

PRODUCT RECOMMENDATION

I. NỘI DUNG BÀI TOÁN

Một ngân hàng trong lĩnh vực bán lẻ và tài chính cá nhân, ngân hàng này hỗ trợ khách hàng từ việc mua căn nhà đầu tiên, tái đầu tư vào tài sản hiện có, đến việc tối ưu hóa nguồn tài chính cá nhân, cung cấp các sản phẩm đa dạng nhằm đáp ứng nhu cầu phong phú của khách hàng.

Tuy nhiên, với hệ thống đề xuất sản phẩm hiện tại, ngân hàng này nhận thấy sự mất cân đối trong việc tiếp cận khách hàng: một số ít khách hàng nhận được nhiều gợi ý sản phẩm liên tục, trong khi nhiều người khác hầu như không được tiếp cận các sản phẩm phù hợp. Điều này dẫn đến trải nghiệm không đồng đều, có thể ảnh hưởng đến sự hài lòng và mức độ gắn kết của khách hàng với ngân hàng.

Để giải quyết vấn đề này, cần sử dụng dữ liệu về hành vi tài chính và xu hướng sử dụng sản phẩm của khách hàng hiện có nhằm xây dựng một hệ thống đề xuất thông minh và toàn diện hơn.

II. MỤC TIÊU PHÂN TÍCH

Tìm ra mô hình dự đoán hiệu quả, giúp ngân hàng hiểu rõ hơn về nhu cầu của từng khách hàng và đưa ra các đề xuất sản phẩm cá nhân hóa một cách chính xác, từ đó củng cố cam kết mang lại sự hài lòng cho mọi khách hàng, bất kể họ đang ở đâu trong hành trình tài chính của mình.

Xây dựng được mô hình dự đoán mạnh mẽ, tận dụng tối đa dữ liệu để giúp ngân hàng hiểu rõ hơn về nhu cầu của từng khách hàng và đưa ra các đề xuất sản phẩm cá nhân hóa một cách chính xác, đảm bảo mọi khách hàng đều được quan tâm và nhận được các gợi ý phù hợp. Hệ thống này không chỉ giúp tăng cường sự hài lòng và lòng trung thành của khách hàng mà còn tối ưu hóa chiến lược bán chéo và bán thêm, từ đó gia tăng doanh thu và nâng cao tính cạnh tranh.

III. YÊU CẦU

1. Data Discovery & Cleaning:

- Sử dụng một số công cụ trực quan để khai phá tập dữ liệu, đưa ra một số nhận định về Khách hàng, Sản phẩm, thời gian, tình hình kinh doanh của bank.
- Làm sạch dữ liệu, xử lý các giá trị bị thiếu, ngoại lệ và không nhất quán để chuẩn bị cho phân tích.

2. Feature Engineering:

- Xây dựng tiêu chí (biến) phái sinh phục vụ mô hình
- Đánh giá mức độ quan trọng, lựa chọn tiêu chí (biến) có tính ảnh hưởng lớn

3. Model Training:

- Phát triển mô hình sử dụng không giới hạn các thuật toán, kỹ thuật (xác suất thống kê, mô hình hồi quy, collaborative filtering – lọc cộng tác, deep learning – học sâu, ensemble model – mô hình kết hợp)

- Đánh giá hiệu suất của mô hình chỉ số Mean Average Precision (MAP).

4. Prediction Making:

Đối với mỗi Khách hàng tại thời điểm cần dự đoán, cần dự đoán một danh sách các sản phẩm nên mua tiếp.

5. Code Optimizing:

Tối ưu hóa tốc độ xử lý và năng lực xử lý khi có nhiều dữ liệu bằng nhiều phương pháp, thể hiện được tư duy tối ưu code.

6. Discovery Presenting:

Trình bày quá trình xử lý tiêu chí, phương pháp luận mô hình, kết quả dự đoán và các nhận định sau khi khai phá dữ liệu bằng cách sử dụng bảng biểu, biểu đồ và kể chuyện dữ liệu một cách hấp dẫn.

IV. MÔ TẢ DỮ LIỆU

Bạn sẽ được cung cấp 1,5 năm dữ liệu hành vi khách hàng của ngân hàng để dự đoán những sản phẩm mới mà khách hàng sẽ mua. Dữ liệu bắt đầu từ ngày 28/01/2015 và ghi nhận hàng tháng các sản phẩm mà một khách hàng đã sở hữu, chẳng hạn như "credit card" (thẻ tín dụng), "saving account" (tài khoản tiết kiệm), v.v. Nhiệm vụ của bạn là dự đoán các sản phẩm tiếp theo mà khách hàng sẽ có vào tháng cuối cùng, ngày 28/05/2016, ngoài những sản phẩm họ đã sở hữu vào ngày 28/04/2016. Các sản phẩm này được biểu thị bằng các cột có tên: ind_(xyz)_ult1, nằm ở cột số 25 đến số 48 trong dữ liệu huấn luyện. Bạn sẽ dự đoán những sản phẩm mà khách hàng sẽ mua thêm so với những sản phẩm họ đã có tại ngày 28/04/2016.

Tập dữ liệu kiểm tra (test set) và huấn luyện (train set) được chia theo thời gian.

