Báo cáo Phân tích Dữ liệu Khám phá (EDA) về Khách hàng Rời bỏ Dịch vụ Ngân hàng

Tháng 6, 2025

# Giới thiệu

Báo cáo này thực hiện phân tích dữ liệu khám phá (EDA) trên tập dữ liệu churn\_customer.csv, chứa thông tin về khách hàng của ngân hàng B, với mục tiêu xác định các yếu tố ảnh hưởng đến việc khách hàng rời bỏ dịch vụ (churn). Tập dữ liệu bao gồm các đặc điểm như điểm tín dụng, quốc gia, giới tính, tuổi, thời gian sử dụng dịch vụ, số dư tài khoản, số lượng sản phẩm, trạng thái thẻ tín dụng, thành viên hoạt động, lương ước tính, và biến mục tiêu là churn (1: rời bỏ, 0: không rời bỏ). Phân tích bao gồm hai phần chính: phân tích mô tả để hiểu cấu trúc dữ liệu và phân phối các biến, và phân tích chẩn đoán để xác định các yếu tố liên quan đến việc rời bỏ dịch vụ. Kết quả chỉ ra rằng tuổi, số dư tài khoản, trạng thái thành viên hoạt động, và quốc gia là những yếu tố có ảnh hưởng đáng kể đến tỷ lệ rời bỏ.

## Mục đích

* Hiểu hành vi khách hàng: Thực hiện phân tích khám phá dữ liệu (EDA) để nhận diện các đặc điểm và xu hướng liên quan đến việc khách hàng rời bỏ dịch vụ, từ đó xác định các yếu tố ảnh hưởng đến quyết định rời bỏ.
* Dự đoán khách hàng có nguy cơ rời bỏ: Phát triển mô hình học máy để dự đoán chính xác những khách hàng có khả năng rời bỏ dịch vụ, hỗ trợ ngân hàng can thiệp kịp thời (ví dụ: cung cấp ưu đãi, cải thiện dịch vụ).
* Hỗ trợ ra quyết định kinh doanh: Cung cấp báo cáo chi tiết và các khuyến nghị dựa trên dữ liệu để tối ưu hóa chiến lược kinh doanh, giảm tỷ lệ churn, và tiết kiệm chi phí so với việc thu hút khách hàng mới.

# Phân tích Mô tả

Phần này cung cấp cái nhìn tổng quan về cấu trúc dữ liệu, phân phối các biến số và biến phân loại, cũng như kiểm tra giá trị thiếu.

## Cấu trúc Dữ liệu

Tập dữ liệu chứa 10,000 bản ghi với 12 cột, bao gồm:

* customer\_id: Mã định danh khách hàng (kiểu số nguyên).
* credit\_score: Điểm tín dụng (kiểu số nguyên, từ 350 đến 850).
* country: Quốc gia (France, Spain, Germany).
* gender: Giới tính (Male, Female).
* age: Tuổi (kiểu số nguyên, từ 18 đến 92).
* tenure: Thời gian sử dụng dịch vụ (năm, từ 0 đến 10).
* balance: Số dư tài khoản (kiểu số thực, đã được chuyển đổi từ chuỗi).
* products\_number: Số lượng sản phẩm sử dụng (từ 1 đến 4).
* credit\_card: Có thẻ tín dụng (1: có, 0: không).
* active\_member: Thành viên hoạt động (1: có, 0: không).
* estimated\_salary: Lương ước tính (kiểu số thực, đã được chuyển đổi từ chuỗi).
* churn: Rời bỏ dịch vụ (1: rời bỏ, 0: không rời bỏ).

Kết quả kiểm tra cho thấy không có giá trị thiếu trong tập dữ liệu, đảm bảo chất lượng dữ liệu cho phân tích.

## Thống kê Mô tả

Bảng 1 tóm tắt các đặc điểm của các biến số:

Bảng 1: Thống kê mô tả các biến số

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Biến | Số lượng | Trung bình | Độ lệch chuẩn | Min | Max | Trung vị |
| credit\_score | 10000 | 650.53 | 96.65 | 350 | 850 | 652 |
| age | 10000 | 38.92 | 10.49 | 18 | 92 | 37 |
| tenure | 10000 | 5.01 | 2.89 | 0 | 10 | 5 |
| balance | 10000 | 76485.89 | 62397.41 | 0 | 250898.09 | 97198.54 |
| products\_number | 10000 | 1.53 | 0.58 | 1 | 4 | 1 |
| estimated\_salary | 10000 | 100090.24 | 57510.49 | 11.58 | 199992.48 | 100193.91 |

Nhận xét:

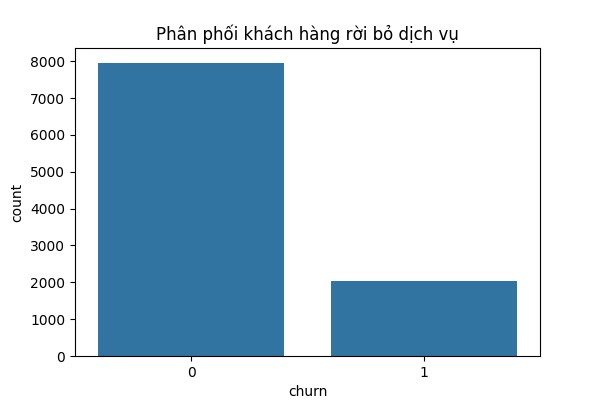
* credit\_score: Phân phối gần giống phân phối chuẩn, với trung bình 650.53 và độ lệch chuẩn 96.65.
* age: Tuổi trung bình là 38.92, nhưng có sự biến động lớn (từ 18 đến 92).
* tenure: Thời gian sử dụng dịch vụ phân bố đều từ 0 đến 10 năm.
* balance: Khoảng 36% khách hàng có số dư bằng 0, trong khi số dư trung bình là

76485.89.

