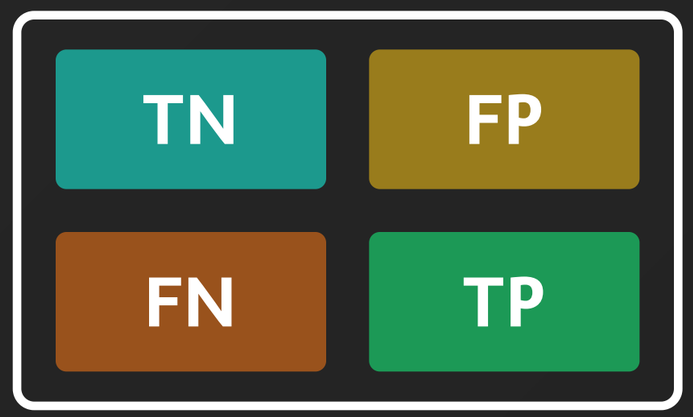
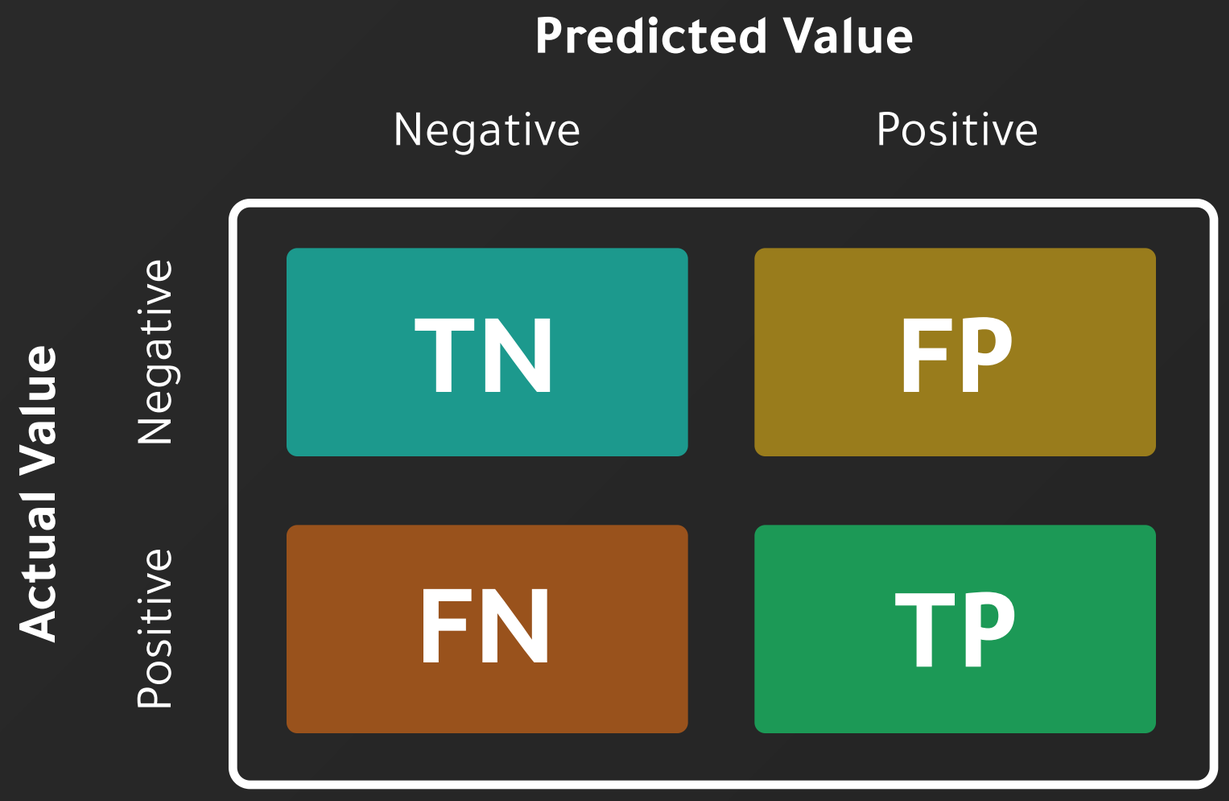
Confusion Matrix



and Classification Evaluation Metrics

Trust is a must when a decision-maker's judgment is critical. To give such trust, we summarize all possible decision outcomes into four categories: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) to serve an outlook of how confused their judgments are, namely, the confusion matrix. From the confusion matrix, we calculate different metrics to measure the quality of the outcomes. These measures influence how much trust we should give to the decision-maker (classifier) in particular use cases. This document will discuss the most common classification evaluation metrics, their focuses, and their limitations in a straightforward and informative manner.

