PRODUCT RECOMMENDATION

Báo cáo của nhóm CNA

Thành viên thực hiện:
Vũ Đức Anh
Phạm Như Ngọc
Chu Quang Cần

Hà Nội, ngày 24 tháng 2 năm 2025

LỜI MỞ ĐẦU

Giới thiệu:

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.

Vấn đề: Thông tin bất cân xứng dẫn đến trải nghiệm tiêu cực của khách hàng

Mục lục:

I. LÀM SẠCH DỮ LIỆU

Dữ liệu gốc có rất nhiều giá trị null và không đồng nhất về kiểu dữ liệu và giá trị (như ở trong notebook nhóm đã trình bày) gồm:

- Có rất nhiều cột (ind_empleado, pais_residencia, fecha_alta,...) có 27734 giá trị null và cứ dòng nào null là phần thông tin khách hàng không có gì cả.
- Dữ liệu ở cột ult_fec_cli_1t null rất nhiều (13622516 giá trị null) bởi vì theo như cột indrel thì hầu hết khách hàng là khách hàng chính.
- Dữ liệu ở cột conyuemp null rất nhiều (13645501 giá trị null) và giá trị lưu là N và S chứ không phải như trong mô tả dữ liệu.
- Các cột số (age, antiguedad,...) có chứa cả dữ liệu dạng số và string.
- Cột age và antiguedad cũng chứa rất nhiều giá trị vô lí như thâm niên khách hàng (antiguedad) là số âm và tuổi (age) khách hàng > 117.
- 2 cột indrel và tiprel_1mes chứa thông tin khá giống nhau.
- Cột tipodom tất cả các giá trị đều là 1
- Cột nomprov thừa vì đã có cod_prov
- Cột pais_residencia lệch vì toàn là người ES và có cột cod_prov chứa thông tin sâu hơn rồi
- Cột cod_prov với những người không ở ES thì giá trị là null

➔ **Giải pháp:**

- Drop hết null ở cột ind_empleado là 1 vài cột khác cũng hết theo
- Drop cột ult_fec_cli_1t vì có quá nhiều giá trị null và thông tin cũng không quan trọng
- Drop cột conyuemp vì dữ liệu không giống với trong mô tả và bị null rất nhiều
- Loại bỏ các giá trị vô lí ở trong cột age và antiguedad
- Chuyển giá trị trong các cột số về int để tránh có dòng dữ liệu được lưu theo kiểu string
- 2 cột indrel và tiprel_1mes gộp chung vào nhau thành cột inti_1mes để có thể điền null dễ hơn. 2 cột này chỉ có 7 cặp gía trị (như đã trình bày trong notebook) gồm:
    - (1, I) : A
    - (1, A): B
    - (2, I) : C
    - (2, A): D
    - (3, P): E
    - (P, R): F và có cặp (P, N) nhưng là do không đồng nhất và gộp chung với (P, R)
    - (4, P): G
- Sau khi gộp thành inti_1mes thì còn 1 ít giá trị null và ở cột này thì A và B chiếm số lượng lớn => Nếu ind_actividad_cliente (Chỉ số hoạt động) là 1 thì lấy giá trị là B còn không thì A
- Bỏ các cột (pais_residencia, fecha_alta, ind_nuevo, canal_entrada, tipodom, nomprov, ind_actividad_cliente) vì:
    - Cột pais_residencia (quốc gia cư trú) đã có thông tin ở cod_prov
    - Cột fecha_alta (ngày đầu kí hợp đồng) không quan trọng

- Cột ind_nuevo (chỉ số khách hàng mới) đã có thông tin chứa trong antiguedad (thâm niên khách hàng)
- Cột canal_entrada (kênh gia nhập) không cần thiết
- Cột tipodom (loại địa chỉ) tất cả đều là 1
- Cột nomprov(tên tỉnh) đã có thông tin trong cod_prov (mã tỉnh)
- Cột ind_actividad_cliente (Chỉ số hoạt động) đã có thông tin trong cột tiprel_1mes

