Regression and Classification Metrics Tutorial

Introduction

Machine learning models need evaluation metrics to measure their performance. The choice of metric depends on your problem type and what you care about most. This tutorial covers the essential metrics for both regression and classification tasks.

Part 1: Regression Metrics

Regression metrics measure how well a model predicts continuous numerical values. They quantify the difference between predicted and actual values.

1. Mean Absolute Error (MAE)

Formula: MAE = $(1/n) \times \Sigma|\text{actual} - \text{predicted}|$

What it means: Average absolute difference between predictions and actual values. Units are the same as your target variable.

Example: If predicting house prices and MAE = $15,000, predictions are off by $15,000 on average.

When to use: When you want an interpretable metric in the original units. Outliers have less impact than in MSE.

---

2. Mean Squared Error (MSE)

Formula: MSE = $(1/n) \times \Sigma(\text{actual} - \text{predicted})^2$

What it means: Average of squared errors. Penalizes larger errors more heavily.

Example: MSE = 225 million (for house prices in dollars squared)

When to use: When larger errors should be penalized more. Mathematically convenient for optimization.

---

3. Root Mean Squared Error (RMSE)

Formula: RMSE = $\sqrt{\text{MSE}}$

What it means: Square root of MSE, bringing the metric back to original units.

Example: RMSE = $15,000 for house prices—directly comparable to MAE but penalizes outliers more.

When to use: Most common metric. Interpretable in original units while penalizing large errors.

---

4. R^2 Score (Coefficient of Determination)

Formula: $R^2 = 1 - (SS_{res} / SS_{tot})$

- $SS_{res} = \Sigma(\text{actual} - \text{predicted})^2$
- $SS_{tot} = \Sigma(\text{actual} - \text{mean})^2$

What it means: Proportion of variance explained by the model. Ranges from 0 to 1 (or negative for very bad models).

Example: $R^2 = 0.85$ means the model explains 85% of the variance in the data.

When to use: When you want a normalized metric that's easy to interpret. Standard in regression reporting.

---

5. Mean Absolute Percentage Error (MAPE)

Formula: $MAPE = (1/n) \times \Sigma|((\text{actual} - \text{predicted}) / \text{actual}) \times 100|$

What it means: Average percentage difference. Useful for comparing across different scales.

Example: MAPE = 5% means predictions are off by 5% on average.

When to use: When the target has different scales or you need percentage-based interpretation. Problematic when actual values are near zero.

Part 2: Classification Metrics

Classification metrics measure how well a model predicts categorical labels. They're based on how predictions compare to actual classes.

Confusion Matrix Foundation

All classification metrics build on the confusion matrix for binary classification:

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

- **TP (True Positive):** Correctly predicted positive cases
- **TN (True Negative):** Correctly predicted negative cases
- **FP (False Positive):** Incorrectly predicted as positive (Type I error)
- **FN (False Negative):** Incorrectly predicted as negative (Type II error)

---

1. Accuracy

Formula: Accuracy = (TP + TN) / (TP + TN + FP + FN)

What it means: Proportion of correct predictions overall.

Example: 95% accuracy means 95 out of 100 predictions are correct.

When to use: When classes are balanced and all errors are equally costly. Misleading with imbalanced classes.

---

2. Precision

Formula: Precision = TP / (TP + FP)

What it means: Of all positive predictions, how many were actually positive? "When we say positive, how often are we right?"

Example: Precision = 0.90 means 90% of spam predictions are actually spam.

When to use: When false positives are costly (spam filtering, loan approvals).

---

3. Recall (Sensitivity)

Formula: Recall = TP / (TP + FN)

What it means: Of all actual positives, how many did we catch? "Did we find the positive cases?"

Example: Recall = 0.85 means we found 85% of actual spam emails.

When to use: When false negatives are costly (disease detection, fraud detection).

---

4. F1 Score

Formula: F1 = 2 × (Precision × Recall) / (Precision + Recall)

What it means: Harmonic mean of precision and recall. Balances both metrics.

