ML Performance Metrics: Complete Interview Guide

CLASSIFICATION METRICS

1. Confusion Matrix (Foundation)

```
                Predicted
              Pos    Neg
Actual  Pos   TP     FN
        Neg   FP     TN
```

- **TP**: Correctly predicted positive
- **FN**: Missed positives (Type II error)
- **FP**: False alarm (Type I error)
- **TN**: Correctly predicted negative

2. Accuracy

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

When to Use: Balanced datasets **When NOT to Use**: Imbalanced data (e.g., fraud detection 99% negative) **Example Issue**: 99% accuracy by predicting all negative in 1% fraud case

3. Precision (Positive Predictive Value)

Formula: TP / (TP + FP)

Meaning: Of all predicted positives, how many are actually positive? **When to Use**: When false positives are costly **Examples**:

- Spam detection (don't want important emails marked spam)
- Medical diagnosis (don't want false cancer positives causing unnecessary treatment)

4. Recall (Sensitivity, True Positive Rate)

Formula: TP / (TP + FN)

Meaning: Of all actual positives, how many did we catch? **When to Use**: When false negatives are costly **Examples**:

- Cancer screening (must catch all cases)
- Fraud detection (can't miss fraudulent transactions)
- Security systems (must detect all threats)

5. F1-Score

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

Meaning: Harmonic mean of precision and recall **When to Use**: Need balance, imbalanced datasets **Range**: 0 to 1 (higher is better) **Why Harmonic Mean**: Penalizes extreme values (if either P or R is low, F1 is low)

6. F-Beta Score

Formula: $(1 + \beta^2) \times (\text{Precision} \times \text{Recall}) / (\beta^2 \times \text{Precision} + \text{Recall})$

- **F0.5**: Weights precision higher (FP costly)
- **F1**: Equal weight
- **F2**: Weights recall higher (FN costly)

7. Specificity (True Negative Rate)

Formula: TN / (TN + FP)

Meaning: Of all actual negatives, how many did we correctly identify? **Use Case**: When correctly identifying negatives matters (healthy patients in medical screening)

8. ROC Curve (Receiver Operating Characteristic)

Plot: TPR (Recall) vs FPR at various thresholds **FPR**: FP / (FP + TN) = 1 - Specificity

Interpretation:

- Diagonal line = random classifier
- Top-left corner = perfect classifier
- Area under curve (AUC-ROC) measures overall performance

When to Use: Compare models, threshold-independent evaluation

9. AUC-ROC (Area Under ROC Curve)

Range: 0 to 1

- **0.5**: Random guessing
- **0.7-0.8**: Fair
- **0.8-0.9**: Good
- **>0.9**: Excellent

Advantages:

- Threshold-independent
- Works for imbalanced data **When NOT to Use**: When you care about performance at specific threshold

10. Precision-Recall Curve

Plot: Precision vs Recall at various thresholds

When to Use: Imbalanced datasets (better than ROC) **Why**: ROC can be overly optimistic with imbalanced data **AUC-PR**: Area under PR curve

11. Log Loss (Cross-Entropy Loss)

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

Meaning: Measures probability estimates, not just predictions **Range**: 0 to ∞ (lower is better) **Use**: When you need calibrated probabilities (not just class labels)

12. Cohen's Kappa

Formula: $(P_o - P_e) / (1 - P_e)$

- Po = observed agreement
- Pe = expected agreement by chance

Range: -1 to 1

- **<0**: Worse than random
- **0**: Random agreement
- **0.4-0.6**: Moderate
- **>0.8**: Strong

Use: Multi-class, accounts for chance agreement

13. Matthews Correlation Coefficient (MCC)

Formula: $(TP \times TN - FP \times FN) / \sqrt{[(TP+FP)(TP+FN)(TN+FP)(TN+FN)]}$

Range: -1 to 1 (1 = perfect, 0 = random) **Advantage**: Works well with imbalanced data, considers all confusion matrix values **Use**: Often considered one of the best single metrics for binary classification

---

REGRESSION METRICS

1. Mean Absolute Error (MAE)

Formula: $(1/n) \times \Sigma|y_i - \hat{y}_i|$

Meaning: Average absolute difference between predictions and actuals **Units**: Same as target variable **Pros**: Robust to outliers, interpretable **Cons**: Doesn't penalize large errors heavily

