Báo cáo môn AI-DS project nhóm 1

Đề tài: Bank Churners Classification

**Thành viên và phân công công việc:**

Cao Đức Duy: Data preprocessing, Modeling.

Trần Phi Ngọc: EDA, phân tích thuật toán.

Trần Thành Trung: Business Understanding.

Nguyễn Hải Anh: Data Understanding.

Nguyễn Tiến Đạt: Model Deployment.

**I. Business Understanding:**

**Hiện trạng:**

Sự rời bỏ của khách hàng là mối quan tâm lớn đối với các ngân hàng và tổ chức tài chính, vì việc giữ chân khách hàng hiện tại thường tiết kiệm chi phí hơn so với việc có được những khách hàng mới. Nhiều yếu tố có thể góp phần khiến khách hàng rời bỏ, bao gồm dịch vụ kém, cạnh tranh gia tăng, nhu cầu và sở thích của khách hàng thay đổi. Để giảm tình trạng rời bỏ khách hàng, các ngân hàng và tổ chức tài chính cần hiểu nhu cầu của khách hàng và cung cấp trải nghiệm đặc biệt cho khách hàng nhằm thúc đẩy lòng trung thành và giảm thiểu tình trạng rời bỏ khách hàng. Theo như một số điều tra gần đây có:

* 69% khách hàng có ý định rời bỏ ngân hàng mình đang sử dụng vì gặp phải dịch vụ kém
* 56% khách hàng sắp rời bỏ cho biết ngân hàng của họ đã không nỗ lực để giữ chân họ
* 80% khách hàng hiện nay coi trải nghiệm của họ với một công ty cũng quan trọng như chính sản phẩm mà công ty đó cung cấp.

Có thể thấy, khách hàng sẽ có xu hướng rời khỏi ngân hàng mình đang theo dùng nếu như gặp phải một số trải nghiệm không vừa lòng. Một khách hàng có tiếp tục sử dụng lại dịch vụ của ngân hàng hay không sẽ gây ảnh hưởng lớn đến nhiều mặt của một ngân hàng như lợi nhuận, khả năng cạnh tranh, thông tin dữ liệu,... Vậy đâu là những

thiệt hại cũng như cơ hội mất đi khi khách hàng ngừng sử dụng các dịch vụ ở một ngân hàng?

* Mất lợi nhuận: Khách hàng sử dụng chính là nguồn thu lớn đối với ngân hàng thông qua các khoản phí dịch vụ, lãi suất, và các sản phẩm và dịch vụ khác. Khách hàng rời đi đồng nghĩa việc ngân hàng sẽ mất đi lợi nhuận và doanh thu liên quan.
* Mất sự tin tưởng và danh tiếng: Nếu một khách hàng quyết định rời bỏ ngân hàng, có thể đây là dấu hiệu cho thấy họ không hài lòng với trải nghiệm hoặc chất lượng dịch vụ của ngân hàng. Điều này có thể ảnh hưởng đến danh tiếng của ngân hàng và làm mất đi sự tin tưởng của các khách hàng hiện tại và tiềm năng.
* Mất cạnh tranh: Khi một khách hàng chuyển sang một ngân hàng khác, ngân hàng sẽ mất đi khách hàng đó cũng như mất cơ hội để cung cấp các dịch vụ và sản phẩm khác cho họ. Điều này có thể làm cho ngân hàng trở nên kém cạnh tranh và mất thị phần trong ngành.
* Mất tiềm năng tăng trưởng: Khách hàng hiện tại có thể có tiềm năng tăng trưởng trong tương lai thông qua việc mở rộng mối quan hệ và sử dụng thêm dịch vụ từ ngân hàng.
* Tìm kiếm khách hàng mới: Khi một khách hàng rời bỏ ngân hàng, ngân hàng sẽ phải đầu tư thêm vào việc tìm kiếm và thu hút khách hàng mới. Điều này có thể liên quan đến chi phí tiếp thị, quảng cáo, và các hoạt động khác để thu hút khách hàng mới và xây dựng mối quan hệ mới.
* Mất thông tin và dữ liệu khách hàng: Khi khách hàng rời đi, ngân hàng có thể mất đi thông tin và dữ liệu quan trọng về khách hàng đó, gây ảnh hưởng đến việc nghiên cứu thị trường, phân tích khách hàng, và việc cung cấp các dịch vụ tương lai dựa trên dữ liệu đó.

**Mục tiêu**: Dựa vào trải nghiệm và thông tin của khách hàng để đưa ra dự đoán xem họ có tiếp tục sử dụng dịch vụ hay không, từ đó có cơ sở để cải thiện dịch vụ.