- Cột cod_prov còn giá trị null là do có người không ở ES thì không có mã tính => điền là 0
- Cột segmento (Phân khúc) còn 1 ít gía trị null nhưng mà ở cột này số lượng 02-Khách hàng cá nhân chiếm đa số nên điền bằng mode của cột

Sau khi xử lý hết các cột thì còn cột renta còn null rất nhiều

➔ **Giải pháp:** Sử dụng KNN để điền null dựa trên những hàng giống với nó nhất

Cần phải encode các dữ liệu dạng string. Các cột có 2 giá trị là sexo, indresi, indext, indfall thì dùng Label encode. Cột nhiều giá trị và không có thứ tự giữa các giá trị là ind_empleado thì dùng one-hot encode. Các cột inti_1mes và segmento thì mapping theo thứ tự quan trọng của các giá trị.

II. TRỰC QUAN HOÁ VÀ PHÂN TÍCH DỮ LIỆU

1) Phân tích sản phẩm và tình hình kinh doanh các loại sản phẩm của ngân hàng

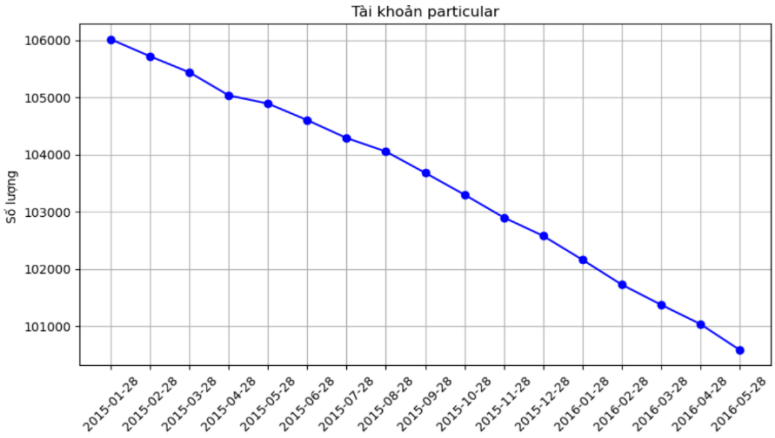
- **Tài khoản tiết kiệm**

Theo cổng thông tin điện tử Bộ tài chính, nhằm tạo đà kích thích tăng trưởng kinh tế, T3/2015 ECB đã quyết định cắt giảm các lãi suất chủ chốt, cụ thể: Giảm lãi suất tái cấp vốn từ 0,05% xuống 0%, lãi suất cho vay thanh khoản từ 0,3% xuống 0,25%, lãi tiền gửi ngân hàng giảm từ -0,3% xuống -0,4%. Việc giảm lãi suất như vậy sẽ làm giảm mạnh nhu cầu gửi tiết kiệm của khách hàng (số liệu lao dốc được thể hiện qua đồ thị bên dưới)

- **Current account**

Số lượng dùng tài khoản hiện tại có xu hướng tăng lên vì đây là loại tài khoản sử dụng để thực hiện các giao dịch hàng ngày với đặc điểm:

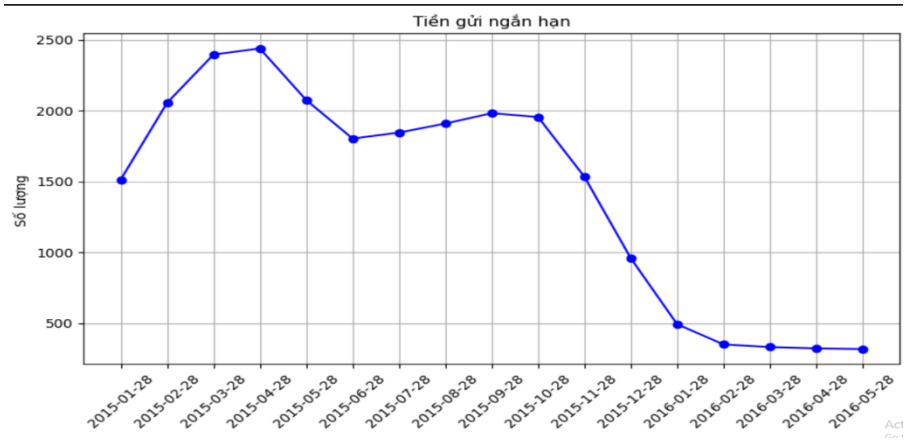
- Không có hoặc có lãi suất rất thấp (khác với tài khoản tiết kiệm).
- Không giới hạn số lần giao dịch: có thể rút tiền, chuyển khoản, thanh toán hóa đơn thường xuyên
- Hỗ trợ các phương thức thanh toán: thẻ ghi nợ, séc, chuyển khoản nhanh.
- **Tài khoản particular**