* products\_number: Phần lớn khách hàng sử dụng 1 hoặc 2 sản phẩm.
* estimated\_salary: Lương ước tính phân bố đều, với trung bình 100090.24.

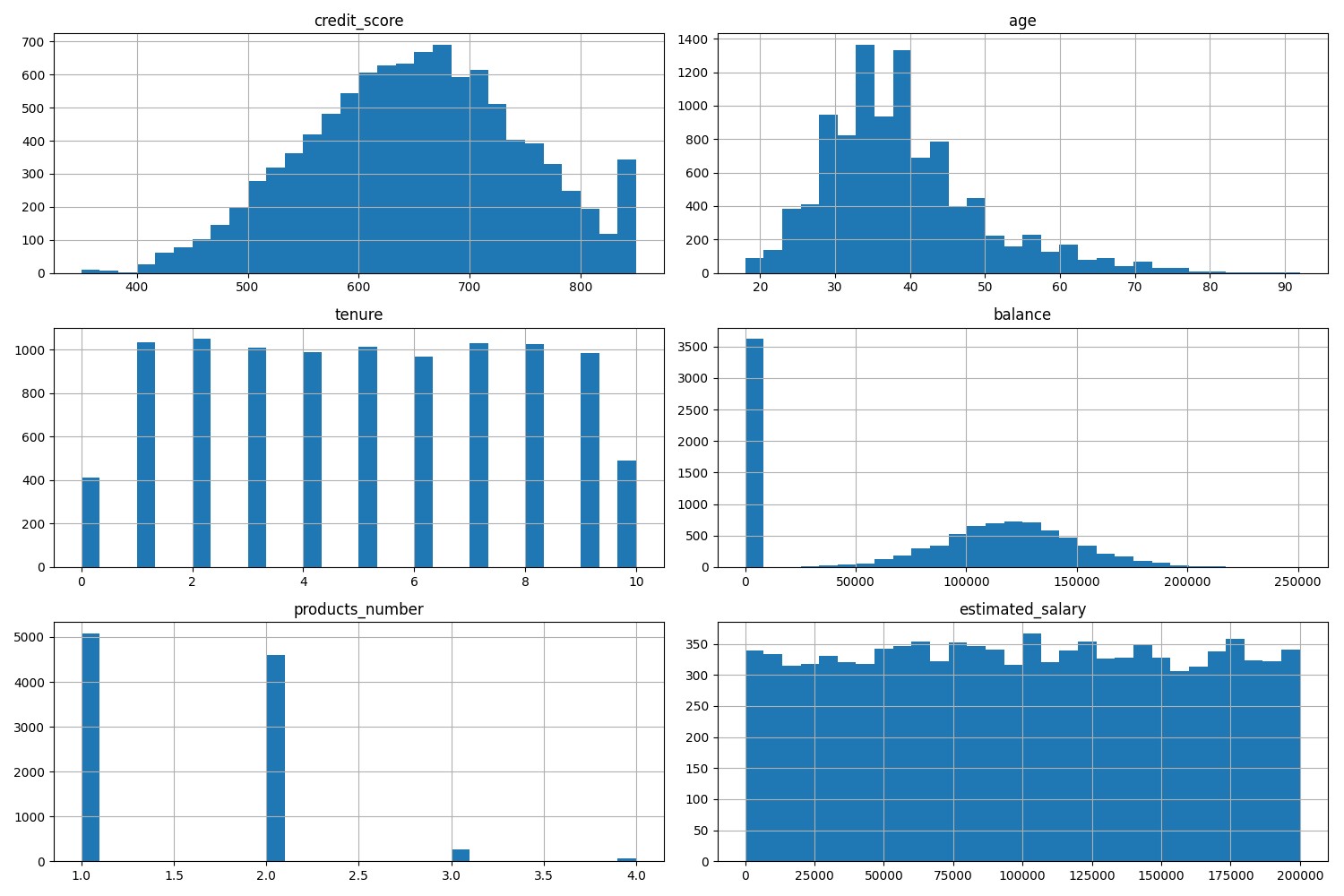
## Phân phối Biến Mục tiêu (Churn)

Hình 1 cho thấy tỷ lệ khách hàng rời bỏ dịch vụ (churn = 1) là khoảng 20.37% (2037 khách hàng), trong khi 79.63% (7963 khách hàng) tiếp tục sử dụng dịch vụ (churn = 0).



