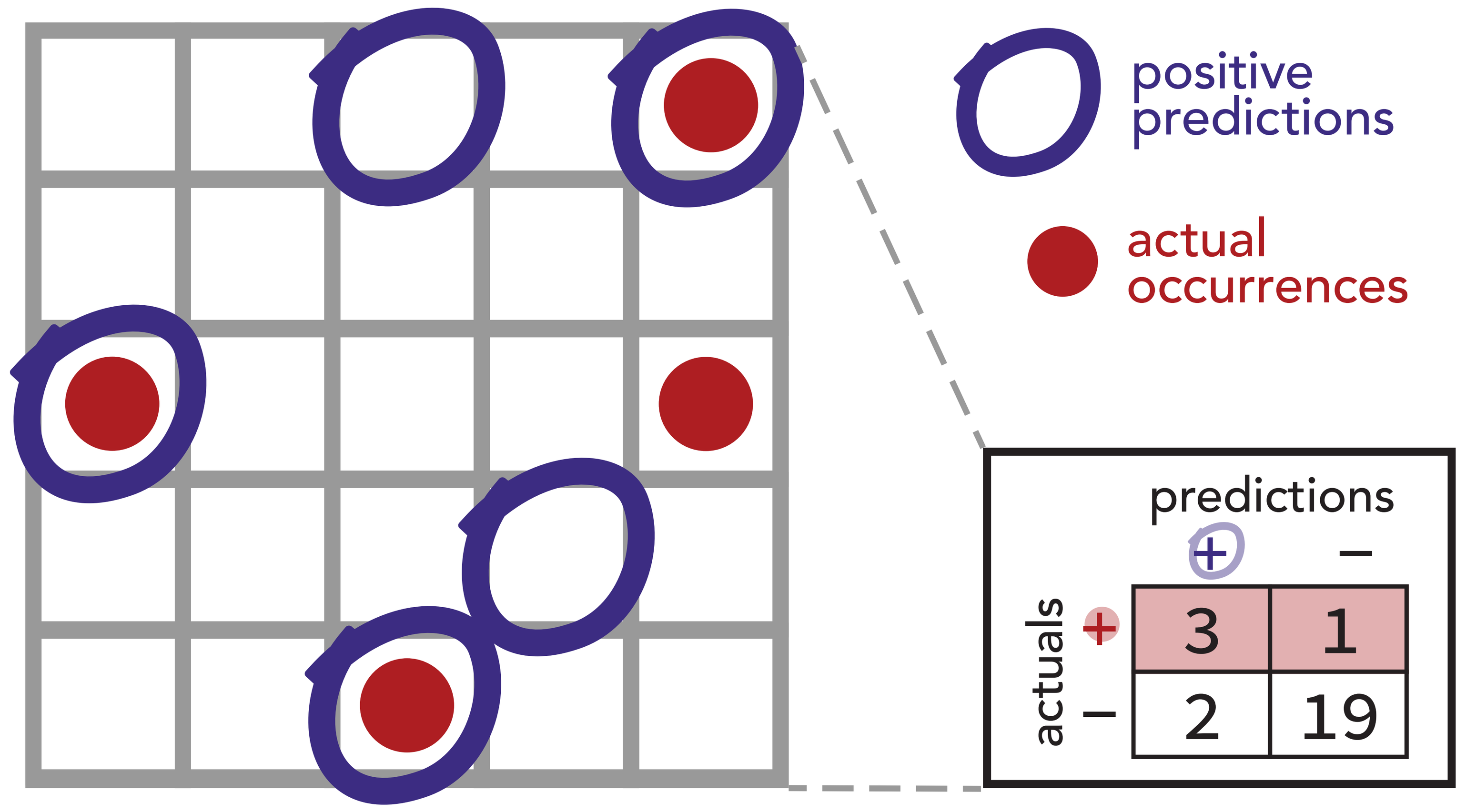
confusion matrix cheat sheet

chrissirico.com



		predictions +	predictions −
actuals	+	3	1
actuals	−	2	19

outcomes

- **T**rue **P**ositives
- **F**alse **N**egatives
- **F**alse **P**ositives
- **T**rue **N**egatives

confusion matrix

		predictions +	predictions −
actuals	+	TP	FN
actuals	−	FP	TN

Each cell in the grid at far left represents an observation. A hypothetical classification model "circles" observations it thinks are positives. Actual positive occurrences show up as red dots and non-occurrences (negatives) as empty cells.

The confusion matrix tallies the model's correct and incorrect positive and negative predictions (true positives, false negatives, true positives and false positives).

airport security analogy

Your goal is to stop smugglers. Search travelers predicted to be smugglers and pass the rest.