**Tại sao lại phải ứng dụng mô hình AI?**

Việc áp dụng AI đang là xu hướng của các ngành công nghiệp, kinh doanh; sử dụng một mô hình AI để dự đoán việc khách hàng rời bỏ ngân hàng là khả thi và có thể mang lại nhiều lợi ích cho ngân hàng như độ chính xác, sự tối ưu…

* Dự đoán chính xác: Mô hình AI có khả năng phân tích và xử lý lượng lớn dữ liệu từ các nguồn khác nhau. Bằng cách sử dụng các thuật toán học máy và mạng nơ-ron, mô hình AI có thể tìm ra các mẫu, xu hướng và thông tin ẩn trong dữ liệu để dự đoán khả năng khách hàng rời bỏ ngân hàng. Điều này giúp đưa ra dự đoán chính xác hơn và đáng tin cậy.
* Xác định chính xác yếu tố ảnh hưởng: Một mô hình AI phức tạp có thể xác định các yếu tố quan trọng và ảnh hưởng nhất định đến quyết định của khách hàng. Các yếu tố này có thể bao gồm các chỉ số chất lượng dịch vụ, tương tác khách hàng, sản phẩm và giá cả, hoặc các yếu tố khác liên quan đến trải nghiệm khách hàng. Bằng cách hiểu rõ những yếu tố này, ngân hàng có thể tập trung vào việc cải thiện các khía cạnh quan trọng để giữ chân khách hàng.
* Tiết kiệm thời gian và tài nguyên: Việc sử dụng mô hình AI tự động hóa quá trình dự đoán và phân tích dữ liệu, giúp tiết kiệm thời gian và tài nguyên của ngân hàng. Thay vì phải phân tích thủ công và xử lý dữ liệu lớn, mô hình AI có thể làm việc nhanh chóng và hiệu quả, giúp ngân hàng tập trung vào các hoạt động chiến lược và cải thiện chất lượng dịch vụ.

Mặc dù việc áp dụng trí tuệ nhân tạo (AI) để dự đoán việc khách hàng rời bỏ ngân hàng có thể mang lại nhiều lợi ích, nhưng cũng có một số điểm yếu cần được xem xét như: cần đầu tư nhiều tài chính, vấn đề đạo đức hoặc quyền riêng tư của khách hàng.

* Với sự vượt trội về ưu điểm các ưu điểm nêu trên, sử dụng một mô hình AI để dự đoán việc khách hàng rời bỏ ngân hàng là rất khả thi và có cơ hội phát triển lớn trong tương lai.

**Giải pháp cải thiện dịch vụ:**

Từ những phân tích và dự đoán có được, các ngân hàng có thể tìm được đâu là những lý do khiến khách hàng không còn muốn sử dụng dịch vụ của họ từ đó đưa ra chiến lược giữ chân khách hàng. Việc này có thể bao gồm việc tạo ra các chương trình khuyến mãi, ưu đãi cá nhân hóa, cải thiện dịch vụ và tương tác khách hàng, hoặc xây dựng các giải pháp tài chính phù hợp với nhu cầu của khách hàng. Mô hình AI giúp ngân hàng đưa ra các quyết định dựa trên dữ liệu và thông tin chính xác, giúp tối ưu hóa các nỗ lực giữ chân khách hàng.

**II. Data Collection:**

Có một số cách để lấy dữ liệu:

* Xin quyền truy cập vào cơ sở dữ liệu khách hàng của các ngân hàng lớn. Điều này có thể được thực hiện một cách thuận tiện vì chính các ngân hàng là khách hàng của các nhà phân tích dữ liệu.
* Lấy nguồn dữ liệu trong cơ sở dữ liệu mở trực tuyến (có thể không chính xác), dùng để tham khảo, so sánh với cơ sở dữ liệu chính.
* Ngoài những thông tin ngân hàng đang nắm giữ, chúng tôi cũng có thể hợp tác với các công ty tài chính hoặc trung tâm mua sắm để nghiên cứu xu hướng sử dụng các dịch vụ banking của khách hàng.

=> Với mục đích học tập và với sự khó khăn trong việc bảo mật về dữ liệu khách hàng, nhóm chúng tôi chọn lấy nguồn dữ liệu trong cơ sở dữ liệu mở trực tuyến (Kaggle) để thực hiện dự án.

**III. DATA UNDERSTANDING:**

**A. Giới thiệu về bộ dữ liệu:**

- Bộ dữ liệu "BankChurners" chứa thông tin về khách hàng của một ngân hàng và trạng thái rời bỏ ngân hàng.

- Bộ dữ liệu gồm 21 cột, bao gồm CLIENTNUM, Customer\_Age, Gender, Dependent\_count, Education\_Level, Marital\_Status, Income\_Category, Card\_Category, Months\_on\_book, Total\_Relationship\_Count, Months\_Inactive\_12\_mon, Contacts\_Count\_12\_mon, Credit\_Limit, Total\_Revolving\_Bal, Avg\_Open\_To\_Buy, Total\_Amt\_Chng\_Q4\_Q1, Total\_Trans\_Amt, Total\_Trans\_Ct, Total\_Ct\_Chng\_Q4\_Q1 và Avg\_Utilization\_Ratio, với cột Attrition\_Flag là cột mục tiêu.

