# **PHẦN 1: NỘI DUNG ĐỀ TÀI**

# **I. Lý do chọn đề tài**

Khách hàng rời bỏ (customer churn) là tình trạng khách hàng dù đã bắt đầu sử dụng sản phẩm hoặc dịch vụ của doanh nghiệp nhưng vì lý do này hay lý do khác, ngừng hoàn toàn và chuyển sang lựa chọn một đối thủ khác. Trong kinh doanh, khi khách hàng không hài lòng với dịch vụ, sản phẩm mà doanh nghiệp cung cấp thì họ sẽ ngừng kết nối, hợp tác với doanh nghiệp. Theo như số liệu thống kê từ nhiều nguồn nghiên cứu trên thế giới:

* Nghiên cứu của Deloitte (2018) chỉ ra rằng khả năng cung cấp giá trị và dịch vụ chất lượng là một trong những yếu tố quan trọng nhất ảnh hưởng đến quyết định của khách hàng về việc ở lại hoặc rời bỏ ngân hàng.
* Nghiên cứu của Accenture (2019) đã nhấn mạnh sự quan trọng của trải nghiệm khách hàng trong quá trình giữ chân khách hàng. Nói chung, khách hàng đánh giá cao các trải nghiệm tích cực và có thể chuyển đổi với các tổ chức có trải nghiệm khách hàng kém.
* Thống kê từ Federal Reserve Bank of St. Louis (2020) cho thấy chi phí và phí liên quan đến tài khoản ngân hàng là một trong những lý do quan trọng khiến kháchhàng đưa ra quyết định rời bỏ ngân hàng.
* Theo Nie et al. (2011), giảm 5% tỷ lệ khách hàng rời bỏ có thể làm tăng lợi nhuậncủa ngân hàng lên tới 85%.

Ta thấy rằng sự rời bỏ hay ở lại của khách hàng là yếu tố quyết định sự phát triển của ngân hàng. Việc tìm kiếm một khách hàng mới sẽ “đắt hơn” rất nhiều so với việc giữ chân một khách hàng hiện có. Thế nên nhu cầu phân tích tỷ lệ khách hàng rời bỏ ngàycàng tăng. Đặc biệt là nhu cầu về mô hình dự đoán được xây dựng trên các phương phápthuộc lĩnh vực khoa học dữ liệu. Nếu ngân hàng có thể dự đoán tỷ lệ rời bỏ của khách hàng, các chiến dịch tiếp thị nhằm giữ chân khách hàng sẽ được cải thiện, mang lại hiệu quả hơn cho doanh nghiệp

# **II. Mô tả bài toán**

## 1. Vấn đề kinh doanh

Một nhà quản lý tại ngân hàng đang lo lắng khi ngày càng có nhiều khách hàng rời bỏ dịch vụ thẻ tín dụng của họ. Họ thực sự đánh giá cao nếu có thể dự đoán khả năng khách hàng rời bỏ dịch vụ để họ có thể chủ động tìm đến khách hàng để cung cấp cho họ dịch vụ tốt hơn và chuyển hướng quyết định của khách hàng theo hướng ngược lại.

## 2. Bối cảnh

Khi một ngân hàng có được khách hàng cho dịch vụ thẻ tín dụng của mình, ba Chỉ số hiệu suất chính (KPI) cần xem xét bao gồm:\

* Customer Acquisition Cost (CAC): Chỉ số này đo lường chi phí liên quan đến việc thu hút từng khách hàng thẻ tín dụng, bao gồm tiếp thị, bán hàng và các chi phí liên quan. CAC thấp hơn phản ánh hiệu quả thu hút khách hàng.
* Customer Lifetime Value (CLV): CLV ước tính tổng doanh thu mà ngân hàng có thể mong đợi tạo ra từ một khách hàng thẻ tín dụng trong suốt mối quan hệ của họ. CLV cao hơn cho thấy giá trị của khách hàng vượt quá chi phí thu hút, đảm bảo lợi nhuận dài hạn.
* Churn Rate: Tỷ lệ mất khách hàng thường được thể hiện dưới dạng phần trăm và biểu thị số lượng khách hàng thẻ tín dụng đã rời đi trong một khoảng thời gian cụ thể chia cho tổng số khách hàng vào đầu khoảng thời gian đó.

Các KPI này giúp ngân hàng đánh giá hiệu quả của các chiến lược thu hút khách hàng thẻ tín dụng và đánh giá lợi ích tài chính tiềm năng dài hạn của các thương vụ mua lại này. Để tối đa hóa lợi nhuận, ngân hàng đặt mục tiêu giảm thiểu CAC và Churn rate trong khi tối đa hóa CLV.

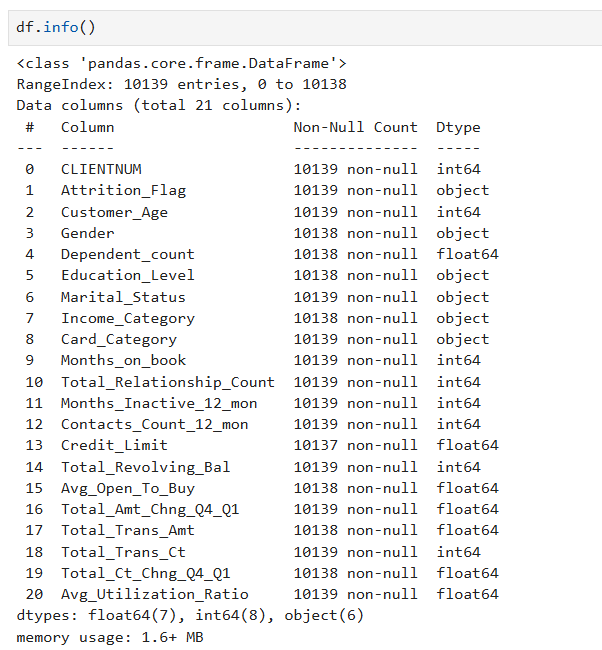
## 3. Mục tiêu của dự án

* Xác định các yếu tố liên quan đến mất khách hàng.
* Xây dựng mô hình có khả năng dự đoán chính xác khả năng mất khách hàng.
* Đề xuất các kế hoạch hành động để ngân hàng giảm mất khách hàng thẻ tín dụng.

# **III. Phân tích dữ liệu**

## 1. Xử lý và khám phá dữ liệu (Exploratory Data Analysis - EDA)

# 



### 1.1. Mô tả dữ liệu

CLIENTNUM: Mã số khách hàng. Mã định danh duy nhất cho khách hàng nắm giữ tài khoản.

Attrition\_Flag: Biến tình trạng rời bỏ- nếu tài khoản bị đóng thì là 1 nếu không thì là 0.

Customer\_Age: Độ tuổi của khách hàng tính theo năm.

Gender: Giới tính - M=Nam, F=Nữ.

Dependent\_count: Số lượng người phụ thuộc.

Education\_Level: Trình độ học vấn của chủ tài khoản (ví dụ: high school, college graduate, v.v.).

Marital\_Status: Tình trạng hôn nhân - Married, Single, Divorced, Unknown.

Income\_Category: Loại thu nhập hàng năm của chủ tài khoản (<40K, 40k - 60K,60k - 80K, 80k - 120K, >).

Card\_Category: Loại thẻ (Blue, Silver, Gold, Platinum).

Months\_on\_book: Thời gian quan hệ với ngân hàng.