- **TP**: searched smugglers (nice work!)
- **FN**: passed smugglers (oops!)
- **FP**: searched, innocent travelers (oops!)
- **TN**: passed, innocent travelers (nice work!)

classification metrics

Accuracy

- the proportion of all predictions that are correct
- appears high when dataset is imbalanced, even if model is no better than naïve (always predicts the majority class)

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

Precision (positive predictive value)

- accuracy of positive predictions
- proportion of searched w/ contraband

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

Recall (sensitivity, true positive rate)

- proportion of actual occurrences correctly predicted positive
- proportion of smugglers caught

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

Specificity (true negative rate)

- proportion of non-occurrences correctly predicted negative
- proportion of innocent travelers passed

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

False Positive Rate (false alarm rate)

- proportion of non-occurrences falsely predicted positive
- proportion of innocent travelers searched

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

False Negative Rate (miss rate)

- proportion of actual occurrences falsely predicted negative
- proportion of smugglers passed

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

F1 Score

- harmonic mean of precision & recall
- balances tradeoff between multiple metrics
- assumes equal value/cost of TP, FP, TN, FN

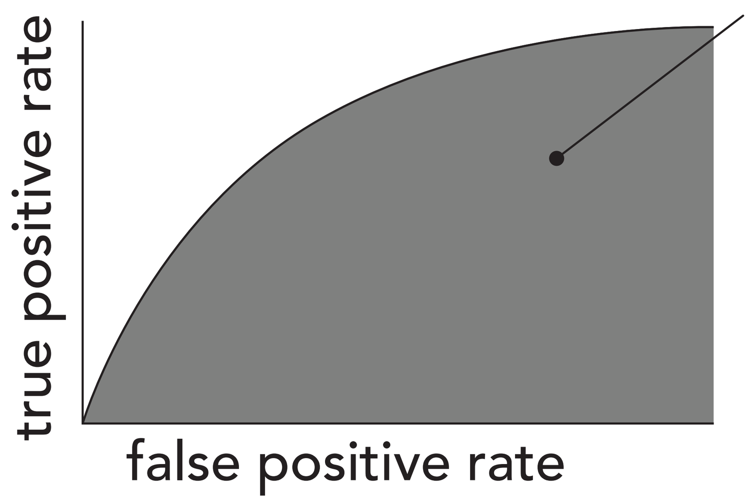
$$2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

threshold-agnostic metrics vs. binary classification metrics

Threshold-agnostic metrics are based on predicted probabilities. They are useful for describing the strength of model signal across all thresholds and for selecting algorithms, features and hyperparameters.
(E.g., log loss, Gini norm, AUC and P-R AUC.)

Classification metrics defined to the left are based on binary (1, 0) predictions. They are useful for selecting a prediction threshold by comparing score tradeoffs resulting from various thresholds.

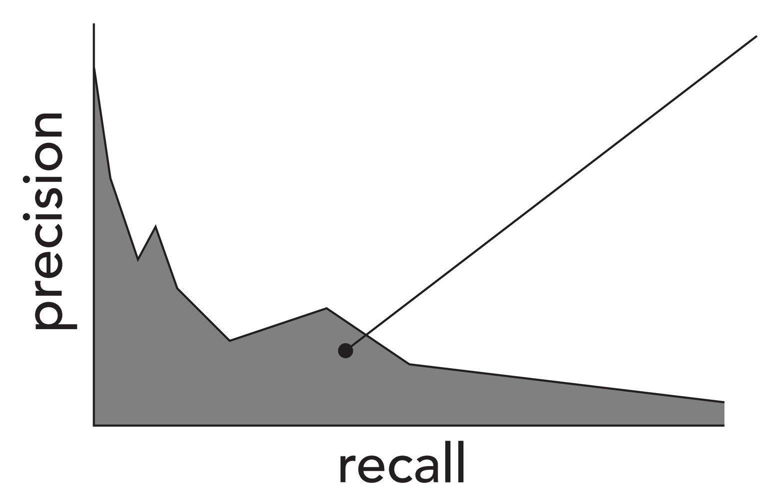
ROC Curve



AUC (area under the ROC curve)

- threshold-agnostic
- shows (false) lift for naïve models on imbalanced datasets

Precision-Recall Curve



P-R AUC (area under precision-recall curve)

- threshold-agnostic
- better for imbalanced datasets with few positive observations

More machine learning resources at **chrissirico.com**:

- algorithm, hyperparameter, feature and threshold selection
- training data setup: group/stratified cross validation and time-based targets
- model bias detection, mitigation and fairness metrics