**Khái niệm của các features chính:**

1. **CLIENTNUM** : Chứa số duy nhất đại diện cho mỗi khách hàng trong dữ liệu
2. **Atrition\_Flag** **:** Cho biết khách hàng đó có tiếp tục sử dụng dịch vụ không
3. **Customer Age** : Tuổi của khách hàng
4. **Gender** : Giới tính
5. **Dependent\_Count** : Số lượng người mà một cá nhân hoặc gia đình cần phải nuôi sống hoặc chăm sóc
6. **Education\_Level**: Trình độ học vấn
7. **Martial\_Status**: Tình trạng hôn nhân
8. **Income\_Category**: Mức độ thu nhập
9. **Card\_Category**: Loại thẻ
10. **Months\_on\_books**: Thời gian hoạt động với ngân hàng
11. **Total\_Relationship\_Count**: Số lượng dịch vụ mà khách hàng đang dùng
12. **Months\_Inactive\_12** : Thời gian khách hàng đã không hoạt dịch hay tương tác với ngân hàng
13. **Contacts\_Count\_12\_mon**: Số lần mà khách hàng đã liên hệ với ngân hàng
14. **Total\_Revolving\_Bal:** Tín dụng xoay vòng
15. **Credit\_Limit**: Hạn mức tín dụng của khách hàng
16. **Avg\_Open\_To\_Buy**: Số hạn mức tín dụng còn lại của khách hàng mỗi tháng (Trung bình 12 tháng)
17. **Total\_Amt\_Chng\_Q4\_Q1**: Tỷ lệ thay đổi giữa số tiền giao dịch trong quý 4 và quý 1
18. **Total\_Trans\_Amt**: Tổng số tiền khách hàng giao dịch trong 12 tháng
19. **Total\_Trans\_Ct**: Tổng số lần giao dịch của khách hàng trong 12 tháng
20. **Total\_Ct\_Chng\_Q4\_Q1**: Tỷ lệ thay đổi giữa số lần giao dịch trong quý 4 và quý 1
21. **Avg\_Utilization\_Ratio**: Tỷ lệ sử dụng thẻ tín dụng trung bình của khách hàng

**B. Mô tả sơ lược về các features:**

Dựa trên dữ liệu mà ta đã thu thập được, tất cả các features (features) trong file dữ liệu "BankChurners" đều có tiềm năng ảnh hưởng đến quyết định rời bỏ ngân hàng của khách hàng. Dưới đây là khái niệm và mối quan hệ của từng features với target.

**CLIENTNUM**: Gồm các số đại diện cho mỗi khách hàng trong dữ liệu.

**Attrition\_Flag:** Đây là cột quan trọng để xác định liệu khách hàng có rời bỏ ngân hàng (churn) hay không. Cột này có thể chứa các giá trị như "Existing Customer" (khách hàng hiện tại) hoặc "Attrited Customer" (khách hàng đã rời bỏ). **Đây là cột mục tiêu (target) trong việc xây dựng mô hình dự đoán.**

1. Customer\_Age: Tuổi của khách hàng có thể ảnh hưởng đến quyết định rời bỏ ngân hàng. Ví dụ, một nhóm khách hàng trẻ có thể có xu hướng chuyển đổi ngân hàng thường xuyên hơn so với nhóm khách hàng lớn tuổi.

2. Gender: Giới tính có thể đóng vai trò trong quyết định rời bỏ ngân hàng. Ví dụ, có thể có sự khác biệt giữa nam và nữ trong mức độ sử dụng các dịch vụ ngân hàng hoặc mức độ hài lòng đối với các sản phẩm và dịch vụ.

3. Dependent\_count: Số lượng người phụ thuộc tài chính có thể ảnh hưởng đến khả năng tài chính của khách hàng và quyết định rời bỏ ngân hàng. Ví dụ, một khách hàng có nhiều người phụ thuộc có thể gặp khó khăn trong việc duy trì một tài khoản ngân hàng và có xu hướng chuyển đổi ngân hàng.

4. Education\_Level: Trình độ học vấn của khách hàng có thể ảnh hưởng đến tư duy tài chính và quyết định rời bỏ ngân hàng. Ví dụ, khách hàng có trình độ học vấn cao có thể có sự hiểu biết và kiến thức tài chính tốt hơn, từ đó làm giảm khả năng rời bỏ ngân hàng.

5. Marital\_Status: Tình trạng hôn nhân có thể ảnh hưởng đến yếu tố tài chính và quyết định rời bỏ ngân hàng. Ví dụ, một khách hàng độc thân có thể có sự linh hoạt tài chính lớn hơn và có xu hướng thay đổi ngân hàng dễ dàng hơn so với một khách hàng đã kết hôn.

6. Income\_Category: Thu nhập của khách hàng có thể ảnh hưởng đến khả năng tài chính và quyết định rời bỏ ngân hàng. Ví dụ, khách hàng có thu nhập thấp có thể gặp khó khăn trong việc duy trì các tài khoản ngân hàng và có xu hướng chuyển đổi ngân hàng nhanh hơn.