Hình 1: Phân phối khách hàng rời bỏ dịch vụ

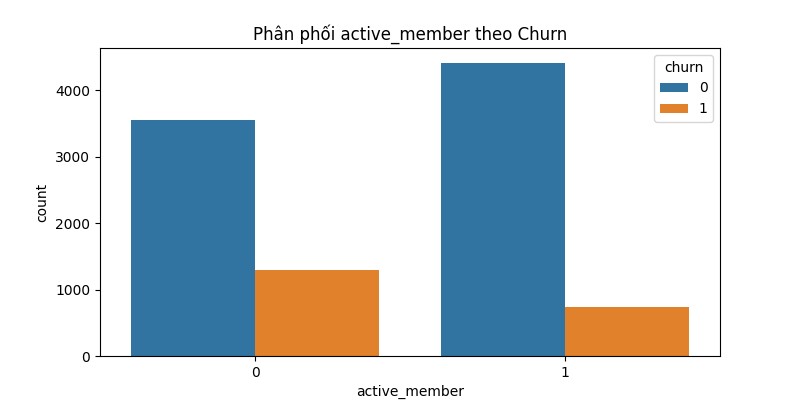
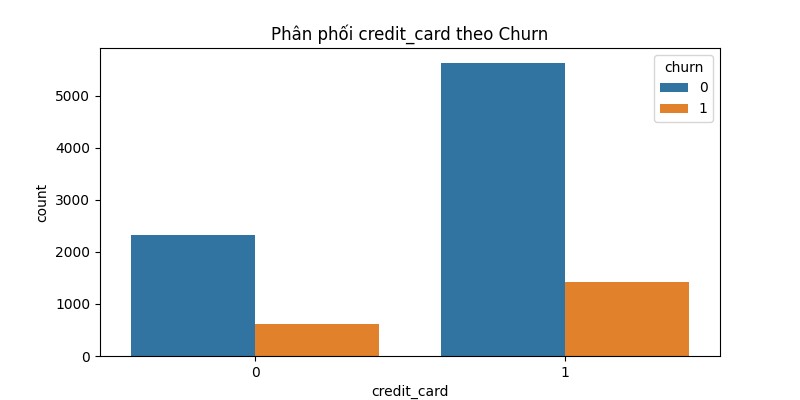
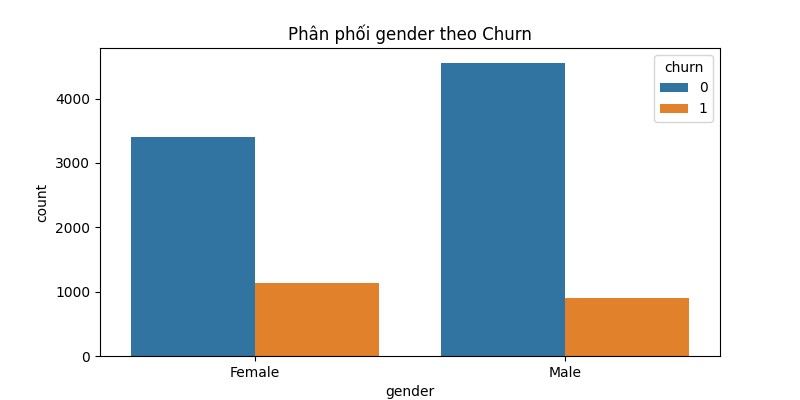
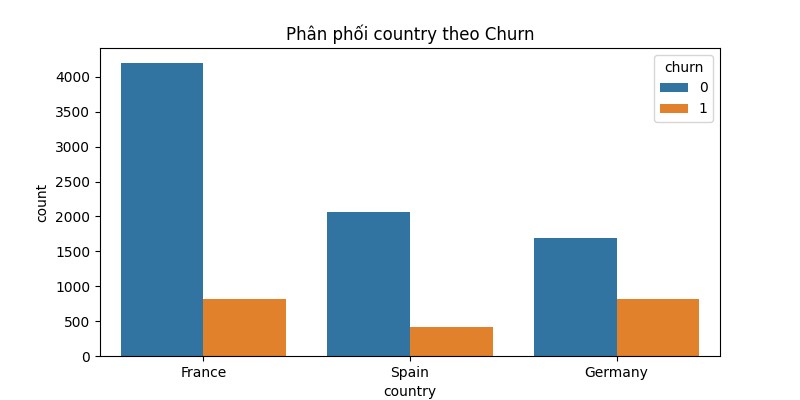
## Phân phối Biến Số



Hình 2: Phân phối các biến số

* credit\_score: Phân phối hơi lệch trái, với một số lượng nhỏ khách hàng có điểm tín dụng thấp.
* age: Phân phối lệch phải, với phần lớn khách hàng từ 30 đến 50 tuổi.
* balance: Có một đỉnh lớn tại giá trị 0, cho thấy nhiều khách hàng không có số dư.
* estimated\_salary: Phân bố gần giống đồng đều.

