**TRƯỜNG ĐẠI HỌC THƯƠNG MẠI**

**KHOA HỆ THỐNG THÔNG TIN KINH TẾ VÀ THƯƠNG MẠI ĐIỆN TỬ**

**----------- 🙥🕮🙧 -----------**



**BÀI THẢO LUẬN**

**HỌC PHẦN: KHAI PHÁ DỮ LIỆU TRONG KINH DOANH**

**ĐỀ TÀI: ỨNG DỤNG KHAI PHÁ DỮ LIỆU TRONG MÔ**

**HÌNH DỰ ĐOÁN VIỆC RỜI BỎ SỬ DỤNG NGÂN HÀNG CỦA KHÁCH HÀNG**

**Nhóm thực hiện : Nhóm 4**

**Lớp học phần : 241\_INFO2111\_03**

**Giảng viên hướng dẫn : PGS.TS. Nguyễn Thị Thu Thủy**

**Hà Nội , 2024**

Danh sách thành viên

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Họ và tên | Nhiệm vụ | Đánh giá |
| 1 | Dương Thọ Ngọc Minh | Nhóm trưởng, chương 3 |  |
| 2 | Nguyễn Bảo Ngọc | Xây dựng cây quyết định |  |
| 3 | Trần Thị Hồng Nhung | Chương 1, Xây dựng mô hình Xgboost với dữ liệu chưa xử lý mất cân bằng, powerpoint |  |
| 4 | Phạm Gia Bảo Phong | Xây dựng mô hình Xgboost với Oversampling, Xgboost kịch bản nâng cao độ chính xác |  |
| 5 | Lê Thị Y Phụng | Xây dựng Dashboard |  |
| 6 | Nguyễn Thị Phượng | Nhận xét Dashboard, Xây dựng mô hình Xgboost với Undersampling, word |  |
| 7 | Nguyễn Thị Quỳnh | Chương 2, Xây dựng mô hình Xgboost với smote |  |
| 8 | Lê Thị Phương Thanh | Xây dựng mô hình hồi quy Logistic |  |

# **MỤC LỤC**

**[MỤC LỤC](#_Toc179410203)** [3](#_Toc179410203)

**[LỜI CẢM ƠN](#_Toc179410204)** [5](#_Toc179410204)

**[LỜI CAM ĐOAN](#_Toc179410205)** [6](#_Toc179410205)

**[LỜI MỞ ĐẦU](#_Toc179410206)** [7](#_Toc179410206)

**[CHƯƠNG I. TỔNG QUAN ĐỀ TÀI](#_Toc179410207)** [8](#_Toc179410207)

**[1. Giới thiệu](#_Toc179410208)** [8](#_Toc179410208)

***[1.1. Xác định vấn đề](#_Toc179410209)*** [8](#_Toc179410209)

***[1.2. Lý do chọn đề tài](#_Toc179410210)*** [9](#_Toc179410210)

**[2. Ý nghĩa và vai trò của đề tài](#_Toc179410211)** [9](#_Toc179410211)

**[CHƯƠNG II. CƠ SỞ LÝ THUYẾT](#_Toc179410212)** [9](#_Toc179410212)

**[1. Khái niệm khai phá dữ liệu](#_Toc179410213)** [9](#_Toc179410213)

**[2. Các bước trong quá trình khai phá](#_Toc179410214)** [11](#_Toc179410214)

**[3. Phân lớp dữ liệu với cây quyết định](#_Toc179410215)** [11](#_Toc179410215)

***[3.1. Phát biểu bài toán phân lớp dữ liệu](#_Toc179410216)*** [11](#_Toc179410216)

***[3.2. Các loại phân lớp](#_Toc179410217)*** [12](#_Toc179410217)

***[3.3. Các thuật toán phân lớp dữ liệu](#_Toc179410218)*** [12](#_Toc179410218)

***[3.4. Phân lớp dữ liệu với cây quyết định](#_Toc179410219)*** [12](#_Toc179410219)

***[3.5. Xác định các luật phân lớp từ cây quyết định](#_Toc179410220)*** [13](#_Toc179410220)

***[3.6. Thuật toán ID3](#_Toc179410221)*** [13](#_Toc179410221)

**[CHƯƠNG III: PHÂN TÍCH VÀ KHAI PHÁ DỮ LIỆU](#_Toc179410222)** [14](#_Toc179410222)

**[1. Cách thức thu thập dữ liệu](#_Toc179410223)** [14](#_Toc179410223)

**[2. Ý nghĩa các thuộc tính và mô tả dữ liệu](#_Toc179410224)** [14](#_Toc179410224)

***[2.1 Ý nghĩa các thuộc tính](#_Toc179410225)*** [14](#_Toc179410225)

***[2.2 Mô tả dữ liệu](#_Toc179410226)*** [14](#_Toc179410226)

**[3. Tiền xử lý dữ liệu](#_Toc179410227)** [14](#_Toc179410227)

**[CHƯƠNG IV: XÂY DỰNG MÔ HÌNH.](#_Toc179410228)** [26](#_Toc179410228)

**[1. MÔ HÌNH HỒI QUY LOGISTIC](#_Toc179410229)** [27](#_Toc179410229)

***[1.1 Xây dựng mô hình](#_Toc179410230)*** [27](#_Toc179410230)

***[1.2 Xử lý mất cân bằng dữ liệu bằng Undersampling](#_Toc179410231)*** [27](#_Toc179410231)

***[1.3 Xứ lý mất cân bằng dữ liệu bằng thay đổi trọng số](#_Toc179410232)*** [28](#_Toc179410232)

***[1.4 Xử lý mất cân bằng dữ liệu với Smote](#_Toc179410233)*** [29](#_Toc179410233)

**[2. MÔ HÌNH CÂY QUYẾT ĐỊNH](#_Toc179410234)** [30](#_Toc179410234)

***[2.1 Xây dựng mô hình](#_Toc179410235)*** [30](#_Toc179410235)

***[2.2 Kịch bản nâng cao độ chính xác của mô hình](#_Toc179410236)*** [37](#_Toc179410236)

**[3. MÔ HÌNH XGBOOST](#_Toc179410237)** [38](#_Toc179410237)

***[3.1 Xây dựng XGboost với Smote](#_Toc179410238)*** [39](#_Toc179410238)

***[3.2. Xây dựng mô hình XGboost với Undersampling](#_Toc179410239)*** [39](#_Toc179410239)

***[3.3 Xây dựng mô hình XGboost với Oversampling](#_Toc179410240)*** [40](#_Toc179410240)

***[3.4 Kịch bản nâng cao độ chính xác của mô hình](#_Toc179410241)*** [41](#_Toc179410241)

***[3.5 Sử dụng mô hình để dự đoán](#_Toc179410242)*** [42](#_Toc179410242)

**[CHƯƠNG VI. XÂY DỰNG DASHBOARD TRÊN POWER BI](#_Toc179410243)** [43](#_Toc179410243)

***[1.Tạo biểu đồ dashboard](#_Toc179410244)*** [43](#_Toc179410244)

***[2. Báo cáo rút ra từ biểu đồ](#_Toc179410245)*** [43](#_Toc179410245)

**[CHƯƠNG VI. TỔNG KẾT](#_Toc179410246)** [52](#_Toc179410246)

**DANH MỤC BẢNG BIỂU, HÌNH ẢNH**

[Hình 1. Quy trình khai phá tri thức từ dữ liệu 11](#_Toc179367189)

[Hình 2. Biểu đồ thể hiện sự phân bố của các giá trị ngoại lai 18](#_Toc179367190)

[Hình 3. Biểu đồ phân bố điểm tín dụng của khách hàng 19](#_Toc179367191)

[Hình 4.. Biểu đồ phân bố độ tuổi của khách hàng 20](#_Toc179367192)

[Hình 5. Biểu đồ phân bố thời gian giữ tài khoản của khách hàng 20](#_Toc179367193)

[Hình 6. Biểu đồ phân bố số dư tài khoản của khách hàng 21](#_Toc179367194)

[Hình 7. Biểu đồ phân bố số sản phẩm dịch vụ khách hàng sử dụng 21](#_Toc179367195)

[Hình 8. Biểu đồ phân bố thẻ tín dụng 22](#_Toc179367196)

[Hình 9. Biểu đồ phân bố hoạt động giao dịch của khách hàng 22](#_Toc179367197)

[Hình 10. Biểu đồ phân bố mức lương ước tính của khách hàng 23](#_Toc179367198)

[Hình 11. Biểu đồ phân bố việc rời đi hay tiếp tục sử dụng của khách hàng 23](#_Toc179367199)

[Hình 12. Biểu đồ phân bố điểm hài lòng của khách hàng 24](#_Toc179367200)

[Hình 13. Biểu đồ heatmap minh họa mối tương quan giữa các thuộc tính 26](#_Toc179367201)

[Hình 14. Biểu diễn cây quyết định 36](#_Toc179367202)

[Hình 15. Biểu diễn cây quyết định với Graphviz 36](#_Toc179367203)

[Hình 16. Dashboard tổng quan dữ liệu dùng cho mô hình dự đoán việc rời bỏ sử dụng ngân hàng của khách hàng 44](#_Toc179367204)

[Hình 17. Tỷ lệ khách hàng rời bỏ 44](#_Toc179367205)

[Hình 18. Tỷ lệ khách hàng có thẻ tín dụng và không có thẻ tín dụng 45](#_Toc179367206)

[Hình 19. Tỷ lệ số dịch vụ của ngân hàng mà mỗi khách hàng sử dụng 46](#_Toc179367207)

[Hình 20. Tỷ lệ khách hàng rời bỏ theo điểm hài lòng khi sử dụng ngân hàng 47](#_Toc179367208)

[Hình 21. Tỷ lệ khách hàng rời bỏ theo số năm khách hàng đã giao dịch với ngân hàng 48](#_Toc179367209)

[Hình 22. Tỷ lệ khách hàng rời bỏ theo độ tuổi 49](#_Toc179367210)

[Hình 23. Tỷ lệ khách hàng rời bỏ theo loại thẻ 50](#_Toc179367211)

[Hình 24. Tỷ lệ khách hàng rời bỏ theo Giới tính 51](#_Toc179367212)

[Hình 25. Tỷ lệ khách hàng rời bỏ theo quốc gia 52](#_Toc179367213)

**LỜI CẢM ƠN**

Lời đầu tiên, nhóm 4 chúng em xin gửi lời cảm ơn đến Trường Đại Học Thương Mại và khoa Hệ thống thông tin kinh tế & Thương mại điện tử đã tạo điều kiện cho chúng em và các bạn sinh viên có cơ hội được học tập trong một môi trường năng động, sáng tạo, tạo điều kiện nỗ lực và phát huy hết tiềm năng của mình. Để hoàn thành báo cáo này, nhóm 4 xin gửi lời cảm ơn chân thành đến: Giảng viên bộ môn Khai phá dữ liệu kinh doanh đã giảng dạy tận tình, chi tiết để nhóm có đủ kiến thức và vận dụng chúng vào bài báo cáo này. Do chưa có nhiều kinh nghiệm làm đề tài cũng như hạn chế về kiến thức, trong bài báo cáo chắc chắn sẽ không tránh khỏi những thiếu sót. Rất mong nhận được sự nhận xét, ý kiến đóng góp từ phía cô để bài báo cáo được hoàn thiện hơn.

Lời cuối cùng, xin kính chúc cô nhiều sức khỏe, thành công và hạnh phúc!

**LỜI CAM ĐOAN**

Chúng em xin cam đoan rằng bài tiểu luận này được thực hiện với sự tận tâm và

sự chân thành cao nhất từ tất cả các thành viên của nhóm chúng em. Tất cả những suy nghĩ, ý kiến, và thông tin được trình bày trong bài tiểu luận đều phản ánh sự đóng góp tích cực của chúng em trong quá trình nghiên cứu và thảo luận.

Chúng em cam đoan rằng mọi nguồn thông tin và tài liệu tham khảo được sử dụng đều được chọn lọc một cách cẩn thận, đồng thời đã được trích dẫn một cách đầy đủ và chính xác. Mọi chi tiết trong bài viết đều tuân thủ các quy định và nguyên tắc nghiêm túc về đạo đức nghiên cứu.

Chúng em hoàn toàn chịu trách nhiệm với nội dung của bài tiểu luận này và sẵn sàng chấp nhận mọi hình phạt kỷ luật nếu có bất kỳ vấn đề nào phát sinh.

# 

# **LỜI MỞ ĐẦU**

Trong môi trường kinh doanh hiện đại, việc giữ chân khách hàng trở thành một yếu tố sống còn đối với các tổ chức tài chính, đặc biệt là các ngân hàng. Khi thị trường ngày càng cạnh tranh và dịch vụ ngày càng đa dạng, việc khách hàng chuyển sang sử dụng dịch vụ của ngân hàng khác, hay còn gọi là "churn," đặt ra thách thức lớn. Để giảm thiểu hiện tượng này, các ngân hàng cần dự đoán sớm khách hàng nào có khả năng rời bỏ nhằm áp dụng các biện pháp can thiệp kịp thời.

Khai phá dữ liệu (data mining) đã nổi lên như một công cụ mạnh mẽ giúp ngân hàng phân tích khối lượng dữ liệu khổng lồ từ hành vi của khách hàng. Bằng cách áp dụng các mô hình và thuật toán tiên tiến, khai phá dữ liệu không chỉ giúp nhận diện xu hướng mà còn đưa ra các dự báo chính xác về nhóm khách hàng có nguy cơ rời bỏ. Điều này không chỉ tăng cường khả năng dự đoán mà còn giúp ngân hàng đưa ra chiến lược phù hợp, cải thiện dịch vụ và tăng cường sự gắn kết của khách hàng.

Nghiên cứu này sẽ tập trung vào việc khám phá các phương pháp khai phá dữ liệu và ứng dụng của chúng trong dự đoán việc khách hàng rời bỏ sử dụng dịch vụ ngân hàng, từ đó giúp các ngân hàng nâng cao hiệu quả hoạt động và gia tăng sự trung thành của khách hàng.

# 

# **CHƯƠNG I. TỔNG QUAN ĐỀ TÀI**

**1. Giới thiệu**

***1.1. Xác định vấn đề***

Vấn đề chính mà đề tài "Ứng dụng khai phá dữ liệu trong dự đoán việc khách hàng rời bỏ sử dụng dịch vụ ngân hàng" muốn giải quyết là việc nhận diện và dự đoán sớm khách hàng có nguy cơ rời bỏ, điều này vô cùng quan trọng trong bối cảnh thị trường tài chính đầy cạnh tranh. Khách hàng không phải lúc nào cũng bộc lộ rõ ràng ý định ngừng sử dụng dịch vụ, vì vậy các ngân hàng cần một công cụ mạnh mẽ để phân tích và phát hiện những dấu hiệu tiềm ẩn từ hành vi, lịch sử giao dịch và mức độ hài lòng của họ. Mặt khác, chi phí để thu hút khách hàng mới cao hơn nhiều so với việc giữ chân khách hàng hiện tại, vì vậy nếu không có biện pháp dự đoán hiệu quả, ngân hàng sẽ gặp khó khăn về tài chính và danh tiếng. Hiện nay, các phương pháp truyền thống không đủ hiệu quả, do đó việc ứng dụng mô hình học máy và khai phá dữ liệu có thể mang lại một phương pháp phân tích chính xác hơn, giúp ngân hàng chủ động trong việc giữ chân khách hàng. Điều này không chỉ giúp tối ưu hóa chiến lược chăm sóc khách hàng, mà còn giảm thiểu thiệt hại về doanh thu và cải thiện đáng kể chất lượng dịch vụ.

***1.2. Lý do chọn đề tài***

Trong ngành ngân hàng, việc duy trì khách hàng lâu dài là một yếu tố sống còn để đảm bảo sự phát triển và ổn định. Tuy nhiên, tình trạng khách hàng rời bỏ dịch vụ ngân hàng (churn) đang trở thành một thách thức lớn đối với các tổ chức tài chính. Nguyên nhân có thể đến từ sự cạnh tranh gay gắt giữa các ngân hàng, sự phát triển của công nghệ tài chính (fintech), hay những trải nghiệm không tốt của khách hàng. Để giải quyết vấn đề này, các ngân hàng cần hiểu rõ những yếu tố dẫn đến việc khách hàng rời bỏ và từ đó đưa ra các giải pháp ngăn chặn kịp thời.

Ứng dụng khai phá dữ liệu (data mining) trong việc dự đoán tỷ lệ khách hàng rời bỏ đang trở nên cấp thiết và có giá trị thực tiễn cao. Việc phân tích dữ liệu khách hàng thông qua các yếu tố như điểm tín dụng, số lượng sản phẩm khách hàng sử dụng, mức độ hoạt động, sự hài lòng, hay thu nhập ước tính sẽ giúp các ngân hàng nhận diện nhóm khách hàng có nguy cơ rời bỏ cao. Dựa trên kết quả dự đoán này, ngân hàng có thể đưa ra các chiến lược giữ chân khách hàng hiệu quả hơn, chẳng hạn như cung cấp các dịch vụ cá nhân hóa, cải thiện chất lượng dịch vụ, hay tăng cường các chương trình khuyến mãi phù hợp. Việc nghiên cứu và khai phá dữ liệu trong lĩnh vực này không chỉ giúp giảm thiểu tổn thất từ việc mất khách hàng mà còn tạo cơ hội gia tăng lợi nhuận và củng cố vị thế của ngân hàng trên thị trường.

Chính vì vậy, nhóm 4 quyết định “ Ứng dụng khai phá dữ liệu trong dự đoán việc khách hàng rời bỏ sử dụng dịch vụ ngân hàng” làm đề tài thảo luận trong bộ môn Khai phá dữ liệu. Việc nghiên cứu này là cần thiết, góp phần nâng cao chất lượng dịch vụ và sự phát triển bền vững của ngành ngân hàng.

**2. Ý nghĩa và vai trò của đề tài**

Đề tài "Ứng dụng khai phá dữ liệu trong dự đoán việc khách hàng rời bỏ sử dụng dịch vụ ngân hàng" mang ý nghĩa quan trọng đối với cả ngân hàng và khách hàng. Với ngân hàng, việc dự đoán chính xác khả năng khách hàng rời bỏ giúp xây dựng các chiến lược giữ chân hiệu quả, giảm thiểu rủi ro mất khách và nâng cao hiệu quả hoạt động. Đối với khách hàng, việc cá nhân hóa dịch vụ giúp cải thiện trải nghiệm và nhận được ưu đãi phù hợp. Đề tài không chỉ là công cụ dự đoán chiến lược giúp ngân hàng phân bổ nguồn lực hiệu quả trong việc giữ chân khách hàng mà còn hỗ trợ các nhà quản lý ra quyết định dựa trên dữ liệu thực tế, từ đó xây dựng chính sách và chiến lược dài hạn nhằm cải thiện dịch vụ và tối ưu hóa lợi nhuận. Việc áp dụng thành công các phương pháp khai phá dữ liệu giúp ngân hàng tăng cường khả năng cạnh tranh trên thị trường bằng cách cung cấp dịch vụ và trải nghiệm khách hàng tốt hơn. Đề tài mang lại giá trị to lớn cho cả ngân hàng, khách hàng và lĩnh vực công nghệ tài chính, góp phần nâng cao hiệu quả hoạt động và tối ưu hóa trải nghiệm khách hàng thông qua các giải pháp dựa trên dữ liệu. 

# **CHƯƠNG II. CƠ SỞ LÝ THUYẾT**

**1. Khái niệm khai phá dữ liệu**

Khai phá dữ liệu (data mining) là quá trình tính toán để tìm ra các mẫu trong các

bộ dữ liệu lớn liên quan đến các phương pháp tại giao điểm của máy học, thống kê và các hệ thống cơ sở dữ liệu. Đây là một lĩnh vực liên ngành của khoa học máy tính.

Mục tiêu tổng thể của quá trình khai thác dữ liệu là trích xuất thông tin từ một bộ dữ

liệu và chuyển nó thành một cấu trúc dễ hiểu để sử dụng tiếp. Ngoài bước phân tích

thô, nó còn liên quan tới cơ sở dữ liệu và các khía cạnh quản lý dữ liệu, xử lý dữ liệu

trước, suy xét mô hình và suy luận thống kê, các thước đo thú vị, các cân nhắc phức

tạp, xuất kết quả về các cấu trúc được phát hiện, hiện hình hóa và cập nhật trực tuyến.

Khai thác dữ liệu là bước phân tích của quá trình "khám phá kiến thức trong cơ sở dữ

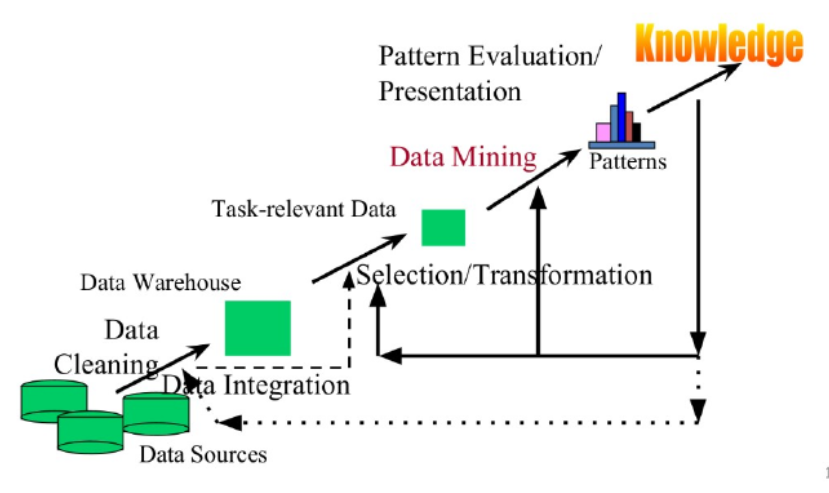
liệu" hoặc KDD.