Total\_Relationship\_Count: Tổng số sản phẩm mà khách hàng nắm giữ. Months\_Inactive\_12\_mon: Số tháng không hoạt động trong 12 tháng qua. Contacts\_Count\_12\_mon: Số lượng liên hệ trong 12 tháng qua. Rời rạc theo số.

Credit\_Limit: Hạn mức tín dụng trên Thẻ tín dụng. Liên tục theo số.

Total\_Revolving\_Bal: Tổng số dư luân chuyển trên thẻ tín dụng.

Avg\_Open\_To\_Buy: Hạn mức tín dụng mở để mua (Trung bình của 12 tháng qua).

Total\_Amt\_Chng\_Q4\_Q1: Thay đổi về số tiền giao dịch (Quý 4 so với Quý 1).

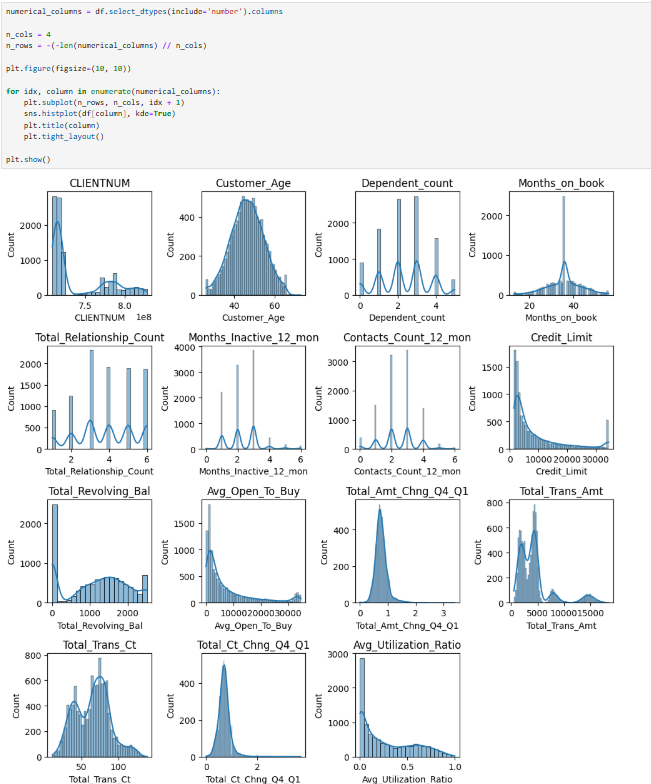
Total\_Trans\_Amt: Tổng số tiền giao dịch (12 tháng qua).

Total\_Trans\_Ct: Tổng số giao dịch (12 tháng qua).

Total\_Ct\_Chng\_Q4\_Q1: Thay đổi về số lượng giao dịch (Quý 4 so với Quý 1).

Avg\_Utilization\_Ratio: Tỷ lệ sử dụng thẻ trung bình.

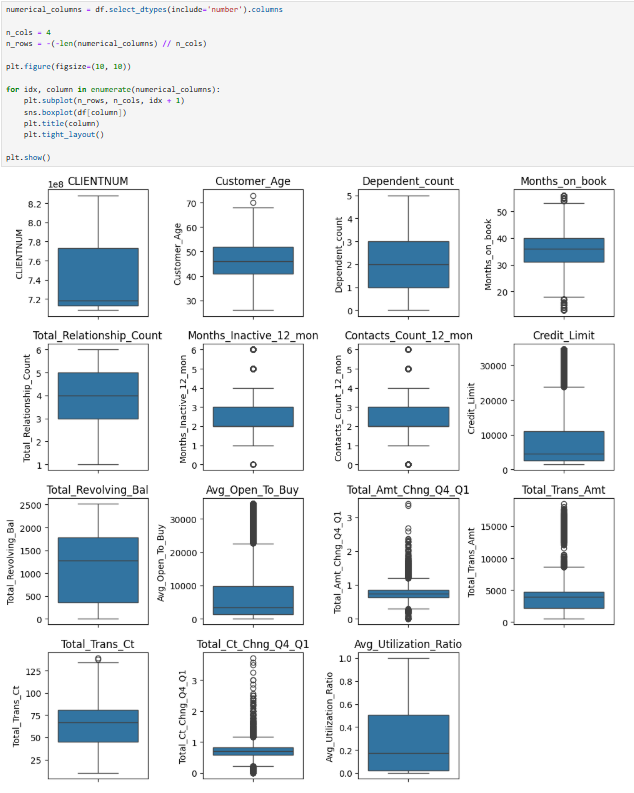
### 1.2. Phân tích dữ liệu khám phá



* Hầu hết khách hàng đều trong độ tuổi từ 40 đến 50.
* Một bộ phận lớn khách hàng đã dành 36 tháng để sử dụng dịch vụ thẻ tín dụng của ngân hàng.
* Hạn mức tín dụng bị lệch phải. Điều này có nghĩa là phần lớn khách hàng có hạn mức tín dụng thấp, trong khi có tương đối ít khách hàng có hạn mức tín dụng cao. Tuy nhiên, có thể thấy rằng một bộ phận đáng kể trong số họ, khoảng 5%, có hạn mức tín dụng cao, khoảng 35.000. Hơn nữa, cả avg\_open\_to\_buy và avg\_utilization\_ratio cũng bị lệch dương.
* Ngân hàng phải cảnh giác với sự phân phối của tỷ lệ sử dụng trung bình. Bị lệch phải có nghĩa là hầu hết khách hàng có tỷ lệ sử dụng dịch vụ thấp. Đặc biệt, một bộ phận lớn trong số họ, khoảng 25%, có tỷ lệ sử dụng bằng không. Do đó, công ty phải lập kế hoạch chiến lược để tăng tỷ lệ sử dụng và do đó là tăng doanh thu.
* Một bộ phận lớn khách hàng, cũng khoảng 25%, có tổng số dư luân chuyển bằng không. Tổng số dư luân chuyển thấp hơn (đặc biệt là số không) có thể chỉ ra khả năng xảy ra tình trạng churn đáng kể. Xem xét tỷ lệ sử dụng trung bình và tổng số dư luân chuyển có cùng tỷ lệ bản ghi bằng không, các biến này có thể có mối tương quan. Tôi sẽ điều tra thêm về điều này trong phần phân tích.
* Tổng số giao dịch và phân phối số tiền dường như có hai đỉnh. Phải có sự khác biệt trong các phân phối này đối với cờ churn, mà tôi sẽ điều tra thêm trong phần phân tích. Hơn nữa, một phần lớn khách hàng đã thực hiện khoảng 80 hoặc 40 giao dịch trong 12 tháng qua. Số lượng giao dịch khoảng 5.000 và 2.500 là phổ biến.
* Khoảng 80% khách hàng nắm giữ ít nhất 3 sản phẩm của ngân hàng.