## Phân phối Biến Phân loại



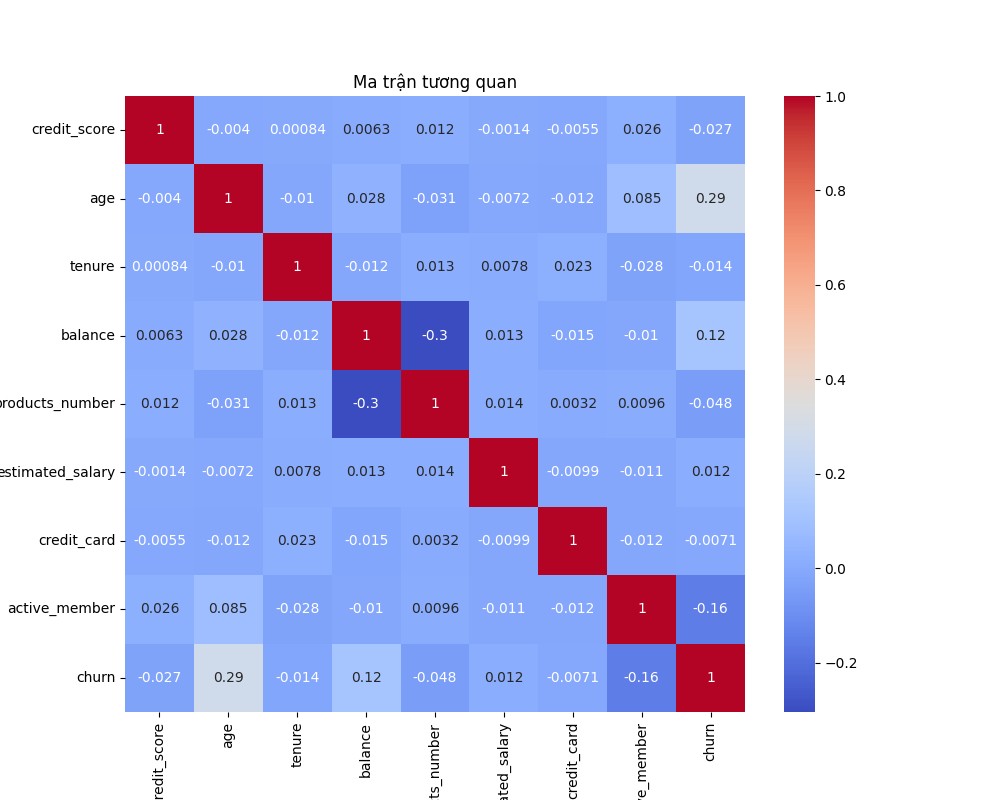
Hình 3: Phân phối các biến phân loại theo Churn

* country: Khách hàng ở Đức có tỷ lệ rời bỏ cao hơn (32.5%) so với Pháp (16.2%) và Tây Ban Nha (16.7%).
* gender: Nữ giới có tỷ lệ rời bỏ cao hơn nam giới (25.1% so với 16.5%).
* Credit\_card: Tỷ lệ rời bỏ không khác biệt đáng kể giữa khách hàng có và không có thẻ tín dụng.

# Phân tích Chẩn đoán

Phần này tập trung vào việc xác định các yếu tố ảnh hưởng đến việc khách hàng rời bỏ dịch vụ thông qua ma trận tương quan và so sánh các biến theo trạng thái churn.

