AWS Certified Machine Learning Engineer - Associate Exam Guide

# Section 2.3 — Analyze Model Performance (In-Depth)

This section covers techniques and tools for evaluating model performance. It emphasizes proper use of evaluation metrics, detecting overfitting/underfitting, and AWS tools for monitoring and explainability.

## 1. Evaluation Metrics

• Classification: Accuracy, Precision, Recall, F1-score, ROC-AUC, confusion matrix

• Regression: MSE, RMSE, MAE, R² (coefficient of determination)

• Clustering: Silhouette score, inertia, Davies–Bouldin index

• Recommendation: Precision@K, Recall@K, NDCG (Normalized Discounted Cumulative Gain)

• Multi-class problems: micro vs macro averaging for precision/recall/F1

## 2. Detecting Overfitting & Underfitting

• Overfitting: training accuracy >> validation accuracy, high variance

• Underfitting: poor performance on both training & validation, high bias

• Use validation curves and learning curves to detect bias/variance tradeoffs

• Apply regularization, early stopping, or feature engineering to mitigate issues

## 3. Baseline & Benchmarking

• Establish baselines: simple models (logistic regression, mean prediction)

• Compare advanced models against baseline to validate improvement

• Use reproducible experiments to ensure consistent results

## 4. AWS Tools

• SageMaker Clarify → bias detection, feature importance (SHAP values), explainability reports

• SageMaker Model Monitor → detect data drift and performance degradation in production

• SageMaker Debugger → detect convergence issues (vanishing gradients, exploding loss)

• SageMaker Experiments → track, compare, and organize multiple training runs

## 5. Shadow Testing

• Shadow mode → deploy new model alongside production without affecting live traffic

• Compare predictions from new and production models

• A/B testing to evaluate real-world performance before rollout

## ✅ Exam Tips

• Always choose the right metric for the problem type (classification vs regression vs clustering)

• Confusion matrix is key for classification understanding

• Use Clarify for fairness & interpretability questions

• Debugger is for convergence and training anomaly detection

• Shadow testing = safe way to validate new models before production rollout