Evaluation Metrics for Supervised ML Models

Model Evaluation Overview

Learning Objectives

By the end of this lesson, students will be able to:

- Explain why evaluating machine learning models is essential for performance assessment and decision-making.
- Identify key factors that influence the choice of evaluation metrics.
- Understand the trade-offs involved in different metrics and how they impact model performance.

Why Model Evaluation Matters

Building an ML model is only half the battle—evaluating its performance is **critical** to ensuring it makes accurate predictions in real-world applications. Without proper evaluation, models may:

- Provide misleading results.
- Overfit or underfit the data.
- Fail to meet business objectives.

Key Considerations in Model Evaluation

Choosing the right evaluation metric depends on:

- **The type of ML task**: Classification vs. regression models require different metrics.
- **The data characteristics**: Class imbalance, outliers, and noise can affect how useful certain metrics are.
- **The business impact**: A false negative in fraud detection is worse than a false positive in a spam filter.

Choosing the Right Evaluation Metric

Different tasks require different approaches to measuring success. Below is a **high-level guide** to common ML tasks and their corresponding evaluation metrics:

ML Task	Common Metrics	When to Use
Classification	Accuracy, Precision, Recall, F1-score	When predicting categories (spam detection, fraud detection)
Regression	Mean Squared Error (MSE), R^2, MAE	When predicting continuous values (house prices, sales forecasts)

Understanding Trade-offs

- **Accuracy vs. Precision/Recall**: Accuracy may not be useful in imbalanced datasets (e.g., detecting rare diseases).
- **MSE vs. MAE**: MSE penalizes large errors more than MAE, which may or may not be desirable.

When Metrics Can Be Misleading

Understanding when metrics might be misleading is crucial for proper model evaluation:

1. **High Accuracy in Imbalanced Datasets**

- **Scenario**: A model detecting a rare disease (1% of cases are positive)
- **Misleading Result**: 99% accuracy by simply predicting “no disease” for every patient
- **Better Metric**: Precision, recall, or F1-score would reveal the model’s true performance

2. **R-squared (R²) in Non-Linear Relationships**

   - **Scenario**: Predicting stock prices with a strong cyclical pattern
   - **Misleading Result**: Low R² despite good predictions due to non-linear patterns
   - **Better Approach**: Consider non-linear metrics or transform data appropriately

3. **Mean Squared Error (MSE) with Outliers**

   - **Scenario**: Predicting house prices with some luxury mansions in the dataset
   - **Misleading Result**: High MSE despite good predictions for typical houses
   - **Better Metric**: Mean Absolute Error (MAE) or robust regression metrics

4. **Perfect Precision but Poor Recall**

   - **Scenario**: Fraud detection system that only flags extremely obvious cases
   - **Misleading Result**: 100% precision but missing most actual fraud cases
   - **Better Metric**: F1-score or balanced accuracy

5. **Cross-Validation Scores on Temporal Data**

   - **Scenario**: Time series prediction with random cross-validation
   - **Misleading Result**: Good CV scores despite using future data to predict past events
   - **Better Approach**: Time-based validation splits

Best Practices to Avoid Misleading Metrics

1. **Always Consider Multiple Metrics**

   - Don’t rely on a single metric
   - Choose metrics that align with business objectives
   - Consider the cost of different types of errors

2. **Understand Your Data Distribution**

   - Check for class imbalance
   - Look for outliers and their impact
   - Consider the temporal nature of data if applicable

3. **Validate Against Business Context**
    - Consult domain experts
    - Compare with baseline models
    - Test on real-world scenarios

Quick Knowledge Check

Question: You are building a model to detect fraudulent transactions. Which metric would be most appropriate to use and why?