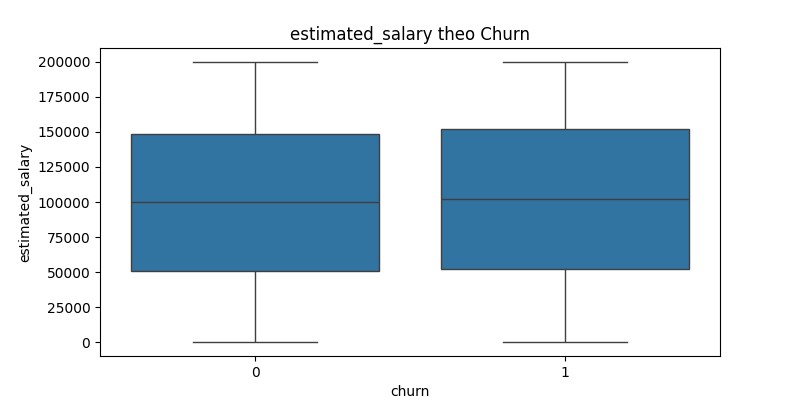
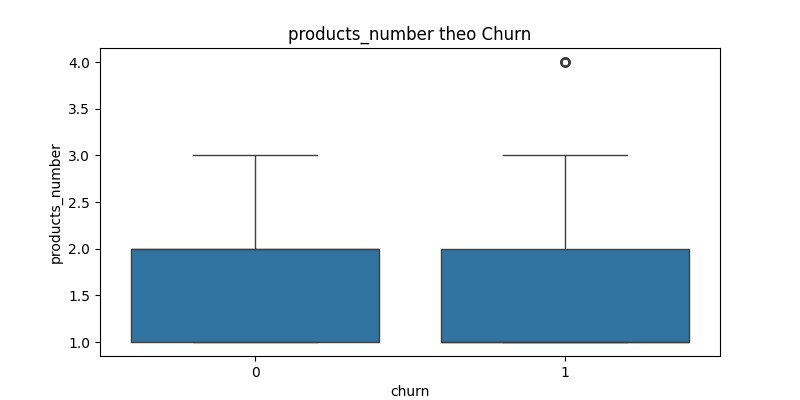
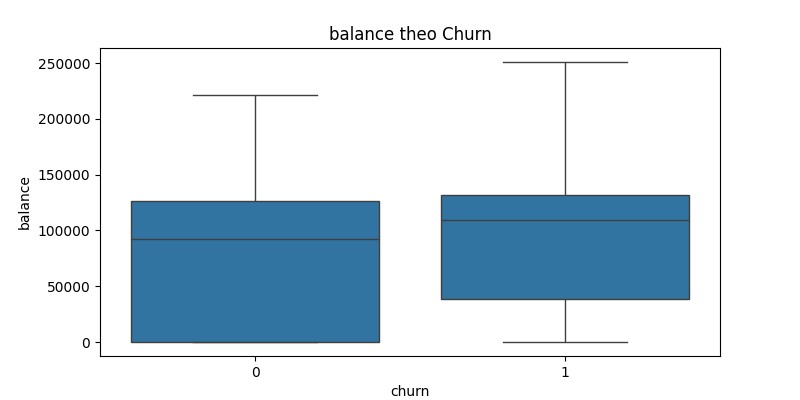
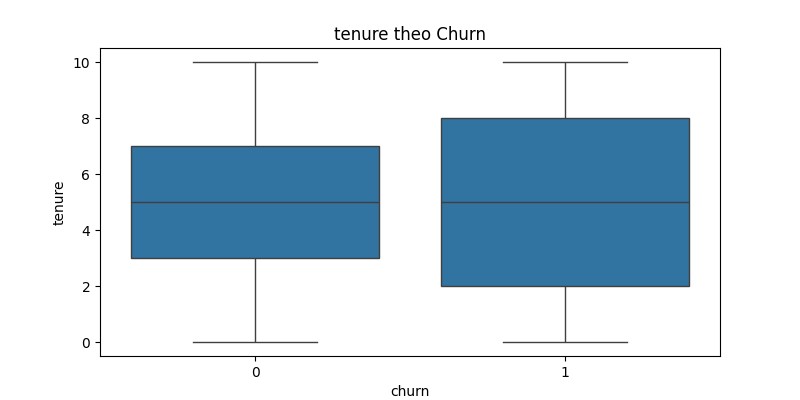
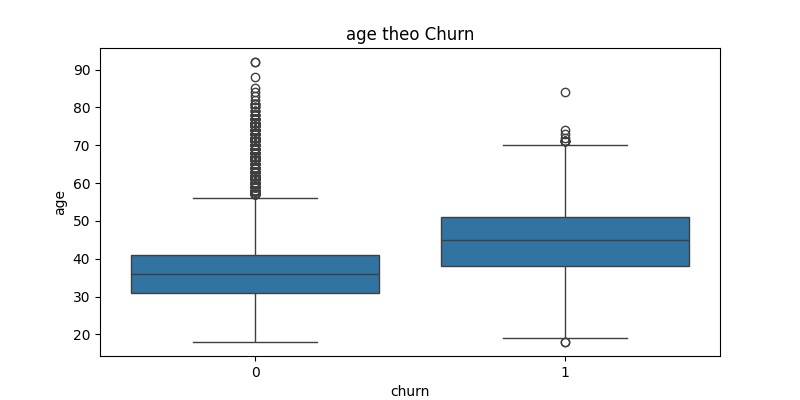
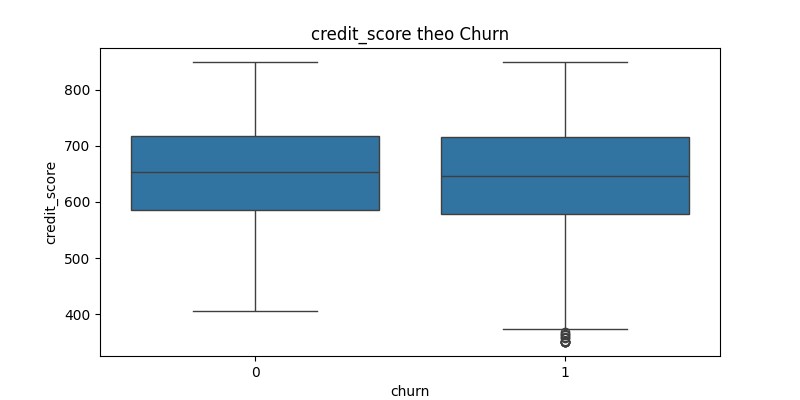
## Ma trận Tương quan



Hình 4: Ma trận tương quan

* age có tương quan dương với churn (0.29), cho thấy khách hàng lớn tuổi hơn có xu hướng rời bỏ nhiều hơn.
* active\_member có tương quan âm với churn (-0.16), cho thấy khách hàng hoạt động ít có khả năng rời bỏ hơn.
* balance có tương quan dương nhẹ với churn (0.12).
* Các biến khác như credit\_score, tenure, estimated\_salary có tương quan rất yếu với churn.