7. Card\_Category: Loại thẻ tín dụng mà khách hàng sử dụng có thể ảnh hưởng đến mức độ ưu đãi, giới hạn tín dụng và yếu tố tài chính khác. Các khách hàng sử dụng các loại thẻ cao cấp có thể có một mức độ tương tác và cam kết cao hơn với ngân hàng, giúp giảm khả năng rời bỏ.

8. Months\_on\_book: Số tháng khách hàng đã sử dụng dịch vụ ngân hàng. Đây có thể là một chỉ số cho thấy mức độ trung thành và cam kết của khách hàng với ngân hàng.

9. Total\_Relationship\_Count: Tổng số mối quan hệ tài chính của khách hàng. Tổng số mối quan hệ tài chính bao gồm: tài khoản tiết kiệm, thẻ tín dụng,các khoản vay mượn và một số dịch vụ tài chính mà khách hàng sử dụng với ngân hàng. Ta có thể đánh giá được mức độ tương tác và cam kết của khách hàng với ngân hàng.

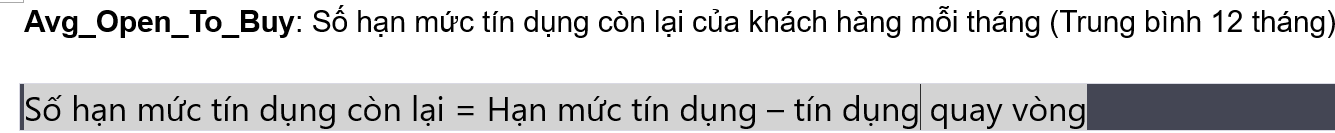
10. Months\_Inactive\_12\_mon: Số tháng mà khách hàng không hoạt động trong vòng 12 tháng. Số này có thể cho thấy mức độ tương tác hoặc sự quan tâm của khách hàng đối với ngân hàng. Khách hàng không hoạt động trong một khoảng thời gian dài có thể có nguy cơ cao hơn về việc rời bỏ ngân hàng.

11. Contacts\_Count\_12\_mon: Số lần trao đổi giữa ngân hàng với khách hàng trong vòng 12 tháng. Số này có thể cho thấy mức độ tương tác và quan tâm của ngân hàng đối với khách hàng. Sự tương tác thường xuyên có thể giữ khách hàng ở lại và giảm nguy cơ rời bỏ.

12. Credit\_Limit: Giới hạn tín dụng của khách hàng. Giới hạn tín dụng có thể phản ánh khả năng tài chính và sự tin tưởng của ngân hàng đối với khách hàng. Mức giới hạn tín dụng cao có thể tạo thuận lợi và độ an toàn cho khách hàng, giảm nguy cơ rời bỏ.

13. Total\_Revolving\_Bal: Tổng số dư quay vòng của khách hàng. Số dư quay vòng thể hiện mức độ sử dụng thẻ tín dụng của khách hàng. Khách hàng có mức sử dụng thẻ cao có thể tạo ra doanh thu cho ngân hàng và có xu hướng giữ chân khách hàng.

14. Avg\_Open\_To\_Buy: Số tiền trung bình có sẵn để sử dụng trên giới hạn tín dụng. Số này có thể liên quan đến khả năng tài chính và sự quản lý tài chính của khách hàng. Khách hàng có số tiền trung bình lớn có thể cho thấy khả năng thanh toán và tư duy tài chính tốt hơn.



15. Total\_Amt\_Chng\_Q4\_Q1: Tổng thay đổi số tiền giao dịch từ quý 4 đến quý 1. Số này có thể phản ánh sự thay đổi trong hành vi và mức độ tương tác của khách hàng trong khoảng thời gian đó. Mức tăng trong số tiền giao dịch có thể cho thấy mức độ quan tâm và tương tác tốt hơn.

16. Total\_Trans\_Amt: Tổng số tiền giao dịch của khách hàng. Số này có thể phản ánh mức độ tương tác và sử dụng dịch vụ của khách hàng. Khách hàng có số tiền giao dịch lớn có thể có cam kết cao hơn và khả năng rời bỏ thấp hơn.

17. Total\_Trans\_Ct: Tổng số lần giao dịch của khách hàng. Số này có thể liên quan đến mức độ tương tác và sử dụng dịch vụ của khách hàng. Khách hàng có số lần giao dịch nhiều có thể có mức độ cam kết cao hơn và khả năng rời bỏ thấp hơn.

18. Total\_Ct\_Chng\_Q4\_Q1: Tổng thay đổi số lần giao dịch từ quý 4 đến quý 1. Số này có thể cho thấy sự thay đổi trong hành vi giao dịch và tương tác của khách hàng trong khoảng thời gian đó.

19. Avg\_Utilization\_Ratio: Tỷ lệ sử dụng trung bình của thẻ tín dụng. Tỷ lệ này thể hiện mức độ sử dụng thẻ tín dụng của khách hàng so với giới hạn tín dụng. Khách hàng có tỷ lệ sử dụng trung bình cao có thể có mức độ sử dụng thẻ cao và khả năng rời bỏ thấp hơn

**Nhận xét:**

Thông tin từ các cột dữ liệu này có thể cung cấp cái nhìn tổng quan về hành vi, tương tác và thông tin tài chính của khách hàng. Đây là những thông tin quan trọng để hiểu rõ hơn về yếu tố ảnh hưởng đến việc rời bỏ ngân hàng và xây dựng mô hình.