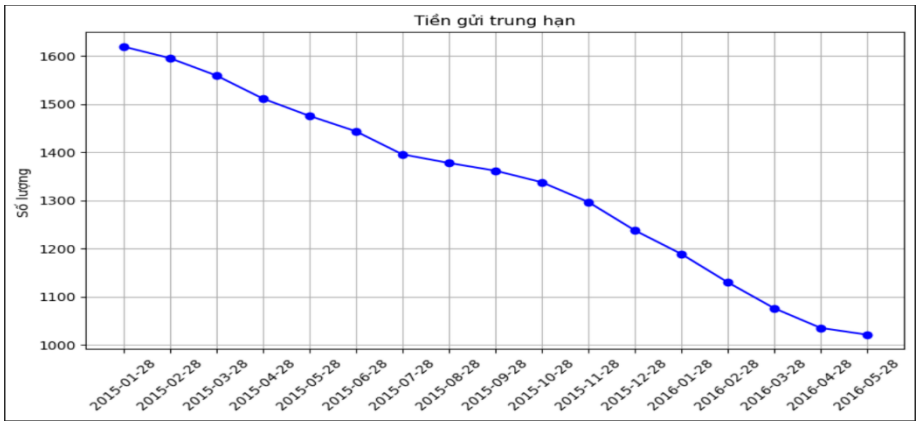


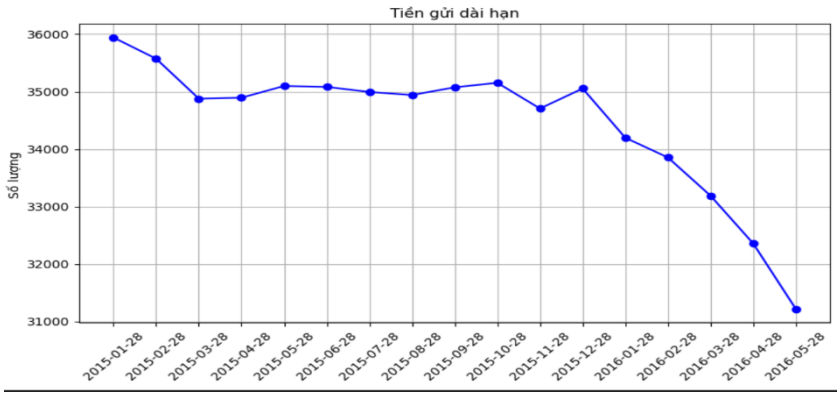
Số lượng tài khoản particular và tài khoản particular plus có xu hướng giảm mạnh đều, tuy nhiên số lượng vẫn ở mức cao hơn hẳn so với các sản phẩm tài khoản khác do:

- Số lượng cá nhân, hộ gia đình có nhu cầu giao dịch vô cùng lớn, đóng vai trò chủ yếu trong nguồn vốn của ngân hàng
- Tuy nhiên nhu cầu sử dụng sản phẩm này có xu hướng giảm vì:
- Chính sách nới lỏng QE của ECB làm cho lãi suất thấp, làm giảm nhu cầu giữ tiền trong ngân hàng.
- Ví điện tử và ngân hàng số phát triển, thay thế ngân hàng truyền thống.