Khai phá cơ sở dữ liệu (database mining) là một nhánh của khai phá dữ liệu, tập

trung vào việc tìm kiếm các mẫu, xu hướng và mối quan hệ trong các tập dữ liệu lớn

được lưu trữ trong cơ sở dữ liệu. Các tập dữ liệu này có thể bao gồm dữ liệu khách

hàng, dữ liệu bán hàng, dữ liệu tài chính, dữ liệu khoa học,…



Hình 1. Quy trình khai phá tri thức từ dữ liệu

Mục tiêu của khai phá cơ sở dữ liệu là thu thập thông tin hữu ích từ các tập dữ

liệu này mà không cần phải phân tích thủ công. Thông tin này có thể được sử dụng để cải thiện hiệu quả hoạt động, đưa ra quyết định tốt hơn hoặc phát triển các sản phẩm và dịch vụ mới.

Có nhiều kỹ thuật khai phá cơ sở dữ liệu khác nhau, mỗi kỹ thuật được sử dụng

để tìm kiếm một loại mẫu cụ thể. Một số kỹ thuật khai phá cơ sở dữ liệu phổ biến bao

gồm:

- Phân loại dữ liệu: Phân loại các mẫu dữ liệu thành các nhóm hoặc lớp. Ví dụ:

một mô hình phân loại có thể được sử dụng để phân loại khách hàng thành các nhóm

"khách hàng tiềm năng" và "khách hàng hiện tại."

- Dự đoán: Dự đoán các giá trị trong tương lai dựa trên các giá trị trong quá khứ. Ví dụ: một mô hình dự đoán có thể được sử dụng để dự đoán doanh số bán hàng

trong tương lai.

- Phân cụm: Xác định mối quan hệ hợp lý trong các sản phẩm và nhóm chúng lại với nhau. Ví dụ như trong một cửa hàng bán lẻ, kem đánh răng và bàn chải đánh răng có thể được nhóm lại.

- Khai phá luật kết hợp: Tìm kiếm các quy tắc xác định mối quan hệ giữa các

thuộc tính trong tập dữ liệu. Ví dụ: một quy tắc kết hợp có thể là "Nếu khách

hàng mua một sản phẩm A, thì họ có nhiều khả năng mua sản phẩm B."

Khai phá cơ sở dữ liệu là một công cụ mạnh mẽ có thể được sử dụng trong nhiều lĩnh vực khác nhau. Nó đang được sử dụng bởi các doanh nghiệp, tổ chức chính phủ và các nhà nghiên cứu để cải thiện hiệu quả, đưa ra quyết định tốt hơn và phát triển các sản phẩm và dịch vụ mới.

**2. Các bước trong quá trình khai phá**

*Bước 1. Thu thập dữ liệu*

Đây là bước đầu tiên của quá trình khai phá dữ liệu. Mục tiêu của bước này là

thu thập dữ liệu cần thiết cho quá trình khai phá. Dữ liệu có thể được thu thập từ nhiều

nguồn khác nhau, bao gồm:

* Thu thập dữ liệu sơ cấp: Thu thập dữ liệu trực tiếp từ nguồn gốc, thường được

thực hiện thông qua các phương pháp như khảo sát, phỏng vấn, quan sát, thực nghiệm,...

* Thu thập dữ liệu thứ cấp: Thu thập dữ liệu từ các nguồn có sẵn, thường được thực hiện thông qua các phương pháp như thu thập dữ liệu từ báo chí, sách, tạp chí, internet,...

*Bước 2. Tiền xử lý dữ liệu*

Bước này bao gồm các hoạt động chuẩn hóa, làm sạch và chuẩn bị dữ liệu để sẵn sàng cho quá trình khai phá. Các hoạt động tiền xử lý dữ liệu bao gồm:

* Chuẩn hóa dữ liệu: Thống nhất các giá trị của dữ liệu để phù hợp với các yêu

cầu của các kỹ thuật khai phá dữ liệu.

* Làm sạch dữ liệu: Xóa bỏ các dữ liệu sai lệch, thiếu sót hoặc không liên quan.
* Chuẩn bị dữ liệu: Chia dữ liệu thành các tập huấn luyện và các tập kiểm tra

*Bước 3. Khai phá dữ liệu*

Đây là bước quan trọng nhất của quá trình khai phá dữ liệu. Mục tiêu của bước này là tìm ra các mẫu hình và xu hướng có ý nghĩa trong dữ liệu. Các kỹ thuật khai phá dữ liệu có thể được sử dụng để tìm ra các mẫu hình và xu hướng này bao gồm:

* Phân tích phân loại: Phân loại các đối tượng vào một hoặc một số lớp cho trước
* Phân tích hồi quy: Dự đoán giá trị của một biến dựa trên các biến khác
* Phân tích hiệp phương sai: Tìm mối quan hệ giữa các biến
* Phân tích cụm: Gộp các đối tượng có đặc điểm tương đồng vào cùng một nhóm

*Bước 4. Phân tích kết quả*

Bước này bao gồm các hoạt động đánh giá, giải thích và trình bày kết quả của quá trình khai phá dữ liệu. Các kết quả này cần được đánh giá để xác định mức độ chính xác và giá trị của chúng. Sau đó, các kết quả này cần được giải thích để hiểu rõ ý nghĩa của chúng. Cuối cùng, các kết quả này cần được trình bày một cách rõ ràng và súc tích để có thể dễ dàng hiểu và sử dụng

Ngoài các bước chính trên, quá trình khai phá dữ liệu còn có thể bao gồm các bước bổ sung như:

* Quản lý dữ liệu: Quản lý dữ liệu một cách hiệu quả để đảm bảo dữ liệu được lưu trữ và sử dụng một cách an toàn và hiệu quả
* An toàn dữ liệu: Bảo vệ dữ liệu khỏi bị truy cập, sử dụng hoặc sửa đổi trái phép
* Moral và đạo đức: Tuân thủ các nguyên tắc đạo đức và pháp lý trong quá trình khai phá dữ liệu

## **3. Phân lớp dữ liệu với cây quyết định**

### ***3.1. Phát biểu bài toán phân lớp dữ liệu***

Phân lớp dữ liệu là quá trình học có giám sát trên một tập dữ liệu đầu vào nhằm xây dựng một mô hình có thể dự đoán xu hướng cho các dữ liệu mới.

*Đầu vào:* Tập các dữ liệu có dạng (x,y) = (x1, x2, …, xn, y)

* x là biến độc lập (Independent variable) mô tả thuộc tính của một đối tượng
* y là biến phụ thuộc (Dependent variable) cần tìm hiểu, phân loại. y còn được gọi là thuộc tính nhãn

*Đầu ra*: Một mô hình có khả năng phân loại đúng đắn cho tập các dữ liệu đầu vào

*Bước học (Bước huấn luyện):* Xây dựng mô hình

* Xác định tập dữ liệu huấn luyện gồm các mẫu đã được gán nhãn y
* Chạy một thuật toán phân lớp trên tập dữ liệu huấn luyện
* Mô hình được biểu diễn dưới dạng các luật phân lớp, các cây quyết định hoặc các công thức toán

*Bước phân loại:* Sử dụng mô hình để gán nhãn thích hợp cho các dữ liệu chưa được gán nhãn.

* Ước lượng độ chính xác của mô hình
* Xác định tập dữ liệu kiểm thử gồm các mẫu đã được gán nhãn y (dữ liệu kiểm thử và dữ liệu huấn luyện phải khác nhau để tránh tình trạng quá khớp over-fitting)
* Chạy mô hình với tập dữ liệu kiểm thử thu được nhãn y’
* So sánh y và y’ để xác định độ chính xác của mô hình

Nếu mô hình chính xác, sử dụng nó để dự đoán nhãn cho các dữ liệu cần gán nhãn

### ***3.2. Các loại phân lớp***

* Phân lớp nhị/đa nhãn
* |C| = 2: phân lớp nhị phân, giá trị của thuộc tính nhãn bao gồm: Có mua hay không mua máy tính?
* |C| > 2: phân lớp đa lớp, giá trị của thuộc tính nhãn bao gồm: Thể thao, chính trị, văn hóa, sức khỏe
* Phân lớp đơn nhãn/đa nhãn
* Đơn nhãn: mỗi mẫu được gán duy nhất vào một lớp, tất cả các mẫu dữ liệu chỉ có một và chỉ một giá trị nhãn
* Đa nhãn: một mẫu có thể được gán nhiều hơn một lớp, một bài báo có thể vừa được gán nhãn Thể thao vừa được gán nhãn Sức khỏe

### ***3.3. Các thuật toán phân lớp dữ liệu***

Nhóm thuật toán dựa trên xác suất và lý thuyết thông tin

* Phân lớp Naïve Bayes.
* Phân lớp cực đại Entropy
* Nhóm thuật toán dựa trên cây
* Phân lớp dựa trên cây quyết định.
* Phân lớp rừng ngẫu nhiên (Random Forest)
* Nhóm thuật toán dựa trên hồi quy
* Hồi quy tuyến tính.
* Hồi quy logistic
* Nhóm thuật toán dựa trên khoảng cách
* Phân lớp K láng giềng gần nhất.
* Phân lớp sử dụng máy vecto hỗ trợ SVM
* Phân lớp Rocchio
* Nhóm thuật toán khác
* Mạng Neural…

### ***3.4. Phân lớp dữ liệu với cây quyết định***

Cây quyết định mô tả một cấu trúc cây, cho phép người dùng dự đoán nhãn của một đối tượng mới dựa trên tập thuộc tính của nó. Trong đó:

* Các lá đại diện cho các nhãn
* Các cành đại diện cho các kết hợp của các thuộc tính dẫn tới phân lớp đó.

*Ưu điểm:*

* Dễ hiểu, người dùng có thể nhanh chóng hiểu được các luật của cây quyết định.
* Có thể xử lý cả dữ liệu có giá trị bằng số và dữ liệu có giá trị là tên thể loại.
* Làm tăng độ tin tưởng của người dùng vào mô hình.
* Có thể xử lý tốt một lượng dữ liệu lớn trong thời gian ngắn…

### ***3.5. Xác định các luật phân lớp từ cây quyết định***

Ta có thể dễ dàng biểu diễn tri thức thu được từ cây quyết định dưới dạng các luật IF-THEN

* Mỗi đường đi từ nút gốc đến nút lá tương ứng với một luật
* Giá trị của các thuộc tính trên một đường đi sẽ tạo thành các điều kiện trong IF
* Nút lá xác định lớp dự đoán trong THEN

Thuật toán ID3: được công bố bởi Quinlan (trường đại học Syney, Australia) vào cuối thập niên 70 của thế kỷ 20.

* ID3 được xem như là một cải tiến của CLS với khả năng lựa chọn thuộc tính tốt nhất để tiếp tục triển khai cây tại mỗi bước.
* ID3 xây dựng cây quyết định theo hướng tiếp cận từ trên - xuống (top-down)

### ***3.6. Thuật toán ID3***

ID3 là thuật toán cơ bản nhất trong lĩnh vực học cây quyết định, hầu hết các thuật toán học cây quyết định cải tiến sau này đều dựa trên nó.

Nhiệm vụ của ID3 là học cây quyết định từ một tập các mẫu huấn luyện gồm:

* Đầu vào: Một tập hợp các mẫu, mỗi mẫu bao gồm các thuộc tính mô tả mộtđối tượng xác định và một thuộc tính nhãn phân lớp giá trị của nó.

Chú ý: Nếu các thuộc tính có giá trị liên tục, ví dụ Tuổi, ta phải tiến hành biến đổi chúng thành các giá trị rời rạc, ví dụ chia tuổi thành 3 loại: Trẻ, Trung niên, Già.

* Đầu ra: Cây quyết định có khả năng phân loại đúng đắn các mẫu trong tập dữ liệu huấn luyện và hy vọng phân loại đúng cho cả các mẫu chưa gặp trong tương lai.
* Đầu vào: Tập mẫu huấn luyện D, thuộc tính phân lớp C, thuộc tính mô tả A.- Đầu ra: Cây quyết định

\* Thuật toán:

Bước 1: Tạo Nút\_gốc cho cây quyết định.

Bước 2: IF tất cả các mẫu huấn luyện đều có giá trị của nhãn C là P, RETURN

cây có một nút duy nhất là Nút\_gốc với nhãn P.

Bước 3: IF A rỗng, RETURN cây có một nút duy nhất là Nút\_gốc với nhãn là

giá trị phổ biến nhất của C trong D.

Bước 4: 4.1: Gọi X là thuộc tính của A phân lớp D tốt nhất.

       4.2:Gán nhãn cho nút gốc với tên thuộc tính C

        4.3: FOREACH giá trị v của X

              4.3.1: Thêm một nhánh mới dưới Nút\_gốc với X=v

              4.3.2: Xác định tập com Dv ứng với X=v

              4.3.3: If Dv rỗng

Thêm một nhánh mới này một nút lá có nhãn là gía trị phổ biến nhất của

thuộc tính quyết định trong D

           ELSE

                    Thêm cây con dưới nhánh này bằng các gọi đệ quy ID3.

Bước 5: RETURN Nút\_gốc

# **CHƯƠNG III: PHÂN TÍCH VÀ KHAI PHÁ DỮ LIỆU**

## **1. Cách thức thu thập dữ liệu**

Bài toán sử dụng nguồn dữ liệu từ Kaggle, một cộng đồng trực tuyến nơi có sẵn nhiều tập dữ liệu đa dạng có thể được sử dụng nghiên cứu và phân tích.

Tập dữ liệu chính: Bank Customer Churn

Nguồn dữ liệu: <https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn>

## **2. Ý nghĩa các thuộc tính và mô tả dữ liệu**

***2.1 Ý nghĩa các thuộc tính***

Tập dữ liệu bao gồm các cột tương ứng với các thuộc tính liên quan đến khách hàng. Ý nghĩa các thuộc tính:

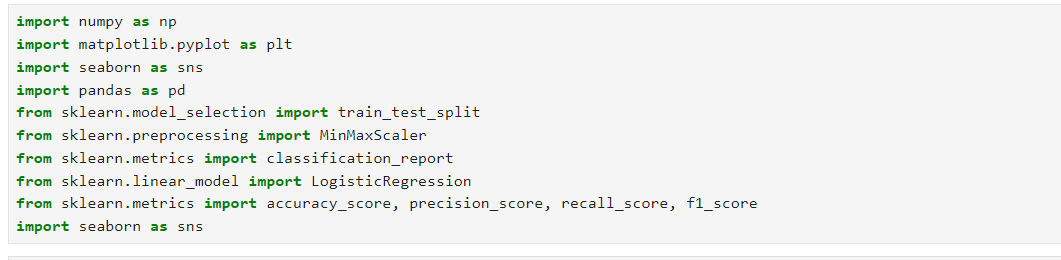
* RowNumber: số hàng
* CustomerID: ID của khách hàng
* Surname: họ của khách hàng
* CreditScore: Điểm tín dụng của khách hàng.
* Geography: Quốc gia của khách hàng (France, Spain, Germany).
* Gender: Giới tính (nam, nữ)
* Age: Tuổi.
* Tenure: Thời gian giữ tài khoản của khách hàng.
* Balance: Số dư tài khoản.
* NumOfProducts: Số sản phẩm của ngân hàng mà khách hàng sử dụng.
* HasCrCard: Khách hàng có thẻ tín dụng không (1 = có, 0 = không).
* IsActiveMember: Khách hàng có hoạt động giao dịch thường xuyên không (1 = có,0 = không).
* EstimatedSalary: Mức lương ước tính của khách hàng.
* Exited: Biến mục tiêu chỉ ra liệu khách hàng đã rời khỏi ngân hàng (1 = đã rời, 0 = ở lại).
* Complain: Khách hàng có phàn nàn về ngân hàng không (1 = có, 0 = không).
* Satisfaction Score: Điểm đánh giá hài lòng của khách hàng
* Card Type: Loại thẻ khách hàng sử dụng (Diamond, Gold, Silver, Platinum).

***2.2 Mô tả dữ liệu***

Tập dữ liệu bao gồm dữ liệu có cấu trúc với dữ liệu ban đầu gồm 10000 dòng và 17 cột. Mỗi hàng đại diện cho một khách hàng riêng lẻ, trong khi đó các cột chứa các thuộc tính và đặc điểm khác nhau liên quan đến khách hàng trong 6 tháng trước đó và biến mục tiêu Exited chứa thông tin rằng sau 6 tháng sau khách hàng còn sử dụng ngân hàng tiếp không hay đã rời đi.

**3. Tiền xử lý dữ liệu**

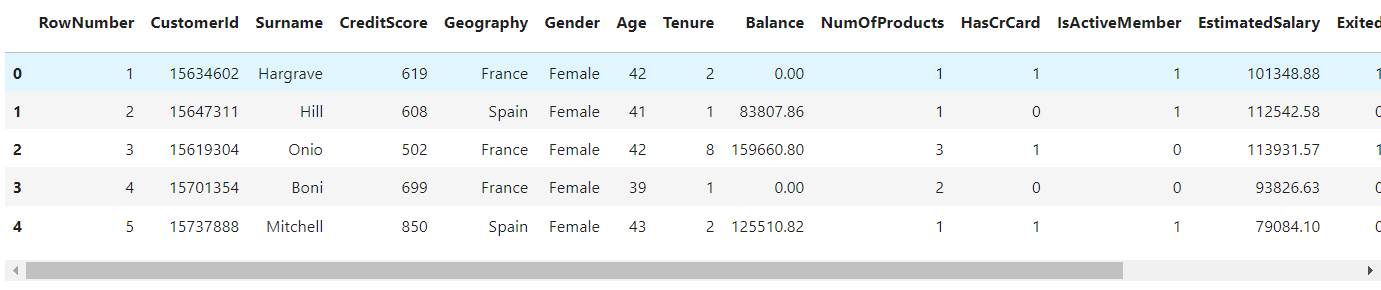
- Khai báo các thư viện cần thiết:



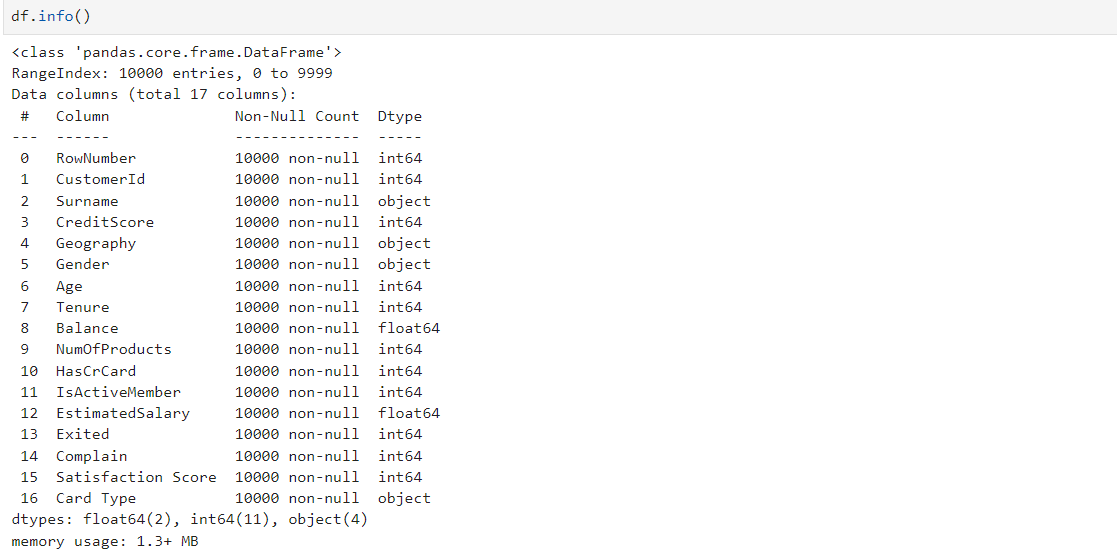
- Đọc dữ liệu và hiển thị dữ liệu:







- Hiển thị thông tin bộ dữ liệu:



- Loại bỏ ba cột ra khỏi tập dữ liệu:

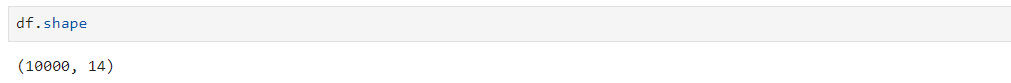
Cột (RowNumber: số hàng, CustomerId: Id của khách hàng, Surname: họ) không ảnh hưởng đến quyết định rời bỏ hay tiếp tục sử dụng dịch vụ ngân hàng của khách hàng, nên loại bỏ đi để giảm dung lượng , làm sạch tập dữ liệu, giúp cho các mô hình học máy không bị ảnh hưởng bởi các yếu tố không cần thiết và cải thiện độ chính xác của mô hình.