### 1.3. Tiền xử lý dữ liệu

* Vì missing data chỉ chiếm 1 phần rất nhỏ không ảnh hưởng đến phân bố dữ liệu nên sẽ bị xóa
* Xóa duplicated data



* Loại bỏ outlier trên các biến liên tục có oulier rất lớn là Total\_Amt\_Chng\_Q4\_Q1, Total\_Ct\_Chng\_Q4\_Q1
* Sau khi loại bỏ outlier cho thấy hiệu suất của mô hình giảm xuống nên sẽ không xử lý outlier cho cột này

## 2. Phân tích mối quan hệ và tầm quan trọng của từng yếu tố

### 2.1. Chuẩn hóa dữ liệu

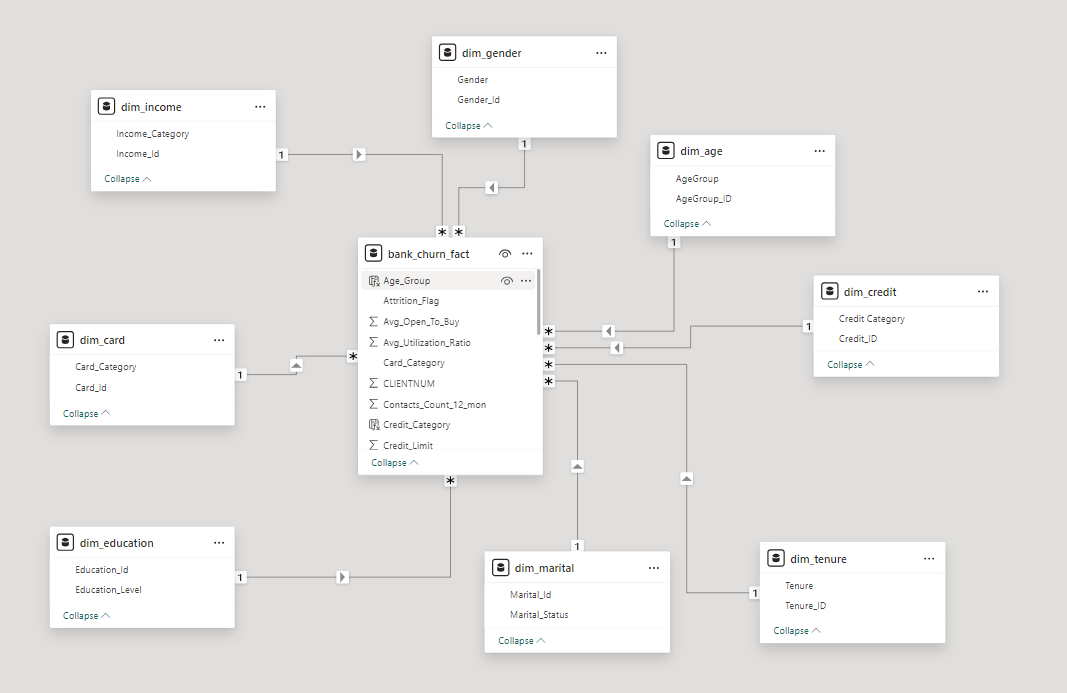
* Tạo các cột category phân loại

Age\_Group = SWITCH(TRUE(), bank\_churn\_fact[Customer\_Age] <= 30, "21-30", bank\_churn\_fact[Customer\_Age] >= 31 && bank\_churn\_fact[Customer\_Age] <= 40, "31-40", bank\_churn\_fact[Customer\_Age] >= 41 && bank\_churn\_fact[Customer\_Age] <= 50, "41-50", bank\_churn\_fact[Customer\_Age] >= 51 && bank\_churn\_fact[Customer\_Age] <= 60, "51-60", bank\_churn\_fact[Customer\_Age] >= 61 && bank\_churn\_fact[Customer\_Age] <= 70, "61-70", ">71")

Credit\_Category = SWITCH(TRUE(), bank\_churn\_fact[Credit\_Limit] <= 10000, "1K-10K", bank\_churn\_fact[Credit\_Limit] > 10000 && bank\_churn\_fact[Credit\_Limit] <= 20000, "10K-20K", bank\_churn\_fact[Credit\_Limit] > 20000 && bank\_churn\_fact[Credit\_Limit] <= 30000, "20K-30K", "30K-35K")

Tenure = SWITCH(TRUE(), bank\_churn\_fact[Months\_on\_book] <= 20, "<20", bank\_churn\_fact[Months\_on\_book] > 20 && bank\_churn\_fact[Months\_on\_book] <= 30, "21-30", bank\_churn\_fact[Months\_on\_book] > 30 && bank\_churn\_fact[Months\_on\_book] <= 40, "31-40", bank\_churn\_fact[Months\_on\_book] > 40 && bank\_churn\_fact[Months\_on\_book] <= 50, "41-50", ">51")

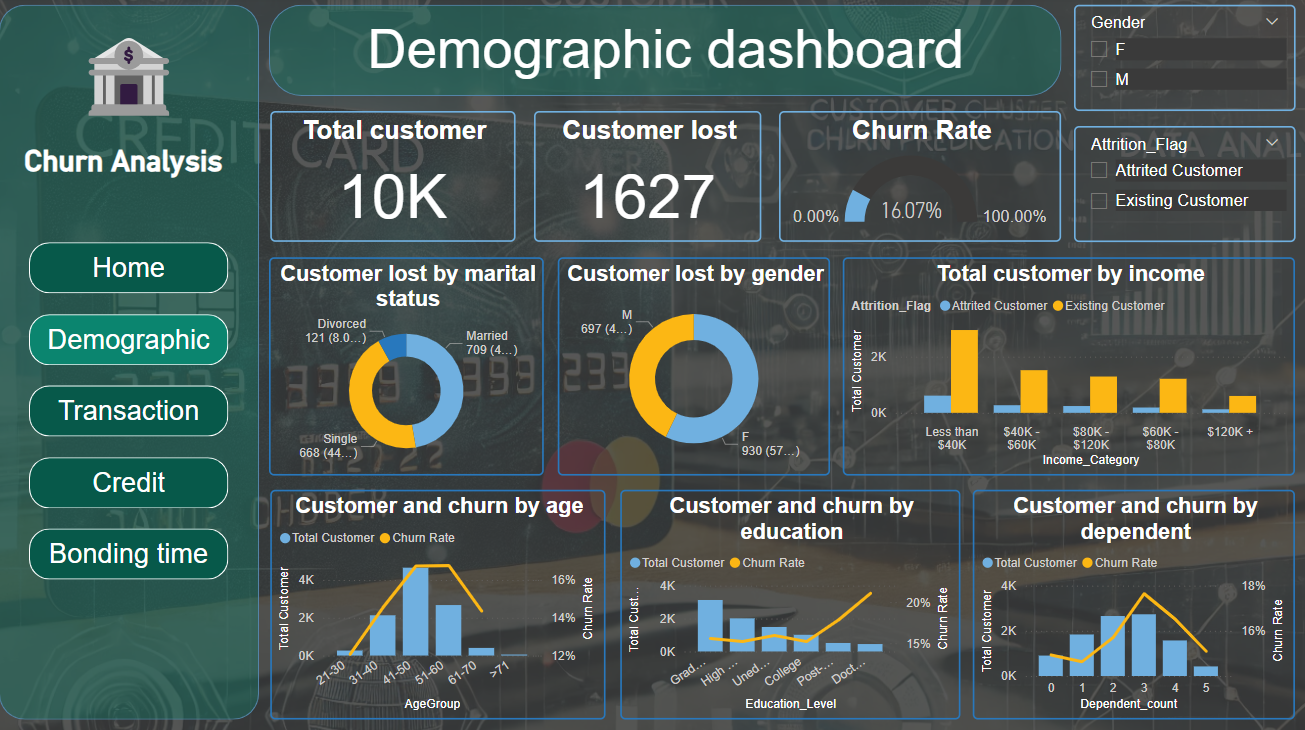
* Tạo 8 bảng dimension từ bảng fact gốc
* Data model: Sử dụng Star Schema giúp cải thiện hiệu suất truy vấn và giúp dữ liệu dễ bảo trì hơn