## So sánh Biến Số theo Churn



Hình 5: Boxplot các biến số theo Churn

* age: Khách hàng rời bỏ có tuổi trung bình cao hơn (44.84 so với 37.41).
* balance: Khách hàng rời bỏ có số dư trung bình cao hơn (91108.54 so với 72745.37).
* products\_number: Khách hàng rời bỏ có xu hướng sử dụng ít sản phẩm hơn (1.47 so với 1.54)

# Kết luận phần EDA

Phân tích EDA trên tập dữ liệu churn\_customer.csv cho thấy các yếu tố chính ảnh hưởng đến việc khách hàng rời bỏ dịch vụ của ngân hàng B:

* Tuổi: Khách hàng lớn tuổi hơn có xu hướng rời bỏ dịch vụ nhiều hơn, với tuổi trung bình của nhóm rời bỏ là 44.84 so với 37.41 của nhóm không rời bỏ.
* Quốc gia: Khách hàng ở Đức có tỷ lệ rời bỏ cao hơn đáng kể (32.5%) so với Pháp và Tây Ban Nha (khoảng 16%).
* Thành viên hoạt động: Khách hàng không hoạt động có tỷ lệ rời bỏ cao hơn (26.8%) so với khách hàng hoạt động (14.3%).
* Số dư tài khoản: Khách hàng có số dư cao hơn có xu hướng rời bỏ nhiều hơn.
* Giới tính: Nữ giới có tỷ lệ rời bỏ cao hơn nam giới.

Các yếu tố như điểm tín dụng, thời gian sử dụng dịch vụ, và lương ước tính không cho thấy mối liên hệ mạnh mẽ với việc rời bỏ. Ngân hàng B nên tập trung vào các chiến lược giữ chân khách hàng lớn tuổi, khách hàng ở Đức, và khách hàng không hoạt động, chẳng hạn như cung cấp các chương trình khuyến khích hoặc cải thiện trải nghiệm khách hàng.

# Xây dựng mô hình

## Mục tiêu bài toán

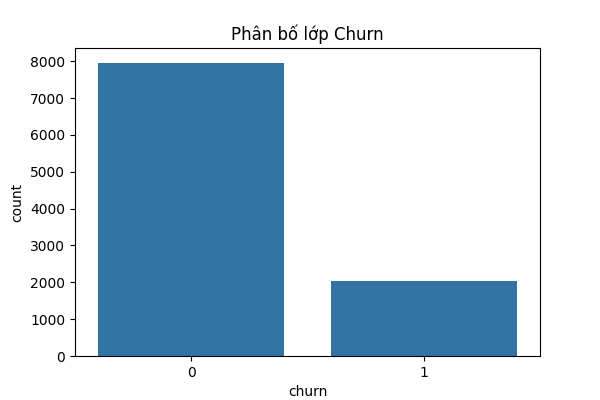
Xây dựng mô hình dự đoán phân loại việc khách hàng rời bỏ dịch vụ ngân hàng B, nhằm phát hiện sớm những khách hàng có nguy cơ rời bỏ (churn) để hỗ trợ ngân hàng triển khai các chiến lược giữ chân hiệu quả.

## Tổng quan về dữ liệu

**Quy mô dữ liệu**: Bộ dữ liệu gồm 10,000 mẫu, với 1593 mẫu không churn (lớp 0) và 407 mẫu churn (lớp 1) trong tập test (tỷ lệ 80:20).

 **Phân bố lớp**:

* Lớp 0 (không churn): 79.63%.
* Lớp 1 (churn): 20.37%.

 **Nhận xét**: Dữ liệu mất cân bằng nghiêm trọng, đòi hỏi các kỹ thuật xử lý như SMOTE, RUS, ADASYN, và SMOTEENN để cải thiện khả năng dự đoán lớp thiểu số (churn).

 **Các đặc trưng**:

* Biến số: credit\_score, age, tenure, balance, products\_number, estimated\_salary.
* Biến phân loại: gender, country, credit\_card, active\_member.
* Biến mục tiêu: churn (0: không churn, 1: churn).

 **Xử lý dữ liệu**:

* Làm sạch: Chuyển đổi dấu phẩy thành dấu chấm cho các cột balance và estimated\_salary.
* Mã hóa: gender bằng LabelEncoder, country bằng One-Hot Encoding.
* Chuẩn hóa: Các biến số bằng StandardScaler.
* Loại bỏ: Cột customer\_id không cần thiết.

## Phương pháp xây dựng mô hình

 **Mô hình sử dụng**:

* **Logistic Regression**: Một mô hình đơn giản, dễ hiểu, và mạnh mẽ trong việc dự đoán churn. Nó cung cấp các **hệ số đặc trưng** rõ ràng, giúp ngân hàng dễ dàng xác định các yếu tố chính ảnh hưởng đến việc khách hàng rời bỏ, như tuổi tác hay quốc gia, từ đó xây dựng chiến lược giữ chân hiệu quả.
* **Random Forest**: Mô hình dựa trên tập hợp nhiều **cây quyết định**, hoạt động như một "đội ngũ chuyên gia" để đưa ra dự đoán chính xác, đặc biệt hiệu quả với dữ liệu phức tạp. Nó phù hợp để phát hiện các mẫu khách hàng churn trong những tình huống có nhiều yếu tố tương tác.
* **XGBoost**: Một mô hình tiên tiến sử dụng **gradient boosting**, được thiết kế để xử lý tốt dữ liệu mất cân bằng. Với khả năng học hỏi thông minh từ dữ liệu, XGBoost giúp phát hiện các khách hàng có nguy cơ rời bỏ với độ chính xác cao, ngay cả khi lớp churn chiếm tỷ lệ nhỏ.