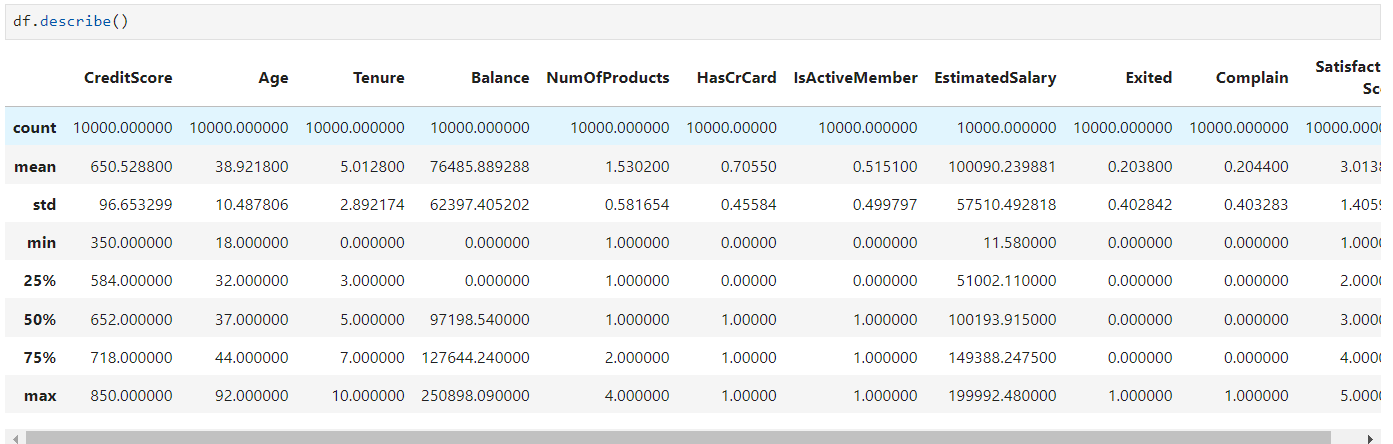


- Hiển thị lại thông tin tập dữ liệu sau khi xóa cột:

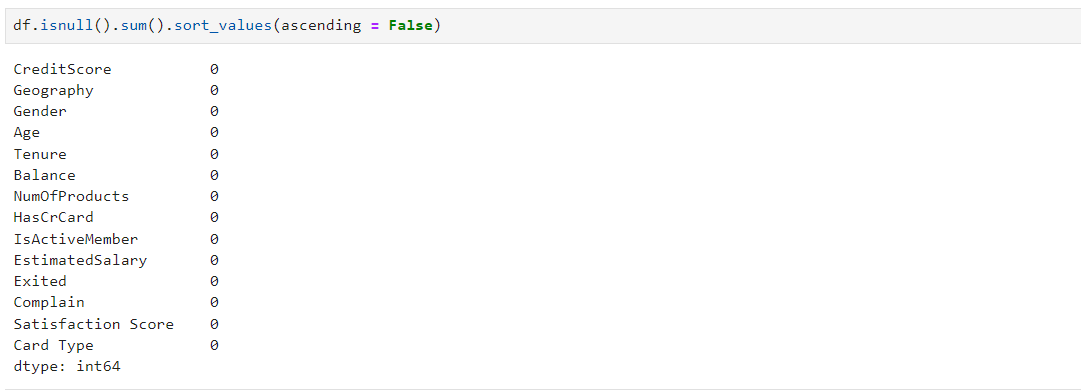


- In ra kích thước và tổng quan bộ dữ liệu:



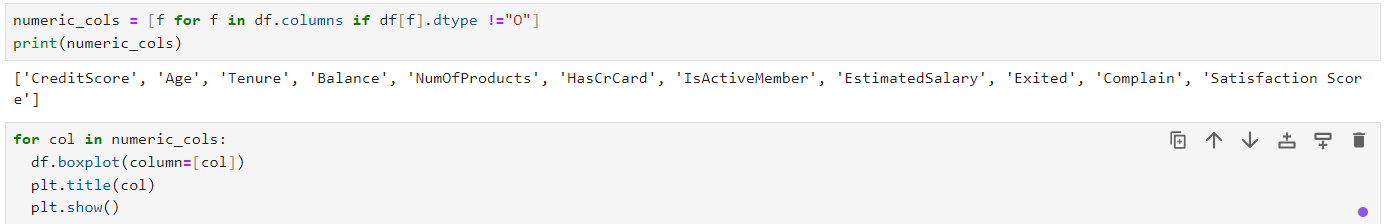


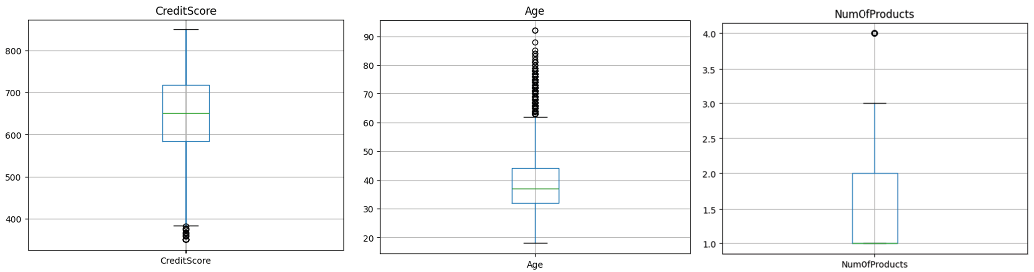
- Kiểm tra giá trị khuyết thiếu:



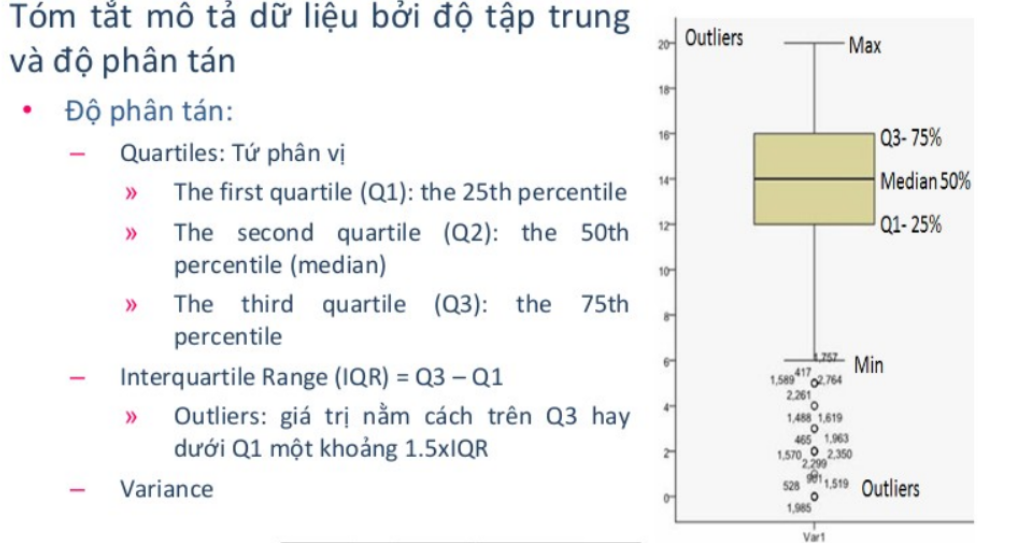
Nhận thấy bộ dữ liệu không bị thiếu giá trị ở bất kì một ô nào trong các cột. Điều này là tốt cho việc phân tích và xây dựng mô hình.

- In ra danh sách các cột dữ liệu dạng số và tạo biểu đồ hộp cho từng cột để kiểm tra giá trị ngoại lai:



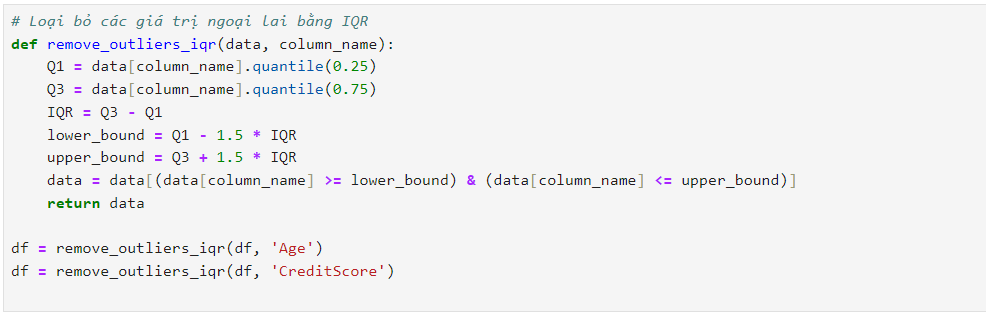


Hình 2. Biểu đồ thể hiện sự phân bố của các giá trị ngoại lai

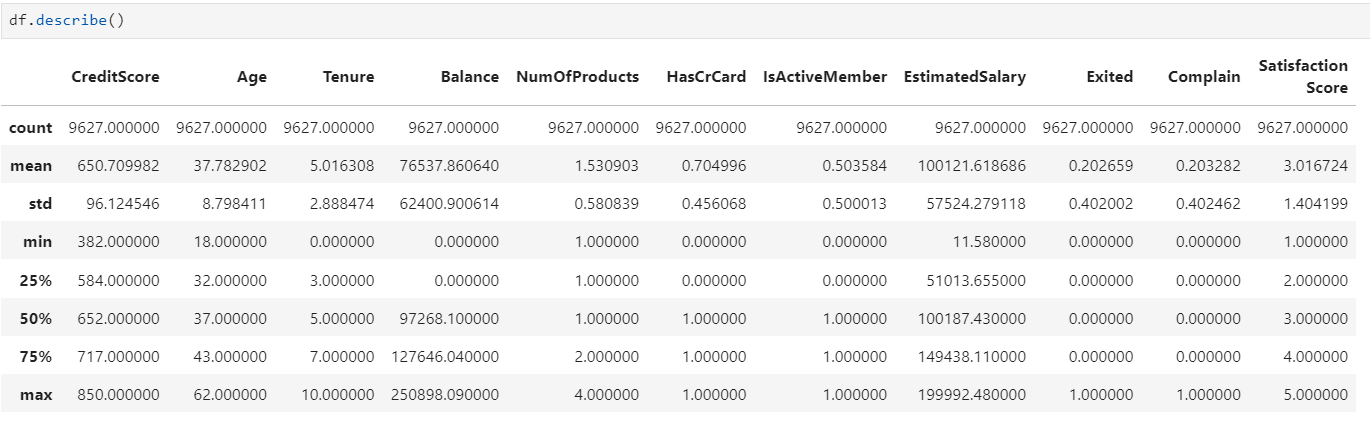


Từ biều đồ hộp của các thuộc tính, có thể thấy trong cột CreditScore, Age có chứa một số lượng giá trị ngoại lai nhất định. Vì vậy cần loại bỏ chúng để tiến hành xây dựng model với độ chính xác là cao nhất.

- Loại bỏ các giá trị ngoại lai:

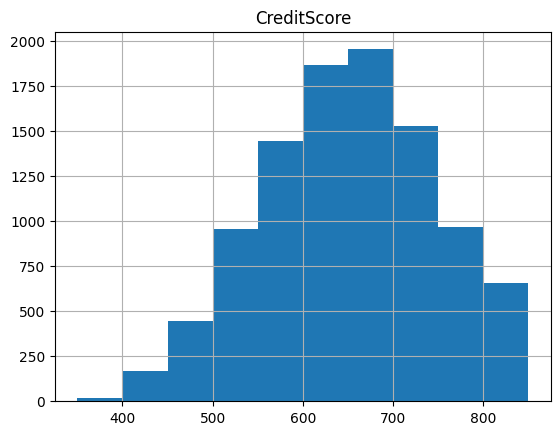


- Dữ liệu sau khi đã loại bỏ các ô giá trị ngoại lai:



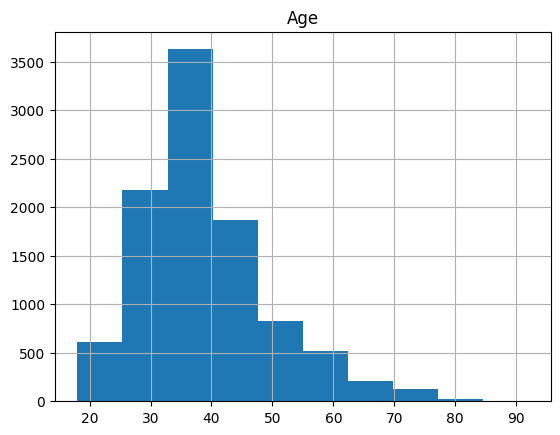
- Tạo biểu đồ histogram trực quan hóa dữ liệu:





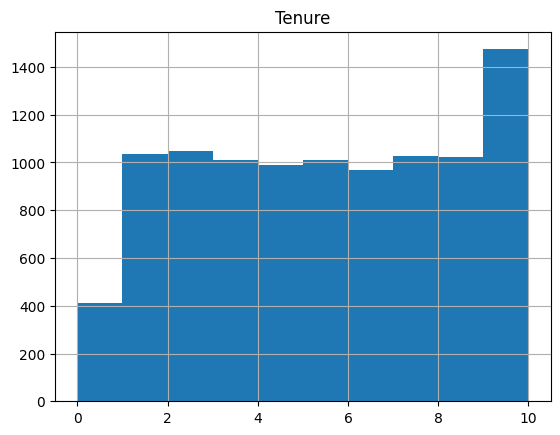
Hình 3. Biểu đồ phân bố điểm tín dụng của khách hàng

Biểu đồ cho thấy rằng phần lớn các điểm tín dụng tập trung trong khoảng từ 600 đến 700, với số lượng giảm dần ở các khoảng giá trị thấp hơn và cao hơn.



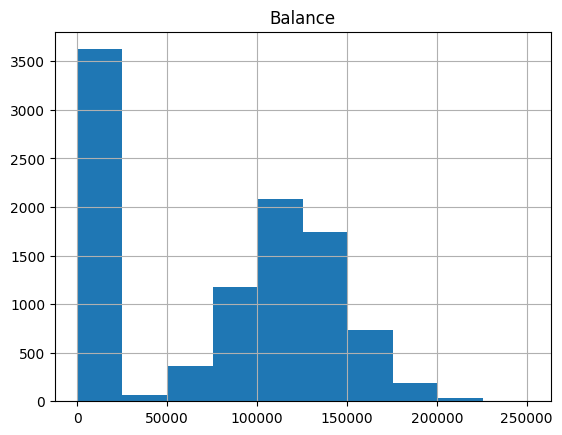
Hình 4.. Biểu đồ phân bố độ tuổi của khách hàng

Biểu đồ cho thấy độ tuổi sử dụng ngân hàng của khách hàng trải dài từ 18 đến hơn 80 tuổi, phần lớn người trong bộ dữ liệu nằm trong khoảng từ 30 đến 40 tuổi,



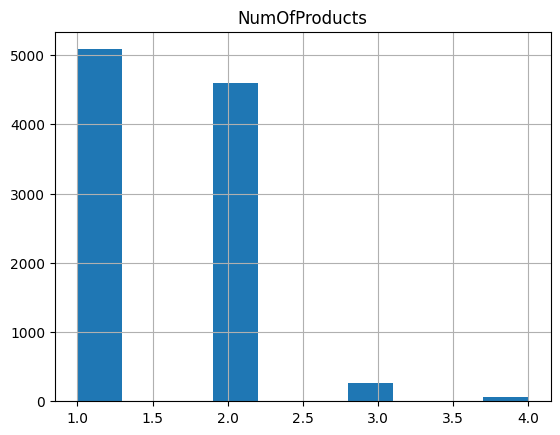
Hình 5. Biểu đồ phân bố thời gian giữ tài khoản của khách hàng

Có thể thấy sự tập trung lớn ở khoảng 8-10 năm cho thấy rằng ngân hàng có một nhóm khách hàng trung thành đáng kể.



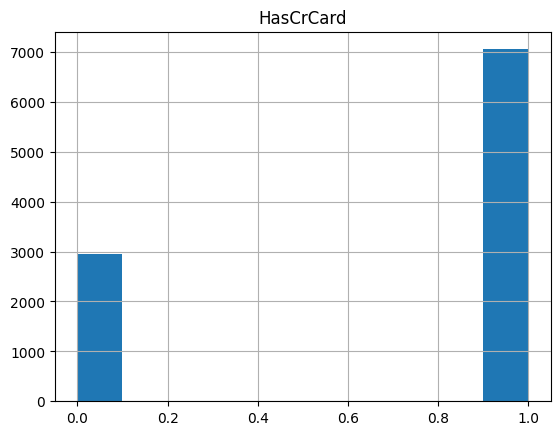
Hình 6. Biểu đồ phân bố số dư tài khoản của khách hàng

Biểu đồ cho thấy số dư không đồng đều, đa số khách hàng có số dư ở mức thấp và trung bình.



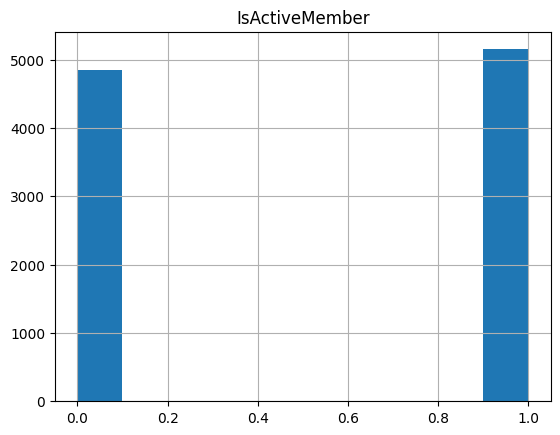
Hình 7. Biểu đồ phân bố số sản phẩm dịch vụ khách hàng sử dụng

Đa số khách hàng sử dụng 1 và 2 sản phẩm, có một lượng rất nhỏ khách hàng sử dụng đến sản phẩm thứ 3 và thứ 4.



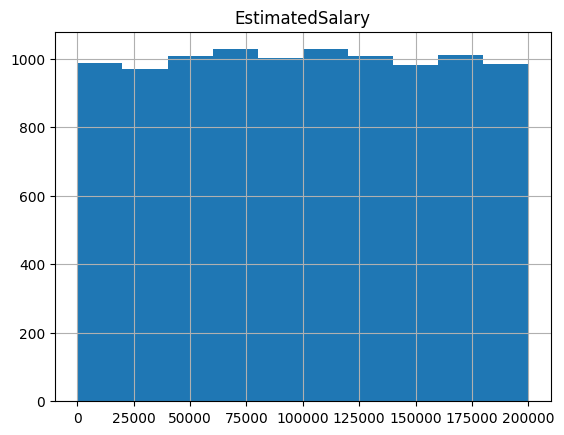
Hình 8. Biểu đồ phân bố thẻ tín dụng

Khoảng 70% khách hàng có thẻ tín dụng của ngân hàng, 30% khách hàng còn lại không kích hoạt thẻ.



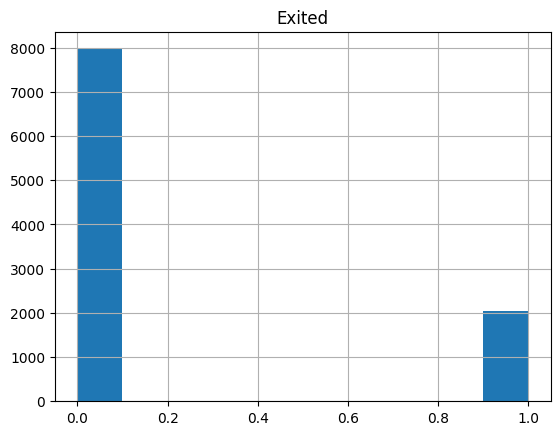
Hình 9. Biểu đồ phân bố hoạt động giao dịch của khách hàng

Có hơn 50% khách hàng thường xuyên giao dịch. Những khách hàng không thường xuyên giao dịch có thể có khả năng rời bỏ ngân hàng cao hơn.



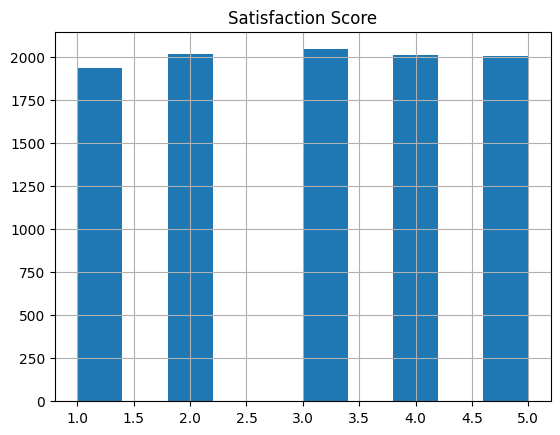
Hình 10. Biểu đồ phân bố mức lương ước tính của khách hàng

Biểu đồ cho thấy sự phân bố khá đồng đều giữa các khoảng giá trị. Không có khoảng nào nổi bật hơn hẳn so với các khoảng khác, cho thấy rằng khách hàng có mức lương ước tính trải đều từ 0 đến 200000.



Hình 11. Biểu đồ phân bố việc rời đi hay tiếp tục sử dụng của khách hàng

Biểu đồ cho thấy con số khá chênh lệch khi số lượng khách hàng rời đi chỉ chiếm khoảng 20%. Qua đó thấy được đa số khách hàng vẫn tiếp tục ở lại với ngân hàng.

