1 Làm sạch và khai phá dữ liệu

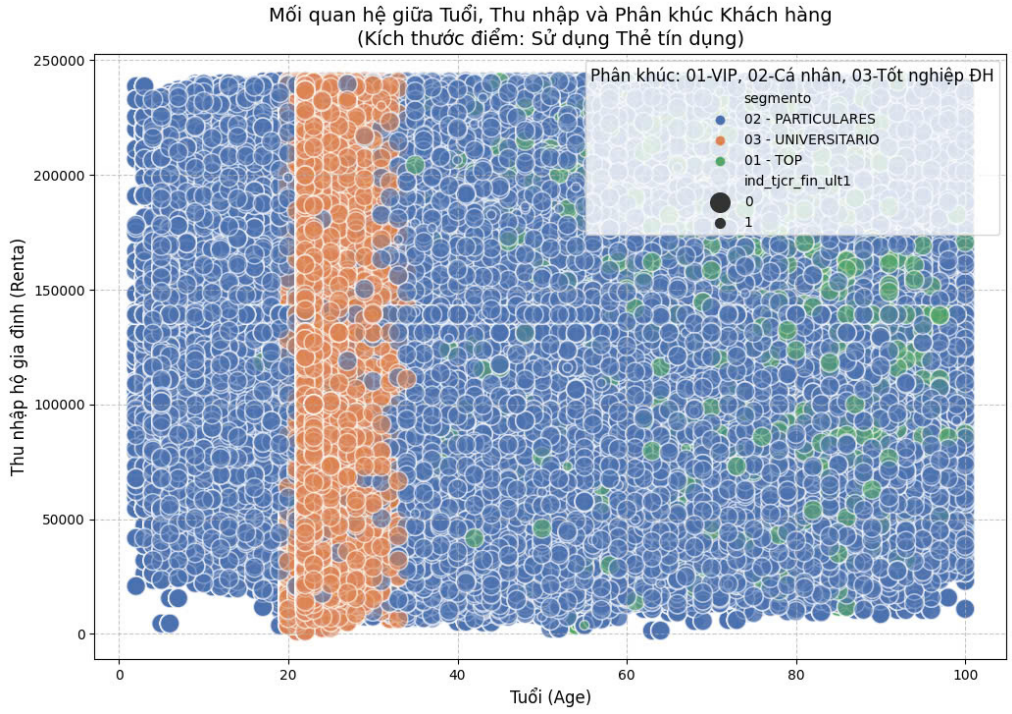
Bộ dữ liệu có kích thướng 2.13 GB với hơn 13 triệu bản ghi và 48 thuộc tính.
Các thao tác xử lý dữ liệu:

- Kiểm tra trùng lặp
- Kiểm tra và xử lý giá trị null
- Loại bỏ hai cột *conyuemp* và *ult_fec_cli_1t* khi có lần lượt 99.986752% và 99.818330% bản ghi mang giá trị null
- Tiếp theo, cột *renta* có 20.475648% bản ghi mang giá trị null ⇒ thay bằng các giá trị trung vị.
- Các cột còn lại đa phần là không có giá trị null hoặc tỉ là giá trị null là rất nhỏ. Những cột này ta sẽ xem xét thay bằng giá trị mode.

Kiểm tra và xử lý giá trị ngoại lai:

- age: Kiểm tra giá trị âm (loại bỏ nếu có), phân nhóm: young (0-18), adult (19-35), middle-age (36-50), senior (51-100).
- antiguedad: Kiểm tra giá trị âm, phân nhóm: New (0-12 tháng), Medium (13-24 tháng), Old (25-60 tháng), Veteran (>60 tháng).
- renta: Dùng Boxplot để xử lý giá trị ngoại lai.

Các thuộc tính sau được loại bỏ vì không mang lại giá trị đáng kể cho việc dự đoán nhu cầu hoặc hành vi khách hàng: ncodpers, fecha_dato, ind_empleado, conyuemp, ult_fec_cli_1t, tipodom, indfall, pairs_residencia.



Ngân hàng nên tập trung tiếp thị thẻ tín dụng cho nhóm VIP và Individuals có thu nhập cao, độ tuổi từ 40-80.
Nhóm Universitarios có thể là mục tiêu cho các sản phẩm tài chính cơ bản như tài khoản hiện tại hoặc tiết kiệm trước khi mở rộng sang thẻ tín dụng.