- **Các sản phẩm tiền gửi**



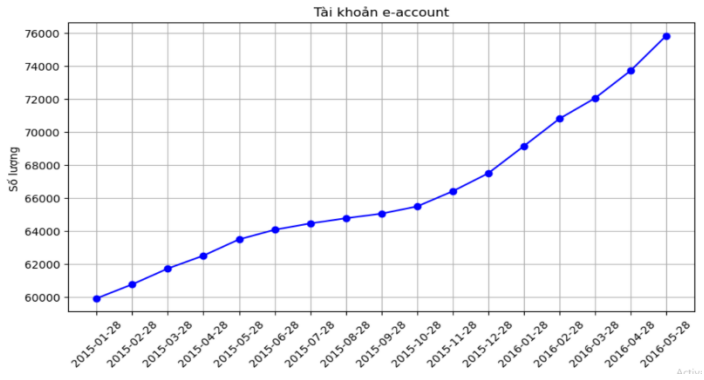




⇒ So sánh 3 sản phẩm tiền gửi ngắn hạn, trung hạn, dài hạn:

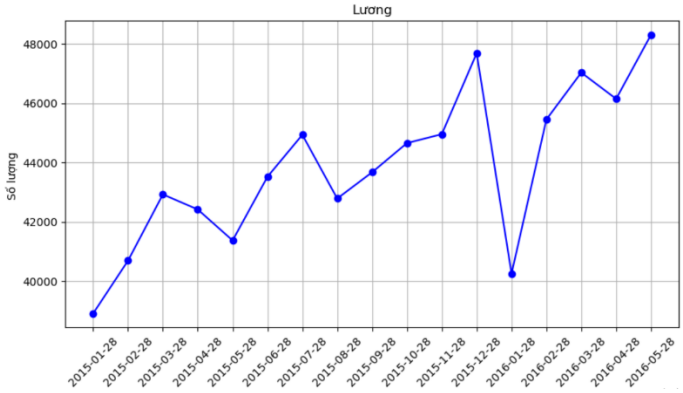
- Cả 3 loại sản phẩm này đều có xu hướng giảm mạnh trong 2 năm 2015-2016
- Tiền gửi trung hạn có xu hướng giảm ổn định nhất và xuống mức rất thấp vào 5/2016
- Tiền gửi ngắn hạn có số lượng người dùng ít nhất vì trong thời gian 2015-2016, nền kinh tế các nước trong Eurozone đang trong giai đoạn phục hồi sau khủng hoảng nên chính sách là thắt chặt chi tiêu nên vòng quay vốn ít hơn. Trong khi đó, tiền gửi ngắn hạn dành cho những người có khả năng quay vòng vốn nhanh, thường là những người chi tiêu nhiều.
- Tiền gửi dài hạn có số lượng người dùng nhiều nhất vì lãi suất loại tiền gửi này là cao nhất, dành cho những người có xu hướng quay vòng vốn ít, muốn ổn định dòng tiền, phù hợp với chính sách thắt chặt chi tiêu trong giai đoạn này.

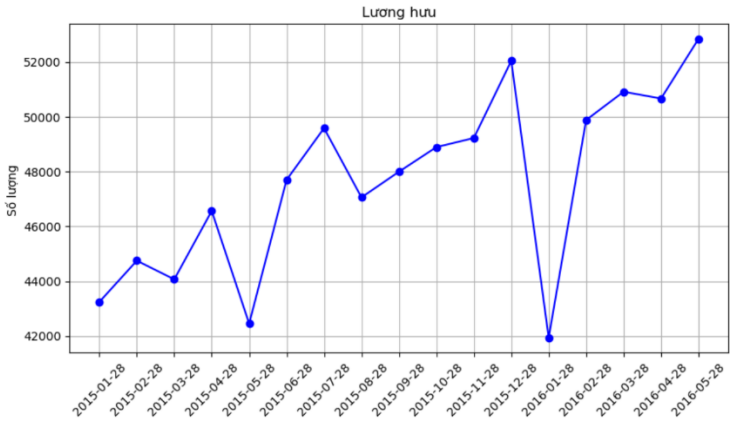
- **Tài khoản e-account**



Số lượng tài khoản e-account có xu hướng tăng mạnh đều và ở mức cao do hệ thống ngân hàng số hóa, hướng tới mục tiêu mọi giao dịch được thực hiện điện tử buộc người dùng phải sở hữu tài khoản này để thực hiện giao dịch của mình hàng ngày