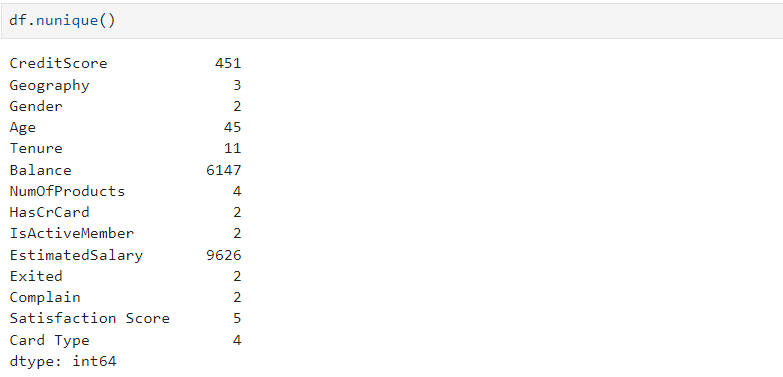


Hình 12. Biểu đồ phân bố điểm hài lòng của khách hàng

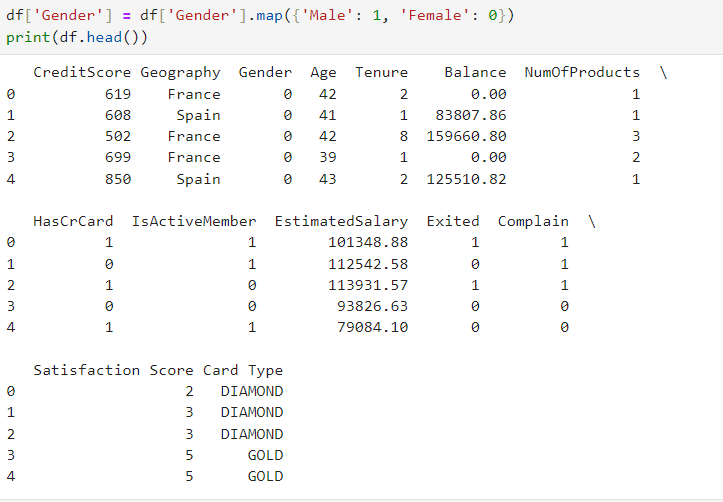
Biểu đồ này cung cấp cái nhìn tổng quan về mức độ hài lòng của khách hàng. Mặc dù có sự phân bố đồng đều, nhưng số lượng khách hàng đánh giá điểm thấp là khá lớn vì vậy việc chú ý đến những khách hàng không hài lòng là rất quan trọng để cải thiện dịch vụ và giữ chân khách hàng.

- Mã hóa dữ liệu:

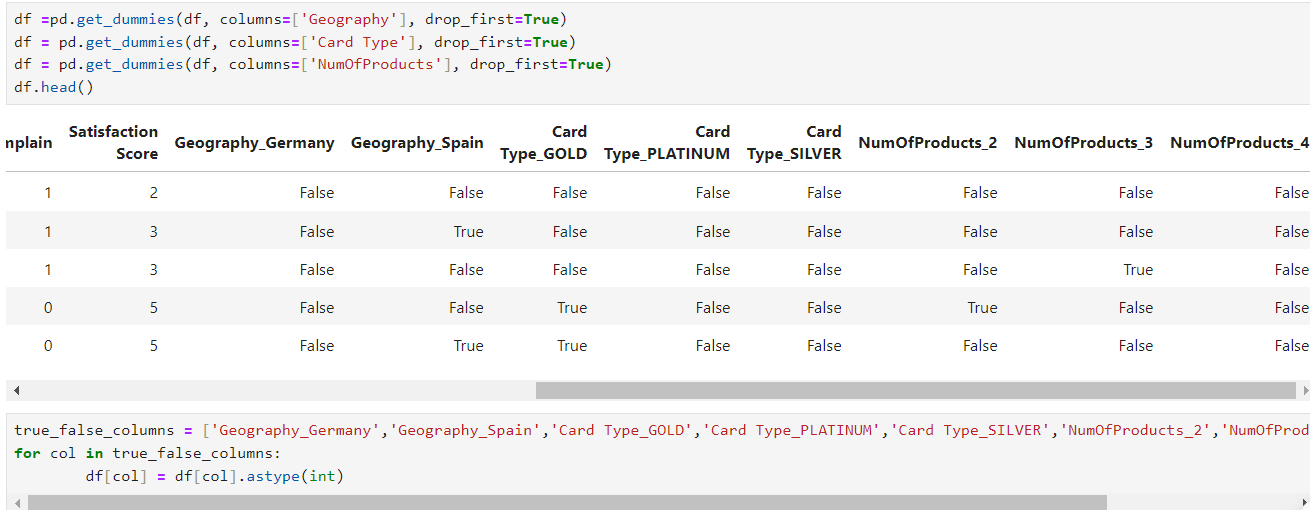
+ In ra giá trị duy nhất của từng thuộc tính:



+ Chuyển đổi dữ liệu cột Gender:

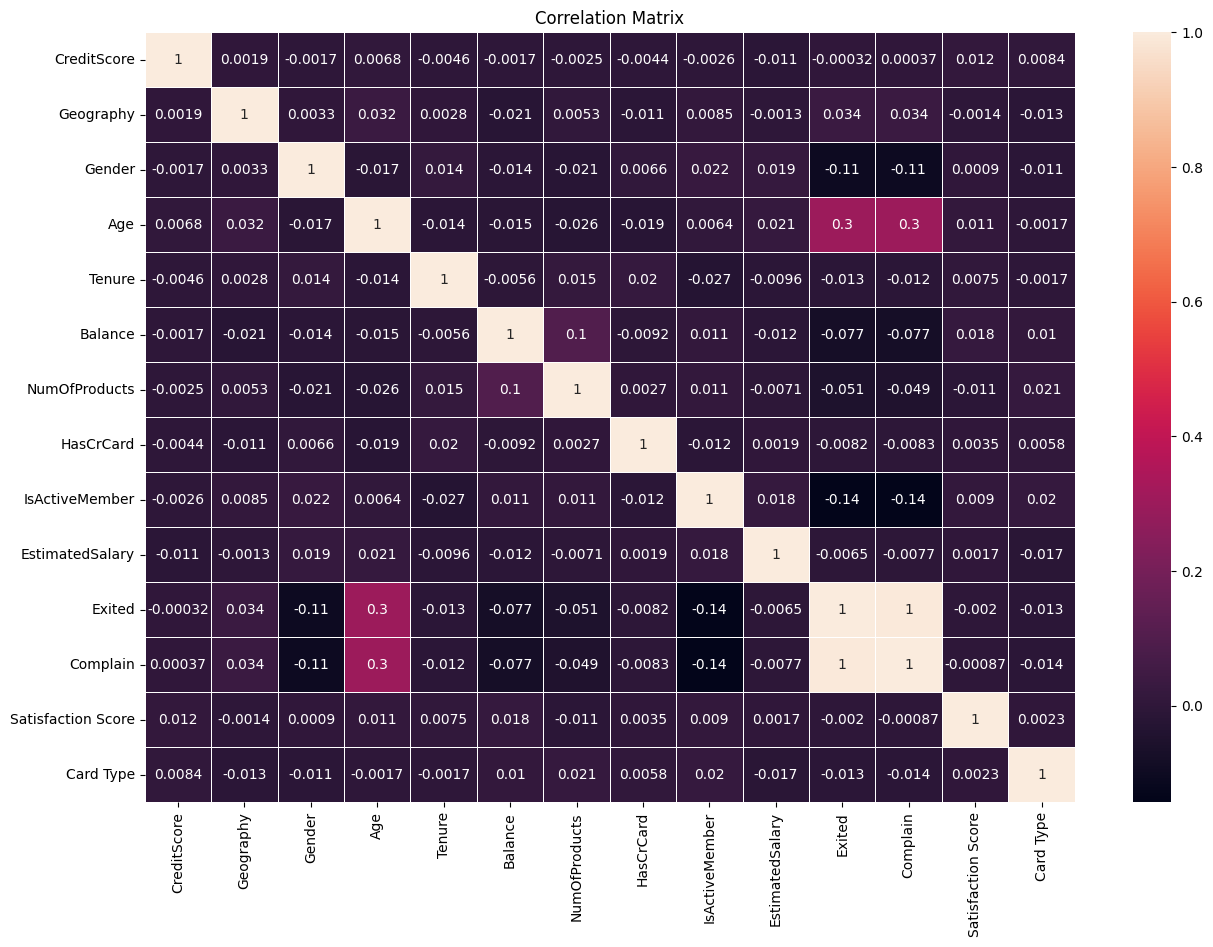


+ Chuyển đổi dữ liệu: One-hot encoding cho các cột ( Geography, Card Type, NumOfProducts )



- Chuẩn hóa dữ liệu: Sử dụng MinmaxScaler cho các biến liên tục (CreditScore, Age, Balance, Estimated Salary, Tenure ). Sau đó in ra dữ liệu để kiểm tra.



- Biểu đồ heatmap minh họa mối tương quan giữa các thuộc tính:   


Hình 13. Biểu đồ heatmap minh họa mối tương quan giữa các thuộc tính

**CreditScore (-0.00032):** Hệ số tương quan gần bằng 0, cho thấy không có mối quan hệ đáng kể giữa điểm tín dụng và việc khách hàng rời bỏ.

**Geography (0.034):** Hệ số tương quan dương rất nhỏ, chỉ ra rằng vị trí địa lý có mối quan hệ yếu và không đáng kể với quyết định rời bỏ của khách hàng.

**Gender (-0.11):** Hệ số tương quan âm nhỏ cho thấy có một mối quan hệ yếu giữa giới tính và khả năng rời bỏ, có thể là phụ nữ có xu hướng rời bỏ nhiều hơn một chút so với nam giới.

**Age (0.3):** Đây là hệ số tương quan cao nhất so với các biến khác, cho thấy tuổi tác có mối quan hệ tương đối mạnh với quyết định rời bỏ. Khách hàng lớn tuổi có khả năng rời bỏ cao hơn.

**Tenure (-0.013):** Hệ số tương quan gần bằng 0, điều này cho thấy thời gian khách hàng ở lại không có ảnh hưởng đáng kể đến quyết định rời bỏ.

**Balance (-0.077):** Hệ số tương quan âm yếu, cho thấy không có mối quan hệ rõ ràng giữa số dư tài khoản và quyết định rời bỏ. Khách hàng với số dư cao hoặc thấp đều có khả năng rời bỏ tương đương.

**NumOfProducts (-0.051):** Hệ số tương quan âm rất nhỏ, chỉ ra rằng số lượng sản phẩm khách hàng sử dụng không ảnh hưởng nhiều đến quyết định rời bỏ.

**HasCrCard (-0.0082):** Hệ số tương quan gần bằng 0, cho thấy việc có thẻ tín dụng hay không không có mối liên hệ đáng kể với khả năng rời bỏ.

**IsActiveMember (-0.14):** Hệ số tương quan âm khá rõ rệt, chỉ ra rằng khách hàng có xu hướng rời bỏ ít hơn nếu họ là thành viên tích cực.

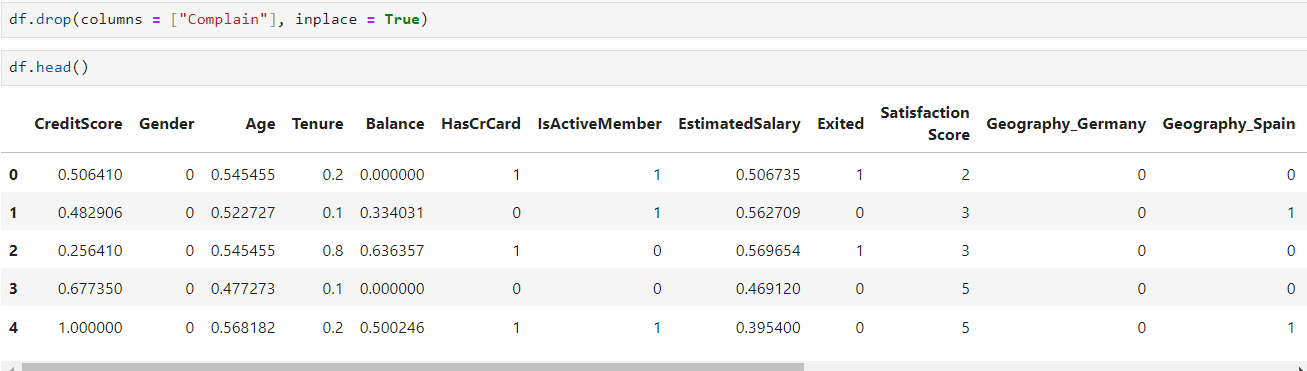
**Complain (1):** Hệ số tương quan bằng 1 cho thấy mối quan hệ hoàn hảo giữa biến này và việc rời bỏ. Điều này có thể là do biến *Complain* trực tiếp hoặc gián tiếp đại diện cho việc khách hàng rời bỏ, có thể do nó biểu thị một dấu hiệu chắc chắn của việc không hài lòng.

**Card Type (-0.013):** Hệ số tương quan rất nhỏ và âm, chỉ ra rằng loại thẻ khách hàng sở hữu không có mối quan hệ đáng kể với quyết định rời bỏ.

Ta thấy biến *Complain* có hệ số tương quan hoàn hảo với biến mục tiêu *Exited*, điều này cho thấy *Complain* có thể là một dấu hiệu trực tiếp của việc khách hàng rời bỏ nên biến này giúp mô hình dự đoán quá dễ dàng mà không cần đến các biến khác.

Để có một mô hình dự đoán thực tế và tổng quát hơn, ta nên loại bỏ biến *Complain*. Loại bỏ biến này sẽ giúp mô hình dựa vào các yếu tố khác để đưa ra dự đoán, giảm nguy cơ overfitting và tăng cường khả năng tổng quát hóa.

- Xóa cột complain:

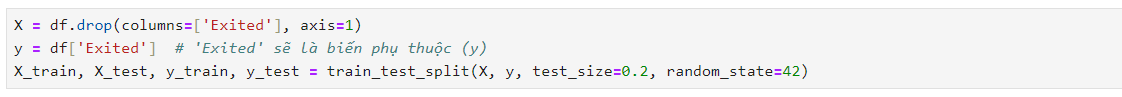


**CHƯƠNG IV: XÂY DỰNG MÔ HÌNH.**

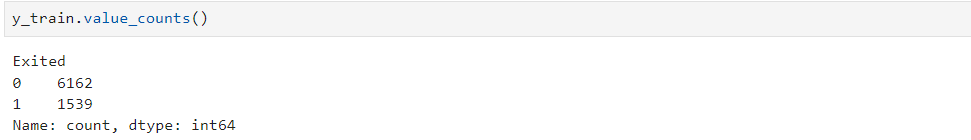
## **1. MÔ HÌNH HỒI QUY LOGISTIC**

### ***1.1 Xây dựng mô hình***

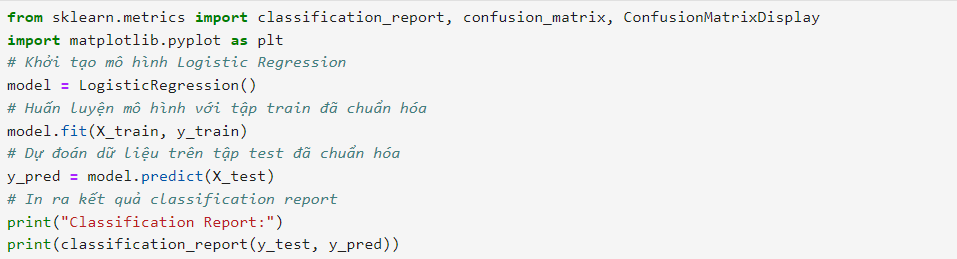
- Chia tập dữ liệu thành tập huấn và kiểm tra:

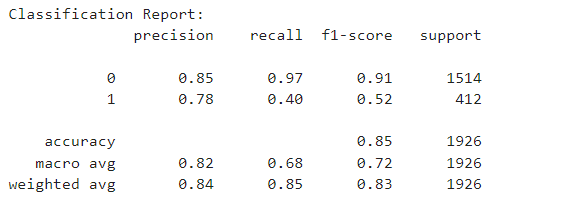


- In ra giá trị trong y\_train:



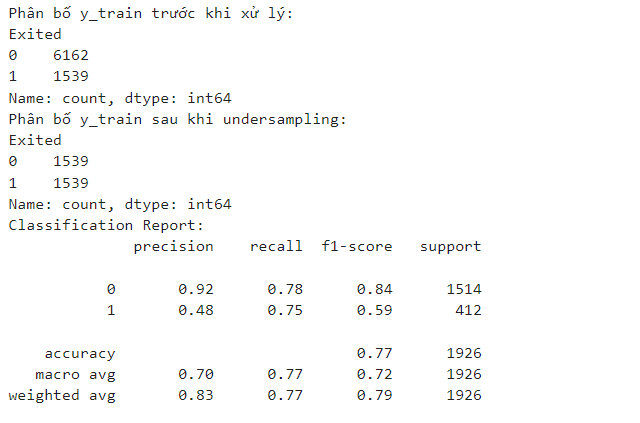
- Xây dựng mô hình:



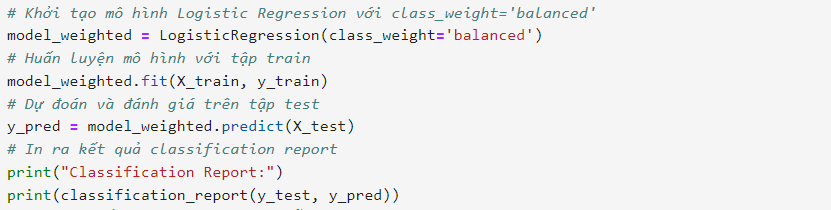


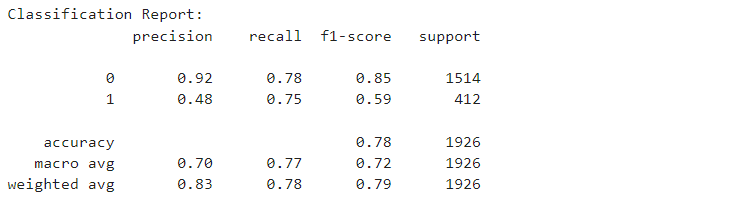
### ***1.2 Xử lý mất cân bằng dữ liệu bằng Undersampling***



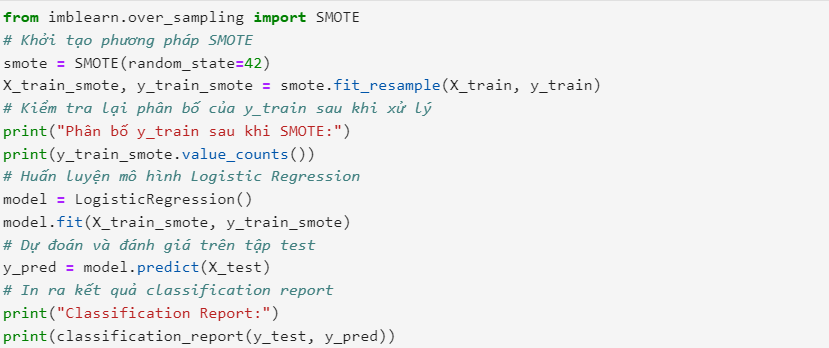


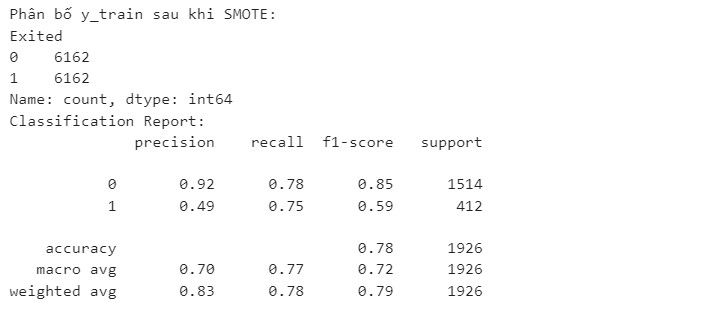
### ***1.3 Xứ lý mất cân bằng dữ liệu bằng thay đổi trọng số***



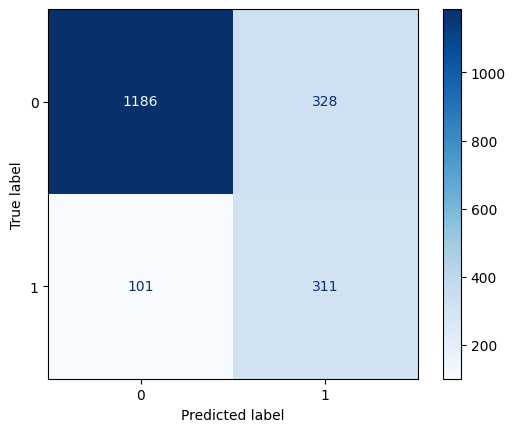


### ***1.4 Xử lý mất cân bằng dữ liệu với Smote***









Sau khi xử lý mất cân bằng dữ liệu và so sánh kết quả, có thể nhận thấy rằng mô hình hồi quy logistic với smote cho kết quả tốt hơn các phương pháp khác. Vì vậy, nhóm nghiên cứu sẽ đi đánh giá kết quả của mô hình này:

Mô hình đạt được độ chính xác khoảng 78%, cho thấy khả năng phân loại tổng quát đúng các trường hợp rời bỏ và ở lại của khách hàng là tương đối tốt.

Recall đạt 75%, cho thấy mô hình có khả năng phát hiện được 75% các trường hợp khách hàng thực sự rời bỏ. Khoảng 25% khách hàng rời bỏ không được mô hình dự đoán.

