1. Thu thập dữ liệu:

Thu thập dữ liệu từ các nguồn khác nhau như hệ thống ghi log, cơ sở dữ liệu khách hàng, hoặc dữ liệu từ các dịch vụ bên ngoài.

Dữ liệu thu thập bao gồm các biến như mã người dùng, lịch sử mua hàng, thời gian trung bình mỗi phiên truy cập, danh mục sản phẩm duyệt qua, đánh giá sản phẩm, tần suất mua hàng và thông tin demographic.

2. Tiền xử lý dữ liệu:

Kiểm tra và xử lý dữ liệu bị thiếu.

Mã hóa các biến categorical như giới tính và vị trí địa lý thành dạng số để có thể sử dụng trong mô hình.

Chuẩn hóa các biến số để đảm bảo chúng có cùng phạm vi.

3. Phân chia tập dữ liệu:

Phân chia tập dữ liệu thành tập huấn luyện và tập kiểm thử để đánh giá hiệu suất của mô hình.

4. Xây dựng mô hình:

Lựa chọn một mô hình dự đoán phù hợp như Random Forest, Gradient Boosting, hoặc Neural Networks dựa trên yêu cầu và tính chất của dữ liệu.

Huấn luyện mô hình trên tập dữ liệu huấn luyện và điều chỉnh siêu tham số để tối ưu hóa hiệu suất.

5. Đánh giá mô hình:

Sử dụng các phương pháp đánh giá như cross-validation, holdout validation hoặc bootstrap để đánh giá độ chính xác của mô hình trên tập kiểm thử.

Đánh giá các độ đo như precision, recall, F1-score để hiểu rõ hơn về hiệu suất của mô hình trên các lớp sản phẩm.

6. Triển khai và theo dõi:

Triển khai mô hình vào môi trường sản xuất và giám sát hiệu suất của nó.

Thực hiện các cập nhật định kỳ và tối ưu hóa mô hình khi cần thiết để đảm bảo tính hiệu quả và tính liên tục của mô hình.