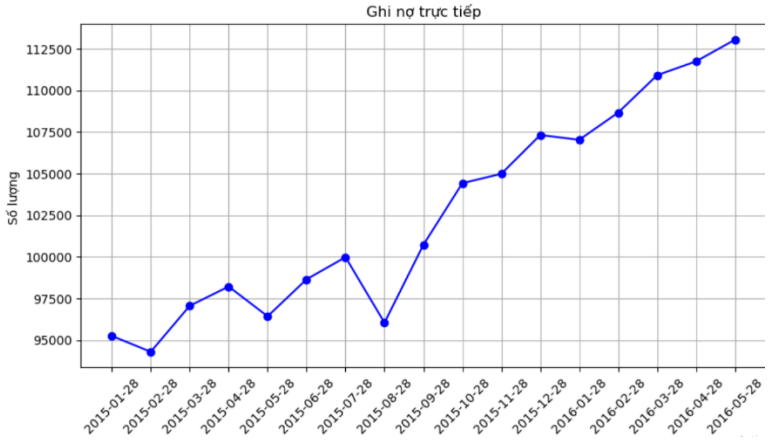
- **Lương & Lương hưu**





Số lượng người dùng tài khoản lương và lương hưu có xu hướng biến động khá tương đồng. Vào đầu năm 2016, tỷ lệ thất nghiệp ở Eurozone vẫn ở mức cao nhưng có dấu hiệu giảm dần so với các năm trước, nhờ vào các chính sách kích thích kinh tế của Ngân hàng Trung ương Châu Âu (ECB) và sự phục hồi chậm của nền kinh tế khu vực. Ngoài ra còn do một số chính sách Cải cách thị trường lao động ở một số quốc gia, đặc biệt là Tây Ban Nha và Bồ Đào Nha. Tài khoản lương và tài khoản lương hưu phản ánh tỷ lệ có việc làm nên điều này lý giải điểm biến động nổi bật nhất tại tháng 1/2016 ở cả 2 tài khoản.

- **Ghi nợ trực tiếp**



Tài khoản ghi nợ trực tiếp có xu hướng tăng mạnh vì 1 số lý do :

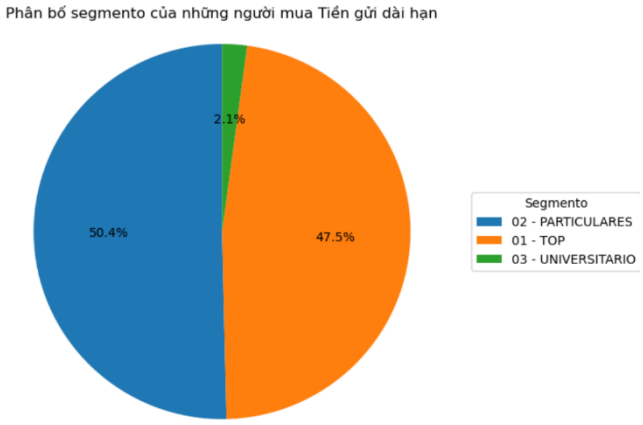
- Lãi suất thấp
- Ngân hàng hiện đại hóa, thúc đẩy các dịch vụ thanh toán không dùng tiền mặt.