F1-Score cho lớp 1 chỉ đạt 59%, cho thấy sự mất cân bằng giữa Precision và Recall.  F1-Score thấp có thể chỉ ra rằng mô hình không thực sự hiệu quả trong việc xác định khách hàng có nguy cơ rời bỏ, điều này làm giảm độ tin cậy của mô hình trong việc dự đoán.

Precision chỉ đạt khoảng 49%, cho thấy rằng gần một nửa số khách hàng mà mô hình dự đoán sẽ rời bỏ thực tế không rời bỏ.

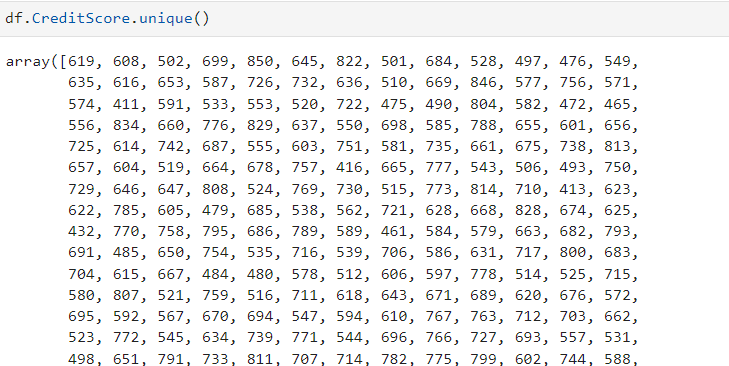
Vì vậy mô hình không thực sự hiệu quả để dự đoán chính xác việc rời bỏ của khách hàng.

## **2. MÔ HÌNH CÂY QUYẾT ĐỊNH**

### ***2.1 Xây dựng mô hình***

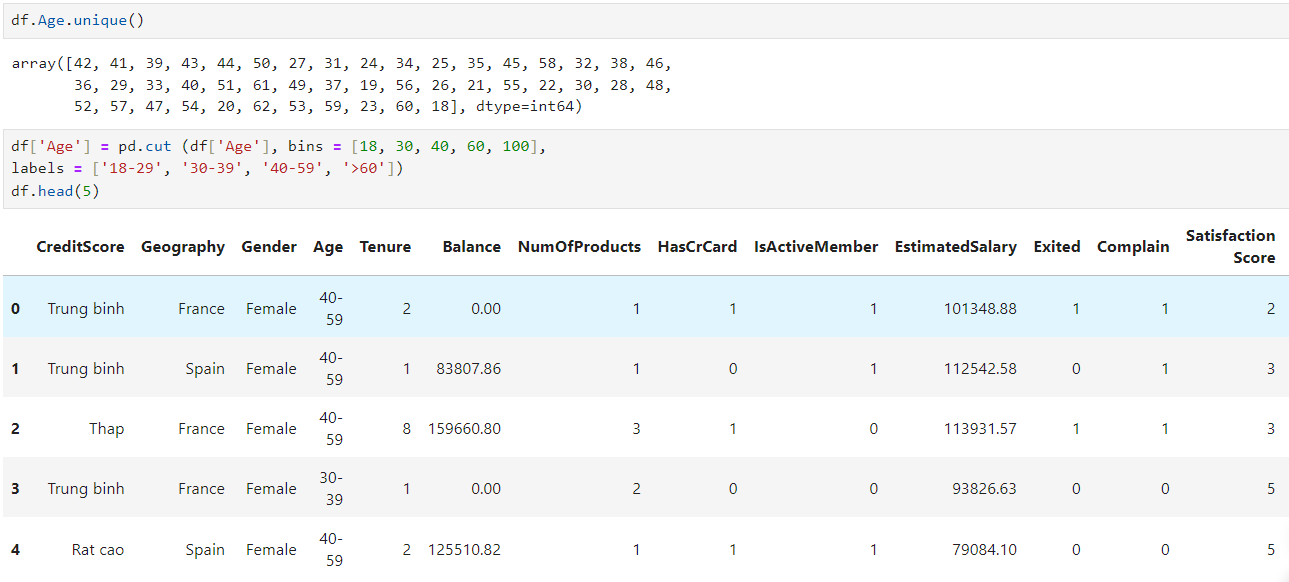
- Trong mô hình cây quyết định, quy trình tiền xử lý dữ liệu cũng tương tự như mô hình hồi quy logistic nhưng có thay đổi cách mã hóa các cột (CreditScore, Age, Balance, EstimatedSalary). Đó là phân chia giá trị ban đầu trong các cột này thành các khoảng giá trị, sau đó gán nhãn cho mỗi khoảng.

+ Phân chia khoảng giá trị cột CreditScore:

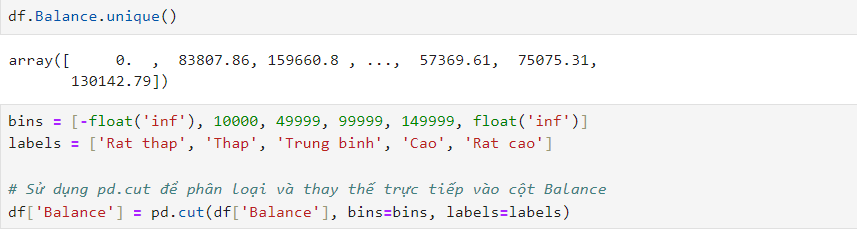




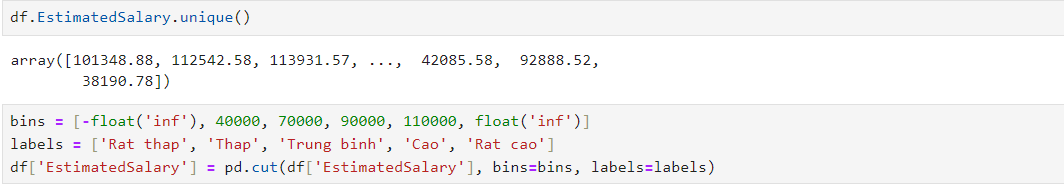
+ Phân chia khoảng giá trị cột Age:



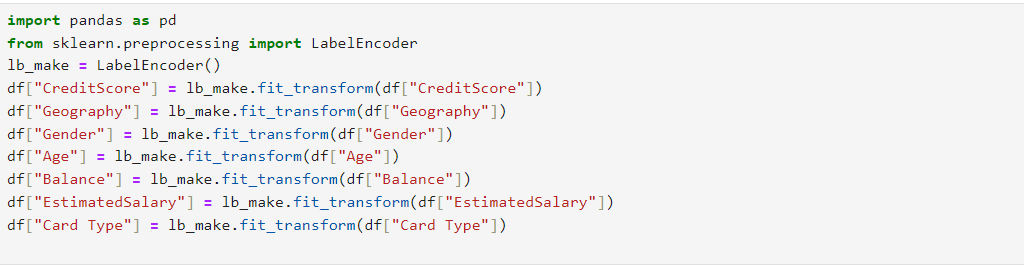
+ Phân chia khoảng giá trị cột Balance:

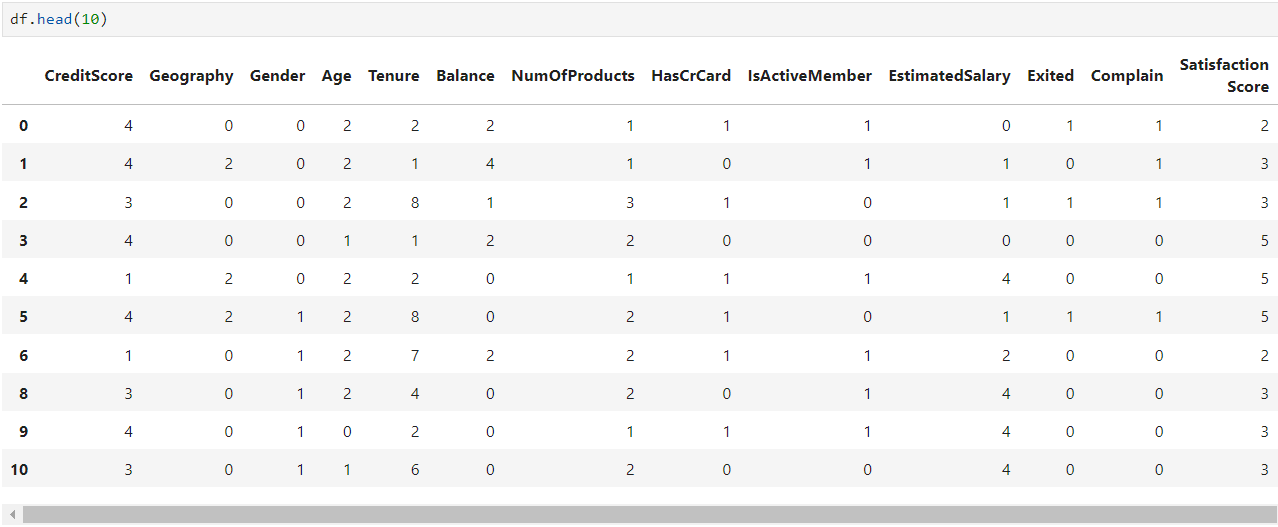


+ Phân chia khoảng giá trị cột EstimatedSalary:



- Mã hóa các cột dữ liệu về kiểu số nguyên:

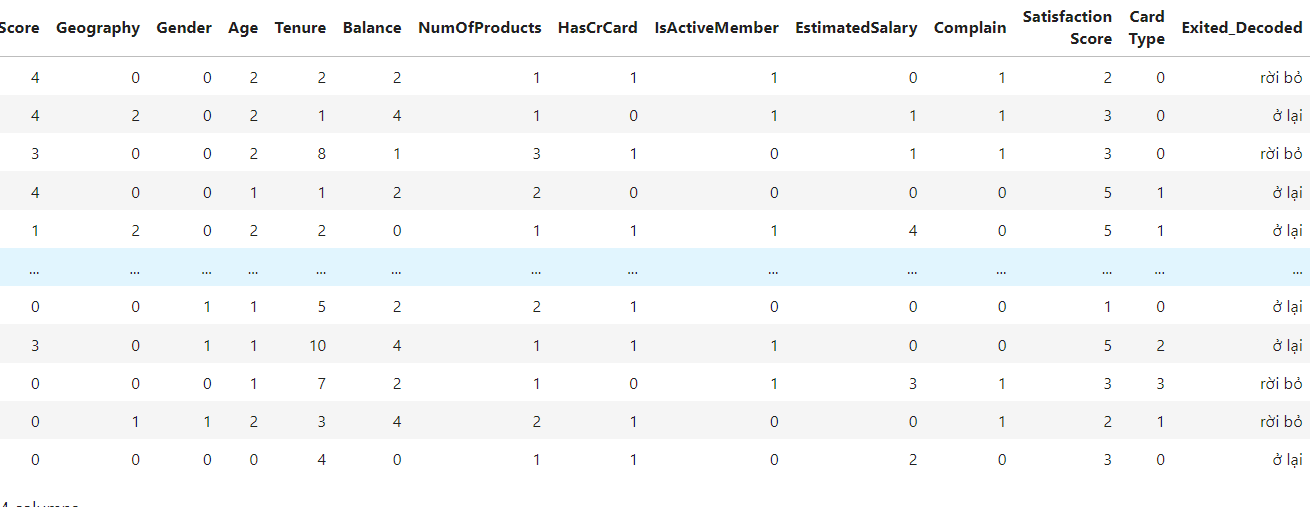




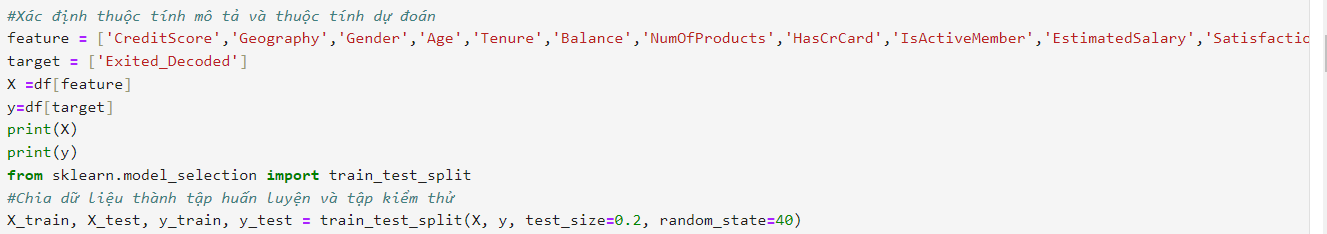
- Loại bỏ cột Complain:



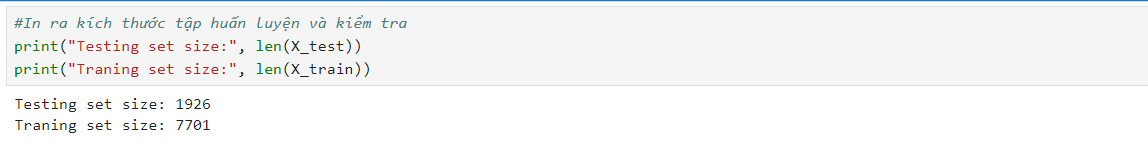
 - Tạo ánh xạ cho biến mục tiêu Exited:



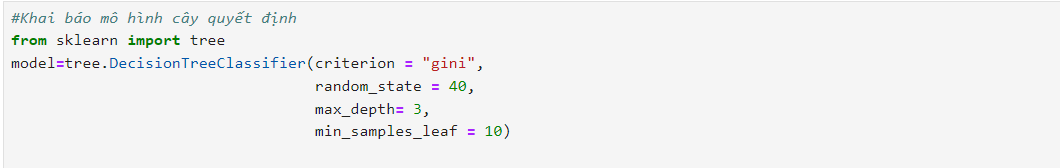
- Chia dữ liệu thành tập huấn luyện và tập kiểm thử:



- In ra kích thước 2 tập dữ liệu:

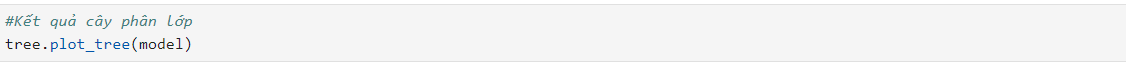


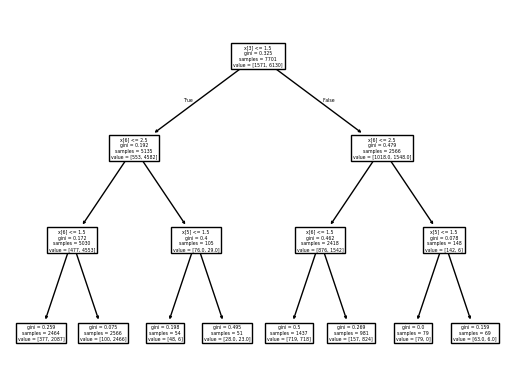
- Xây dựng mô hình:



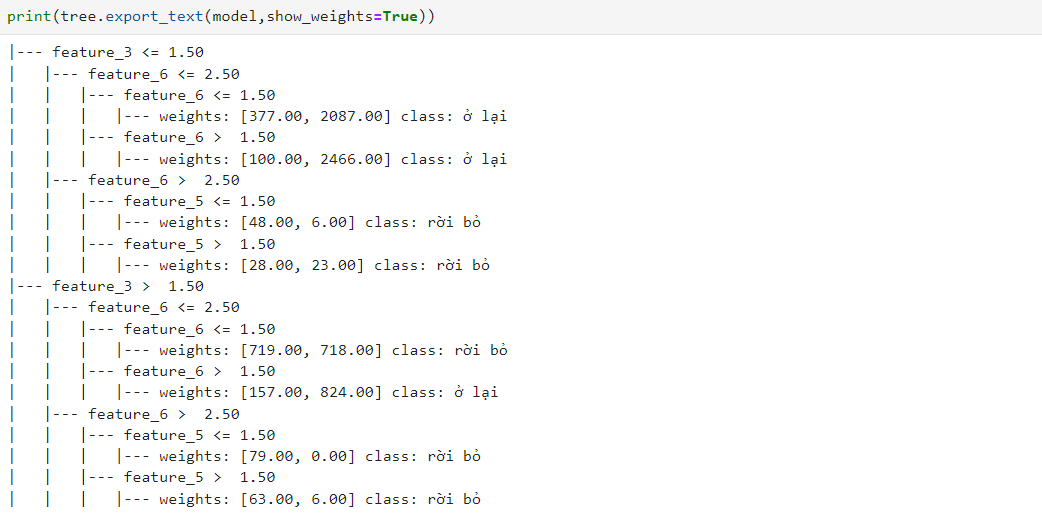
* Criterion = “gini”: Tham số này xác định phương pháp đo lường sự không chắc chắn khi chia một nút trong quá trình xây dựng cây. “gini” sử dụng độ mất mát thông tin dựa trên gini
* random\_state = 40: Đây là một Đây là một giá trị seed được sử dụng trong việc tạo các giá trị ngẫu nhiên. Đặt giá trị này giúp việc lặp lại quá trình đào tạo mô hình khi cần đảm bảo kết quả giống nhau trong mỗi lần chạy.
* max\_depth = 3: Xác định độ sâu tối đa của cây quyết định. Điều này giúp kiểm soát việc mô hình có thể trở nên quá phức tạp và có thể giúp tránh tình trạng quá mức điều chỉnh (overfitting) dữ liệu đào tạo.
* min\_samples\_leaf = 10: Số lượng mẫu tối thiểu mà một lá của cây quyết định cần phải có. Nếu số lượng mẫu tại một lá nhỏ hơn giá trị này, thì việc tách nút sẽ không được thực hiện

- Thực thi mô hình:  

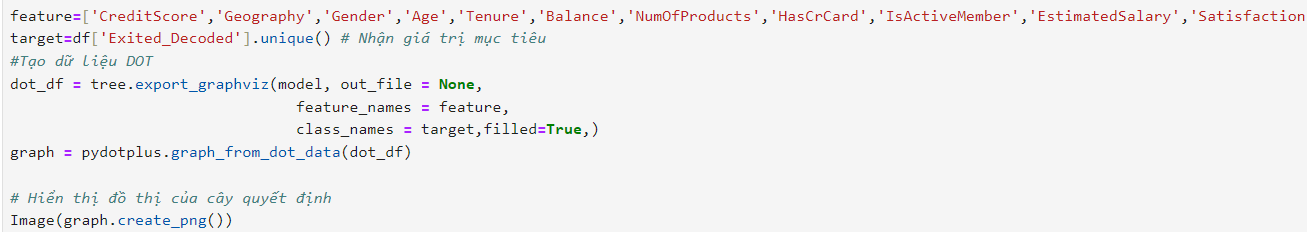


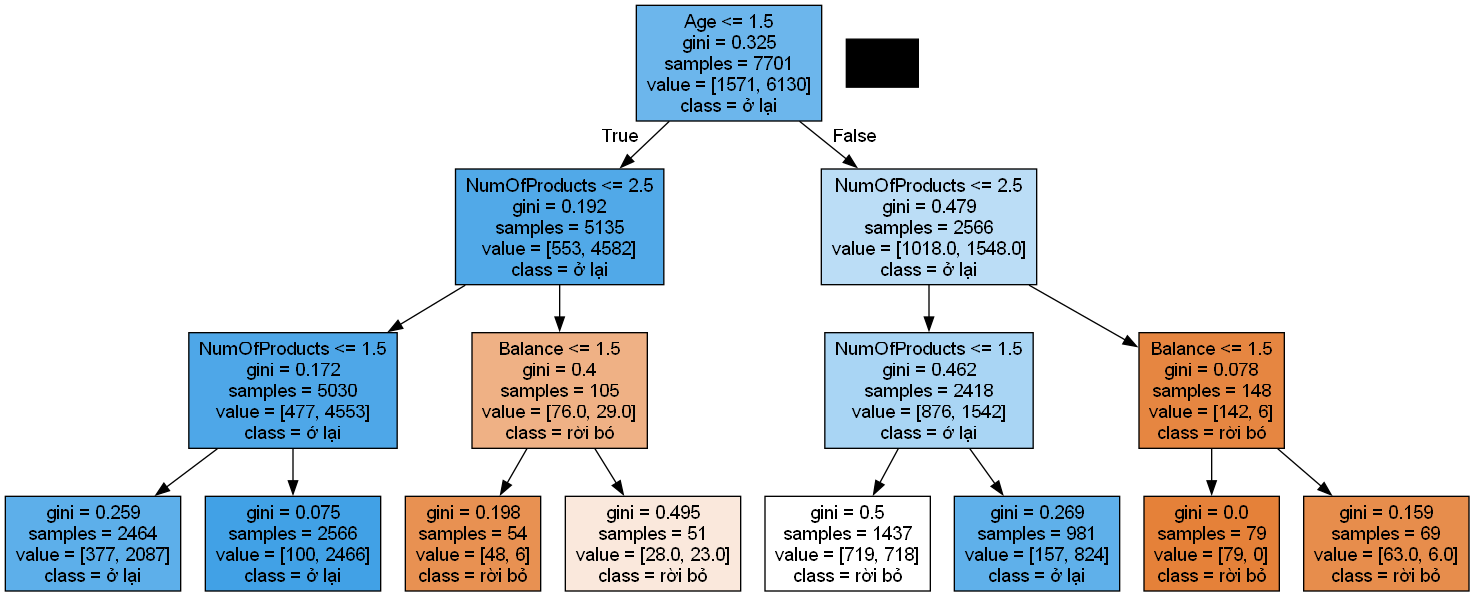
- Biểu diễn cây quyết định:



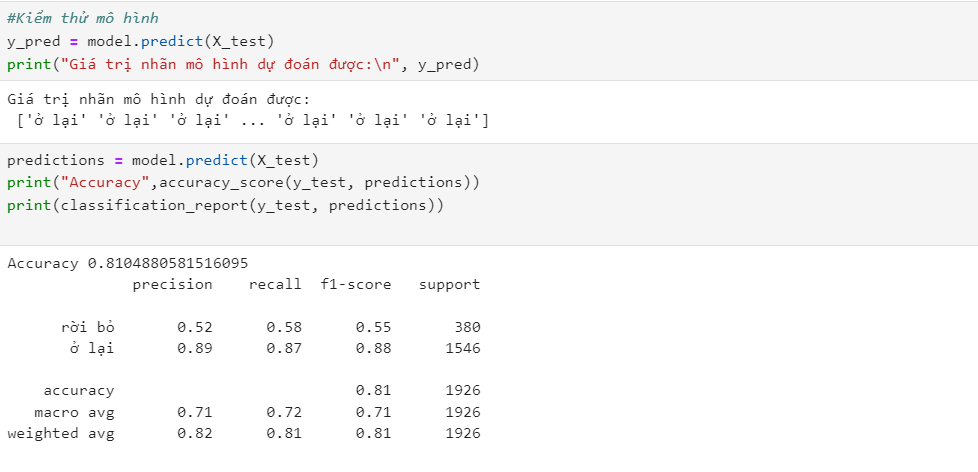
Hình 14. Biểu diễn cây quyết định

- Biểu diễn cây quyết định với Graphviz:





Hình 15. Biểu diễn cây quyết định với Graphviz

- Kết quả mô hình:  


Mô hình đạt được độ chính xác khoảng 81%, cho thấy khả năng phân loại đúng các trường hợp rời bỏ và ở lại của khách hàng là tương đối tốt.