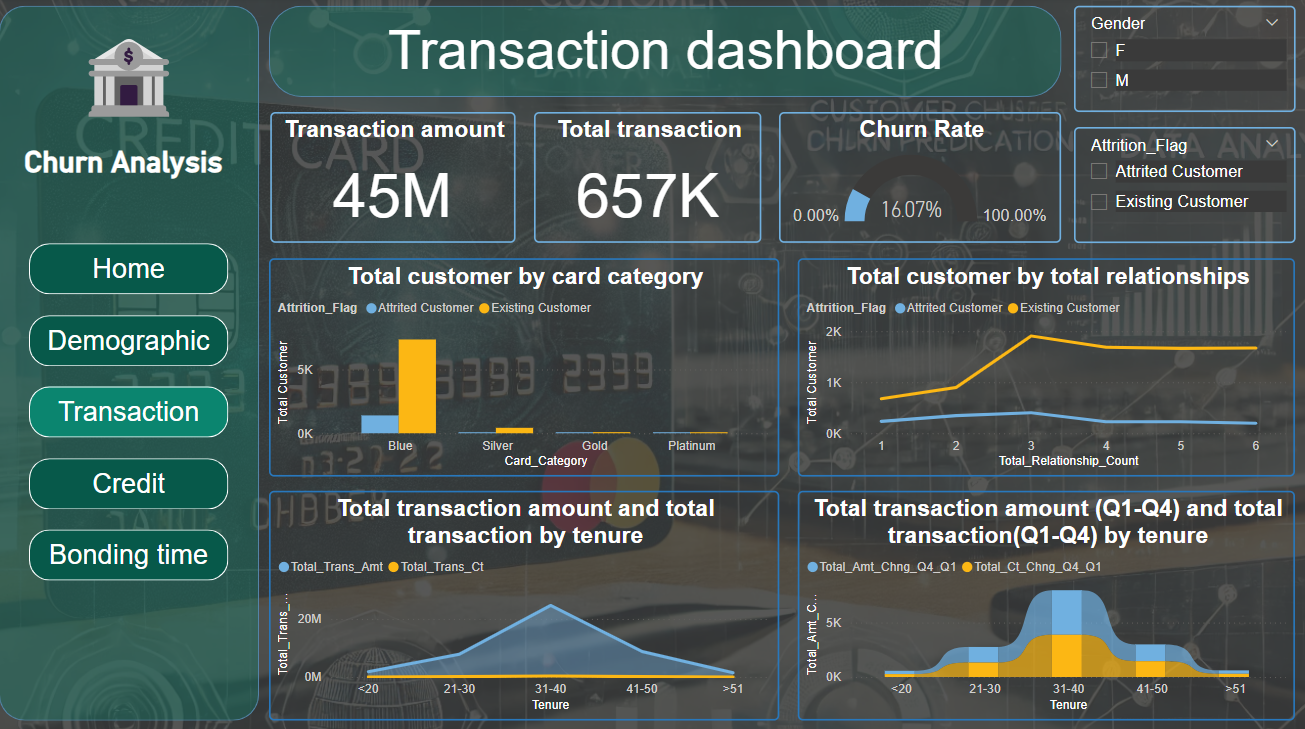
### 2.2. Phân tích dữ liệu

#### 2.2.1. Phân tích từng nhóm yếu tố

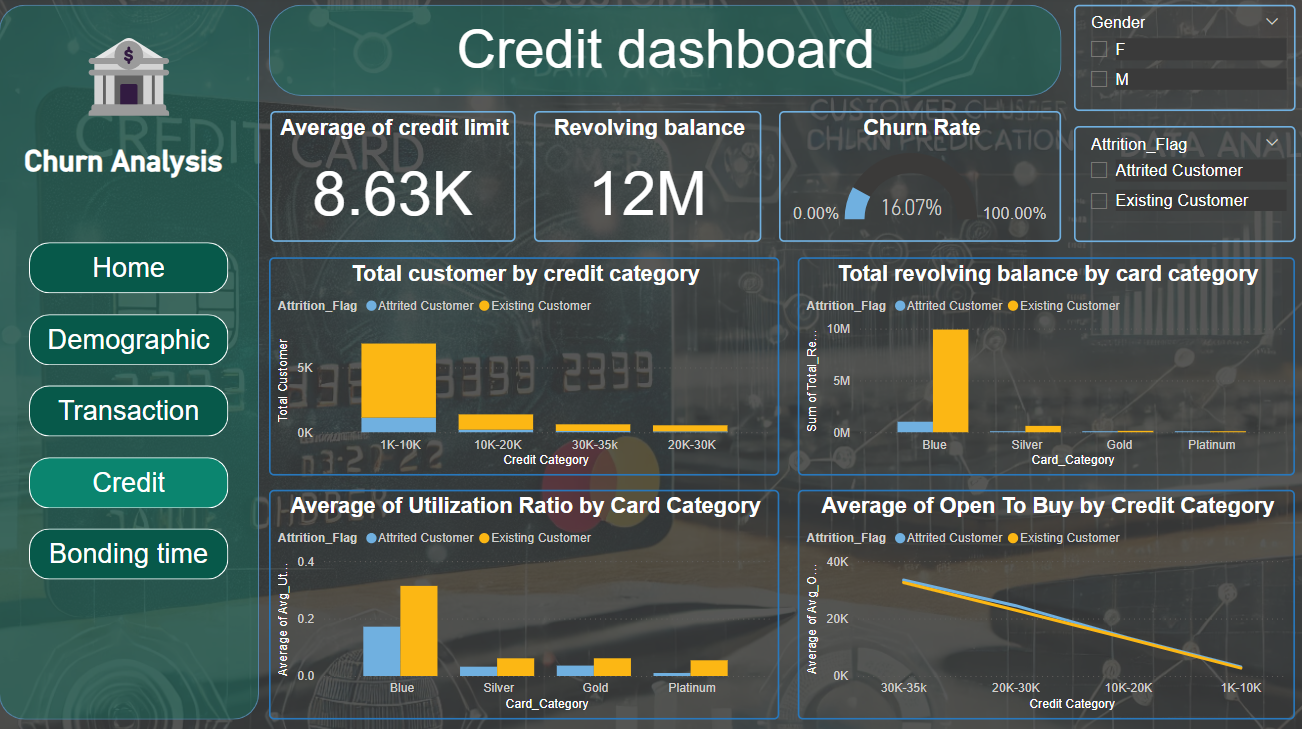
* Nhóm nhân khẩu học:



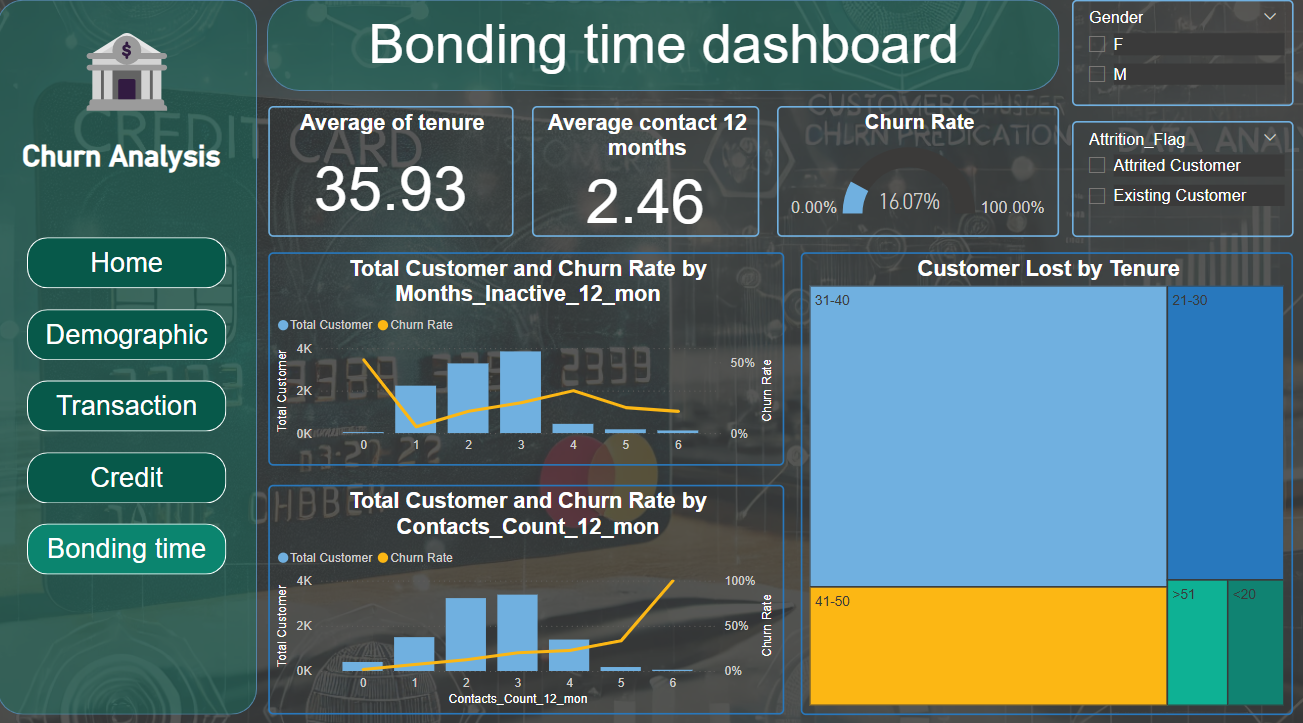
* + - Tỷ lệ churn 16.07% là khá cao, cho thấy doanh nghiệp đang đối mặt với vấn đề giữ chân khách hàng nghiêm trọng. Đây là tín hiệu cần thiết để ưu tiên các chiến lược cải thiện sự trung thành của khách hàng.
    - Khách hàng nữ có tỷ lệ churn cao hơn đáng kể so với nam (57.2% so với 42.8%). Điều này cho thấy cần điều tra sâu hơn về lý do tại sao khách hàng nữ rời bỏ nhiều hơn, có thể liên quan đến trải nghiệm dịch vụ hoặc các ưu đãi chưa phù hợp.
    - Khách hàng kết hôn và độc thân chiếm phần lớn tỷ lệ churn (tương ứng 43.6% và 41.0%), trong khi khách hàng ly hôn chỉ chiếm tỷ lệ nhỏ (8.0%). Tình trạng hôn nhân dường như không phải yếu tố quyết định lớn đến churn, nhưng cần lưu ý đến nhóm khách hàng kết hôn và độc thân vì đây là các nhóm đông đảo.
    - Nhóm có 3 người phụ thuộc có số khách hàng cao nhất (2,732) và tỷ lệ churn cao nhất (17.64%). Nhóm không có người phụ thuộc (0) và nhóm 5 người phụ thuộc có số khách hàng thấp nhất (904 và 424), với churn rate lần lượt là 14.93% và 15.09%. Nhóm có 2 người phụ thuộc có churn rate thấp nhất (14.64%). Kết luận: Khách hàng có 3-4 người phụ thuộc có xu hướng rời bỏ nhiều hơn, có thể do áp lực tài chính hoặc nhu cầu dịch vụ cao hơn.
    - Nhóm Graduate có số khách hàng cao nhất (3,128) nhưng churn rate ở mức trung bình (15.57%). Tỷ lệ churn tăng dần theo trình độ học vấn, đạt đỉnh ở nhóm Doctorate (21.06%), dù số lượng khách hàng thấp (451). Các nhóm học vấn thấp hơn (High School, Uneducated, College) có churn rate tương đối ổn định (15.20%-15.94%). Kết luận: Khách hàng có học vấn cao hơn (Post-Graduate, Doctorate) có xu hướng rời bỏ nhiều hơn, có thể do kỳ vọng cao hoặc dễ chuyển sang đối thủ cạnh tranh.
    - Nhóm 41-50 tuổi có số khách hàng lớn nhất (4,700) và tỷ lệ churn cao nhất (16.75%), là nhóm cần chú ý nhất. Nhóm trẻ (21-30) có ít khách hàng (300) và churn rate thấp hơn (12.08%). Nhóm lớn tuổi (61-70) không có churn (0%), cho thấy sự trung thành cao, nhưng số lượng khách hàng thấp (400). Kết luận: Khách hàng trung niên (41-50) là nhóm quan trọng nhất để cải thiện giữ chân do vừa đông vừa có tỷ lệ churn cao.
    - Nhóm thu nhập thấp nhất (dưới $40K) có số khách hàng hiện tại cao nhất (2,949) nhưng cũng có số lượng khách hàng rời bỏ đáng kể (612), dẫn đến tỷ lệ churn cao (17.2%). Nhóm thu nhập cao nhất (trên $120K) có ít khách hàng hơn (601 hiện tại, 126 rời bỏ) nhưng tỷ lệ churn cũng cao (17.3%). Các nhóm thu nhập trung bình ($40K-$80K) có tỷ lệ churn thấp hơn, dao động từ 14.1% đến 15.8%. Kết luận: Khách hàng ở hai cực thu nhập (thấp và cao) có xu hướng rời bỏ nhiều hơn, có thể do nhóm thấp nhạy cảm với giá, còn nhóm cao kỳ vọng dịch vụ tốt hơn.
* Nhóm liên quan đến sản phẩm và giao dịch:



* + - Hạng thẻ Blue là hạng phổ biến nhất, có số lượng khách hàng hiện tại (1,399 khách hàng đã churn và 7,354 khách hàng hiện tại, chiếm số lượng lớn nhất trong tất cả các hạng thẻ.). Tỷ lệ churn của các hạng thẻ Blue (15.98%), Silver (15.31%), và Gold (16.35%) khá tương đồng, dao động trong khoảng 15-16%. Hạng Platinum có tỷ lệ churn cao nhất (27.78%), nhưng tổng số khách hàng rất ít (18 người), nên con số này có thể không đại diện cho xu hướng chung mà chỉ phản ánh biến động ngẫu nhiên do mẫu nhỏ.
    - Tỷ lệ churn cao nhất xảy ra ở nhóm khách hàng nắm giữ ít sản phẩm: 1 sản phẩm (25.60%) và 2 sản phẩm (27.79%). Từ 3 sản phẩm trở lên, tỷ lệ churn giảm đáng kể: 17.35% (3 sản phẩm), 11.77% (4 sản phẩm), 12.00% (5 sản phẩm), và 10.50% (6 sản phẩm). Khách hàng nắm giữ càng nhiều sản phẩm thì tỷ lệ churn càng thấp, đặc biệt từ 3 sản phẩm trở lên, tỷ lệ churn giảm mạnh và ổn định ở mức 10-12%.
    - Lượng giao dịch (Blue Area): Đạt đỉnh 4.1K ở nhóm tenure 31-40, giảm xuống 1.4K (21-30), 1.5K (41-50), và 0.3K (<20 và >51). Số giao dịch (Yellow Area): Đạt đỉnh 3.9K ở nhóm tenure 31-40, với các giá trị tương tự thấp hơn ở các nhóm khác. Nhóm tenure 31-40 có cả lượng và số giao dịch cao nhất, cho thấy đây là nhóm tích cực nhất.
* Nhóm tín dụng:



* + - Hạng thẻ Blue chiếm phần lớn tổng số dư quay vòng, đặc biệt ở nhóm khách hàng hiện tại với 9.909 triệu USD, cao gấp gần 10 lần so với khách hàng đã churn (1.016 triệu USD). Điều này cho thấy thẻ Blue được sử dụng rất nhiều trong tín dụng. Các hạng thẻ cao cấp (Silver, Gold, Platinum) có tổng số dư quay vòng rất thấp, gần như không đáng kể ở cả hai nhóm khách hàng. Điều này có thể phản ánh rằng các thẻ này ít được sử dụng hoặc có ít khách hàng sở hữu hơn. Khách hàng hạng thẻ Blue là nhóm chủ lực trong việc duy trì số dư tín dụng, trong khi các hạng thẻ cao cấp dường như không đóng góp nhiều vào hoạt động tín dụng.
    - Phần lớn khách hàng tập trung ở khoảng tín dụng thấp (1K-10K), với tổng cộng 6.85 nghìn khách hàng (5.72K hiện tại + 1.13K đã churn), chiếm ưu thế tuyệt đối so với các khoảng khác. Các khoảng tín dụng cao hơn (10K-20K, 20K-30K, 30K-35K) có số lượng khách hàng giảm dần, đặc biệt số khách hàng đã churn ở các mức này rất thấp (từ 20 đến 200). Khách hàng có hạn mức thấp (1K-10K) không chỉ đông nhất mà còn có tỷ lệ churn cao nhất (1.13K), cho thấy nhóm này dễ rời bỏ dịch vụ hơn các nhóm hạn mức cao.
    - Hạng thẻ Blue có tỷ lệ sử dụng tín dụng trung bình cao nhất, đặc biệt ở khách hàng hiện tại (31.4%), gần gấp đôi so với khách hàng đã churn (17.2%). Điều này cho thấy nhóm khách hàng hiện tại ở hạng Blue sử dụng tín dụng rất tích cực. Các hạng thẻ cao cấp (Silver, Gold, Platinum) có tỷ lệ sử dụng thấp (dưới 6.2%), đặc biệt ở khách hàng đã churn (1.0%-3.6%). Điều này có thể do họ ít sử dụng tín dụng hoặc có hạn mức cao nhưng không tận dụng. Hạng thẻ Blue nổi bật về mức độ sử dụng tín dụng, trong khi các hạng thẻ cao cấp có xu hướng sử dụng tín dụng rất hạn chế.
    - Số tiền có thể vay thêm giảm dần từ khoảng tín dụng cao (30K-35K) xuống thấp (1K-10K), điều này hợp lý vì hạn mức tín dụng cao hơn sẽ có tiềm năng vay thêm lớn hơn. Trong mỗi khoảng, khách hàng đã churn thường có số tiền có thể vay thêm cao hơn một chút so với khách hàng hiện tại (ví dụ: 33.41K vs 32.58K ở 30K-35K). Điều này có thể chỉ ra rằng khách hàng đã churn không sử dụng hết hạn mức tín dụng trước khi rời bỏ dịch vụ. Khách hàng đã churn để lại nhiều dư địa tín dụng chưa sử dụng, đặc biệt ở các hạn mức cao, có thể do họ không hài lòng với dịch vụ hoặc không có nhu cầu sử dụng thêm.
* Nhóm thời gian và tương tác:



* + - Nhóm tenure 31-40 có số lượng khách hàng churn cao nhất (811), theo sau là nhóm 41-50 (319) và 21-30 (258). Nhóm >51 có số lượng churn thấp nhất (57). Điều này cho thấy khách hàng trong các nhóm tenure trung bình (21-50) có xu hướng churn nhiều hơn so với nhóm khách hàng lâu năm (>51). Có thể khách hàng lâu năm đã hình thành sự trung thành hoặc thói quen sử dụng dịch vụ.
    - Số lượng khách hàng không hoạt động cao nhất ở 3 tháng (3.85K), sau đó giảm mạnh ở các tháng tiếp theo. Tỷ lệ churn cao nhất ở khách hàng không hoạt động 0 tháng (51.72%), sau đó giảm mạnh ở 1 tháng (4.48%) nhưng lại tăng dần từ 2 đến 4 tháng, đạt đỉnh ở 4 tháng (29.89%), rồi giảm nhẹ ở 5-6 tháng. Khách hàng không hoạt động ngay lập tức (0 tháng) có nguy cơ churn rất cao, có thể do họ không hài lòng hoặc không có nhu cầu sử dụng dịch vụ. Tỷ lệ churn tăng dần ở khách hàng không hoạt động từ 2 đến 4 tháng, đặc biệt ở 4 tháng. Công ty cần có biện pháp kích hoạt lại khách hàng trong khoảng thời gian này.
    - Số lượng khách hàng cao nhất ở nhóm có 3 lần liên hệ (3,380), sau đó giảm dần khi số lần liên hệ tăng. Tỷ lệ churn tăng dần theo số lần liên hệ: từ 1.75% (0 lần) đến 100% (6 lần). Khách hàng có nhiều lần liên hệ (đặc biệt từ 4 lần trở lên) có nguy cơ churn rất cao, thậm chí tất cả khách hàng có 6 lần liên hệ đều churn. Số lần liên hệ là một yếu tố quan trọng ảnh hưởng đến churn. Có thể khách hàng liên hệ nhiều lần do gặp vấn đề hoặc không hài lòng với dịch vụ hỗ trợ. Công ty cần cải thiện chất lượng hỗ trợ để giảm thiểu churn ở nhóm này.

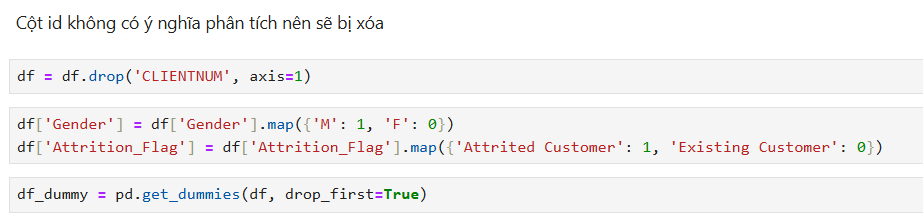
#### 2.2.2. Các insight chính

* Nữ giới, khách hàng đã kết hôn, và nhóm tuổi 41-50 có nguy cơ rời bỏ cao, cần chú ý cải thiện trải nghiệm cho các nhóm này.
* Khách hàng có 3 người phụ thuộc và trình độ học vấn cao (như Tiến sĩ) cũng có tỷ lệ rời bỏ đáng kể, phản ánh nhu cầu hoặc kỳ vọng đặc thù.
* Thẻ Blue là loại thẻ chủ lực với cơ sở khách hàng lớn và hoạt động giao dịch cao.
* Nhóm khách hàng 31-40 tuổi là nhóm giá trị nhất, cần được ưu tiên trong các chiến lược giữ chân và tăng cường tương tác.
* Các hạng tín dụng cao có ít khách hàng và thấp mức sử dụng, là cơ hội để mở rộng hoặc đánh giá lại sản phẩm.
* Khách hàng rời bỏ với tín dụng chưa sử dụng cho thấy khả năng không hài lòng hoặc có lựa chọn tốt hơn.
* Không hoạt động là yếu tố quan trọng ảnh hưởng đến churn, đặc biệt ở giai đoạn đầu và khi kéo dài.
* Số lần liên hệ cao tương quan với tỷ lệ rời bỏ cao, có thể phản ánh vấn đề về dịch vụ khách hàng.