 **Tối ưu hóa**:

* Sử dụng **GridSearchCV** để tìm siêu tham số tốt nhất (ví dụ: C cho Logistic Regression, n\_estimators, max\_depth cho Random Forest và XGBoost).
* Thử nghiệm ngưỡng phân loại (0.2, 0.25, 0.3, 0.4, 0.5) để tối ưu Recall lớp 1.

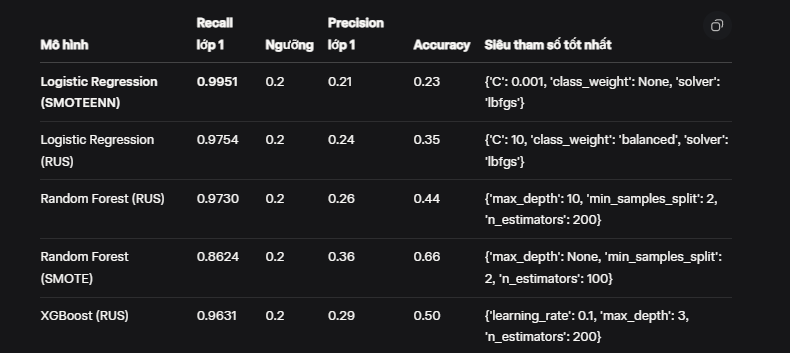
 **Kỹ thuật xử lý mất cân bằng**:

* **SMOTE**: Tạo mẫu tổng hợp cho lớp thiểu số, đạt tỷ lệ 50:50.
* **RUS**: Giảm mẫu lớp đa số ngẫu nhiên, đạt tỷ lệ 50:50.
* **ADASYN**: Tạo mẫu tổng hợp tập trung vào các mẫu khó phân loại, tỷ lệ gần 50:50.
* **SMOTEENN**: Kết hợp SMOTE và loại bỏ nhiễu (ENN), dẫn đến tỷ lệ lớp 1 cao hơn (57.46%).

 **Chỉ số đánh giá**:

* Ưu tiên **Recall lớp 1** để phát hiện tối đa khách hàng churn.
* Sử dụng **Confusion Matrix** và **Classification Report** để đánh giá Recall, Precision, và Accuracy.
* Precision và Accuracy là chỉ số phụ, chấp nhận thấp để ưu tiên Recall.

## Kết quả của mô hình



Mô hình tốt nhất: Logistic Regression (SMOTEENN):

Recall lớp 1: 0.9951 (phát hiện 405/407 trường hợp churn, chỉ bỏ sót 2 trường hợp).

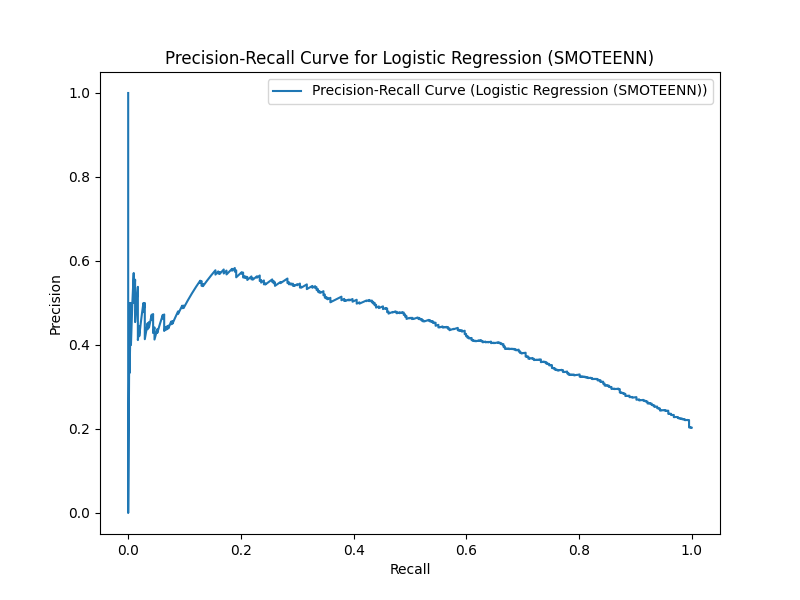
Precision lớp 1: 0.21 (chỉ 21% dự đoán churn là đúng, do 1547 False Positives).

Accuracy: 0.23 (thấp do dự đoán quá nhiều mẫu là lớp 1).

 **Đánh đổi**

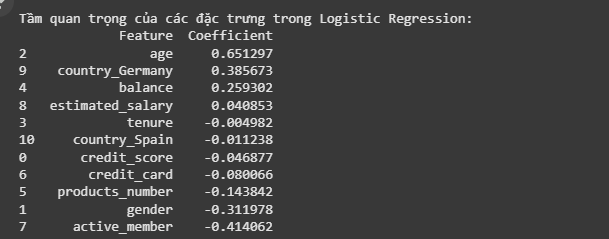
* Recall cao (0.9951) là lý tưởng để phát hiện gần hết khách hàng churn, phù hợp với mục tiêu đề bài.
* Precision thấp (0.21) dẫn đến nhiều khách hàng không churn bị dự đoán sai (1547 FP), làm tăng chi phí giữ chân không cần thiết.