2) Phân tích khách hàng

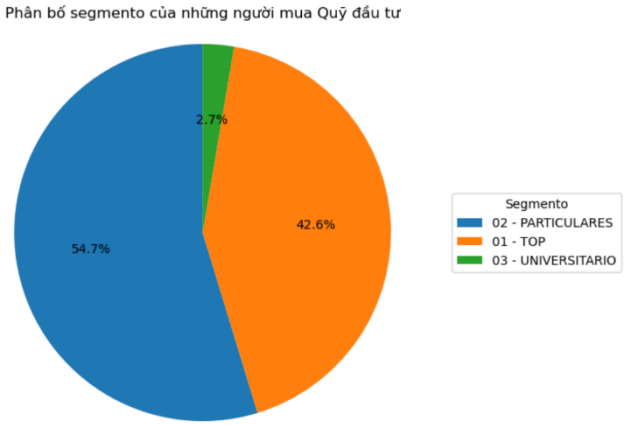
2.1. Phân khúc khách hàng

- **TOP:**

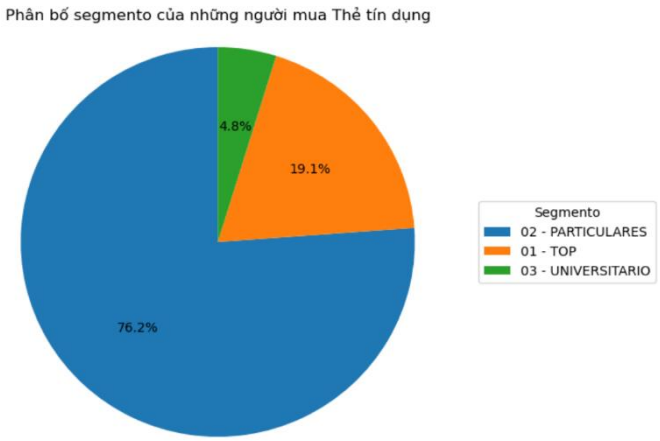
- Tỷ lệ người dùng thuộc phân khúc TOP cho sản phẩm tiền gửi là cao và đồng đều nhất (ở mức 30-50%),



- Tiếp theo là các loại sản phẩm đầu tư (Quỹ đầu tư, chứng khoán hay Derivada) ở mức trung bình 30% mỗi loại sản phẩm



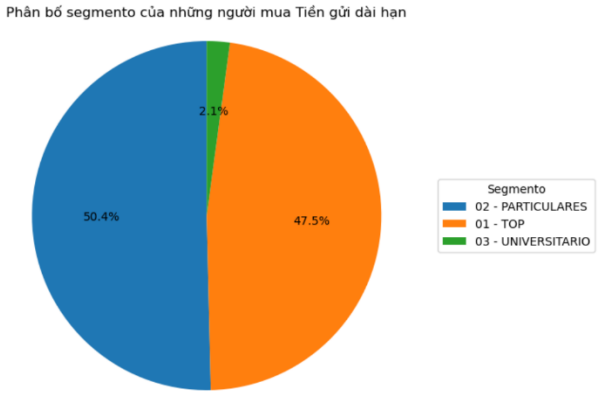
- Các sản phẩm còn lại ở mức trung bình 10-20% mỗi loại sản phẩm



➔Phân khúc này là những người có thu nhập cao, ổn định nên họ tập trung vào những sản phẩm vừa cần có lượng vốn tương đối, vừa sinh lời, vừa ổn định như các sản phẩm tiền gửi và đầu tư (chứng khoán, phái sinh)

- **PARTICULARES**

- Chiếm đa số với tỷ lệ người dùng các sản phẩm trên thuộc phân khúc Particulares thấp nhất khoảng 50% và cao nhất là hơn 99%



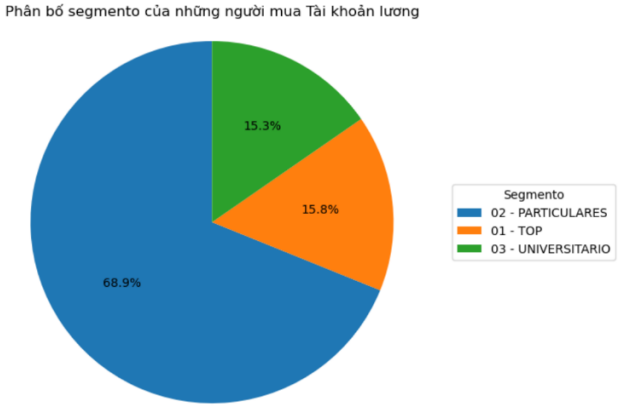
➔ Phân khúc này là những người có thu nhập tầm trung, có khẩu vị rủi ro cao, nhu cầu giao dịch nhiều và thường xuyên và họ mong muốn trải nghiệm được trải nghiệm nhiều loại sản phẩm của ngân hàng để quản lý kế hoạch tài chính cá nhân 1 cách hiệu quả nhất

