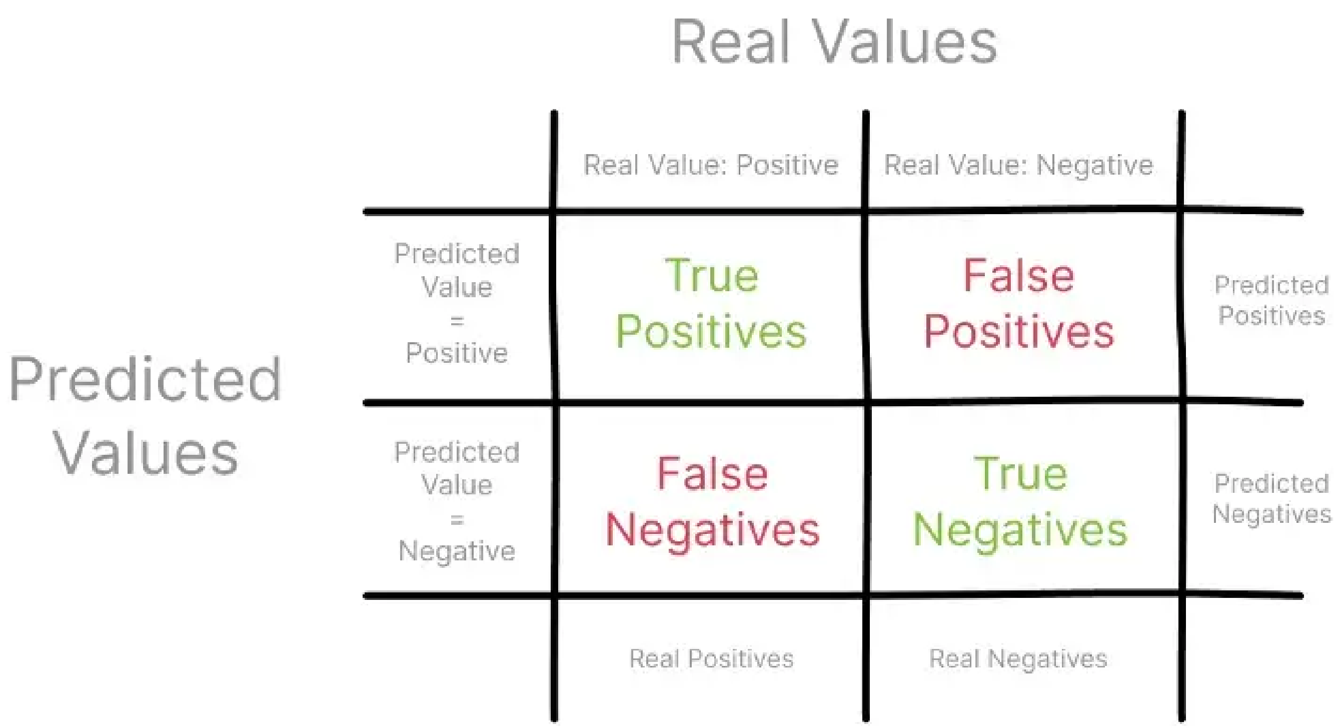
Classification metrics

The confusion matrix



- Accuracy = (TP + TN) / (TP + FP + FN + TN)
  - use this when the dataset classes are well balanced
- When the dataset is not well balanced, we can use a set of metrics that view the classification problem from different angles. Use the following table as a guide to tell the 4

main metrics apart:

Conditioned on actual true values	Conditioned on actual false values
$\frac{TP}{TP+FN}$ = Recall / Sensitivity / True Positive rate	$\frac{TN}{TN+FP}$ = Specificity (true negative rate)
$\frac{FN}{TP+FN}$ = False negative rate	$\frac{FP}{TN+FP}$ = False positive rate
Conditioned on predicted true values	**Conditioned on predicted false values**
$\frac{TP}{TP+FP}$ = Precision / Positive Predictive Value	$\frac{TN}{TN+FN}$ = Negative Predictive Value
$\frac{FP}{TP+FP}$ = False discovery rate	$\frac{FN}{TN+FN}$ = False omission rate

- Precision / positive predictive value
  - use this when false positives are unacceptable
  - use this when you want the model to be as certain as possible when it makes a true prediction
  - examples: credit default, crime prediction, spam detection, etc.
  - If a model is optimized for precision, lots of true positives will fall through the cracks, but those that we do catch we can be more certain that it's not a false positive

$$\text{Precision} = \frac{\text{true positives}}{\#\text{ of precdicted positives}} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- Recall / recall / hit rate / true positive rate / sensitivity
  - use this when false negatives are unacceptable
  - use this when you want the model to pick up as many of the positive samples as possible, regardless of how many negatives it picks up by accident
  - examples: cancer diagnosis
  - If a model is optimized for recall, you may get lots of false alarms, but you'll be very likely to capture all the real emergencies

$$\text{Recall} = \frac{\text{true positives}}{\#\text{ of actual positives}} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- Specificity / true negative rate
  - use this when false positives are unacceptable
  - use this when you want the model to pick up as many of the negative samples as possible, regardless of how many positives it picks up by accident
  - examples: detecting a very rare but severe disease, where a false positive could lead to

an invasive and risky procedure
- If a model is optimized for specificity, you may get lots of false negatives, but you'll be very likely to capture all the true negatives

$$\text{specificity} = \frac{\text{true negatives}}{\text{\# of actual negatives}} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