Recall đạt 58%, cho thấy mô hình có khả năng phát hiện được 58% các trường hợp khách hàng thực sự rời bỏ. Khoảng 42% khách hàng rời bỏ không được mô hình dự đoán.

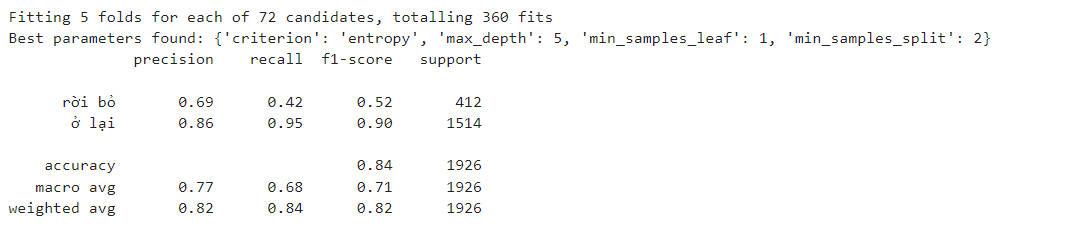
F1-Score cho lớp 1 chỉ đạt 55%, cho thấy sự mất cân bằng giữa Precision và Recall. F1- Score thấp có thể chỉ ra rằng mô hình không thực sự hiệu quả trong việc xác định khách hàng có nguy cơ rời bỏ, điều này làm giảm độ tin cậy của mô hình trong việc dự đoán.

Precision chỉ đạt khoảng 52%, cho thấy rằng gần một nửa số khách hàng mà mô hình dự đoán sẽ rời bỏ thực tế không rời bỏ.

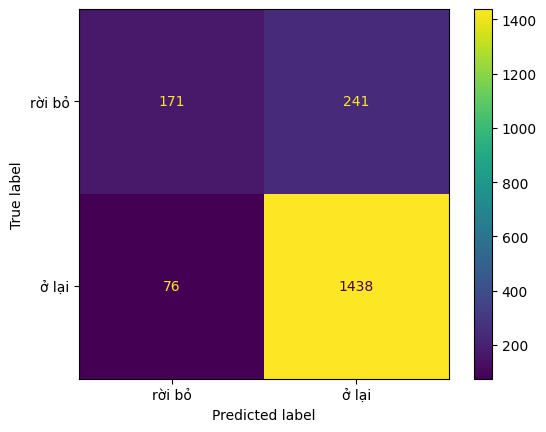
Các chỉ số dự báo lớp rời bỏ chỉ ở mức trung bình. Điều này cho thấy rằng mô hình cây quyết định hiện tại chưa thực sự tối ưu để phát hiện chính xác các khách hàng có ý định rời bỏ.

### ***2.2 Kịch bản nâng cao độ chính xác của mô hình***







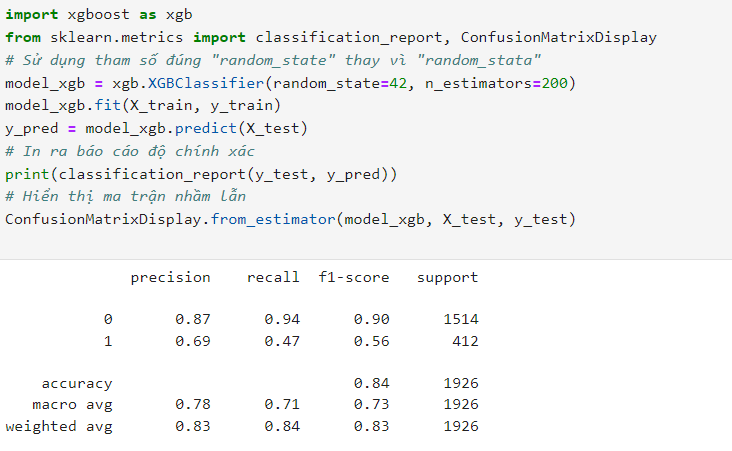


Độ chính xác tăng lên 69% cho thấy trong các dự đoán về khách hàng rời bỏ, có 69% là đúng. Điều này cho thấy rằng mô hình đã đạt được một mức độ hiệu quả nhất định,

Tuy nhiên recall chỉ đạt 42% thể hiện số lượng lớn khách hàng rời bỏ mà mô hình không phát hiện được. F1-Score vẫn ở mức thấp. Mô hình cần được cải thiện thêm hoặc cần áp dụng mô hình khác để tối ưu kết quả hơn.

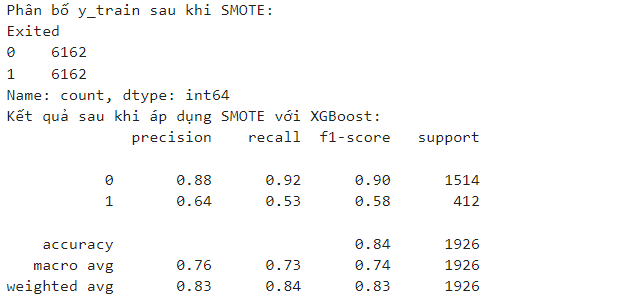
## **3. MÔ HÌNH XGBOOST**

- Xây dựng mô hình Xgboost:



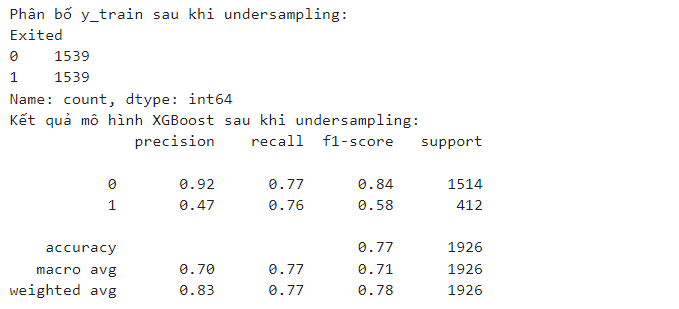
### ***3.1 Xây dựng XGboost với Smote***





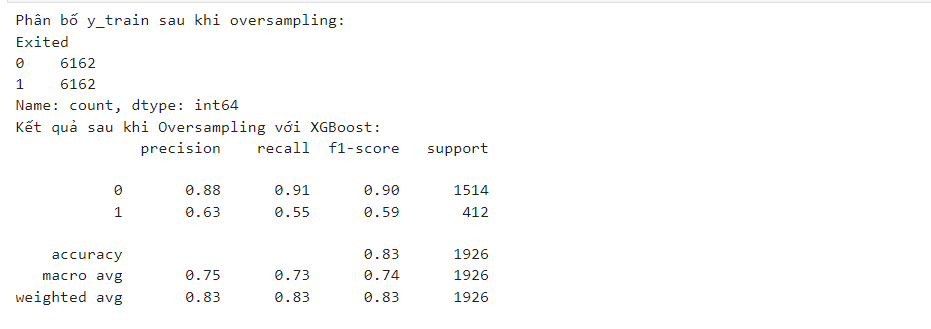
### ***3.2. Xây dựng mô hình XGboost với Undersampling***



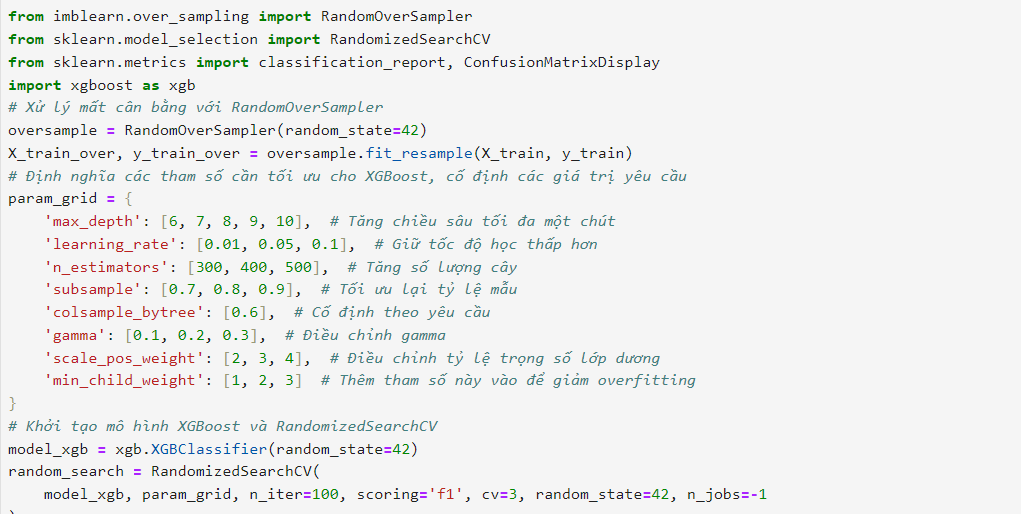


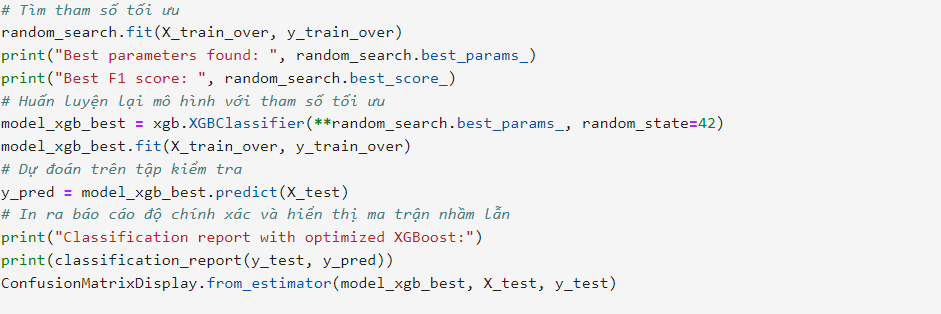
### ***3.3 Xây dựng mô hình XGboost với Oversampling***

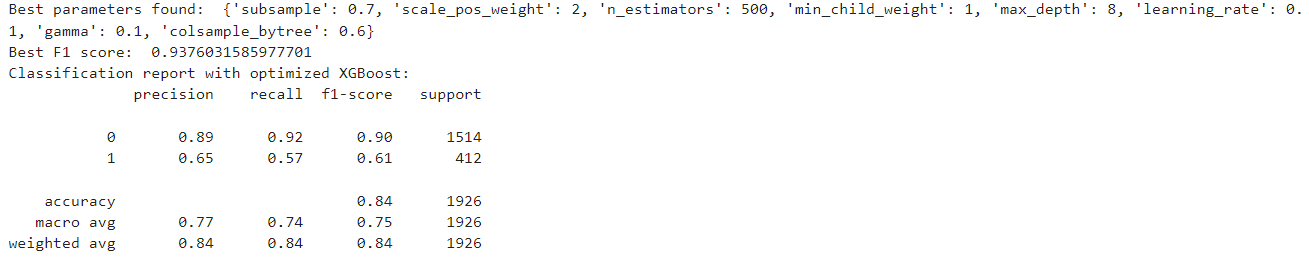




### ***3.4 Kịch bản nâng cao độ chính xác của mô hình***





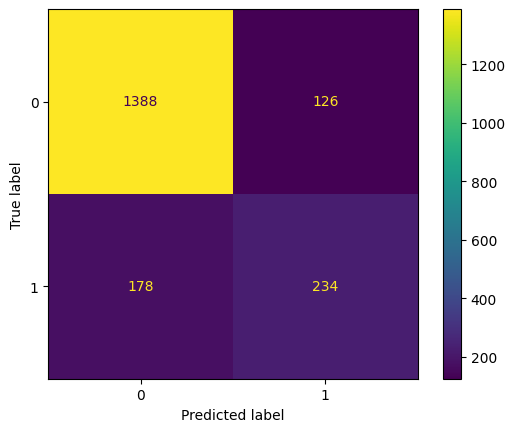


Độ chính xác (Precision): Mô hình đạt được độ chính xác 65% cho lớp 1, cho thấy trong số các mẫu dự đoán là lớp 1, có 65% mẫu thực sự thuộc lớp này. Mặc dù không cao, nó vẫn cho thấy khả năng dự đoán hợp lý.

Tỉ lệ hồi phục (Recall): Tỉ lệ hồi phục cho lớp 1 chỉ đạt 57%. Điều này cho thấy mô hình có xu hướng bỏ lỡ một số mẫu thực sự thuộc lớp 1. Tỉ lệ này thấp, cần cải thiện để tăng khả năng phát hiện các mẫu của lớp thiểu số.

F1-Score: Chỉ số F1-Score cho lớp 1 là 61%, cho thấy mô hình đã đạt được sự cân bằng nhất định giữa độ chính xác và khả năng phát hiện, nhưng vẫn chưa quá tốt.

- Ma trận nhầm lẫn:



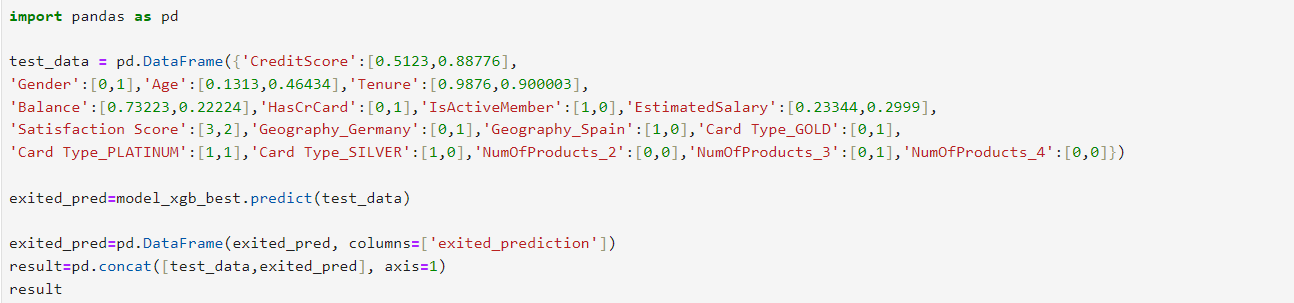
True Negatives: 1388

False Positives: 126

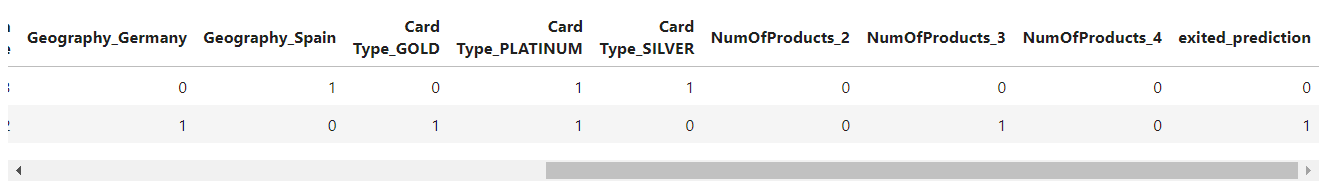
False Negatives: 178

True Positives: 234

***3.5 Sử dụng mô hình để dự đoán***

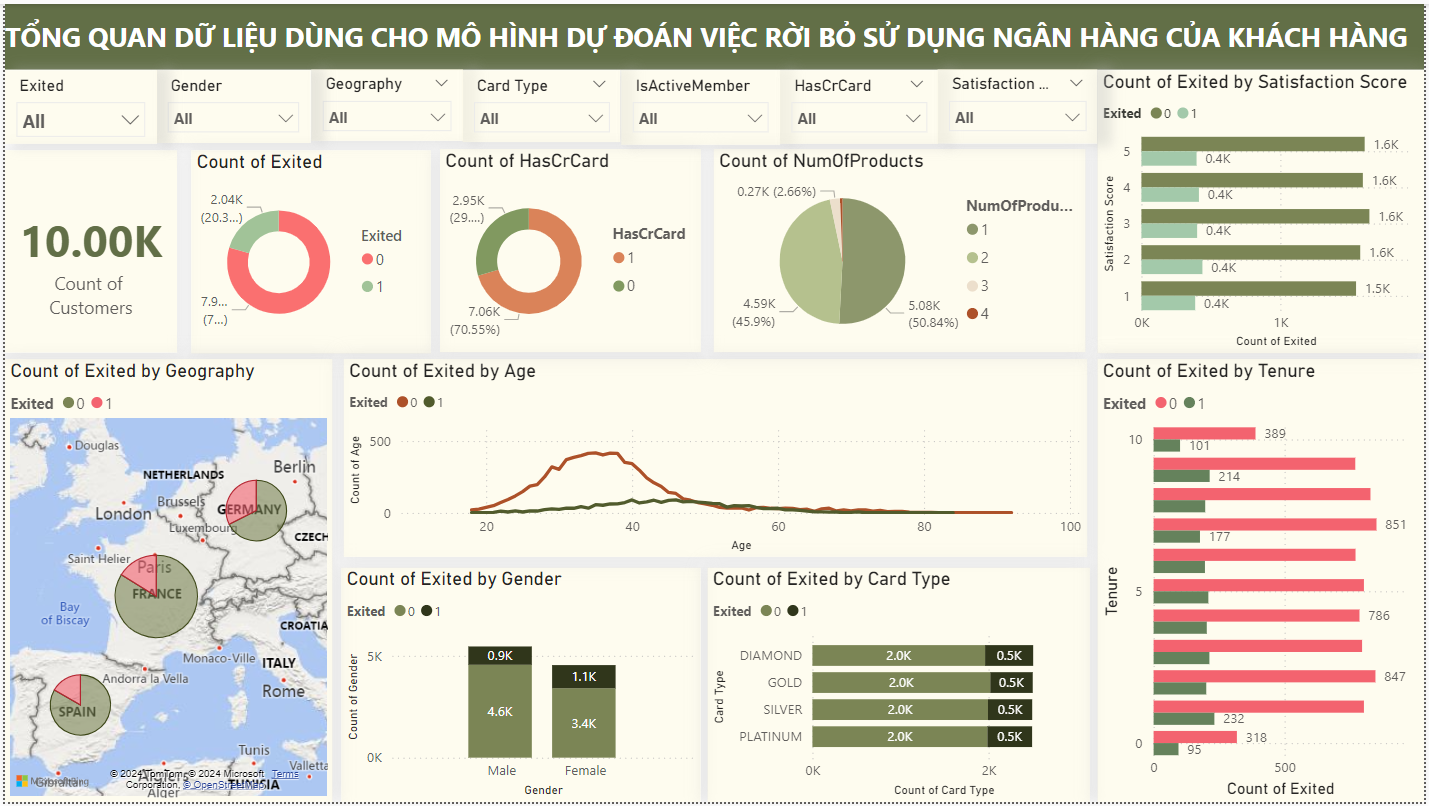


Kết quả dự đoán:



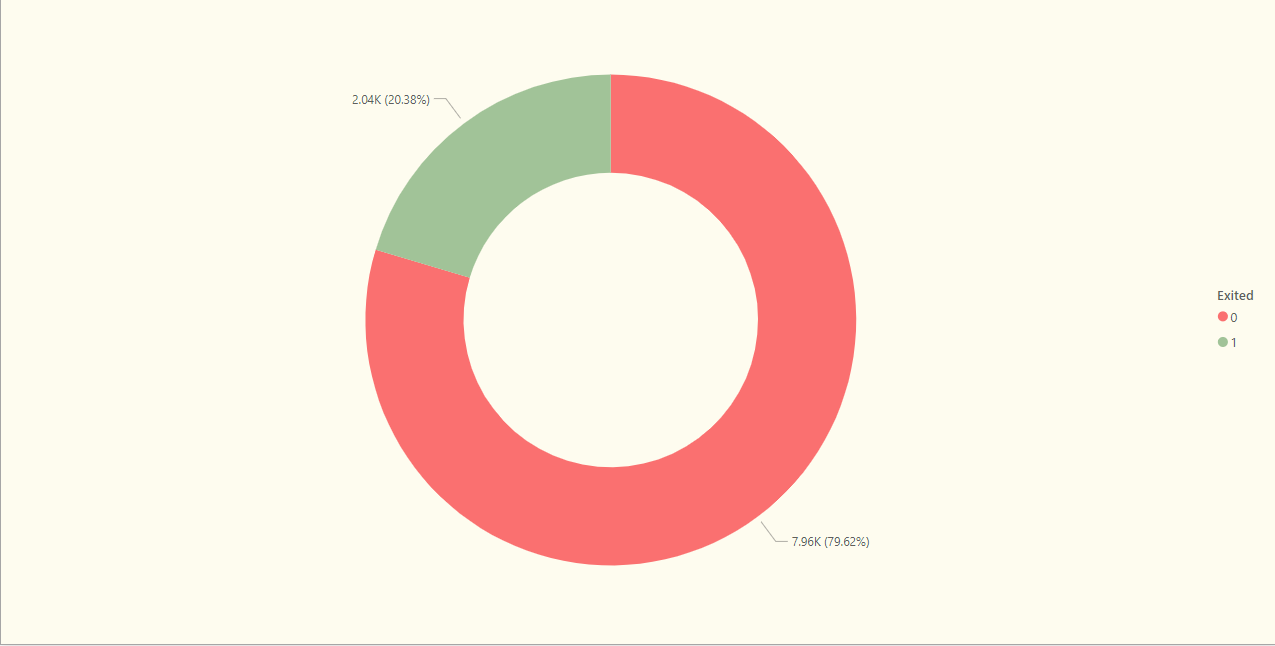
# **CHƯƠNG VI. XÂY DỰNG DASHBOARD TRÊN POWER BI**

## ***1.Tạo biểu đồ dashboard***



Hình 16. Dashboard tổng quan dữ liệu dùng cho mô hình dự đoán việc rời bỏ sử dụng ngân hàng của khách hàng

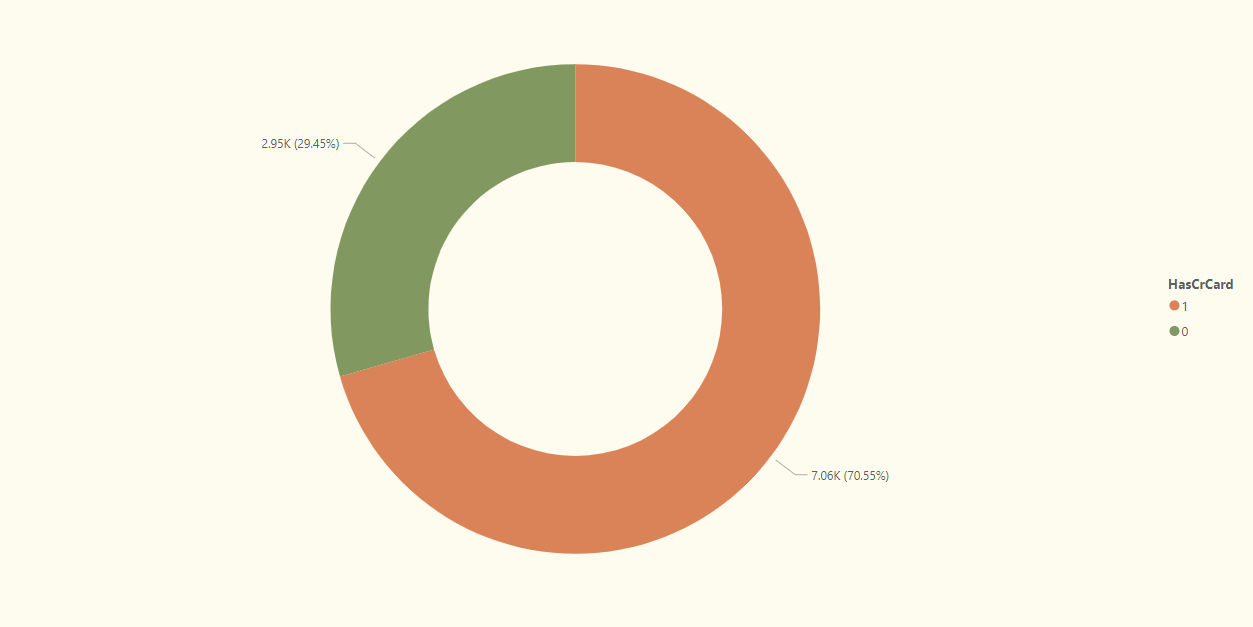
## ***2. Báo cáo rút ra từ biểu đồ***



Hình 17. Tỷ lệ khách hàng rời bỏ

Biểu đồ này thể hiện tổng quan về tỷ lệ khách hàng rời bỏ dịch vụ của ngân hàng so với tổng số khách hàng:

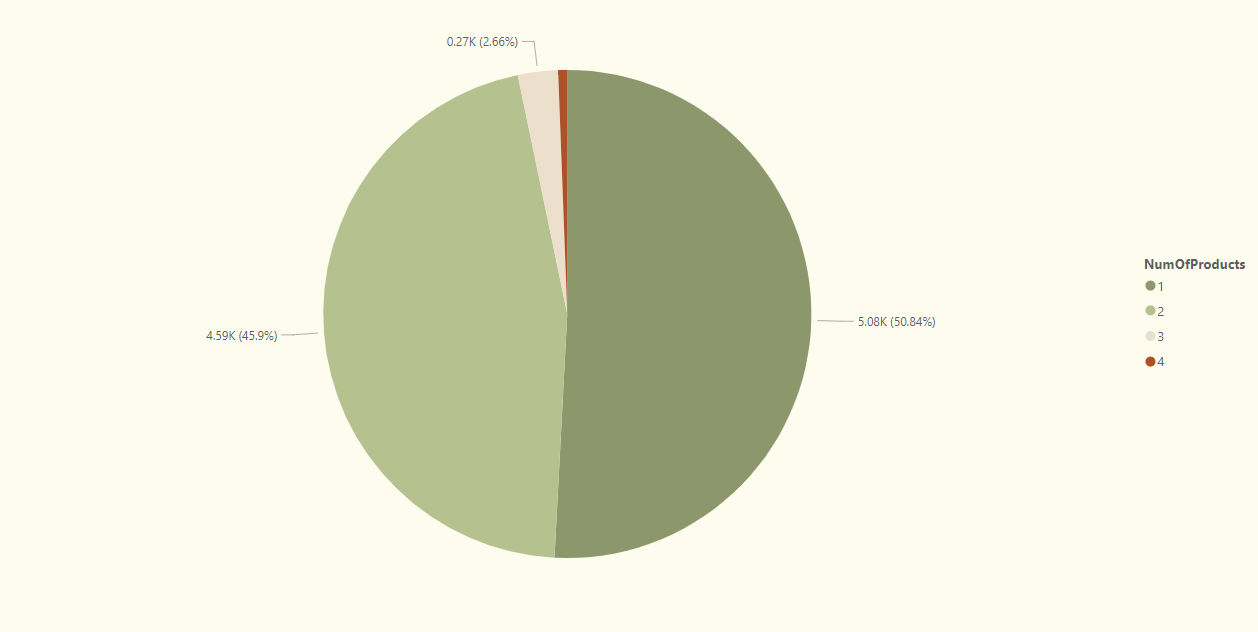
* **Khách hàng không rời bỏ (79.62%)**: Tỷ lệ này chiếm phần lớn tổng số khách hàng, cho thấy ngân hàng đang thực hiện tốt việc giữ chân khách hàng. Một tỷ lệ cao như vậy có thể phản ánh sự hài lòng của khách hàng đối với dịch vụ, sản phẩm, và trải nghiệm tổng thể mà ngân hàng cung cấp. Để duy trì và củng cố tỷ lệ này, ngân hàng cần tiếp tục cải thiện chất lượng dịch vụ, lắng nghe phản hồi từ khách hàng, và triển khai các chương trình chăm sóc khách hàng hiệu quả.
* **Khách hàng rời bỏ (20.38%)**: Ngân hàng nên phân tích nguyên nhân tại sao một phần khách hàng lại quyết định rời bỏ, có thể là do sự không hài lòng với dịch vụ, thiếu sản phẩm phù hợp, hoặc cạnh tranh từ các ngân hàng khác. Việc hiểu rõ nguyên nhân sẽ giúp ngân hàng đưa ra các giải pháp hợp lý để giảm thiểu tỷ lệ rời bỏ, chẳng hạn như cải thiện trải nghiệm khách hàng hoặc cung cấp các sản phẩm mới hấp dẫn hơn.



Hình 18. Tỷ lệ khách hàng có thẻ tín dụng và không có thẻ tín dụng

Biểu đồ này cho thấy tỷ lệ khách hàng sở hữu thẻ tín dụng và không sở hữu thẻ:

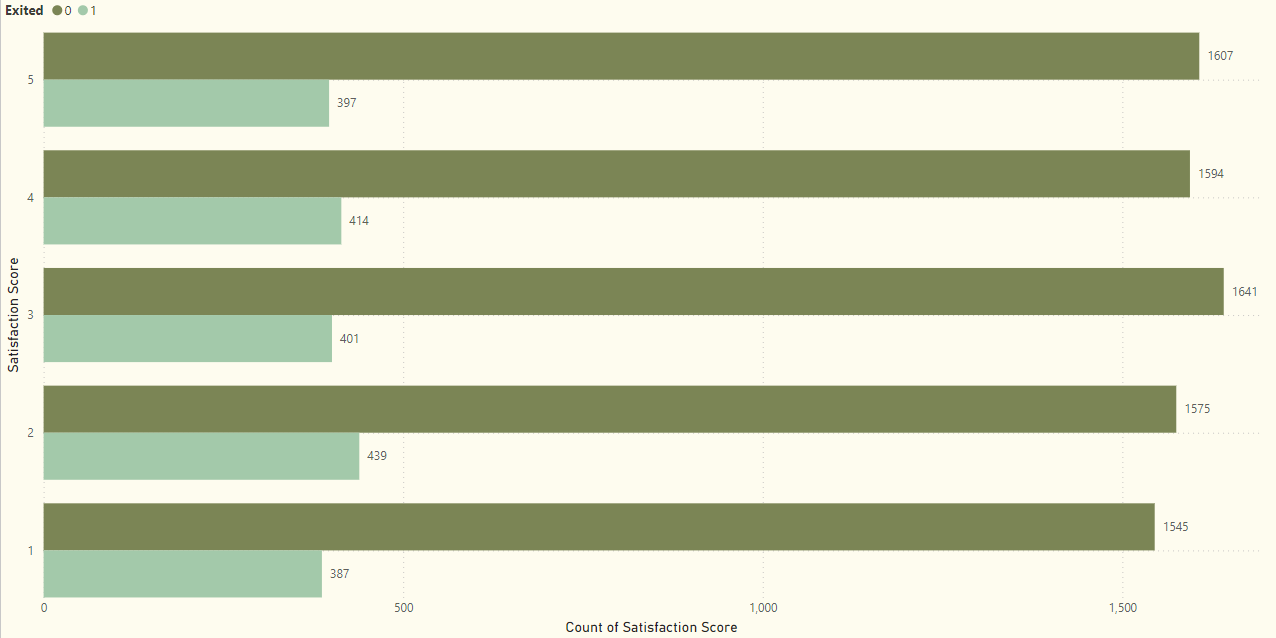
* **Khách hàng có thẻ tín dụng (70.55%)**: Tỷ lệ này cho thấy phần lớn khách hàng đã sử dụng thẻ tín dụng, điều này chỉ ra rằng ngân hàng đã thành công trong việc tiếp cận và khuyến khích khách hàng mở thẻ. Sự phổ biến của thẻ tín dụng có thể phản ánh tính năng hữu ích và các lợi ích mà ngân hàng đã cung cấp, như chương trình ưu đãi, điểm thưởng, và các dịch vụ hỗ trợ khách hàng. Ngân hàng nên tiếp tục duy trì và phát triển các chương trình này để củng cố sự trung thành của khách hàng.
* **Khách hàng không có thẻ tín dụng (29.45%)**: Mặc dù tỷ lệ này thấp hơn nhiều so với nhóm khách hàng có thẻ, nhưng vẫn cần chú ý đến nhóm này. Ngân hàng nên nghiên cứu lý do vì sao một phần khách hàng không sử dụng thẻ tín dụng. Có thể họ chưa nhận thức được lợi ích của việc sở hữu thẻ hoặc có những lo ngại về phí duy trì hoặc lãi suất. Ngân hàng có thể phát triển các chương trình khuyến mãi, chẳng hạn như ưu đãi khi mở thẻ lần đầu hoặc miễn phí năm đầu, để khuyến khích nhóm này tham gia.



Hình 19. Tỷ lệ số dịch vụ của ngân hàng mà mỗi khách hàng sử dụng

**Ý nghĩa**: Biểu đồ này thể hiện tỷ lệ khách hàng sử dụng 1, 2, 3, hoặc 4 dịch vụ của ngân hàng. Nó giúp ngân hàng hiểu rõ mức độ thâm nhập và sự đa dạng trong việc sử dụng các sản phẩm dịch vụ của khách hàng:

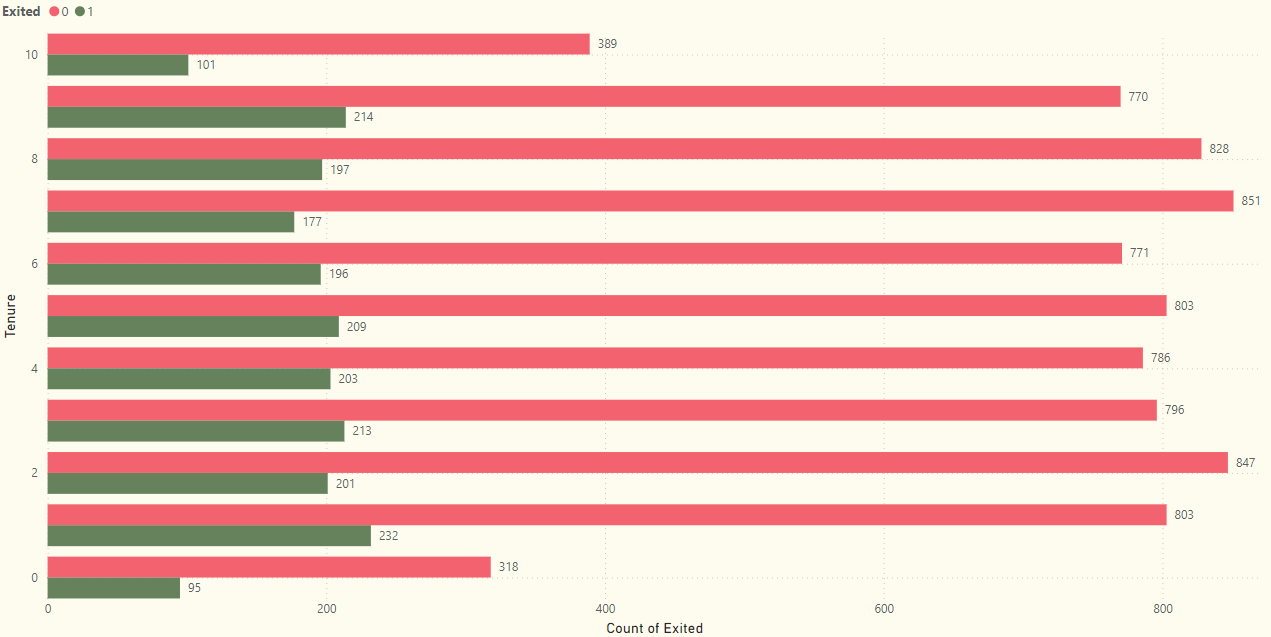
* **Sử dụng 1 dịch vụ (50.84%)**: Tỷ lệ này chiếm hơn một nửa tổng số khách hàng, cho thấy nhiều khách hàng chỉ sử dụng một sản phẩm duy nhất. Điều này có thể chỉ ra rằng ngân hàng cần tăng cường quảng bá và khuyến khích khách hàng khám phá các dịch vụ khác. Các chiến dịch truyền thông và ưu đãi có thể giúp nâng cao nhận thức và khuyến khích họ mở thêm các sản phẩm như thẻ tín dụng hoặc khoản vay.
* **Sử dụng 2 dịch vụ (45.9%)**: Tỷ lệ này gần với số lượng khách hàng chỉ sử dụng một dịch vụ, cho thấy rằng một phần lớn khách hàng đã bắt đầu khai thác thêm các sản phẩm khác. Ngân hàng nên tận dụng cơ hội này để tiếp cận những khách hàng này, khuyến khích họ mở thêm dịch vụ thứ ba hoặc thứ tư, chẳng hạn thông qua các gói kết hợp hoặc ưu đãi hấp dẫn.
* **Sử dụng 3,4 dịch vụ (3.26%):** Tỷ lệ này rất thấp, chỉ ra rằng số lượng khách hàng khai thác đa dạng dịch vụ còn hạn chế. Ngân hàng cần tìm hiểu nguyên nhân và cải thiện trải nghiệm để làm cho các sản phẩm bổ sung trở nên hấp dẫn hơn. Việc cung cấp thông tin rõ ràng về lợi ích của các dịch vụ này có thể giúp tăng số lượng khách hàng sử dụng nhiều sản phẩm hơn.



Hình 20. Tỷ lệ khách hàng rời bỏ theo điểm hài lòng khi sử dụng ngân hàng

Biểu đồ này phân tích tỷ lệ khách hàng rời bỏ dựa trên điểm hài lòng mà họ đánh giá về dịch vụ của ngân hàng. Mục đích của biểu đồ là giúp ngân hàng nhận diện mối liên hệ giữa mức độ hài lòng của khách hàng và quyết định rời bỏ, từ đó có cái nhìn sâu sắc hơn về tác động của trải nghiệm dịch vụ đến sự trung thành của khách hàng.

* Tỷ lệ rời bỏ không giảm đều theo mức độ hài lòng. Cụ thể, mức hài lòng 2 có tỷ lệ rời bỏ cao nhất, với 21.8%. Trong khi đó, mức hài lòng 5 (hài lòng cao nhất) vẫn có tỷ lệ rời bỏ đáng kể, chiếm 19.8%. Điều này cho thấy rằng mức độ hài lòng thấp (1 và 2) không đồng nghĩa với tỷ lệ rời bỏ cao nhất. Mức hài lòng 1 có tỷ lệ rời bỏ là 20.03%
* Mức hài lòng 2 lại ghi nhận tỷ lệ rời bỏ cao nhất với 21.8%, có thể do khách hàng ở mức hài lòng này không đủ hài lòng để duy trì sử dụng nhưng cũng không quá bất mãn ngay từ đầu, dẫn đến quyết định rời bỏ sau một thời gian.
* Mặt khác, nhóm khách hàng hài lòng vừa (mức 3 và 4) cũng có tỷ lệ rời bỏ khá cao, lần lượt là 19.6% và 20.6% .Điều này cho thấy ngân hàng cần chú trọng hơn trong việc cải thiện trải nghiệm cho các nhóm khách hàng hài lòng trung bình, vì họ vẫn chưa đủ trung thành và có khả năng rời bỏ.



Hình 21. Tỷ lệ khách hàng rời bỏ theo số năm khách hàng đã giao dịch với ngân hàng

Biểu đồ này phản ánh tỷ lệ khách hàng rời bỏ dựa trên số năm họ đã giao dịch với ngân hàng (Tenure), giúp ngân hàng hiểu được mối liên hệ giữa thời gian gắn bó của khách hàng với tổ chức và khả năng rời bỏ dịch vụ.

