**Choosing the Right Metric for Evaluating classification models**

For classification models we have the following metrics:

ROC-AUC score

F1 score

Accuracy

Log-Loss

Confusion matrix can also be used to verify how well a classifier is performing

**ROC-AUC score**

The probabilistic interpretation of ROC-AUC score is that if you randomly choose a positive case and a negative case, the probability that the positive case outranks the negative case according to the classifier is given by the AUC.

Mathematically, it is calculated by area under curve of sensitivity (TPR) vs.   
FPR(1-specificity). Ideally, we would like to have high sensitivity & high specificity, but in real-world scenarios, there is always a tradeoff between sensitivity & specificity.

ROC-AUC score is independent of the threshold set for classification because it only considers the rank of each prediction and not its absolute value. The same is not true for F1 score which needs a threshold value in case of probabilities output. Also, F1 score is sensitive to threshold.

**Which metric to choose and when?**

In situation where we have balanced classes in the target:

* If you care for absolute probabilistic difference, go with log-loss
* If you care only for the final class prediction and you don’t want to tune threshold, go with AUC score
* F1 score is sensitive to threshold and you would want to tune it first before comparing the models

In situation where we have imbalanced classes in the target:

* log-loss function is symmetric and does not differentiate between classes.
* If you care for a class which is smaller in number independent of the fact whether it is positive or negative, go for ROC-AUC score.
* F1 cares only for the misclassification of positive labels. So in case of large positive labels the F1 score might give similar results for different models.

**When will you prefer F1 over ROC-AUC?**

* When you have a small positive class, then F1 score makes more sense. This is the common problem in fraud detection where positive labels are few.

**Which metric should you use for multi-classification?**

There are three types of non-binary classification:

Multi-class: For example classifying the target variable into more than two classes like orange, apple, and mango. ROC-AUC score in skLearn does not work for multi-class classification problems.

Multi-label: Example: document classification

Hierarchical: For situations involving questions like: How can we train a system to be both general enough to pick out a meaningful category from a broad variety of documents and specialized enough to potentially dive deeper into each of these categories? This involves neural network like LSTM(Long Short Term Memory)-A, GRU(Gated Recurrent Unit)-A(A stands for attention).