Machine Learning

Bộ môn Khoa học dữ liệu
Khoa Công nghệ thông tin
Trường Đại học Công nghiệp thành phố Hồ Chí Minh-IUH

Cho bộ dữ liệu ở file đính kèm. Bộ dữ liệu mô tả hành vi sử dụng của khoảng 9000 chủ thẻ tín dụng đang hoạt động trong 6 tháng với 18 hành vi. Nhiệm vụ chính của bạn là xác định phân khúc khách hàng để xác định các chiến lược tiếp thị cho phù hợp. Mô tả chi tiết về bộ dữ liệu này như sau:

CUST_ID : ID của chủ thẻ tín dụng
BALANCE: Số dư còn lại trong tài khoản để mua hàng
BALANCE_FREQUENCY : Tần suất cập nhật Số dư, điểm từ 0 đến 1 (1 = cập nhật thường xuyên, 0 = không cập nhật thường xuyên)
PURCHASES: Số lượng mua hàng được thực hiện từ tài khoản
ONEOFF_PURCHASES : Số tiền mua tối đa được thực hiện trong một lần
INSTALLMENTS_PURCHASES : Số tiền mua trả góp
CASH_ADVANCE : Tiền mặt người dùng ứng trước
PURCHASES_FREQUENCY : Tần suất mua hàng được thực hiện, điểm từ 0 đến 1 (1=mua thường xuyên, 0=không mua thường xuyên)
ONEOFFPURCHASESFREQUENCY: Tần suất mua hàng diễn ra một lần (1=mua thường xuyên, 0=không mua thường xuyên)
PURCHASESINSTALLMENTSFREQUENCY: Tần suất mua hàng trả góp được thực hiện như thế nào (1= thực hiện thường xuyên, 0=không thực hiện thường xuyên)
CASHADVANCEFREQUENCY : Tần suất thanh toán tiền mặt trước
CASHADVANCETRX : Số lượng giao dịch được thực hiện bằng "Cash in Advanced"
PURCHASES_TRX : Số lượng giao dịch mua hàng được thực hiện
CREDIT_LIMIT : Hạn mức thẻ tín dụng cho người dùng
PAYMENTS: Số tiền thanh toán được thực hiện bởi người dùng
MINIMUM_PAYMENTS : Số tiền thanh toán tối thiểu do người dùng thực hiện
PRCFULLPAYMENT : Phần trăm số tiền thanh toán đầy đủ do người dùng thanh toán
TENURE : Thời hạn sử dụng dịch vụ thẻ tín dụng của người sử dụng

hãy thực hiện các yêu cầu sau:

1. Sử dụng thống kê mô tả, mô tả về bộ dữ liệu trên với min, max, std, avg,…
2. Trực quan hóa dữ liệu với các biểu đồ grid line, box, histogram, và scatter matrix
3. Cho biết những dữ liệu còn thiếu và đề xuất cách xử lý dữ liệu thiếu đó một cách tự động, lưu bộ dữ liệu đã xử lý dữ liệu thiếu sử dụng cho các ý tiếp theo.

4. Hãy xử lý các ngoại lệ (Outliers).
5. Chuẩn hóa dữ liệu đầu vào (Normalizing).
6. Khai phá dữ liệu trên và đề xuất các đặc trưng quan trọng để thực hiện gom nhóm phân khúc khách hàng.
7. Đề xuất các nhóm phân khúc khách hàng và giải thích lý do lựa chọn này.
8. Sử dụng mô hình Kmean để gom nhóm phân khúc khách hàng, trực quan hóa kết quả đạt được.
9. Sử dụng mô hình Kmean++ để gom nhóm phân khúc khách hàng, trực quan hóa kết quả đạt được.
10. Sử dụng mô hình BFR để gom nhóm phân khúc khách hàng, trực quan hóa kết quả đạt được.
11. Sử dụng mô hình CURE để gom nhóm phân khúc khách hàng, trực quan hóa kết quả đạt được.
12. Nếu yêu cầu sử dụng mô hình Hierarchical để gom nhóm phân khúc khách hàng, bạn hãy đề xuất cách thực hiện và thực hiện yêu cầu này, trực quan hóa kết quả đạt được.