Trong học máy, khi làm việc với các tập dữ liệu không cân bằng (skewed datasets), nơi một lớp chiếm đa số so với lớp khác, việc sử dụng các Error Metrics (đánh giá hiệu suất) phù hợp là quan trọng để đảm bảo rằng mô hình được đánh giá một cách chính xác và không bị ảnh hưởng quá mức bởi sự mất cân bằng trong dữ liệu. Dưới đây là một số Error Metrics thường được sử dụng trong trường hợp tập dữ liệu không cân bằng:

**1. Accuracy (Độ Chính Xác):**

* **Ưu điểm:** Dễ hiểu và tính toán.
* **Nhược điểm:** Không phản ánh chính xác trên dữ liệu không cân bằng khi một lớp chiếm đa số.

**2. Precision (Độ Chính Xác Cao):**

* **Ưu điểm:** Đánh giá khả năng của mô hình trong việc không làm giả mạo.
* **Nhược điểm:** Có thể cao khi số lượng dự đoán đúng dương (True Positives) nhỏ so với tổng số dự đoán dương (True Positives + False Positives).

**3. Recall (Tỉ Lệ Nhận Dạng):**

* **Ưu điểm:** Đánh giá khả năng của mô hình trong việc phát hiện tất cả các trường hợp dương thực sự.
* **Nhược điểm:** Có thể thấp khi số lượng dự đoán đúng dương nhỏ so với tổng số thực sự dương (True Positives + False Negatives).

**4. F1 Score (Điểm F1):**

* **Ưu điểm:** Kết hợp cả Precision và Recall.
* **Nhược điểm:** Còn phụ thuộc vào tình hình cụ thể (có thể không phản ánh tốt trên cả hai thước đo nếu một trong chúng thấp).

**5. Area Under the Receiver Operating Characteristic (ROC-AUC):**

* **Ưu điểm:** Đánh giá khả năng phân biệt giữa các lớp.
* **Nhược điểm:** Có thể không phản ánh chính xác trên dữ liệu không cân bằng nếu một lớp chiếm đa số.

**6. Precision-Recall Curve:**

* **Ưu điểm:** Đánh giá Precision và Recall ở nhiều ngưỡng khác nhau.
* **Nhược điểm:** Đặc biệt hữu ích khi một lớp chiếm đa số và cần tập trung vào việc phát hiện trường hợp dương, như trong các tình huống không cân bằng.

Khi làm việc với tập dữ liệu không cân bằng, quan trọng để chọn metrics phù hợp với mục tiêu cụ thể của bạn và không dựa quá mức vào Accuracy, vì nó có thể bị nghiệt.