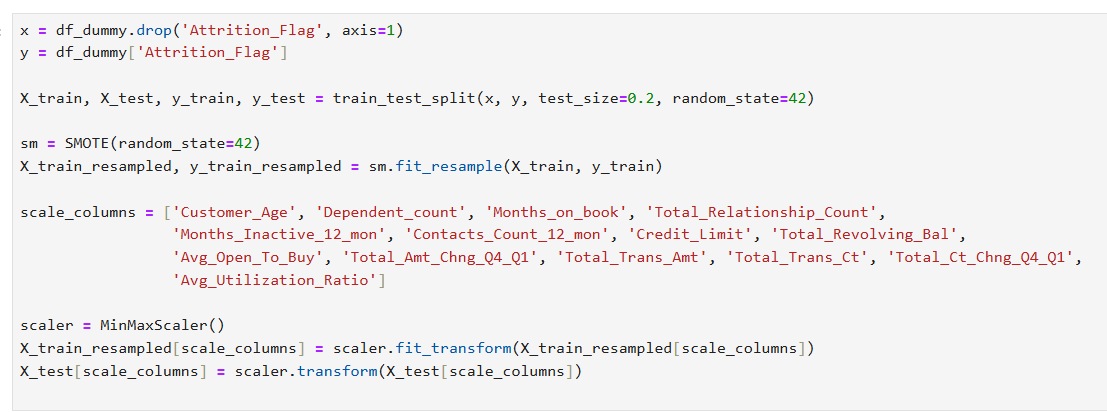
## 3. Xây dựng mô hình dự đoán

### 3.1. Xử lý dữ liệu

* Mã hóa biến phân loại (one-hot encoding).

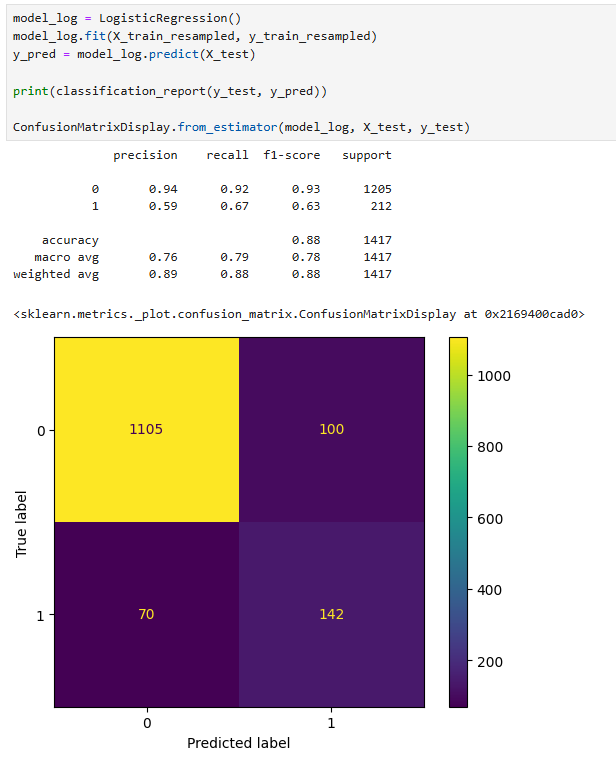


* Chuẩn hóa biến liên tục (min-max scaling) và xử lý imbalance(smote)

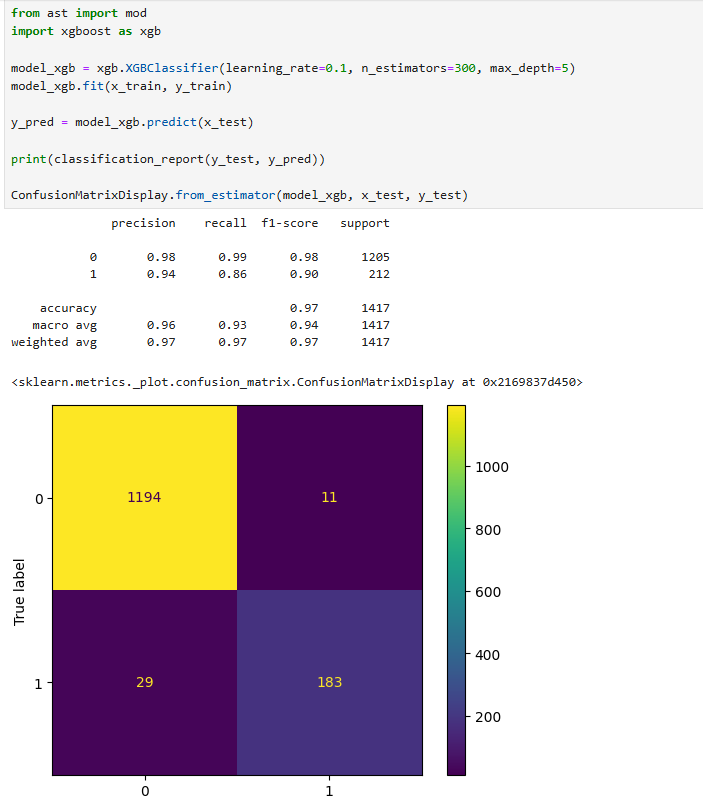


### 3.2. Chọn và đánh giá mô hình

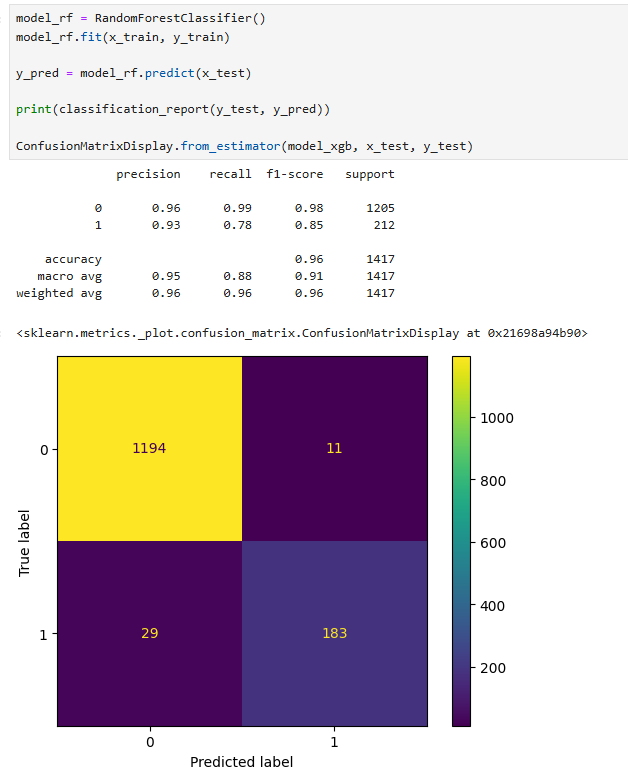
* Logistic regression

****

* Mô hình cho độ chính xác khá cao nhưng vì đây là bài toán với imbalance data nên chúng ta phải xét nhiều yếu tố khác.
* Cả precision, recall, f1 score rất thấp chỉ từ 0.59 - 0.67 ở nhãn 1.
* Xgboost