2. Mean Squared Error (MSE)

Formula: $(1/n) \times \Sigma(y_i - \hat{y}_i)^2$

Pros: Penalizes large errors more (squared) **Cons**: Not in original units, sensitive to outliers **Use**: When large errors are particularly bad

3. Root Mean Squared Error (RMSE)

Formula: $\sqrt{MSE}$

Pros: Same units as target, penalizes large errors **Cons**: Sensitive to outliers **Most Common**: Standard choice for many regression tasks

4. R^2 (R-Squared / Coefficient of Determination)

Formula: $1 - (SS\_res / SS\_tot)$

- $SS\_res = \Sigma(y_i - \hat{y}_i)^2$
- $SS\_tot = \Sigma(y_i - \bar{y})^2$

Range: $-\infty$ to 1

- **1**: Perfect predictions
- **0**: Model as good as mean baseline
- **<0**: Worse than predicting mean

Meaning: Proportion of variance explained **Issue**: Always increases with more features (even irrelevant ones)

5. Adjusted R^2

Formula: $1 - [(1-R^2)(n-1)/(n-p-1)]$

- n = samples
- p = features

Advantage: Penalizes adding irrelevant features **Use**: Feature selection, model comparison

6. Mean Absolute Percentage Error (MAPE)

Formula: $(100/n) \times \Sigma|((y_i - \hat{y}_i)/y_i)|$

Pros: Scale-independent, interpretable as percentage **Cons**:

- Undefined when $y_i = 0$
- Asymmetric (penalizes over-predictions more)
- Biased towards under-predictions

7. Symmetric MAPE (SMAPE)

Formula: $(100/n) \times \Sigma|(y_i - \hat{y}_i)|/((|y_i| + |\hat{y}_i|)/2)$

Advantage: More symmetric than MAPE **Range**: 0% to 200%

8. Huber Loss

Formula:

- $|y - \hat{y}|^2/2$ if $|y - \hat{y}| \leq \delta$
- $\delta(|y - \hat{y}| - \delta/2)$ otherwise

Use: Robust to outliers (MSE for small errors, MAE for large) **Parameter**: δ controls transition point

---

RANKING METRICS

1. Mean Average Precision (MAP)

Use: Information retrieval, recommendation systems **Meaning**: Average precision across multiple queries **Formula**: (1/Q) × Σ AP(q) for all queries

2. Normalized Discounted Cumulative Gain (NDCG)

Use: Search ranking, recommendations **Meaning**: Measures ranking quality considering position **Key**: Higher-ranked relevant items contribute more

3. Mean Reciprocal Rank (MRR)

Formula: (1/Q) × Σ(1/rank_of_first_relevant_item) **Use**: Search engines (how quickly do we show relevant result)

CLUSTERING METRICS

1. Silhouette Score

Range: -1 to 1

- **Close to 1**: Well-clustered
- **0**: Overlapping clusters
- **Negative**: Wrong cluster assignment

Formula: (b - a) / max(a, b)

- a = avg distance within cluster
- b = avg distance to nearest cluster

2. Davies-Bouldin Index

Lower is better Meaning: Average similarity between each cluster and its most similar cluster

3. Calinski-Harabasz Index (Variance Ratio)

Higher is better Meaning: Ratio of between-cluster to within-cluster variance

4. Inertia (Within-Cluster Sum of Squares)

Lower is better Use: Elbow method for choosing K in K-means **Issue**: Always decreases with more clusters

---

PROBABILISTIC METRICS

1. Brier Score

Formula: $(1/n) \times \Sigma(p_i - y_i)^2$ **Range**: 0 to 1 (lower is better) **Use**: Measures calibration of probability predictions

2. Expected Calibration Error (ECE)

Meaning: Difference between confidence and accuracy **Use**: Assessing probability calibration

---

ADVANCED/SPECIALIZED METRICS

1. Lift

Formula: (TP rate in model) / (TP rate in baseline) **Use**: Marketing, campaign effectiveness

2. Gini Coefficient

Formula: 2 × AUC - 1 **Range**: 0 to 1 **Use**: Credit scoring, ranking quality

3. Kolmogorov-Smirnov (KS) Statistic

Formula: max(TPR - FPR) across all thresholds **Use**: Credit scoring, comparing distributions

4. Business Metrics

- **Cost-Benefit Analysis**: Assign costs to FP/FN
- **Revenue Impact**: Direct business value
- **Customer Lifetime Value**: Long-term impact

INTERVIEW SCENARIOS

Q: Imbalanced dataset (1% positive class). Which metric?