Dữ liệu đề bài có thể được lấy tại đây:

+ Google Drive: [PRODUCT RECOMMENDATION](#)

+ Kaggle: [DATAFLOW2025 - PRODUCT RECOMMENDATION](#)

Mô tả các cột của bộ dữ liệu:

Tên cột	Description	Diễn giải
fecha_data	The table is partitioned for this column	Bảng được phân vùng theo cột này
ncodpers	Customer code	Mã khách hàng
ind_empleado	Employee index: A active, B ex employed, F filial, N not employee, P pasive	Chỉ số nhân viên: A (đang làm việc), B (nghỉ việc), F (chi nhánh), N (không phải nhân viên), P (thụ động)

pais_residencia	Customer's Country residence	Quốc gia cư trú của khách hàng
sexo	Customer's sex	Giới tính của khách hàng
age	Age	Tuổi
fecha_alta	The date in which the customer became as the first holder of a contract in the bank	Ngày khách hàng trở thành người đứng tên chính đầu tiên trên hợp đồng tại ngân hàng
ind_nuevo	New customer Index. 1 if the customer registered in the last 6 months.	Chỉ số khách hàng mới. Bằng 1 nếu khách hàng đăng ký trong 6 tháng qua.
antiguedad	Customer seniority (in months)	Thâm niên của khách hàng (tính bằng tháng)
indrel	1 (First/Primary), 99 (Primary customer during the month but not at the end of the month)	1 (Khách hàng chính), 99 (Khách hàng chính trong tháng nhưng không phải vào cuối tháng)
ult_fec_cli_1t	Last date as primary customer (if he isn't at the end of the month)	Ngày cuối cùng là khách hàng chính (nếu không phải cuối tháng)
indrel_1mes	Customer type at the beginning of the month ,1 (First/Primary customer), 2 (co-owner) ,P (Potential),3 (former primary), 4(former co-owner)	Loại khách hàng đầu tháng: 1 (Khách hàng chính), 2 (Đồng sở hữu), P (Tiềm năng), 3 (Khách hàng chính cũ), 4 (Đồng sở hữu cũ)
tiprel_1mes	Customer relation type at the beginning of the month, A (active), I (inactive), P (former customer), R (Potential)	Loại quan hệ khách hàng đầu tháng: A (Hoạt động), I (Không hoạt động), P (Khách hàng cũ), R (Tiềm năng)
indresi	Residence index (S (Yes) or N (No) if the residence country is the same than the bank country)	Chỉ số cư trú: S (Có) hoặc N (Không) nếu quốc gia cư trú giống với quốc gia của ngân hàng
indext	Foreigner index (S (Yes) or N (No) if the customer's birth country is different than the bank country)	Chỉ số người nước ngoài: S (Có) hoặc N (Không) nếu quốc gia sinh của khách hàng khác quốc gia của ngân hàng
conyuemp	Spouse index. 1 if the customer is spouse of an employee	Chỉ số vợ/chồng: 1 nếu khách hàng là vợ/chồng của nhân viên
canal_entrada	Channel used by the customer to join	Kênh khách hàng sử dụng để gia nhập
indfall	Deceased index. N/S	Chỉ số đã mất: N/S
tipodom	Address type. 1, primary address	Loại địa chỉ. 1, địa chỉ chính
cod_prov	Province code (customer's address)	Mã tỉnh (địa chỉ khách hàng)
nomprov	Province name	Tên tỉnh
ind_actividad_cliente	Activity index (1, active customer; 0, inactive customer)	Chỉ số hoạt động (1, khách hàng hoạt động; 0, khách hàng không hoạt động)

renta	Gross income of the household	Thu nhập hộ gia đình trước thuế
segmento	Segmentation: 01 - VIP, 02 - Individuals 03 - college graduated	Phân khúc: 01 - VIP, 02 - Cá nhân, 03 - Tốt nghiệp đại học
ind_ahor_fin_ult1	Saving Account	Tài khoản tiết kiệm
ind_aval_fin_ult1	Guarantees	Đảm bảo
ind_cco_fin_ult1	Current Accounts	Tài khoản hiện tại
ind_cder_fin_ult1	Derivada Account	Tài khoản Derivada
ind_cno_fin_ult1	Payroll Account	Tài khoản lương
ind_ctju_fin_ult1	Junior Account	Tài khoản Junior
ind_ctma_fin_ult1	Más particular Account	Tài khoản Más particular
ind_ctop_fin_ult1	Particular Account	Tài khoản particular
ind_ctpp_fin_ult1	Particular Plus Account	Tài khoản particular Plus
ind_deco_fin_ult1	Short-term deposits	Tiền gửi ngắn hạn
ind_deme_fin_ult1	Medium-term deposits	Tiền gửi trung hạn
ind_dela_fin_ult1	Long-term deposits	Tiền gửi dài hạn
ind_ecue_fin_ult1	E-account	Tài khoản e-account
ind_fond_fin_ult1	Funds	Quỹ đầu tư
ind_hip_fin_ult1	Mortgage	Thế chấp
ind_plan_fin_ult1	Pensions	Lương hưu
ind_pres_fin_ult1	Loans	Khoản vay
ind_reca_fin_ult1	Taxes	Thuế
ind_tjcr_fin_ult1	Credit Card	Thẻ tín dụng
ind_valo_fin_ult1	Securities	Chứng khoán
ind_viv_fin_ult1	Home Account	Tài khoản nhà
ind_nomina_ult1	Payroll	Lương
ind_nom_pens_ult1	Pensions	Lương hưu
ind_recibo_ult1	Direct Debit	Ghi nợ trực tiếp

V. TIÊU CHÍ ĐÁNH GIÁ

- Chất lượng mô hình thông qua chỉ số Mean Average Precision (MAP) - Độ chính xác trung bình tổng quát (25%)
- Làm sạch và khai phá dữ liệu (10%)
- Lựa chọn tiêu chí – feature engineering (15%)

- Chất lượng code và tối ưu trong lập trình (15%)
- Chất lượng báo cáo (15%)
- Trình bày hiệu quả (20%)