Example: F1 = 0.87 provides a single score for comparison.

When to use: When you need a balanced metric and don't want to choose between precision and recall. Standard for imbalanced datasets.

---

5. Specificity (TNR - True Negative Rate)

Formula: Specificity = TNR = TN / (TN + FP)

What it means: True negative rate—of actual negatives, how many did we correctly identify?

Example: Specificity = 0.95 means we correctly identified 95% of non-spam emails.

When to use: When you care about correctly identifying negative cases, especially with imbalanced data.

---

6. TPR (True Positive Rate) / Sensitivity

Formula: TPR = TP / (TP + FN)

What it means: Same as Recall. The proportion of actual positive cases that were correctly identified. Also called Sensitivity.

Example: TPR = 0.85 means we caught 85% of the actual positive cases.

When to use: When you want to know how well the model finds positive instances. Critical in medical diagnosis, fraud detection, and disease screening where missing cases is costly.

Important: TPR is plotted on the y-axis of the ROC curve.

---

7. FPR (False Positive Rate)

Formula: FPR = FP / (FP + TN)

What it means: Proportion of actual negative cases that were incorrectly classified as positive. The inverse of Specificity.

Relationship: FPR = 1 - Specificity

Example: FPR = 0.05 means we incorrectly flagged 5% of negative cases as positive.

When to use: When you want to measure false alarms. Important for evaluating the cost of incorrect positive predictions.

Important: FPR is plotted on the x-axis of the ROC curve.

---

8. ROC-AUC (Area Under the Receiver Operating Characteristic Curve)

What it means: Probability that the model ranks a random positive example higher than a random negative example. Ranges from 0 to 1.

Example: AUC = 0.92 indicates excellent discrimination between classes.

When to use: When you want a threshold-independent metric. Good for imbalanced datasets. Standard in binary classification tasks.

---

9. Precision-Recall Curve and AP (Average Precision)

What it means: Area under the precision-recall curve. Focuses on the positive class performance.

Example: AP = 0.88 summarizes precision-recall tradeoffs.

When to use: With imbalanced datasets where you care more about positive class. Better than AUC for imbalanced problems.

---

10. Log Loss (Cross-Entropy)

Formula: Log Loss = $-(1/n) \times \Sigma[y \times \log(p) + (1-y) \times \log(1-p)]$

What it means: Penalizes confident incorrect predictions heavily. Measures probability calibration.

Example: Log Loss = 0.25 (lower is better, 0 is perfect).

When to use: When you want to evaluate probability predictions, not just class predictions. Important for probabilistic models.

---

11. Confusion Matrix for Multi-class

For problems with 3+ classes, calculate precision, recall, and F1 for each class using one-vs-rest approach. Then average them:

- **Macro average:** Simple average across all classes
- **Weighted average:** Average weighted by class frequency
- **Micro average:** Calculate metrics globally by counting total TP, FN, FP

Choosing the Right Metric

For Regression:

- **Want interpretability?** → Use MAE
- **Want to penalize outliers?** → Use RMSE
- **Want normalized comparison?** → Use R^2
- **Want percentage-based?** → Use MAPE

For Classification:

- **Balanced classes, all errors equal?** → Use Accuracy
- **False positives costly?** → Use Precision
- **False negatives costly?** → Use Recall
- **Need one balanced metric?** → Use F1 or ROC-AUC
- **Imbalanced dataset?** → Use F1, ROC-AUC, or Precision-Recall Curve
- **Probabilistic predictions matter?** → Use Log Loss

Summary Table

Metric	Type	Range	Interpretation
MAE	Regression	0 to ∞	Lower is better
RMSE	Regression	0 to ∞	Lower is better
R^2	Regression	-∞ to 1	Higher is better (1 is perfect)
Accuracy	Classification	0 to 1	Higher is better
Precision	Classification	0 to 1	Higher is better
Recall	Classification	0 to 1	Higher is better
F1	Classification	0 to 1	Higher is better
ROC-AUC	Classification	0 to 1	Higher is better (0.5 is random)