Answer: Not accuracy! Use:

- Precision-Recall curve and AUC-PR
- F1-score or F2-score (if FN costly)
- MCC (Matthews Correlation Coefficient) **Why not ROC-AUC**: Can be misleading with extreme imbalance

Q: Medical diagnosis. Which metric?

Answer: Recall (must catch all diseases, FN very costly) **Trade-off**: Accept more false positives (FP) **Also consider**: Precision-Recall trade-off, F2-score

Q: Spam detection. Which metric?

Answer: Precision (can't mark important emails as spam) **Trade-off**: Some spam might get through (FN acceptable) **Also consider**: F0.5 score

Q: Why not always use accuracy?

Answers:

1. Misleading with imbalanced data
2. Doesn't distinguish between FP and FN costs

3. Threshold-dependent
4. Doesn't reflect business value

Q: Precision vs Recall trade-off

High Threshold → High Precision, Low Recall (conservative) **Low Threshold** → Low Precision, High Recall (aggressive) **Balance**: Use F1-score or choose based on cost of errors

Q: When to use MAE vs RMSE?

MAE:

- Robust to outliers
- All errors treated equally **RMSE**:
- Penalize large errors more
- When outliers matter **Rule**: If outliers are errors, use MAE. If outliers are important signals, use RMSE.

Q: R^2 is 0.95. Is the model good?

Not necessarily!

- Could be overfit
- Check on validation/test set
- Could be data leakage
- Need to check residual plots
- Compare with baseline

Q: How to choose metric for business problem?

Process:

1. Understand business cost of FP vs FN
2. Consider data imbalance
3. Determine if you need probabilities or classes
4. Align metric with business KPI
5. Use multiple metrics (never rely on one)

METRIC SELECTION GUIDE

Classification - Balanced Data

Primary: Accuracy, F1-Score **Secondary**: ROC-AUC, Precision, Recall

Classification - Imbalanced Data

Primary: F1-Score, Precision-Recall AUC, MCC **Secondary**: Class-weighted metrics **Avoid**: Accuracy

Classification - Cost-Sensitive

Primary: Custom cost matrix **Secondary**: Precision (FP costly) or Recall (FN costly)

Regression - General

Primary: RMSE (if outliers matter) or MAE (robust) **Secondary**: R^2, Adjusted R^2

Regression - Scale-Independent

Primary: MAPE (if no zeros) **Secondary**: SMAPE

Ranking/Recommendations

Primary: NDCG, MAP **Secondary**: MRR, Precision@K

Clustering

Primary: Silhouette Score **Secondary**: Davies-Bouldin, Calinski-Harabasz **Use with**: Domain knowledge (metrics alone insufficient)

KEY FORMULAS TO MEMORIZE

1. **Precision**: TP / (TP + FP)
2. **Recall**: TP / (TP + FN)
3. **F1**: 2PR / (P + R)
4. **Accuracy**: (TP + TN) / Total
5. **MAE**: $(1/n)\ \Sigma|y - \hat{y}|$
6. **MSE**: $(1/n)\ \Sigma(y - \hat{y})^2$
7. **RMSE**: $\sqrt{MSE}$
8. **R^2**: 1 - SS_res/SS_tot

COMMON MISTAKES TO AVOID

❌ Using accuracy for imbalanced data ❌ Ignoring business context when choosing metric ❌ Relying on single metric ❌ Not checking calibration for probability predictions ❌ Comparing metrics across different datasets/scales ❌ Optimizing for wrong metric (doesn't align with goal) ❌ Not considering class weights in imbalanced scenarios ❌ Using training metrics instead of validation/test

✅ Always use validation/test set for evaluation ✅ Consider multiple metrics ✅ Align metrics with business objectives ✅ Check metric assumptions (e.g., MAPE needs $y \neq 0$) ✅ Use cross-validation for robust estimates ✅ Plot curves (ROC, PR) not just summary statistics