Các features trong bộ dữ liệu "BankChurners.csv" như tuổi, giới tính, số người phụ thuộc, trình độ học vấn, tình trạng hôn nhân, phân loại thu nhập và loại thẻ tín dụng có thể ảnh hưởng đến quyết định rời bỏ ngân hàng của khách hàng. Thông tin về thời gian sử dụng dịch vụ, mối quan hệ khách hàng, tương tác với ngân hàng và thông tin tài chính khác cũng cung cấp cái nhìn quan trọng về hành vi và yếu tố ảnh hưởng đến hành vi của khách hàng.

Tuy nhiên, để xây dựng một mô hình dự đoán chính xác về Bank churners, cần tiếp tục phân tích, đánh giá bao gồm việc khám phá sự tương quan, xử lý dữ liệu thiếu, khử nhiễu, và chọn phương pháp phù hợp để xây dựng mô hình.

**C. Mối quan hệ giữa các feature:**

Dưới đây là một số mối quan hệ có thể tồn tại giữa các features trong bộ dữ liệu "Bank Churners":

1. **Mối quan hệ giữa giới tính (Gender) và trình độ học vấn (Education\_Level) và nguồn thu (Income\_Category)**: Mối quan hệ giữa giới tính và trình độ học vấn của khách hàng. Ví dụ, có thể có tỷ lệ cao hơn của khách hàng nam có trình độ học vấn cao hơn so với khách hàng nữ điều đó suy ra được khả năng tạo ra tài sản của nam giới trung bình cao hơn phụ nữ.
2. **Income\_Category (Phân loại thu nhập) và Credit\_Limit (Giới hạn tín dụng):** Có thể có mối quan hệ giữa mức thu nhập và giới hạn tín dụng. Khách hàng có thu nhập cao có thể được cấp một giới hạn tín dụng cao hơn, trong khi khách hàng có thu nhập thấp có thể có giới hạn tín dụng thấp hơn.
3. **Total\_Trans\_Amt (Tổng số giao dịch) và Total\_Trans\_Ct (Tổng số giao dịch):** Khả năng cao sẽ có mối quan hệ mạnh giữa số lượng giao dịch và tổng số giao dịch. Ví dụ, khách hàng có số lượng giao dịch nhiều có thể có tổng số giao dịch cao hơn.
4. **Avg\_Utilization\_Ratio (Tỷ lệ sử dụng trung bình) và Total\_Revolving\_Bal (Tổng số dư quay vòng**): Có thể có mối quan hệ giữa tỷ lệ sử dụng trung bình của thẻ tín dụng và tổng số dư quay vòng. Tỷ lệ sử dụng trung bình cao có thể gợi ý rằng khách hàng sử dụng thẻ tín dụng nhiều và có tổng số dư quay vòng lớn.
5. **Education\_Level (Trình độ học vấn) và Marital\_Status (Tình trạng hôn nhân):** Có thể có mối quan hệ giữa trình độ học vấn và tình trạng hôn nhân. Ví dụ, nhóm khách hàng có trình độ học vấn cao có thể có tỷ lệ kết hôn cao hơn so với nhóm khách hàng có trình độ học vấn thấp.

**IV. METHODOLOGY:**

Trong bước này, chúng tôi sẽ tiến hành các bước nhằm xây dựng 1 machine learning model và đánh giá. Chi tiết từng phần và cách thực hiện sẽ có trong file jupyter notebook đính kèm.

**I. Data Wrangling:**

1. Drop cột Clientnum.

2. Data Formatting: Không cần format vì dữ liệu được lấy từ 1 nguồn duy nhất.

3. Handling missing value: Sử dụng KNN impute cho 3 feature chứa giá trị thiếu Education\_Level, Martial\_Status, Income\_Category.

4. Encoding các categorical features => numerical feature:

Binary encoding: Attrition\_Flag, Gender.

Ordinal encoding: Card\_Category, Education\_Level, Income\_Category.

One-hot encoding: Martial\_Status.

5. Data types modify.

**II. Exploratory Data Analysis:**

**1. Pairplot:**

Vì đây là bài toán classification, với target là category variable 0/1, đầu tiên chúng tôi sẽ vẽ pairplot (scatterplot) để xét sự phân phối của các feature dựa trên target.

**A group of blue and orange dots

Description automatically generated**

**2. Mối tương quan giữa các tính năng định lượng - mục tiêu sử dụng tương quan Point-biserial**

Để xét rõ hơn về loại này thì chúng tôi sẽ sử dụng point-biserial correlation coefficient để tính correlation giữa quantitive variable và a dichotomous variable (category 0/1). Xét về công thức toán học, point-biserial correlation chính là Pearson's correlation khi target chỉ nhận 2 giá trị 0 và 1.