2 Quy Trình Feature Engineering

2.1 Xử Lý Đặc Trưng Hiện Có

Quá trình xử lý đặc trưng ban đầu giúp loại bỏ những cột không cần thiết, mã hóa dữ liệu để phù hợp với mô hình và lọc ra các sản phẩm ít phổ biến:

- **Loại bỏ cột không cần thiết:** Một số cột như `fecha_dato`, `ind_empleado`, `pais_residencia` bị loại bỏ do chúng không đóng góp đáng kể vào việc dự đoán hành vi mua sản phẩm mới. Những cột này có thể chứa thông tin trùng lặp hoặc không có sự thay đổi theo thời gian, dẫn đến việc không có ý nghĩa trong việc huấn luyện mô hình.
- **Mã hóa biến phân loại:** Các cột như `sexo`, `tiprel_1mes`, `segmento` được chuyển đổi sang dạng số bằng cách sử dụng phương pháp `cat.codes`. Điều này giúp mô hình hiểu và xử lý dữ liệu dạng phân loại một cách hiệu quả hơn so với dữ liệu dạng chuỗi.
- **Loại bỏ sản phẩm ít phổ biến:** Một số sản phẩm như `ind_aval_fin_ult1` và `ind_ahor_fin_ult1` có tỷ lệ sở hữu rất thấp (dưới 1%) nên bị loại bỏ khỏi tập dữ liệu. Điều này giúp giảm nhiễu và tăng độ tập trung của mô hình vào các sản phẩm phổ biến hơn.

2.2 Tạo Đặc Trưng Mới

Ngoài việc xử lý dữ liệu hiện có, việc tạo ra các đặc trưng mới giúp mô hình có thể nắm bắt tốt hơn hành vi của khách hàng:

- **Product Cumulative Sum (`product_cumsum`):**
  - Đo lường số tháng liên tục mà khách hàng đã sở hữu một sản phẩm cụ thể.
  - Giúp phản ánh mức độ trung thành của khách hàng với sản phẩm đó và có thể chỉ ra những khách hàng có xu hướng giữ sản phẩm trong thời gian dài.
  - Ví dụ: Nếu khách hàng sở hữu thẻ tín dụng trong 5 tháng liên tiếp, giá trị của `product_cumsum` sẽ là 5.
- **Product Change (`product_change`):**
  - Ghi nhận sự thay đổi của khách hàng đối với từng sản phẩm giữa hai tháng liên tiếp:
    - Nếu khách hàng mua thêm sản phẩm, giá trị là `1`.
    - Nếu khách hàng dừng sử dụng sản phẩm, giá trị là `-1`.
    - Nếu không có thay đổi, giá trị là `0`.
  - Giúp mô hình phát hiện xu hướng mua hoặc từ bỏ sản phẩm theo thời gian.
- **Product Transaction (`product_transaction`):**
  - Đo lường số lần giao dịch hoặc số lần khách hàng đã tương tác với sản phẩm trong khoảng thời gian huấn luyện.
  - Những khách hàng có số lần giao dịch cao có thể có nhu cầu cao hơn đối với một sản phẩm cụ thể.

2.3 Chuẩn Bị Dữ Liệu

Sau khi xử lý và tạo đặc trưng, dữ liệu được chuẩn bị để đưa vào mô hình huấn luyện:

- **Dữ liệu đầu vào (`X`):**
  - Gồm các đặc trưng được thu thập từ tháng 1/2015 đến tháng 3/2016 để làm tập huấn luyện.

  - Tháng 4/2016 được sử dụng làm dữ liệu đầu vào để dự đoán sản phẩm khách hàng sẽ mua trong tháng 5/2016.

- **Biến mục tiêu (`Y`):**
  - Biến này là nhãn nhị phân cho từng sản phẩm, biểu thị liệu khách hàng có kích hoạt sản phẩm trong tháng tiếp theo hay không.
  - Mục tiêu của mô hình là dự đoán những sản phẩm mới mà khách hàng có thể sẽ mua thêm trong tháng 5/2016.

3 Mô Hình Hóa với XGBoost

Sau khi hoàn thành quá trình Feature Engineering, mô hình XGBoost được lựa chọn để dự đoán sản phẩm mới mà khách hàng có thể mua. Lý do sử dụng XGBoost thay vì các mô hình khác là:

- **Hiệu suất cao:** XGBoost được tối ưu hóa để làm việc với dữ liệu lớn và có thể khai thác tối đa thông tin từ các đặc trưng được tạo ra.
- **Khả năng xử lý bài toán nhiều nhãn:** Với sự hỗ trợ của `MultiOutputClassifier`, XGBoost có thể dự đoán nhiều sản phẩm cùng lúc cho từng khách hàng.
- **Mô hình mạnh mẽ:** XGBoost có khả năng tự động điều chỉnh trọng số của các đặc trưng quan trọng, giúp tối ưu hóa kết quả dự đoán.

3.1 Quy Trình Huấn Luyện và Dự Đoán

1. **Huấn luyện mô hình trên tập dữ liệu đã qua xử lý:** Tận dụng các đặc trưng như `product_cumsum`, `product_change` và `product_transaction` để giúp mô hình học tốt hơn về hành vi khách hàng.
2. **Dự đoán sản phẩm mới cho tháng 5/2016:** Sử dụng dữ liệu tháng 4/2016 để đưa ra dự đoán về những sản phẩm mà khách hàng sẽ mua trong tháng tiếp theo.
3. **Đánh giá hiệu suất mô hình:** Độ chính xác của mô hình được đo bằng chỉ số Mean Average Precision (MAP), giúp đánh giá mức độ phù hợp giữa danh sách sản phẩm được đề xuất và thực tế.