- Negative predictive value
  - use this when false negatives are unacceptable
  - use this when you want the model to be as certain as possible when it makes a negative prediction

$$\text{Negative predictive value} = \frac{\text{true negatives}}{\text{\# of predicted negatives}} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

- F1 score
  - balance between precision and recall

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

- weighted F1 score
  - use β paramerter to describe how much more importance to give to recall vs. precision

$$\text{F}_\beta = (1 + \beta^2) * \frac{\text{precision} * \text{recall}}{(\beta^2 * \text{precision}) + \text{recall}}$$

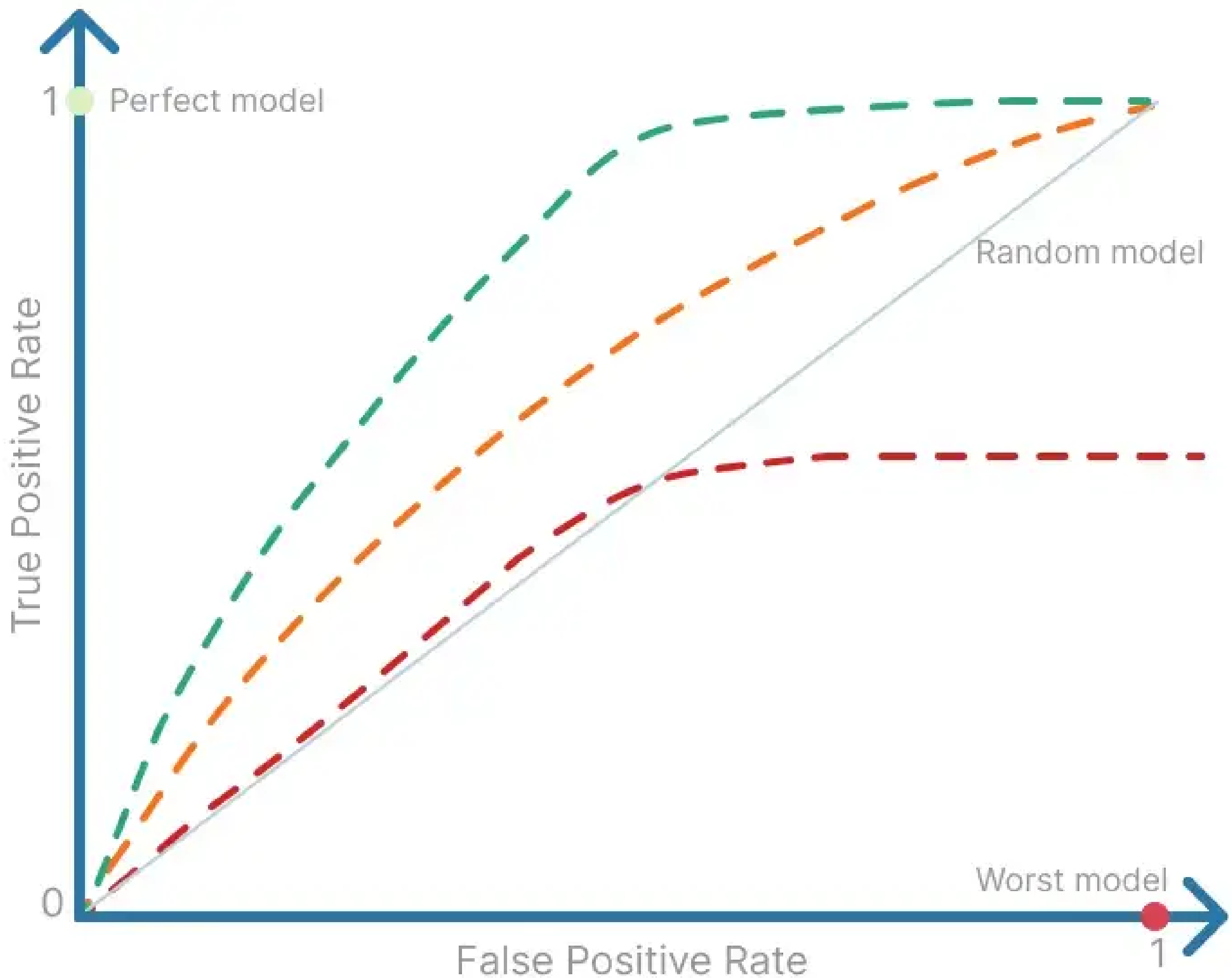
ROC and Precision-Recall Curves

Using these 4 (8) metrics in isolation may provide a limited view on how the model is performing. Another way to score models is by using the predicted probabilities instead of predicted classes and varying the threshold for what constitutes a positive prediction to evaluate how the metrics respond.

ROC curve

- Plot the False positive rate (X axis) against the true positive rate (y axis)
- **False Positive Rate = 1 - Specificity**
- **True positive rate = Recall**

When you make the plot you get something like this:



Where green > orange > red

Precision-Recall Curve

- Plot recall (x axis) against precision (y axis)

When to use ROC vs Precision-Recall curves?

- ROC: balanced classes
- Precision-Recall curve: moderate - large class imbalance

AUC

to sum up the results of the ROC analysis, use AUC (Area under the curve) (also called a **c-statistic**)

higher AUC generally means a better model

References

- https://towardsdatascience.com/the-5-classification-evaluation-metrics-you-must-know-aa97784ff226
- https://en.wikipedia.org/wiki/Sensitivity_and_specificity
- https://towardsdatascience.com/roc-analysis-and-the-auc-area-under-the-curve-404803b694b9
- https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/