A graph of numbers and symbols

Description automatically generatedBiểu đồ thể hiện độ tương quan với target

**3**. **Boxplot**

Ta tiếp tục vẽ box plot cho 3 features có correlation với target nhất để nhìn rõ hơn phân bố của chúng.

A graph of a number of different colored boxes

Description automatically generated

**4. Correlation between quantitive features (and quantitive features - encoding categorical feature) using Pearson's correlation**

**A colorful squares with black and red squares

Description automatically generated**

**5. Categorical variables' correlation using Chi-square test**

**A chart with different colored squares

Description automatically generated**

**III. Modeling**

**1. Handling outliers and noise**

Outliers là những điểm mà giá trị của nó có sự khác biệt đáng kể với phần lớn các điểm dữ liệu khác trong bộ dữ liệu. Outliers có thể là valid data points hoặc cũng có thể là noise. Tuy phần lớn outliers là noise nhưng nó cũng chứa các thông tin thú vị.

Noise là các điểm data phân phối không theo xu hướng chung, mislabeled, hoặc có thể coi là lỗi. Noise xuất hiện trong những feature có độ tương quan cao với target sẽ ảnh hưởng lớn đến bài toán phân loại. Việc giảm nhiễu sẽ có lợi ích lớn cho performance của model sau này.

Mục tiêu chính ở đây của chúng ta là giảm nhiễu qua việc giảm ảnh hưởng của outliers (đánh bừa còn hơn bỏ sót).

A diagram of a box plot

Description automatically generatedSố lượng outliers còn lại sau khi xử lý

Độ tương quan của 3 feature quan trọng nhất sau khi loại bỏ outliers:

A chart with different colors

Description automatically generated

**2. Feature Selection**

* Loại bỏ các feature có độ tương quan thấp với target: Customer\_Age, Dependent\_count, Months\_on\_book, Avg\_Open\_To\_Buy, Education\_Level, Card\_Category.
* Loại bỏ các feature có độ tương quan cao với nhau: Gender (Income\_category), Total\_Trans\_Amt (Total\_Trans\_Ct).

**3. Resampling**

Positive class chỉ chiếm 15% trong bộ data nên bộ dữ liệu được sử dụng không cân bằng.

Bộ dữ liệu không cân bằng sẽ gây ảnh hưởng đến machine learning models, vì nó có thể bias về lớp đa số và không thể hiện tốt trên lớp thiểu số.

* Tiến hành resampling sử dụng phương pháp SMOTE cho tập train.

Tỉ lệ giữa positive class với toàn bộ là 50%

A black screen with white text

Description automatically generated

Ở trong project, chúng tôi sẽ sử dụng 3 model cho vấn đề về classification: KNN, SVM và Logistics Regression và lựa chọn sử dụng mô hình dựa vào performance của model và ưu điểm trong thực tế.

Về cách xây dựng, chúng tôi sẽ sử dụng phương pháp GridSearch và Cross Validation để tìm ra hyperparameters tối ưu cho mô hình.

**KNN Model:**

Chúng tôi sẽ bắt đầu với mô hình đơn giản nhất.

Kết quả của KNN Model:

1. Confusion matrix:

A comparison of a bar chart

Description automatically generated

2. Precision, Recall và F1-score:

A black background with white numbers

Description automatically generated

**SVM Model:**

Kết quả của SVM Model:

1. Confusion matrix:

**A comparison of a graph

Description automatically generated**

2. Precision, Recall và F1-score:

**A black background with white numbers

Description automatically generated**

**Logistics Regression Model:**

Kết quả của Logistics Regression Model:

1. Confusion matrix:

**A comparison of a graph

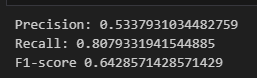
Description automatically generated**

2. ROC:

**A graph of a function

Description automatically generated**

3. Precision, Recall và F1-score:



**Nhận xét:**

Dựa vào f1-score của cả 3 mô hình, ta thấy không có sự khác biệt quá lớn. Nhưng khi xét trên bài toán thực tế, Logistics regression có những ưu điểm đáng cân nhắc để sử dụng.

Tính chất khiến cho Logistics regression vượt lên so với các mô hình xét ở trên là ở khả năng trả về giá trị là xác suất xảy ra của sự kiện positive. Từ đó, có thể tùy ý chỉnh threshold theo nhu cầu và ưu tiên của người sử dụng.

* Sử dụng mô hình Logistics Regression để áp dụng vào trong thực tế.

**Trong bài toán phân loại Bank churners, việc chỉnh threshold của Logistic regression có thể mang lại các ưu điểm sau:**

* Điều chỉnh độ nhạy và độ chính xác: Threshold quyết định ngưỡng để phân loại một mẫu vào lớp positive hoặc negative. Bằng cách thay đổi threshold, ta có thể điều chỉnh độ nhạy và độ chính xác của mô hình. Nếu muốn tăng độ nhạy (tìm nhiều churners hơn), ta có thể giảm threshold. Ngược lại, nếu muốn tăng độ chính xác (giảm số lượng false positive), ta có thể tăng threshold.
* Đáp ứng yêu cầu kinh doanh: Việc chỉnh threshold cho phép tùy chỉnh mô hình để đáp ứng yêu cầu kinh doanh cụ thể. Ví dụ, nếu việc giữ chân khách hàng hiện tại là ưu tiên hàng đầu, có thể điều chỉnh threshold để tìm ra nhiều churners potential nhất có thể. Ngược lại, nếu việc giảm số lượng false positive (khách hàng bị phân loại nhầm là churners) là quan trọng, người sử dụng mô hình có thể tăng threshold để tăng độ chính xác.