Confusion Matrix



Guess	→	Fact	
This car is NOT red	→	This car is NOT red	TN
This car is red	→	This car is red	TP
This car is NOT red	→	This car is red	FN
This car is red	→	This car is NOT red	FP

Positive Predictive Value

Precision

$$\frac{TP}{TP + FP}$$

Sensitivity or True Positive Rate

Recall

$$\frac{TP}{TP + FN}$$

True Negative Rate

Specificity

$$\frac{TN}{TN + FP}$$

(Negative Predictive Value)

NPV

$$\frac{TN}{TN + FN}$$

Accuracy

$$\frac{TP + TN}{FP + TP + TN + FN}$$

F1-Score

$$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Balanced Accuracy

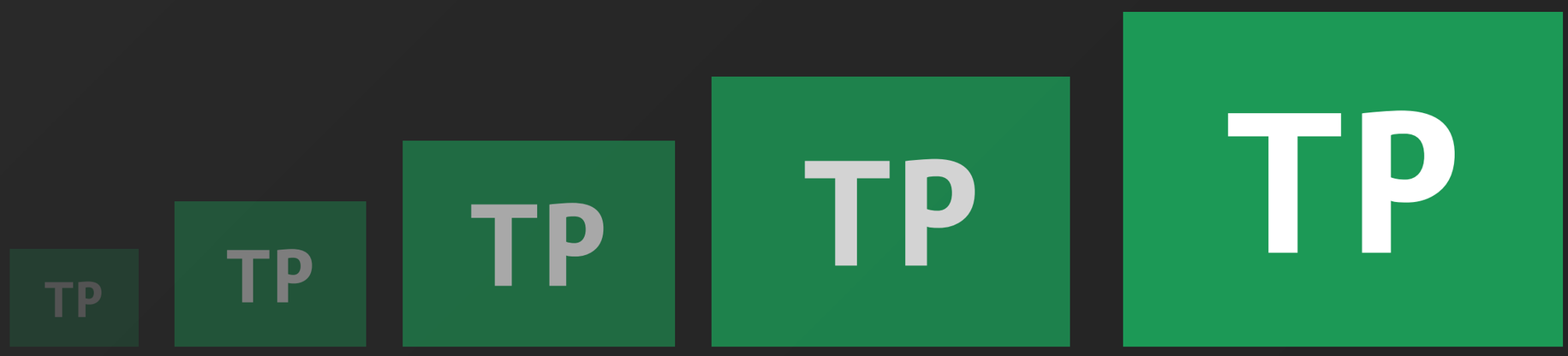
$$\frac{\text{Sensitivity} + \text{Specificity}}{2}$$

Matthews Correlation Coefficient (MCC)

$$\frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

Precision & Recall

Common Goal



We use both metrics when actual negatives are less relevant. For example, googling "Confusion Matrix" will have trillions of unrelated (negative) web pages, such as the "Best Pizza Recipe!" web page. Accounting for whether we have correctly predicted the latter webpage and alike as negative is impractical.

Precision Goal

This customer loves steak! — No, this customer is vegan. FP

Bad Product Recommendation → Less Conversion → Decrease in Sales

Recall Goal

The product has no defects. — A customer called... He's angry. FN

Bad Defect Detector → Bad Quality → Customer Dissatisfaction

Specificity & NPV

Common Goal

We use both metrics when actual positives are less relevant. In essence, we aim to rule out a phenomenon. For example, we want to know how many healthy people (no disease detected) there are in a population. Or, how many trustworthy websites (not fraudulent) is someone visiting.

Specificity Goal

This person is a criminal. — They were detained for no reason. FP

Bad Predictive Policing → Injustice

NPV Goal

They don't have cancer. — No, they should be treated! FN

Bad Diagnosis → No Treatment → Consequences