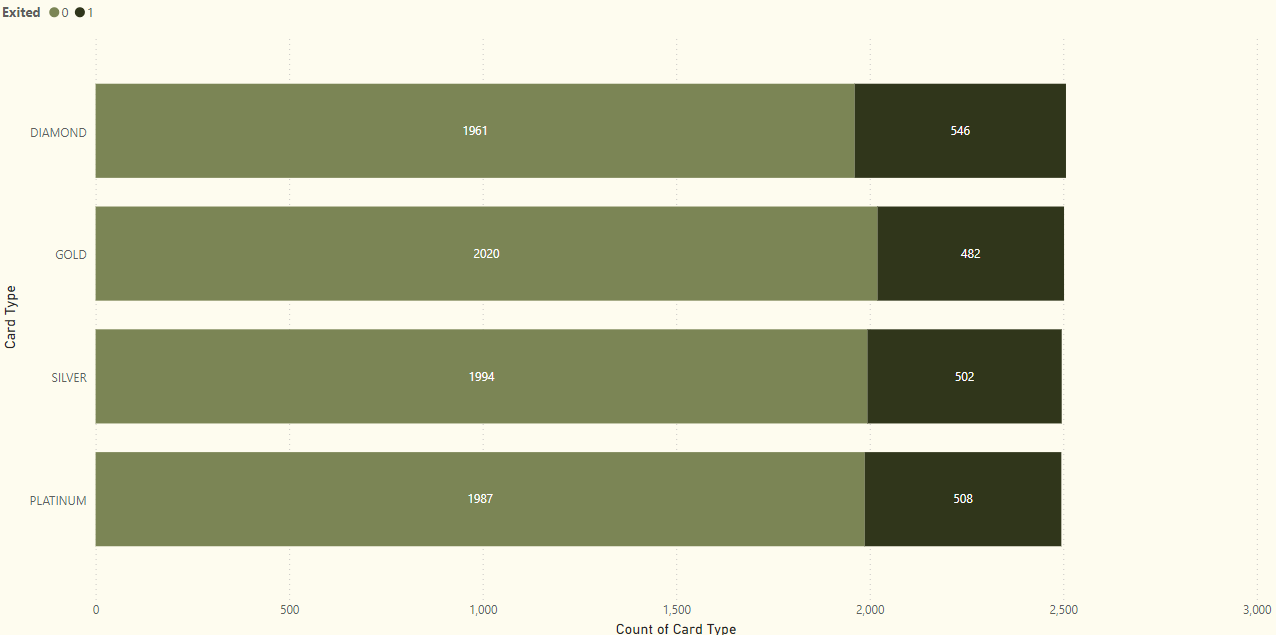
* Tỷ lệ rời bỏ giảm dần theo thời gian sử dụng dịch vụ: Khách hàng mới trong 2 năm đầu có tỷ lệ rời bỏ rất cao, nhóm khách hàng mới có tỷ lệ rời bỏ cao nhất chiếm 77%, nhóm khách hàng đã giao dịch 1 năm là 70.95%, 2 năm là 51.4%. Trong khi khách hàng giao dịch từ 3 năm trở lên có xu hướng rời bỏ thấp hơn (48.6%) nhưng vẫn đáng kể.
* Khách hàng lâu năm vẫn có khả năng rời bỏ: Mặc dù khách hàng có thời gian giao dịch lâu dài 6 -10 năm có tỷ lệ rời bỏ giảm nhẹ, dao động từ 19.6% đến 20%.Họ vẫn có khả năng rời bỏ khi các dịch vụ không còn phù hợp hoặc khi họ bị thu hút bởi các ngân hàng khác.
* Tóm lại, các ngân hàng cần phải chú trọng hơn trong việc cải thiện trải nghiệm và dịch vụ cho khách hàng mới trong giai đoạn 1-2 năm đầu, vì đây là thời điểm khách hàng dễ rời bỏ nhất. Các chiến lược chăm sóc và duy trì lòng trung thành nên được triển khai sớm và liên tục trong giai đoạn này, đồng thời củng cố dịch vụ với nhóm khách hàng lâu năm.



Hình 22. Tỷ lệ khách hàng rời bỏ theo độ tuổi

Biểu đồ này xem xét tỷ lệ rời bỏ giữa các nhóm tuổi khác nhau, giúp ngân hàng hiểu rõ hơn về mối quan hệ giữa độ tuổi và hành vi rời bỏ dịch vụ. Việc phân tích này rất quan trọng để ngân hàng có thể nhận diện các nhóm tuổi có xu hướng rời bỏ cao hơn, từ đó xây dựng các chiến lược duy trì khách hàng phù hợp với từng độ tuổi.

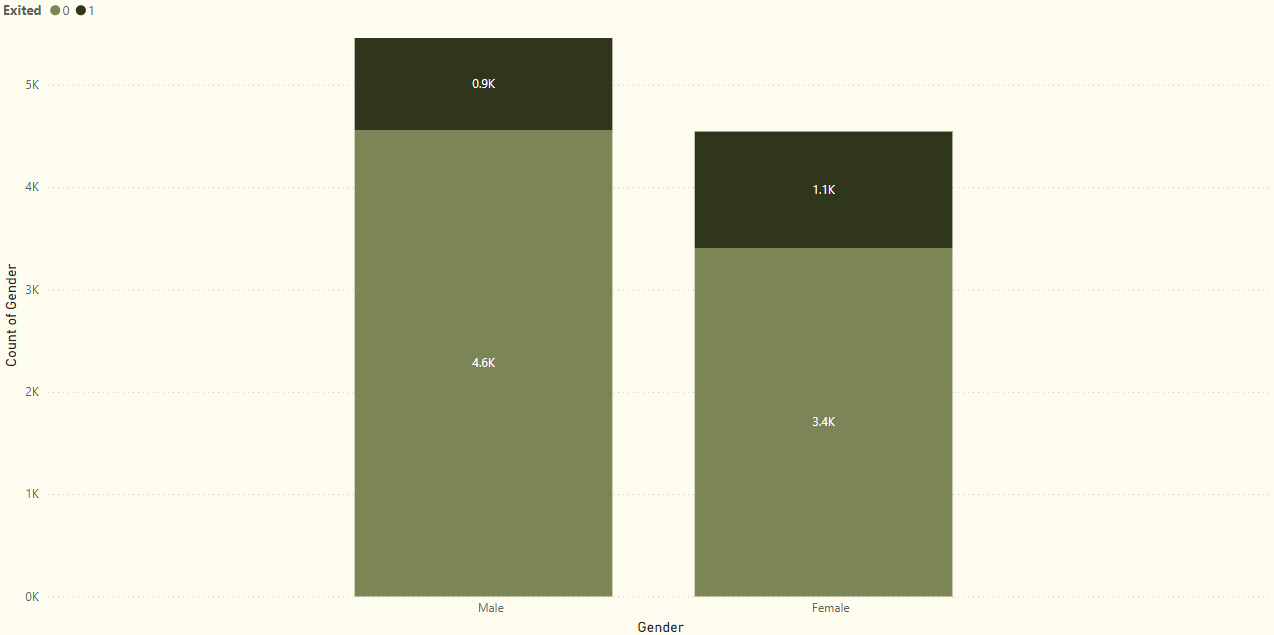
* Tỷ lệ khách hàng rời bỏ theo độ tuổi :Tỷ lệ khách hàng rời bỏ bắt đầu tăng lên từ độ tuổi khoảng 20 và đạt đỉnh ở khoảng độ tuổi 46. Sau độ tuổi này, tỷ lệ khách hàng rời bỏ giảm dần và trở nên rất thấp ở độ tuổi trên 60. Điều này chỉ ra rằng khách hàng trong độ tuổi trung niên (khoảng 40-50 tuổi) có tỷ lệ rời bỏ cao nhất. Đây là giai đoạn mà họ có thể có nhiều thay đổi về nhu cầu tài chính, sự nghiệp, hoặc sự hài lòng với dịch vụ hiện tại.
* Tỷ lệ khách hàng không rời bỏ theo độ tuổi: có xu hướng tăng từ độ tuổi khoảng 20, tăng nhanh trong khoảng 30-40 tuổi và sau đó giảm dần khi tuổi tác tăng lên, vẫn duy trì ổn định ở mức cao hơn so với số lượng rời bỏ sau độ tuổi 60, cho thấy sự ổn định và trung thành của nhóm khách hàng lớn tuổi.
* Tóm lại, nhóm khách hàng trung niên (40-50 tuổi) là nhóm có nguy cơ rời bỏ cao nhất và cần tập trung vào các chiến lược giữ chân khách hàng, bằng cách cung cấp các dịch vụ cá nhân hóa và trải nghiệm tốt hơn. Nhóm khách hàng lớn tuổi (trên 60 tuổi) cho thấy sự trung thành cao và cần duy trì quan hệ tốt với họ. Nhóm khách hàng trẻ tuổi không có nhiều sự biến động, nhưng doanh nghiệp nên xem xét các chiến lược dài hạn để xây dựng lòng trung thành với nhóm này từ sớm.



Hình 23. Tỷ lệ khách hàng rời bỏ theo loại thẻ

Biểu đồ này cho thấy sự phân bổ tỷ lệ rời bỏ giữa các loại thẻ mà khách hàng sử dụng, giúp ngân hàng phân tích mối liên hệ giữa loại thẻ và mức độ gắn bó của khách hàng.

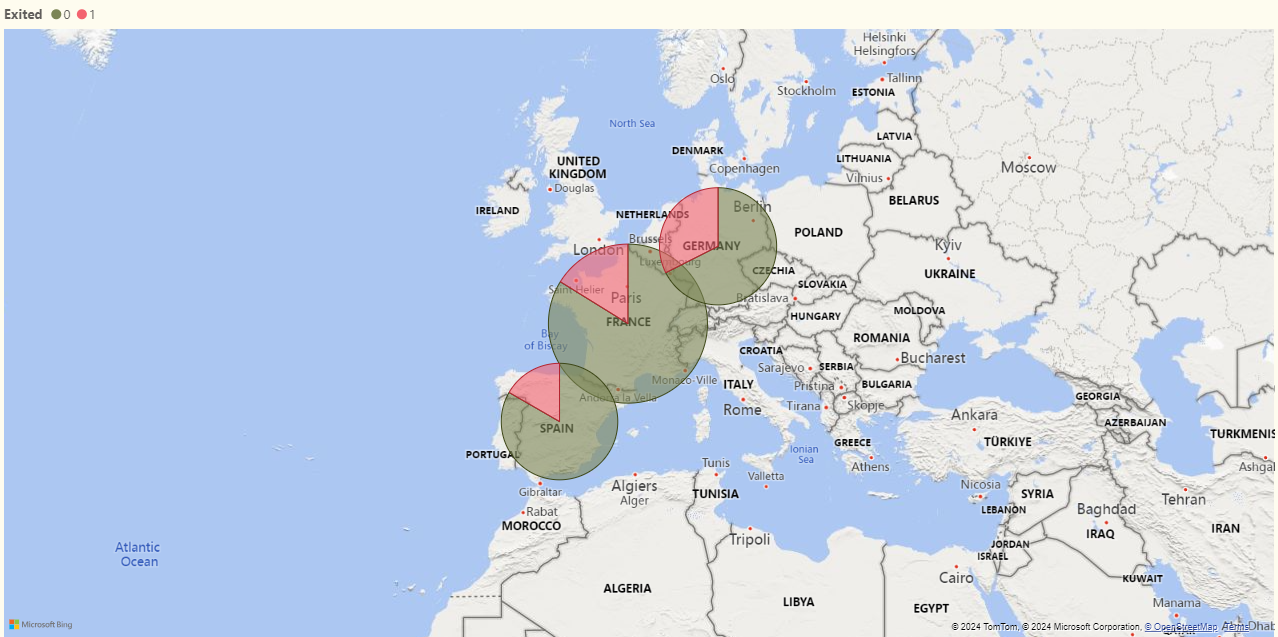
* Thẻ Platinum: có tỷ lệ rời bỏ 20.34% cần được cải thiện. Ngân hàng nên xem xét cung cấp thêm các ưu đãi hoặc cải tiến sản phẩm để giữ chân khách hàng.
* Thẻ Silver: Tương tự như thẻ Platinum, thẻ Silver cũng có tỷ lệ rời bỏ 20.05%. Ngân hàng có thể điều chỉnh các chương trình ưu đãi để tăng cường sự trung thành của khách hàng sử dụng thẻ này.
* Thẻ Gold: có tỷ lệ rời bỏ thấp nhất trong số các loại thẻ (19.25%). Điều này cho thấy rằng khách hàng sử dụng thẻ Gold có xu hướng trung thành hơn. Ngân hàng nên tiếp tục phát triển các sản phẩm và dịch vụ cho nhóm này để củng cố lòng trung thành.
* Thẻ Diamond: có tỷ lệ rời bỏ cao nhất trong số các loại thẻ (21.76%) , điều này cho thấy khách hàng có thể chưa nhận thấy đủ giá trị từ sản phẩm này. Ngân hàng cần xem xét lại các lợi ích và ưu đãi dành cho thẻ Diamond để tăng cường sự hấp dẫn và giữ chân khách hàng.
* Tóm lại, cả 4 loại thẻ đều có tỉ lệ rời bỏ khá cao nên ngân hàng cần chú ý để tìm hiểu nguyên nhân và cải thiện các ưu đãi, nhất là thẻ Diamond. Đồng thời, việc duy trì và phát triển các sản phẩm cho thẻ Gold có thể giúp tăng cường sự trung thành và gắn kết của khách hàng.



Hình 24. Tỷ lệ khách hàng rời bỏ theo Giới tính

Biểu đồ này phản ánh tỷ lệ rời bỏ của khách hàng theo giới tính (nam và nữ), giúp ngân hàng nhận diện rõ xu hướng rời bỏ dịch vụ giữa các nhóm giới tính. Việc phân tích tỷ lệ rời bỏ dựa trên giới tính cung cấp thông tin quan trọng để ngân hàng hiểu rõ hơn về hành vi của khách hàng nam và nữ, từ đó đưa ra các giải pháp phù hợp nhằm giảm thiểu tỷ lệ rời bỏ cho từng giới.

* Nữ giới có tỷ lệ rời bỏ cao hơn đáng kể so với nam giới (24.44% so với 16.36%). Điều này cho thấy rằng nữ giới có xu hướng rời bỏ dịch vụ của ngân hàng nhiều hơn. Ngân hàng cần phân tích và tìm hiểu nguyên nhân khiến khách hàng nữ có tỷ lệ rời bỏ cao hơn, chẳng hạn như có sự không hài lòng về dịch vụ, sản phẩm không phù hợp, hoặc các yếu tố khác.
* Sự khác biệt về tỷ lệ rời bỏ giữa hai giới cho thấy rằng cần có các chính sách chăm sóc khách hàng và chiến lược tiếp thị riêng biệt, nhằm tăng cường sự hài lòng và lòng trung thành của từng nhóm giới tính.



Hình 25. Tỷ lệ khách hàng rời bỏ theo quốc gia

Biểu đồ này phản ánh tỷ lệ rời bỏ của khách hàng theo từng quốc gia, giúp ngân hàng nhận diện những thị trường có tỷ lệ rời bỏ cao và phân tích nguyên nhân tại sao khách hàng trong một số khu vực địa lý có xu hướng rời bỏ nhiều hơn.

* **Tây Ban Nha (16.67%)**: Tỷ lệ rời bỏ tại Tây Ban Nha khá thấp, chỉ 16.67%. Điều này cho thấy phần lớn khách hàng ở Tây Ban Nha cảm thấy hài lòng với dịch vụ của ngân hàng và ít có xu hướng rời bỏ. Ngân hàng có thể tiếp tục duy trì các chương trình dịch vụ và chăm sóc khách hàng tại đây, đồng thời xem xét mở rộng thêm các dịch vụ để giữ chân khách hàng hiện tại và thu hút khách hàng mới.
* **Pháp (16.17%)**: Tương tự như Tây Ban Nha, tỷ lệ khách hàng rời bỏ tại Pháp là 16.17%, một mức khá thấp. Điều này cho thấy thị trường Pháp đang khá ổn định về mặt giữ chân khách hàng. Tuy nhiên, ngân hàng vẫn có thể cải thiện các chương trình chăm sóc khách hàng để tiếp tục giảm thiểu tỷ lệ rời bỏ và nâng cao lòng trung thành của khách hàng tại đây.
* **Đức (32.44%)**: Tỷ lệ rời bỏ ở Đức là 32.44%, cao nhất trong số ba quốc gia. Đây là một tín hiệu đáng lo ngại cho ngân hàng, cho thấy rằng khách hàng ở Đức có xu hướng rời bỏ dịch vụ nhiều hơn so với các quốc gia khác. Ngân hàng cần điều tra và phân tích kỹ lưỡng lý do dẫn đến tỷ lệ rời bỏ cao tại Đức, chẳng hạn như sự cạnh tranh từ các ngân hàng khác, sự không hài lòng về dịch vụ, hoặc nhu cầu chưa được đáp ứng. Các giải pháp có thể bao gồm cải thiện trải nghiệm khách hàng, tăng cường các chương trình ưu đãi, hoặc điều chỉnh sản phẩm để phù hợp hơn với thị trường Đức.
* Tóm lại, ngân hàng cần tập trung vào thị trường Đức để tìm hiểu nguyên nhân của vấn đề và cải thiện dịch vụ nhằm giảm thiểu sự mất khách hàng, đồng thời duy trì các chiến lược hiệu quả tại Tây Ban Nha và Pháp để giữ chân khách hàng.

# **CHƯƠNG VI. TỔNG KẾT**

KẾT LUẬN:  
Trong bài nghiên cứu này, nhóm 4 đã áp dụng ba mô hình hồi quy logistic, cây quyết định, và XGBoost để dự đoán khả năng rời bỏ ngân hàng của khách hàng. Mục tiêu chính là tối ưu hóa việc nhận diện lớp "rời bỏ" để ngân hàng có thể triển khai các biện pháp giữ chân khách hàng hiệu quả hơn.

**1. Hồi quy Logistic:**

* Mô hình này đạt độ chính xác 78%, nhưng precision cho lớp "rời bỏ" chỉ ở mức 0.48, mặc dù recall là 0.75. Điều này có nghĩa là mô hình đã nhận diện được phần lớn khách hàng có khả năng rời bỏ nhưng lại thiếu chính xác.
* Do đó, mô hình hồi quy logistic có thể tạo ra nhiều dự đoán sai cho lớp "rời bỏ," khiến ngân hàng có thể phải tập trung không cần thiết vào các khách hàng không có nguy cơ thực sự.

**2. Cây Quyết Định:**

* Độ chính xác tổng thể của cây quyết định đạt 84%, nhưng F1-score cho lớp "rời bỏ" vẫn thấp, chỉ ở mức 0.52. Precision là 0.69, nhưng recall giảm xuống còn 0.42, cho thấy mô hình này cẩn trọng hơn trong việc xác định khách hàng rời bỏ, nhưng vẫn bỏ lỡ nhiều trường hợp có khả năng rời bỏ.
* Cây quyết định có thể giúp ngân hàng xác định các yếu tố ảnh hưởng đến hành vi của khách hàng, nhưng khả năng dự đoán của nó cho lớp "rời bỏ" chưa đáp ứng kỳ vọng.

**3. XGBoost:**

* XGBoost cho kết quả tốt nhất với độ chính xác tổng thể 84%. Tuy nhiên, precision cho lớp "rời bỏ" đạt 0.65 và recall là 0.57, cho thấy mô hình này có mức độ chính xác và khả năng nhận diện tốt hơn hai mô hình trước, nhưng vẫn còn hạn chế. Với F1-score chỉ đạt 0.61, XGBoost có mức độ cân bằng giữa precision và recall, nhưng hiệu suất chưa thực sự tối ưu đối với lớp "rời bỏ."
* Mặc dù XGBoost là lựa chọn tốt nhất trong ba mô hình, ngân hàng vẫn có thể cân nhắc cải thiện hoặc thử nghiệm các phương pháp khác, như kỹ thuật tăng cường (ensemble) hoặc mô hình phức tạp hơn để nâng cao độ chính xác cho lớp "rời bỏ."

Kết quả của cả ba mô hình cho thấy XGBoost mang lại hiệu suất tốt nhất, nhưng độ chính xác cho lớp "rời bỏ" vẫn chưa thực sự cao, điều này có thể gây ra thách thức trong việc xác định đúng đối tượng khách hàng có nguy cơ rời bỏ. Ngân hàng có thể tiếp tục cải thiện hiệu suất mô hình bằng cách điều chỉnh thêm tham số, thử các kỹ thuật tăng cường dữ liệu hoặc kết hợp nhiều mô hình khác nhau để đạt được độ chính xác cao hơn. Tóm lại, trong ngắn hạn, XGBoost có thể sử dụng như một giải pháp dự đoán với hiệu suất khá, nhưng ngân hàng cần xem xét cải tiến mô hình để đạt hiệu quả tốt hơn trong việc giữ chân khách hàng.

Một số thông tin tiềm ẩn rút ra được:

- Nhóm khách hàng trung niên (40-50 tuổi) là nhóm có nguy cơ rời bỏ cao nhất và cần tập trung vào các chiến lược giữ chân khách hàng, bằng cách cung cấp các dịch vụ cá nhân hóa và trải nghiệm tốt hơn. Nhóm khách hàng lớn tuổi (trên 60 tuổi) cho thấy sự trung thành cao. Các ngân hàng nên xem xét các dịch vụ tài chính và chương trình ưu đãi hướng đến khách hàng lớn tuổi, giúp duy trì sự hài lòng và gắn bó của nhóm khách hàng này.

- Khách hàng từ Đức có tỷ lệ rời bỏ cao hơn so với các khu vực khác. Điều này có thể phản ánh các yếu tố văn hóa, sự khác biệt về yêu cầu dịch vụ, hoặc sự không hài lòng chung của khách hàng từ khu vực này đối với các sản phẩm hoặc dịch vụ của ngân hàng. Ngân hàng nên xem xét các yếu tố địa phương và cá nhân hóa dịch vụ để phù hợp hơn với nhu cầu của từng khu vực.

- Khách hàng sử dụng nhiều sản phẩm (3 hoặc 4 sản phẩm) có xu hướng rời bỏ cao hơn. Điều này có thể cho thấy rằng việc sử dụng quá nhiều sản phẩm có thể dẫn đến sự quá tải hoặc cảm giác không hài lòng. Ngân hàng có thể cần phân tích sâu hơn các sản phẩm khách hàng đang sử dụng để xác định xem có sản phẩm nào cần cải thiện không, hoặc điều chỉnh chiến lược bán chéo (cross-selling) để đảm bảo rằng khách hàng thực sự thấy hữu ích với các sản phẩm bổ sung.

- Nữ giới có tỷ lệ rời bỏ cao hơn đáng kể so với nam giới (24.44% so với 16.36%). Sự khác biệt về tỷ lệ rời bỏ giữa hai giới cho thấy rằng cần có các chính sách chăm sóc khách hàng và chiến lược tiếp thị riêng biệt, nhằm tăng cường sự hài lòng và lòng trung thành của từng nhóm giới tính.