**V. Phân tích Logistics Regression**

1. **Giới thiệu**

Thuật toán Logistic Regression là một thuật toán học máy được sử dụng cho bài toán phân loại. Nó dựa trên khái niệm của hồi quy tuyến tính và sử dụng hàm sigmoid để ước lượng xác suất xảy ra của một sự kiện.

1. **Nguyên lý hoạt động**
2. **Ý tưởng**

Logistic Regression được sử dụng để dự đoán xác suất xảy ra của một sự kiện thuộc hai lớp phân loại. Trong dự án này là dự đoán xác suất khách hàng có rời bỏ ngân hàng hay không.

Trong Logistic Regression, sử dụng hàm Sigmoid để biến đổi kết quả của một phép tính tuyến tính thành một giá trị xác suất nằm trong khoảng từ 0 đến 1.

1. **Hàm Sigmoid**

Hàm sigmoid  được định nghĩa là:



Trong đó, z là tổng trọng số tuyến tính của các biến đầu vào:



 là các biến đầu vào,  là các trọng số tương ứng. Hàm sigmoid có đặc điểm là giá trị đầu ra nằm trong khoảng từ 0 đến 1. Nó chuyển đổi đầu vào tuyến tính thành một giá trị xác suất.

**3. Mô hình dự đoán**

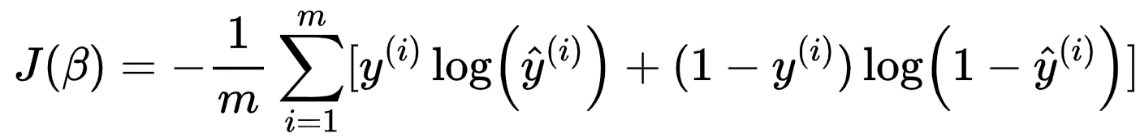
Để dự đoán xác suất xảy ra của sự kiện, chúng ta sử dụng công thức sau:



Trong đó  là xác suất sự kiện xảy ra khi có các giá trị  của biến đầu vào. Chúng ta có thể chọn một ngưỡng xác suất (threshold) để phân loại dựa trên kết quả dự đoán. Ví dụ, nếu xác suất dự đoán lớn hơn ngưỡng, chúng ta có thể dự đoán lớp 1, ngược lại, nếu xác suất nhỏ hơn ngưỡng, chúng ta dự đoán lớp 0.

**4. Hàm mất mát (Loss function)**

Trong logistic regression, chúng ta sử dụng hàm mất mát Cross-Entropy (hay Log-Loss) để đánh giá độ chính xác của mô hình và điều chỉnh các trọng số  để tối ưu hóa mô hình. Hàm mất mát Cross-Entropy được định nghĩa như sau:



Trong đó,  là số lượng điểm dữ liệu,  là nhãn thực tế của điểm dữ liệu thứ i,  là xác suất dự đoán tương ứng với điểm dữ liệu thứ i. Mục tiêu là tìm các trọng số  sao cho hàm mất mát này đạt giá trị nhỏ nhất, thông qua các phương pháp tối ưu hóa như gradient descent.

**5. Gradient Descent**

Quá trình huấn luyện logistic regression bao gồm tối ưu hóa các trọng số  giảm thiểu hàm mất mát. Phương pháp phổ biến để tối ưu hóa là sử dụng gradient descent, trong đó chúng ta cập nhật các trọng số theo đạo hàm của hàm mất mát.

Để tối ưu hóa, chúng ta cần tính đạo hàm riêng của hàm mất mát theo từng trọng số  Sau đó, cập nhật các trọng số bằng cách di chuyển theo hướng ngược với gradient của hàm mất mát, với một tốc độ học (learning rate) đã được định trước.

1. **Training**

Quá trình huấn luyện logistic regression bao gồm các bước sau:

**1. Chuẩn bị dữ liệu:** Tiền xử lý dữ liệu, bao gồm việc chuẩn hóa và mã hóa các biến đầu vào nếu cần thiết.

**2. Khởi tạo trọng số:** Khởi tạo các trọng số  ban đầu

**3. Tính toán dự đoán:** Tính toán giá trị dự đoán sử dụng hàm sigmoid với trọng số hiện tại

**4. Tính toán gradient:** Tính toán gradient của hàm mất mát để điều chỉnh các trọng số.

**5. Cập nhật trọng số:** Sử dụng thuật toán tối ưu hóa (ví dụ: gradient descent) để cập nhật các trọng số  theo hướng giảm mất mát.