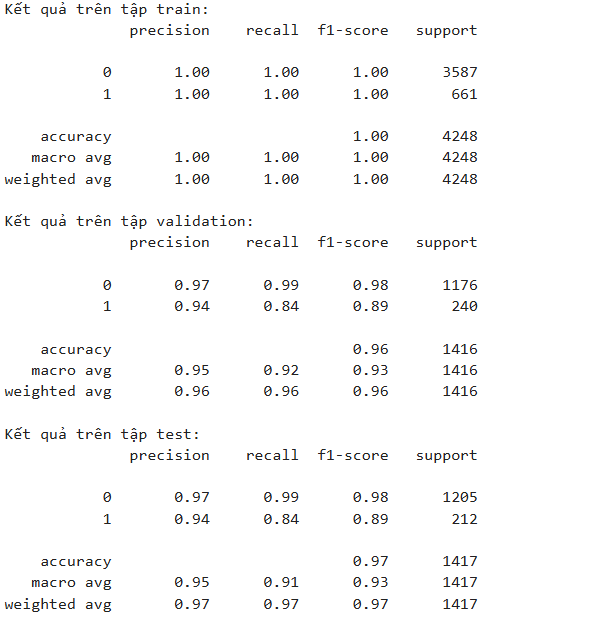
* Độ chính xác mô hình đã lên đến 0.97 và các điểm precision, recall, f1 đều khá cao lần lượt là 0.94, 0.86 và 0.90 ở nhãn 1.
* Mô hình khá tốt và có thể sử dụng.
* Random Forest

****

* Độ chính xác mô hình đã giảm xuống 0.96 và các điểm precision, recall, f1 đều giảm lần lượt là 0.93, 0.78 và 0.85 ở nhãn 1.
* Mô hình khá tốt và có thể sử dụng.

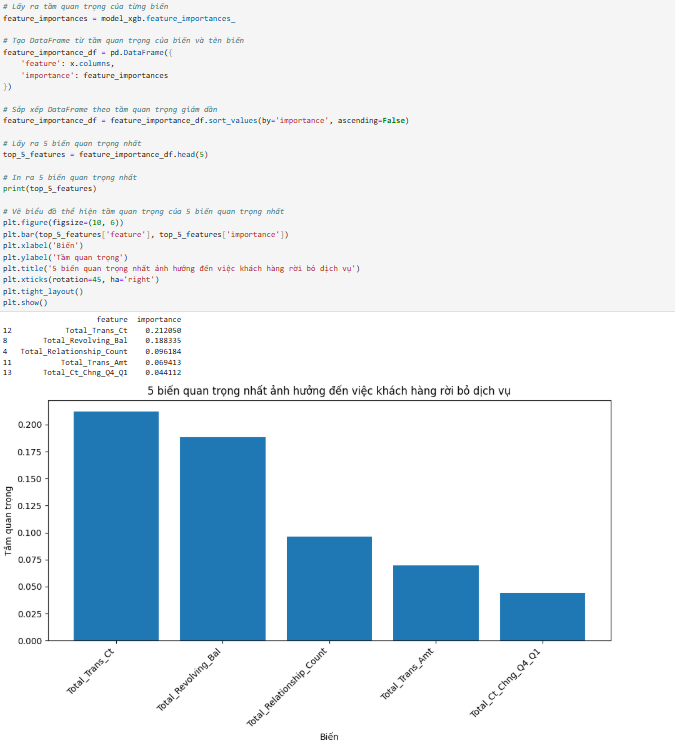






* Hiệu suất trên tập train: Mô hình đạt được hiệu suất rất cao trên tập train, với các chỉ số gần như hoàn hảo. Điều này là bình thường vì mô hình được huấn luyện trực tiếp trên tập dữ liệu này.
* Hiệu suất trên tập validation: Hiệu suất giảm nhẹ so với tập train, nhưng vẫn ở mức khá tốt. Điều này cho thấy mô hình đã bắt đầu tổng quát hóa được dữ liệu.
* Hiệu suất trên tập test: Hiệu suất trên tập test tương đồng với tập validation. Đây là một dấu hiệu tốt, cho thấy mô hình không bị overfitting quá nghiêm trọng.

### 3.3. Giải thích kết quả



* Số lượng giao dịch (Total\_Trans\_Ct) là yếu tố quan trọng nhất (21.2%). Khách hàng ít giao dịch có nguy cơ rời bỏ cao hơn.
* Số dư xoay vòng (Total\_Revolving\_Bal) đóng vai trò lớn thứ hai (18.8%). Khách hàng không sử dụng tín dụng nhiều dễ rời bỏ hơn.
* Mức độ gắn kết (Total\_Relationship\_Count) ảnh hưởng đáng kể (9.6%). Khách hàng chỉ sử dụng 1-2 sản phẩm/dịch vụ có nguy cơ churn cao hơn.
* Giá trị giao dịch (Total\_Trans\_Amt) và sự thay đổi tần suất giao dịch (Total\_Ct\_Chng\_Q4\_Q1) có tác động thấp hơn nhưng vẫn quan trọng trong việc dự đoán churn.

## 4. Đưa ra khuyến nghị

* Chiến Lược Giữ Chân:
* Tập trung vào nhóm khách hàng 31-40 năm, nữ giới, đã kết hôn, và có 3 người phụ thuộc bằng các chương trình khuyến khích hoặc cải thiện dịch vụ.
* Giảm thiểu thời gian không hoạt động bằng cách triển khai các chiến dịch tương tác chủ động.
* Cải Thiện Sản Phẩm và Dịch Vụ:
* Nâng cao giá trị của các thẻ cao cấp (Silver, Gold, Platinum) và hạng tín dụng cao để tăng mức sử dụng và hài lòng.
* Tùy chỉnh sản phẩm/dịch vụ cho nhóm khách hàng có học vấn cao để giảm tỷ lệ rời bỏ.
* Tăng Cường Tương Tác Khách Hàng:
* Giảm số lần liên hệ không cần thiết và cải thiện chất lượng dịch vụ để tránh làm khách hàng khó chịu.
* Xây dựng chiến lược giữ chân lâu dài nhằm tận dụng sự trung thành của khách hàng >51 năm.

## **PHẦN 2: KẾT LUẬN**

Trong bối cảnh ngành ngân hàng ngày càng cạnh tranh, việc giảm thiểu tỷ lệ khách hàng rời bỏ (churn) đóng vai trò thiết yếu trong việc duy trì lợi thế cạnh tranh và gia tăng lợi nhuận. Nghiên cứu này đã cung cấp một phân tích về hiện tượng churn thông qua việc xem xét dữ liệu từ các khía cạnh giao dịch, thời gian gắn bó, tín dụng và nhân khẩu học. Dự án không chỉ làm sáng tỏ nguyên nhân và xu hướng churn trong lĩnh vực ngân hàng mà còn đưa ra các khuyến nghị khả thi nhằm giảm thiểu churn, tối ưu hóa giá trị khách hàng và thúc đẩy sự phát triển bền vững. Việc áp dụng những giải pháp này hứa hẹn sẽ mang lại hiệu quả lâu dài cho tổ chức tài chính trong bối cảnh thị trường đầy thách thức.