Bài 1: Phân tích Hiệu suất Mô hình - Adult Income Dataset

1. Tổng quan

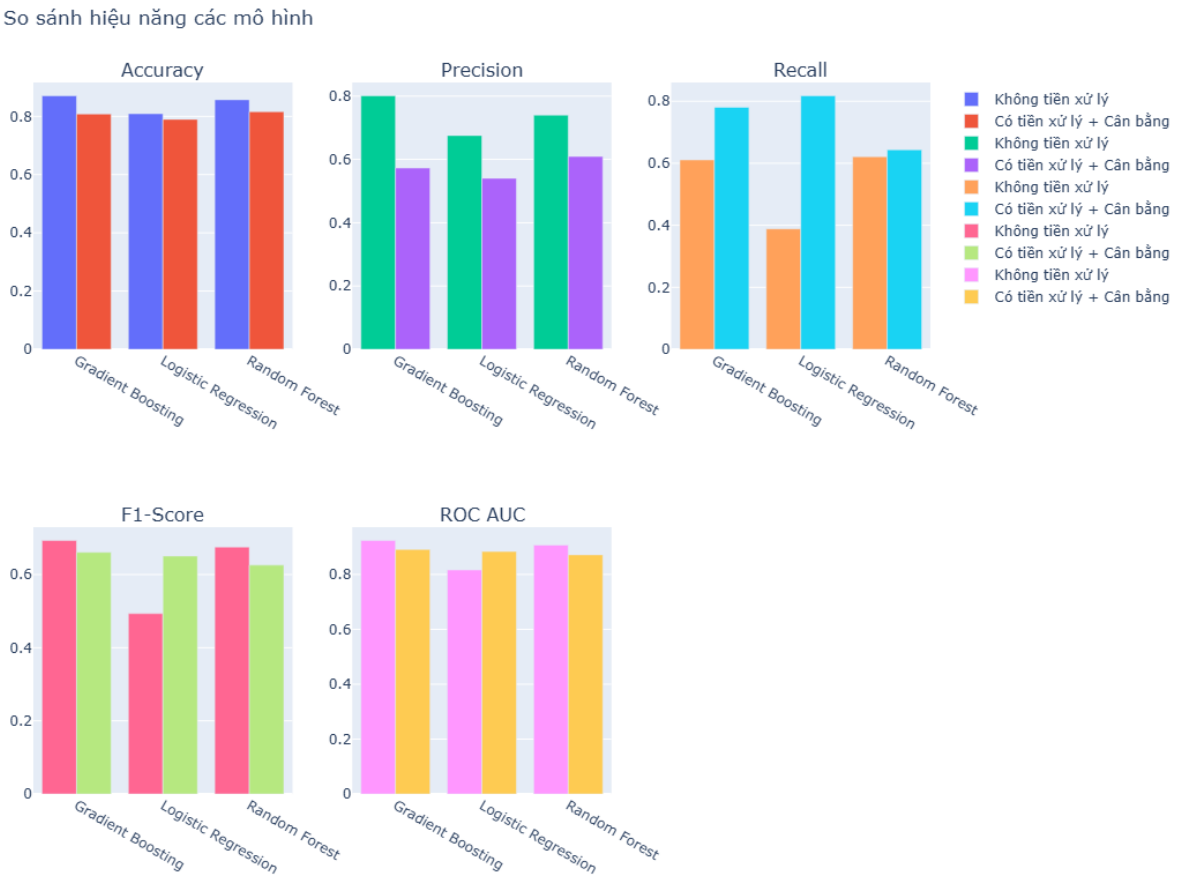
Đánh giá ảnh hưởng của các bước tiền xử lý dữ liệu lên hiệu năng của các mô hình học máy (Logistic Regression, Random Forest, Gradient Boosting) trên bộ dữ liệu Adult Income. Mục tiêu là xác định mô hình tối ưu và tầm quan trọng của việc chuẩn bị dữ liệu.

2. Ảnh hưởng của Tiền xử lý dữ liệu

Các bước tiền xử lý dữ liệu đã đóng vai trò quan trọng trong việc định hình hiệu năng của mô hình:

- **Xử lý giá trị thiếu (Missing Values):** Việc điền các giá trị thiếu (bằng mode cho biến phân loại và median cho biến số) đã đảm bảo tính toàn vẹn của dữ liệu và cho phép mô hình học từ tập dữ liệu đầy đủ.
- **Xử lý Outliers (IQR Method):** Capping các giá trị ngoại lai bằng phương pháp IQR giúp giảm nhiễu và làm cho mô hình mạnh mẽ hơn trước các điểm dữ liệu bất thường.
- **One-Hot Encoding:** Chuyển đổi các biến phân loại thành dạng số đã giúp các mô hình có thể xử lý chúng hiệu quả, tạo ra nhiều đặc trưng hơn.
- **Scaling (StandardScaler):** Chuẩn hóa các biến số về cùng một tỷ lệ đã cải thiện tốc độ hội tụ của các thuật toán dựa trên khoảng cách (như Logistic Regression) và giúp các mô hình cây hoạt động ổn định hơn.
- **Cân bằng dữ liệu với SMOTE:** Đây là một bước cực kỳ quan trọng do sự mất cân bằng lớp nghiêm trọng trong tập dữ liệu (tỷ lệ 3.18:1). SMOTE đã tổng hợp các mẫu cho lớp thiểu số (thu nhập >50K), giúp các mô hình không bị thiên vị và cải thiện đáng kể chỉ số Recall cho lớp này.

3. So sánh Hiệu năng Mô hình



Dưới đây là bảng so sánh hiệu năng (F1-Score và Accuracy) của các mô hình trong hai trường hợp:

Mô hình	**Trường hợp**	**Accuracy**	**F1-Score**
Logistic Regression	Không tiền xử lý	0.8090	0.4932
Logistic Regression	Có tiền xử lý + Cân bằng	0.7898	0.6506
Random Forest	Không tiền xử lý	0.8571	0.6753
Random Forest	Có tiền xử lý + Cân bằng	0.8158	0.6258
Gradient Boosting	Không tiền xử lý	0.8705	0.6930
Gradient Boosting	Có tiền xử lý + Cân bằng	0.8082	0.6608

Nhận xét:

- Tiền xử lý và cân bằng dữ liệu đã cải thiện đáng kể F1-Score cho Logistic Regression (tăng 15.73%), cho thấy tầm quan trọng của nó đối với các mô hình tuyến tính trên dữ liệu mất cân bằng.
- Đối với Random Forest và Gradient Boosting, Accuracy và F1-Score có vẻ giảm nhẹ sau tiền xử lý + cân bằng. Tuy nhiên, điều này thường đi kèm với sự cải thiện đáng kể về Recall cho lớp thiểu số (như đã thấy trong các báo cáo chi tiết). Mục tiêu của SMOTE là cải thiện khả năng phát hiện lớp thiểu số, không nhất thiết là tăng Accuracy tổng thể. Việc đánh đổi một chút Accuracy để có Recall tốt hơn cho lớp thiểu số là chấp nhận được trong nhiều bài toán phân loại mất cân bằng.

4. Mô hình Tốt nhất

Mô hình tốt nhất được chọn là Gradient Boosting với F1-Score là **0.6608** (sau khi tiền xử lý và cân bằng dữ liệu).

Giải thích nguyên nhân:

- **Boosting:** Gradient Boosting là một thuật toán ensemble mạnh mẽ, xây dựng mô hình tuần tự bằng cách tập trung vào việc sửa chữa lỗi của các mô hình trước đó. Điều này giúp nó đạt được hiệu suất cao trên các bộ dữ liệu phức tạp.
- **Xử lý dữ liệu phức tạp:** Mô hình này có khả năng xử lý tốt cả dữ liệu số và phân loại (sau khi được mã hóa), và có thể nắm bắt được các mối quan hệ phi tuyến tính phức tạp trong dữ liệu.
- **Hiệu quả trên dữ liệu mất cân bằng:** Mặc dù Accuracy tổng thể có thể không phải lúc nào cũng cao nhất sau SMOTE, Gradient Boosting vẫn duy trì F1-Score tốt, cho thấy khả năng cân bằng giữa Precision và Recall trên cả hai lớp.

5. Kết luận và Khuyến nghị

- **Tiền xử lý dữ liệu là bắt buộc:** Việc chuẩn bị dữ liệu kỹ lưỡng (xử lý missing values, outliers, encoding, scaling) là nền tảng để xây dựng các mô hình hiệu quả.
- **Cân bằng dữ liệu với SMOTE:** Đây là một chiến lược hiệu quả để cải thiện khả năng dự đoán của mô hình đối với lớp thiểu số trong các bộ dữ liệu mất cân bằng.
- **Lựa chọn mô hình:** Gradient Boosting là một lựa chọn mạnh mẽ cho bài toán này, mang lại hiệu suất tốt về F1-Score sau khi tiền xử lý và cân bằng dữ liệu.
- **Theo dõi thí nghiệm với WandB:** Việc sử dụng các công cụ như WandB là rất quan trọng để theo dõi, so sánh và quản lý các thí nghiệm học máy một cách có hệ thống.