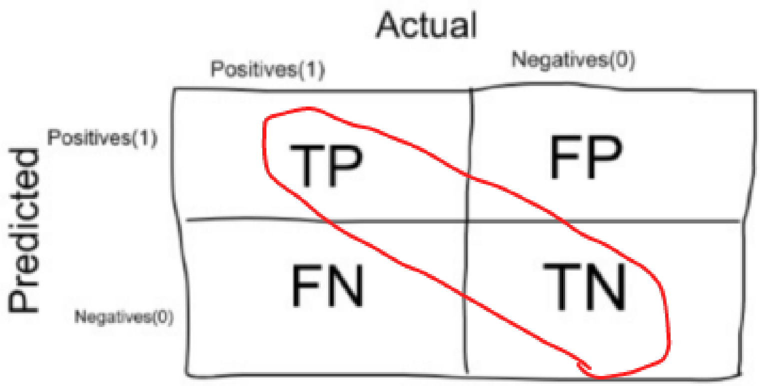


Accuracy

A **classification model's accuracy** is defined as the **percentage of predictions** it got right.

However, it's **important to understand** that **it becomes less reliable when the probability of one outcome is significantly higher than the other one, making it less ideal as a stand-alone metric.**

The expression used to calculate accuracy is as follows



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

Detection rate

This **metric basically shows** the **number of correct positive class predictions** made **as a proportion of all of the predictions made**.

$$\text{Detection Rate} \quad DR = \frac{TP}{TP+FN} \times 100\%$$

Logarithmic loss

Also **known as log loss**, **logarithmic loss basically functions by penalizing all false/incorrect classifications. **

The **classifier must assign a specific probability** to **each class for all samples while working with this metric.**

The **formula** for **calculating log loss** is as follows:

$$logloss = -\frac{1}{N}\sum_{i}^{N}\sum_{j}^{M} y_{ij}\log(p_{ij})$$

- N is the number of rows
- M is the number of classes

In a nutshell, **the range of log loss varies from 0 to infinity (∞)**.

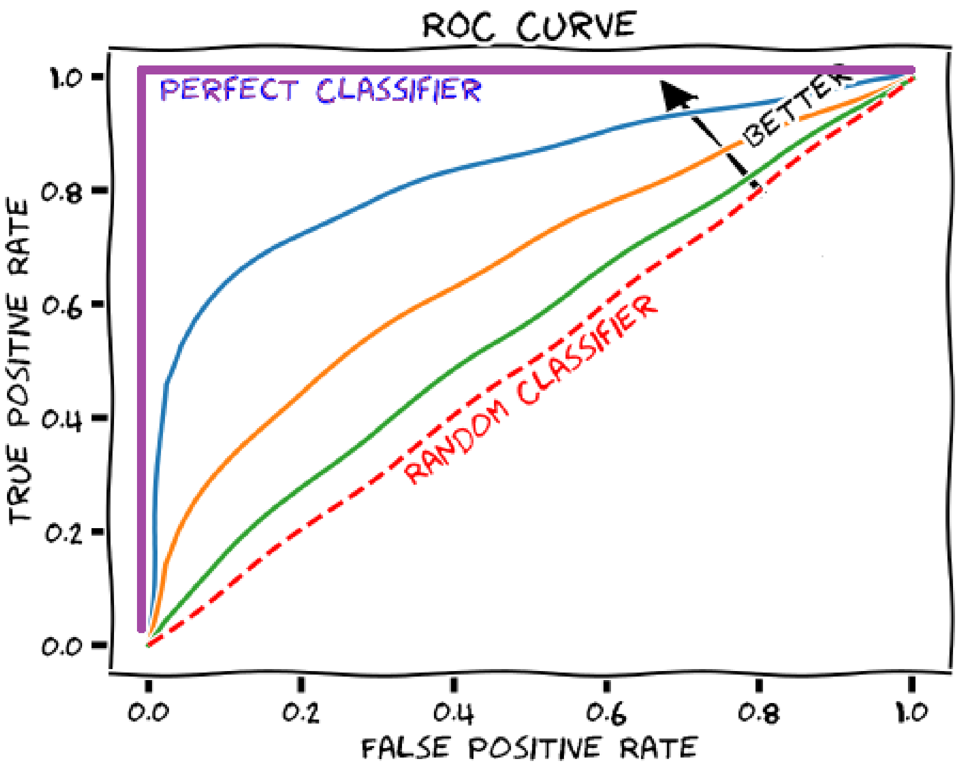
The **closer it is to 0**, the **higher the prediction accuracy. Minimizing it is a top priority.**

Receiver operating characteristic curve (ROC) / area under curve (AUC) score

The **ROC curve** is basically a graph that displays the **classification model’s performance at all thresholds. **

As the name suggests, the AUC is the entire area below the two-dimensional area below the ROC curve.

This curve basically generates two important metrics: sensitivity and specificity.



Sensitivity (true positive rate)

The true positive rate, also known as sensitivity, corresponds to the **proportion of positive data points that are correctly considered as positive, with respect to all positive data points.**

$$\begin{aligned}\text{sensitivity} &= \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}} \\ &= \frac{\text{number of true positives}}{\text{total number of sick individuals in population}} \\ &= \text{probability of a positive test given that the patient has the disease}\end{aligned}$$

Specificity (false positive rate)

False positive rate, also known as specificity, corresponds to the **proportion of negative data points that are mistakenly considered as positive**, with respect to all negative data points.

$$\begin{aligned}\text{specificity} &= \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}} \\ &= \frac{\text{number of true negatives}}{\text{total number of well individuals in population}} \\ &= \text{probability of a negative test given that the patient is well}\end{aligned}$$

Please note that **both FPR and TPR** have values **in the range of 0 to 1**

Precision

This **metric is the number of correct positive results** divided by the **number of positive results predicted by the classifier.**

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

Recall

Recall is the number of correct positive results divided b**y the number of all samples that should have been identified as positive.**

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

F1 score

The F1 score is **basically the harmonic mean between precision and recall**.

It is **used to measure the accuracy of tests** and is a **direct indication of the model's performance.**

The range of the **F1 score is between 0 to 1**, with the **goal being to get as close as possible to 1**. It is calculated as per

$$F1 = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{recall}}$$

Micro-, Macro-, Weighted-averaged Precision

Let's take the previous example, and draw the true values and the predictions as a table:

Label	Predicted
cat	cat
cat	cat
cat	cat
cat	cat
dog	dog
dog	dog
dog	cat
bird	dog
bird	bird

Micro-averaged Precision is **calculated** as **precision of Total values**:

$$\text{Micro-averaged Precision} = \frac{TP_{total}}{TP_{total} + FP_{total}} = \frac{7}{7+2} = 0.7777$$

Weighted-averaged Precision is also **calculated based on Precision per class but takes into account the number of samples of each class in the data**:

$$\text{Weighted-averaged Precision} = \frac{Precision_{birds} * N_{birds} + Precision_{cats} * N_{birds} + Precision_{dogs} * N_{birds}}{\text{Total number of samples}}$$

Micro-averaged: **all samples equally contribute** to **the final averaged metric**

Macro-averaged: **all classes equally contribute** to **the final averaged metric**

Weighted-averaged: each **classes's contribution** to **the average is weighted by its size**

```
Report :
              precision    recall  f1-score   support

           0       0.97      0.88      0.92     21011
           1       0.76      0.93      0.84      8915

    accuracy                           0.89     29926
   macro avg       0.86      0.90      0.88     29926
weighted avg       0.91      0.89      0.90     29926
```