Sự kết hợp giữa Feature Engineering và mô hình XGBoost, hệ thống đề xuất sản phẩm có thể đưa ra các gợi ý chính xác hơn, góp phần nâng cao trải nghiệm khách hàng và tối ưu hóa chiến lược kinh doanh của ngân hàng.

4 Thực nghiệm và Đánh giá

4.1 Kết quả Thực Nghiệm

Với Phương pháp trên thực nghiệm trên dữ liệu nhận được kết quả classification report (báo cáo phân loại) như sau:

	precision	recall	f1-score	support
ind_cco_fin_ult1	0.85	0.88	0.87	2648
ind_cder_fin_ult1	0.00	0.00	0.00	1
ind_cno_fin_ult1	0.85	0.52	0.65	1613
ind_ctju_fin_ult1	1.00	1.00	1.00	22
ind_ctma_fin_ult1	0.56	0.09	0.16	160
ind_ctop_fin_ult1	0.65	0.37	0.47	169
ind_ctpp_fin_ult1	0.68	0.40	0.51	119
ind_deco_fin_ult1	0.56	0.29	0.39	78
ind_deme_fin_ult1	0.00	0.00	0.00	12
ind_dela_fin_ult1	0.54	0.17	0.26	607
ind_ecue_fin_ult1	0.85	0.54	0.66	1142
ind_fond_fin_ult1	0.55	0.07	0.12	169
ind_hip_fin_ult1	0.00	0.00	0.00	4
ind_plan_fin_ult1	0.00	0.00	0.00	33
ind_pres_fin_ult1	0.50	0.17	0.25	6
ind_reca_fin_ult1	0.54	0.11	0.18	412
ind_tjcr_fin_ult1	0.84	0.72	0.78	3375
ind_valo_fin_ult1	0.75	0.23	0.35	242
ind_viv_fin_ult1	1.00	0.20	0.33	5
ind_nomina_ult1	0.85	0.68	0.76	3372
ind_nom_pens_ult1	0.87	0.73	0.80	3902
ind_recibo_ult1	0.83	0.86	0.84	6858
micro avg	0.84	0.71	0.77	24949
macro avg	0.60	0.37	0.43	24949
weighted avg	0.82	0.71	0.75	24949
samples avg	0.75	0.74	0.74	24949

Hình 1: Bảng báo cáo phân loại

Thực nghiệm với mAP(Mean Average Precision) ta thu được kết quả:

```
Mean Average Precision (MAP): 0.31513171562633735
```

Hình 2: Kết quả đánh giá chỉ số mAP

4.2 Đánh giá

Có thể thấy đối với classification report đạt kết quả tốt với những sản phẩm phổ biến và hạn chế trong các sản phẩm hiếm do việc hạn chế dữ liệu. Điều này do về dữ liệu mất cân bằng lớp, một hạn chế thường gặp trong bài toán "Product recommendation)
Kết quả mAP = 0.31 có thể xem là một kết quả rất tốt đối với một bài toán gợi ý sản phẩm, đặc biệt khi hệ thống đối mặt với dữ liệu rời rạc và đa nhãn.Nghĩa là, trung bình, khoảng 31% các đề xuất ở vị trí hàng đầu thật sự phù hợp với nhu cầu của khách hàng.
Tuy nhiên việc kết quả đặt được vẫn chứng tỏ bài toán có thể cải thiện thêm về hiệu suất chương trình như cần tập trung vào xử lý mất cân bằng lớp, tối ưu hóa đặc trưng đầu vào, và điều chỉnh hyperparameters của mô hình, sử dụng ensemble method hay hyperparameters.

5 Lời cảm ơn

Team 2 xin gửi lời cảm ơn chân thành đến Khoa Toán Cơ tin học, CLB Hamic và các công ty đồng hành đã tổ chức và hỗ trợ cuộc thi "Data Flow"năm 2025. Nhờ sự nỗ lực và tâm huyết của quý vị, cuộc thi đã diễn ra thành công, mang đến cho sinh viên cơ hội quý báu để thể hiện và phát triển kỹ năng của mình trong sự nghiệp phân tích dữ liệu của bản thân và cũng như công việc sau này. Team cũng xin cảm ơn các đội thi khác đã tham gia nhiệt tình, góp phần tạo nên một sân chơi sôi động và ý nghĩa. Hy vọng rằng, những hoạt động như thế này sẽ tiếp tục được tổ chức và phát triển trong tương lai, góp phần nâng cao chất lượng giáo dục và đào tạo và trình độ của con người Việt.