**6. Lặp lại quá trình:** Lặp lại các bước trên cho đến khi đạt được tiêu chí dừng (ví dụ: đạt được số lần lặp tối đa, hội tụ mất mát đủ nhỏ).

1. **Đánh giá và dự đoán**

Sau khi hoàn thành quá trình huấn luyện, chúng ta có thể đánh giá mô hình logistic regression bằng cách sử dụng các phép đo đánh giá như độ chính xác (accuracy), precision, Confusion matrix, F1-score và đường cong ROC. Bên cạnh đó, chúng ta có thể sử dụng mô hình đã huấn luyện để dự đoán nhãn cho các điểm dữ liệu mới.

1. **Ưu điểm và hạn chế**

**Ưu điểm:**

* **Dễ triển khai và hiểu:** Logistic Regression là một trong những thuật toán đơn giản nhất trong học máy. Nó dễ triển khai và thực hiện, và không yêu cầu nhiều kiến thức chuyên sâu về toán học hay lập trình.
* **Tính toán nhanh chóng:** Vì Logistic Regression là một thuật toán tuyến tính, việc tính toán và huấn luyện mô hình nhanh chóng, đặc biệt là với tập dữ liệu có kích thước lớn.
* **Dự đoán xác suất:** Logistic Regression cho phép dự đoán xác suất xảy ra của một sự kiện, giúp hiểu rõ hơn về mức độ chắc chắn của kết quả dự đoán.
* **Khả năng xử lý dữ liệu phân loại:** Logistic Regression thích hợp cho các bài toán phân loại nhị phân (hai lớp) và phân loại đa lớp (nhiều lớp).
* **Không yêu cầu dữ liệu chuẩn hóa:** Logistic Regression không nhạy cảm với dữ liệu không được chuẩn hóa, nhưng việc chuẩn hóa dữ liệu vẫn có thể cải thiện hiệu suất của mô hình.

**Hạn chế**:

* **Khả năng xử lý tương quan và phi tuyến:** Logistic Regression không xử lý tốt với các mối quan hệ tương quan cao giữa các biến đầu vào và các tương tác phi tuyến. Điều này có thể dẫn đến hiệu suất dự đoán không tốt trong trường hợp các biến đầu vào có sự phụ thuộc phức tạp lẫn nhau.
* **Dễ bị ảnh hưởng bởi nhiễu:** Logistic Regression dễ bị ảnh hưởng bởi các điểm ngoại lai hoặc các điểm dữ liệu không cân bằng, đặc biệt trong các bài toán có tỷ lệ mất cân bằng giữa các lớp.
* **Phụ thuộc vào đặc trưng tuyến tính:** Logistic Regression yêu cầu các biến đầu vào có mối quan hệ tuyến tính với biến mục tiêu. Trong trường hợp các biến không có mối quan hệ tuyến tính, kết quả dự đoán có thể không chính xác.
* **Không thích hợp cho dữ liệu phức tạp:** Logistic Regression không phù hợp cho các bài toán có dữ liệu phức tạp và không thể phân loại tuyến tính.
* **Hạn chế với dữ liệu lớn và không gian đặc trưng lớn:** Khi số lượng đặc trưng lớn, Logistic Regression có thể gặp khó khăn trong việc xử lý dữ liệu và huấn luyện mô hình.

VI. **MODEL DEPLOYMENT**

1. Triển khai:

Dưới đây là một số phương pháp để triển khai mô hình dự đoán chuyển đổi khách hàng trong dự án này:

* API: Xây dựng một API (Application Programming Interface) để cung cấp các dự đoán từ mô hình cho các ứng dụng khác sử dụng. API này có thể được tích hợp vào các ứng dụng hiện có hoặc bên thứ ba, cho phép gửi dữ liệu và nhận lại dự đoán chuyển đổi khách hàng.
* Tích hợp với hệ thống hiện có: Nếu công ty đã có hệ thống CRM, mô hình dự đoán chuyển đổi khách hàng có thể được tích hợp vào hệ thống này để tự động phát hiện khách hàng có nguy cơ chuyển đổi và thực hiện các biện pháp để giữ chân khách hàng.
* Tích hợp mô hình vào hệ thống POS (Point of Sale): Tích hợp mô hình dự đoán chuyển đổi vào hệ thống POS của cửa hàng. Khi khách hàng thực hiện giao dịch tại cửa hàng, dữ liệu giao dịch sẽ được gửi đến mô hình để dự đoán khả năng chuyển đổi.

Tuy nhiên, trong quá trình triển khai, cần lưu ý các vấn đề về quyền riêng tư và bảo mật dữ liệu của khách hàng. Đảm bảo tuân thủ các quy định và chính sách liên quan đến bảo vệ thông tin cá nhân của khách hàng.

2. Giám sát và cập nhật mô hình:

Sau khi triển khai mô hình, quan trọng để giám sát hiệu suất và chất lượng của mô hình. Theo dõi các chỉ số đánh giá và thực hiện các cập nhật hoặc tái huấn luyện mô hình khi cần thiết để duy trì tính chính xác và hiệu quả của mô hình.