- **UNIVERSITARIO**

- Tỷ lệ khách hàng thuộc phân khúc này sử dụng tài khoản lương và ghi nợ trực tiếp là cao nhất, ở mức 15%

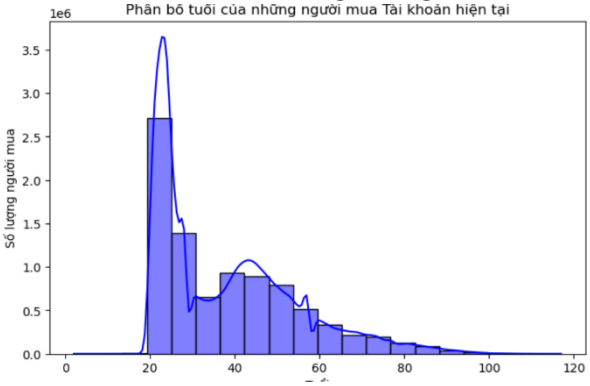
- Tỷ lệ khách hàng thuộc phân khúc này sử dụng tài khoản Junior là thấp nhất, ở mức gần như 0%

➔ Phân khúc này là những người vừa tốt nghiệp đại học hoặc mới bước chân vào thị trường lao động, họ chưa có nhiều vốn nên tập trung chủ yếu vào tài khoản lương.

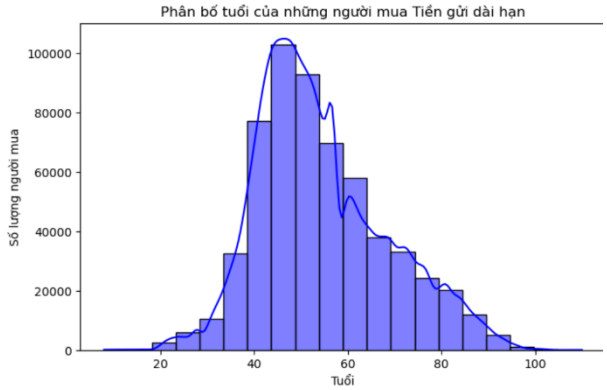


2.2. Độ tuổi

- 5 - 20 tuổi: dành cho tài khoản Junior
- 20 - 40 tuổi: Số lượng dùng tài khoản hiện tại ở độ tuổi này là nhiều nhất



- 40 - 60 tuổi: Khách hàng chính của hầu hết các loại sản phẩm
- > 60 tuổi: Tập trung chủ yếu các loại sản phẩm đầu tư an toàn, ổn định như tài khoản tiết kiệm, các loại tiền gửi và tài khoản lương hưu



2.3. **Thu nhập:**

Tất cả các loại sản phẩm đều tập trung nhiều vào nhóm khách hàng có thu nhập trung bình từ €100000 EUR - €200000 EUR

2.4. **Thời gian:**

Tình hình kinh doanh nhìn chung đều đi ngang giai đoạn nửa đầu năm 2015 và bắt đầu tăng trưởng từ cuối năm 2015 đến năm 2016.

III. MÔ HÌNH

1) Model

Trước khi vào phần model, nhóm chúng tôi đã tách ra tập train và test. Trong đó input sẽ là toàn bộ thông tin khách hàng và sản phẩm đã mua ở tháng trước, và output là sản phẩm mà khách hàng sẽ mua thêm trong tháng sau. Nếu khách hàng không mua thêm hoặc bỏ bớt sản phẩm thì không capture lại vào tập dữ liệu này. Và tập test tách dữ liệu ra theo ngày 28/4/2016, tức là input để predict ra sản phẩm khách hàng sẽ mua thêm vào 28/5/2016.