 **Precision-Recall Curve**:



cho thấy trade-off giữa Recall và Precision, với Recall gần 1 tại ngưỡng thấp.

## Phân tích đặc trưng quan trọng



* **Mô hình**: Logistic Regression (SMOTEENN).
* **Hệ số đặc trưng**:
* **Nhận xét**:
  + **age** (0.6513): Khách hàng lớn tuổi có nguy cơ churn cao hơn.
  + **country\_Germany** (0.3857): Khách hàng ở Đức dễ churn hơn so với các quốc gia khác.
  + **balance** (0.2593): Số dư tài khoản cao liên quan đến khả năng churn cao.
  + **active\_member** (-0.4141): Thành viên tích cực ít có khả năng churn.
  + **products\_number** (-0.1438): Sử dụng nhiều sản phẩm giảm nguy cơ churn.
* **Ý nghĩa kinh doanh**: Ngân hàng nên tập trung vào khách hàng lớn tuổi, ở Đức, hoặc có số dư cao để triển khai chiến lược giữ chân.

## Lý do chọn mô hình

 **Logistic Regression (SMOTEENN)** được chọn vì:

* **Recall lớp 1 cao nhất (0.9951)**: Phát hiện gần hết khách hàng churn, giảm thiểu thiệt hại do bỏ sót (False Negatives).
* **Tính đơn giản và diễn giải**: Cung cấp hệ số đặc trưng rõ ràng, giúp ngân hàng hiểu các yếu tố ảnh hưởng đến churn.
* **Phù hợp với yêu cầu đề bài**: Ưu tiên Recall và khả năng diễn giải.

 **Trade-off**:

* Precision thấp (0.21) dẫn đến nhiều False Positives (1547 khách hàng không churn bị dự đoán sai), có thể làm tăng chi phí giữ chân.
* Tuy nhiên, trong bài toán churn, chi phí bỏ sót khách hàng (FN) thường cao hơn chi phí giữ chân sai (FP), nên Precision thấp là chấp nhận được.

## Kết luận

Kết quả từ việc xây dựng và đánh giá mô hình dự đoán khách hàng rời bỏ dịch vụ ngân hàng B cho thấy **Logistic Regression kết hợp với SMOTEENN** là mô hình hiệu quả nhất, đạt **Recall lớp 1 = 0.9951** tại ngưỡng phân loại 0.2. Điều này đồng nghĩa với việc mô hình phát hiện được gần như toàn bộ khách hàng có nguy cơ rời bỏ (405/407 trường hợp churn), đáp ứng hoàn hảo mục tiêu tối ưu hóa Recall để giảm thiểu thiệt hại do bỏ sót khách hàng (False Negatives). Mô hình này không chỉ hiệu quả mà còn đơn giản, dễ diễn giải, cung cấp thông tin giá trị về các yếu tố ảnh hưởng đến hành vi churn, như:

* **Tuổi tác**: Khách hàng lớn tuổi có nguy cơ churn cao hơn (hệ số 0.6513).
* **Quốc gia Đức**: Khách hàng ở Đức dễ rời bỏ hơn (hệ số 0.3857).
* **Số dư tài khoản**: Số dư cao liên quan đến khả năng churn cao (hệ số 0.2593).
* **Thành viên tích cực**: Khách hàng tích cực ít có khả năng churn (hệ số -0.4141).

Tuy nhiên, mô hình có **Precision lớp 1 thấp (0.21)**, dẫn đến nhiều trường hợp khách hàng không churn bị dự đoán sai thành churn (1547 False Positives). Đây là trade-off chấp nhận được trong bối cảnh bài toán ưu tiên phát hiện tối đa khách hàng churn, vì chi phí bỏ sót khách hàng thường cao hơn chi phí áp dụng chiến lược giữ chân không cần thiết. Accuracy tổng thể thấp (0.23) cũng phản ánh việc mô hình ưu tiên lớp 1 để đạt Recall cao, phù hợp với mục tiêu đề bài.

**Khuyến nghị áp dụng**:

* **Triển khai mô hình**: Sử dụng Logistic Regression (SMOTEENN) với ngưỡng 0.2 để phát hiện sớm khách hàng có nguy cơ churn, đặc biệt là các nhóm rủi ro cao như khách lớn tuổi, ở Đức, hoặc có số dư tài khoản cao. Ngân hàng có thể áp dụng các chiến lược giữ chân như ưu đãi cá nhân hóa, tăng cường tương tác với khách hàng tích cực.
* **Cân nhắc chi phí kinh doanh**: Đánh giá chi phí của False Positives (giữ chân sai, ví dụ: 10 USD/khách hàng) so với False Negatives (mất khách hàng, ví dụ: 1000 USD/khách hàng). Với 1547 FP và 2 FN, chi phí ước tính là 1547×10+2×1000=17,470USD cho thấy mô hình hiện tại phù hợp nếu chi phí khách hàng rời bỏ cao hơn nhiều.