Nhóm chúng tôi đã sử dụng 2 phương pháp là KNeighborsClassifier và XGBoost để dự đoán ra sản phẩm khách hàng sẽ mua tiếp vào tháng sau.

1.1. KNeighborsClassifier:

a. Tiền xử lí:

- Tất cả các cột string còn lại được label
- Bỏ cột fecha_dato vì sau khi visualize không thấy 1 trend cụ thể nào theo tháng hay theo năm
- Tất cả các dữ liệu được scale về khoảng 0-1 với MinMaxScaling

b. Chọn tham số:

Chúng tôi cho duyệt qua các tham số và in ra đồ thị độ chính xác thì tham số k = 60 có độ chính xác cao nhất.

1.2. Ensemble Model (XGBoost + RandomForest)

a. Tiền xử lý dữ liệu:

- Tương tự như KNeighborsClassifier, chúng tôi cũng thực hiện các công việc tiền xử lý dữ liệu như đánh nhãn bằng số cho các cột dạng string và thực hiện scale giá trị về khoảng 0 - 1.
- Sau khi tiền xử lý dữ liệu, chúng tôi chia dữ liệu thành hai dataframe là X_train, Y_train và X_test, Y_test. Dataframe Y chỉ bao gồm 1 cột là cột ‘output’ - danh sách sản phẩm mà người dùng sẽ mua.
- Để thực hiện truyền dữ liệu vào mô hình, chúng tôi thực hiện thêm bước cuối cùng là đánh nhãn cho cột sản phẩm sẽ mua - ‘output’ bằng LabelEncoder().

b. Khởi tạo và huấn luyện mô hình:

- Nhóm quyết định khởi tạo hai mô hình XGBoost và RandomForest với các tham số mặc định, và sử dụng thư viện GridSearch để tìm ra tham số tối ưu nhất

- Sử dụng tham số tối ưu ở quá trình trên, khởi tạo lại hai mô hình với bộ tham số tối ưu như đã đề cập trong code.
- Để kết hợp hai mô hình lại, chúng tôi sử dụng thư viện VotingClassifier, cho phép kết hợp dự đoán của hai mô hình và đưa ra kết quả chính xác nhất.

2) Metrics

Về metrics đánh giá thì nhóm nhận thấy là việc khách hàng sử dụng thêm sản phẩm gì thì không quan trọng về thứ tự được dự đoán ra. Nên nhóm chúng tôi sử dụng accuracy để đánh giá model. Model của chúng tôi chỉ dự đoán ra 1 sản phẩm và nếu sản phẩm đó đúng là khách hàng sẽ mua thêm thì sẽ là đúng, còn không là sai. Sau đó chia số dự đoán đúng cho tổng số sản phẩm dự đoán để tính accuracy. Thì qua đó nhóm nhận được kết quả của model KNeighborsClassifier là 73.92% và XGBoost + RandomForest là 74.12%

Sau đó nhóm chúng tôi có thử dùng KNeighborsClassifier dự đoán top các sản phẩm dựa trên số lần xuất hiện của id khách hàng trong cột ncodpers. Ví dụ 1 khách hàng A mua thêm 3 sản phẩm thì id khách hàng đó sẽ xuất hiện 3 lần trong tập test với 3 sản phẩm khác nhau thì chúng tôi cũng đoán top 3 sản phẩm có xác suất cao nhất. Sau đó cũng tính accuracy theo cách trên thì kết quả được 75.5% với k = 70

IV. GỢI Ý VÀ KẾT LUẬN

Sau khi thử trên tập test thì nhóm chúng tôi thực hiện dự đoán trên toàn bộ người dùng. Với cách cắt dữ liệu như này, KNN và XGBoost hoạt động khá tốt mặc dù chỉ gợi ý ra 1 sản phẩm. Nếu gợi ý top 3 sản phẩm thì kết quả sẽ cao hơn. Kết quả gợi ý sản phẩm cho tất cả customer của tháng 4/2016 là Product Recommendations