ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT

KHOA TOÁN KINH TẾ

BÀI TIỂU LUẬN CHƯƠNG III

MÔN HỌC: THỐNG KÊ TRONG QUẢN TRỊ KINH DOANH VÀ MARKETING

ĐỀ TÀI : SỬ DỤNG MÔ HÌNH LOGISTIC ĐÁNH GIÁ KHẢ NĂNG RỜI BỎ TÍN DỤNG CỦA KHÁCH HÀNG TRONG LĨNH VỰC NGÂN HÀNG

Giảng viên hướng dẫn: Nguyễn Đình Uông
Sinh viên thực hiện: Vũ Thị Hạnh Dung MSSV: K204131871

I. Giới thiệu **3**
II. Tổng quan các nghiên cứu đi trước **4**
III. Phương pháp nghiên cứu **5**
3.1. Hồi quy Logistic - mô hình Logit 5
3.2. Dữ liệu 6
3.2.1 Mô tả dữ liệu 6
3.2.2 Đề xuất mô hình nghiên cứu 6
IV. Kết quả nghiên cứu **9**
4.1 Thống kê mô tả 9
4.2 Mô hình hồi quy Logistic 15
4.2.1 Quá trình training và testing mô hình logit 15
4.2.2 Kết quả mô hình logit 15
V. Kết luận **21**
Tài liệu tham khảo **23**

I. Giới thiệu

Mặc dù các công ty lớn cố gắng thu hút khách hàng mới, nhưng họ cũng muốn giữ chân khách hàng cũ của mình. Do đó, phân tích khách hàng rời bỏ là rất quan trọng để xác định khách hàng cũ mà không bị mất đi và phát triển các sản phẩm mới và đưa ra các quyết định chiến lược mới để giữ chân khách hàng.Xác định khách hàng rời bỏ trong ngân hàng sẽ giúp quản lý phân loại những người có khả năng rời bỏ sớm và tập trung vào khách hàng sử dụng chương trình khuyến mãi, cũng như cung cấp thông tin về những yếu tố nào nên được xem xét khi giữ chân khách hàng.

Chuyện khách hàng rời bỏ là một vấn đề nghiêm trọng trong thời đại đầy cạnh tranh ngày nay khi thị trường quá đầy đủ. Theo nhiều nghiên cứu, chi phí để tìm kiếm khách hàng mới chỉ bằng 1/5 chi phí giữ chân khách hàng cũ (Athanassopoulos 2000). Vì vậy, các công ty thường ưu tiên giữ chân khách hàng hiện tại hơn là tìm kiếm khách hàng mới và áp dụng các chính sách nhắm tới mục tiêu đó. Một trong những yếu tố quan trọng nhất trong các chiến lược giảm thiểu hoặc ngăn chặn khách hàng rời bỏ là dữ liệu hành vi khách hàng trong cơ sở khách hàng hiện tại (Ganesh et al. 2000). Do đó, trong khuôn khổ kế hoạch chiến lược nhắm tới giảm thiểu việc khách hàng rời bỏ, việc khám phá và khai thác các khách hàng có nguy cơ cao muốn rời bỏ tổ chức, hay dự đoán việc khách hàng rời bỏ, là rất quan trọng (Blattberg et al. 2008).

Các tổ chức phát hành thẻ tín dụng thường dựa vào các đặc điểm về nhân khẩu học của khách hàng để xây dựng chiến lược khách hàng mục tiêu.Đây thường là những yếu tố thuộc đặc điểm của khách hàng như: giới tính, độ tuổi, tình trạng hôn nhân, trình độ, nghề nghiệp, thu nhập. Những yếu tố này có những ảnh hưởng nhất định đến những đánh giá của khách hàng đối với sản phẩm, dịch vụ, từ đó dẫn đến quyết định có tiêu dùng hay không. Trong đó, tập trung chủ yếu là các cá nhân thường có thu nhập cao, trình độ tương đối và có nghề nghiệp được đánh giá cao trong giai đoạn phát triển của thị trường thẻ tín dụng.

II.Tổng quan các nghiên cứu đi trước

Mutanen (2006) đã thực hiện một nghiên cứu về khả năng khách hàng rời bỏ ngân hàng bán lẻ dựa trên mô hình hồi quy logistic. Naveen et al. (2009) đã tiến hành một nghiên cứu chi tiết về khách hàng rời bỏ sử dụng thẻ tín dụng với các kỹ thuật khai thác dữ liệu. Bilal (2016) sử dụng giới tính, độ tuổi, thu nhập trung bình hàng tháng, tình trạng tiêu dùng (nghỉ hưu, sinh viên, đã đi làm, chưa đi làm), và việc khách hàng có sử dụng hai sản phẩm ngân hàng trở lên như các biến kiểm soát trong mô hình mạng nơ-ron. Theo Bilal, khách hàng sử dụng nhiều sản phẩm ngân hàng có khả năng rời bỏ thấp hơn. Keramati et al. (2016) sử dụng mô hình cây quyết định (DT) để điều tra khách hàng rời bỏ trong ngân hàng điện tử (ngân hàng internet, ngân hàng điện thoại, ngân hàng di động, máy ATM). Họ phát hiện ra rằng sự bất mãn của khách hàng (thời gian tương tác với khách hàng, số lần khiếu nại của khách hàng), việc sử dụng dịch vụ (tổng số lần sử dụng và số tiền giao dịch), và các biến số nhân khẩu học (tuổi, giới tính, tình trạng việc làm, trình độ giáo dục) ảnh hưởng đến khả năng khách hàng rời bỏ.

Xinyu Miao và Haoran Wang (2022) đã kiểm tra độ quan trọng của các đặc trưng trong tập dữ liệu và phát hiện rằng tổng số tiền giao dịch trong 12 tháng qua, tổng số lần giao dịch trong 12 tháng qua và số dư xoay vòng trên thẻ tín dụng có ảnh hưởng đáng kể đến dự đoán của mô hình. Điều này cho thấy rằng khách hàng sử dụng thẻ tín dụng càng thường xuyên, họ càng ít có khả năng rời bỏ, do đó, các quản lý ngân hàng có thể điều chỉnh dịch vụ thẻ tín dụng dựa trên điều này để chống lại việc khách hàng rời bỏ và tăng tỷ lệ giữ chân khách hàng. Việc tăng tỷ lệ giữ chân mang lại sự tăng trưởng lợi nhuận lớn hơn. Bằng cách sử dụng mô hình này, họ có đủ thời gian trước khi thực hiện các hành động để giữ chân khách hàng, ví dụ như thông qua các chương trình khuyến mãi, cung cấp phiếu giảm giá để khuyến khích người sử dụng thẻ tín dụng và khuôn khổ hành vi sử dụng của họ

Bằng cách sử dụng hồi quy logistic và cây quyết định, Abbas và cộng sự đã tìm thấy rằng thời gian quan hệ với khách hàng, độ tuổi khách hàng, giới tính và số giao dịch ngân hàng di động ảnh hưởng đến việc khách hàng rời bỏ

- Khoảng trống nghiên cứu

Các nghiên cứu khác đã phát triển các mô hình để dự đoán khách hàng rời bỏ mà không sử dụng các biến quan trọng. Để khắc phục vấn đề này, đã được đề xuất rằng các biến phân loại được hợp nhất thành một biến. Do đó, khoảng trống nghiên cứu này đã thúc đẩy các tác giả tìm kiếm một mô hình phù hợp để dự đoán khách hàng rời bỏ.

III. Phương pháp nghiên cứu

3.1. Hồi quy Logistic - mô hình Logit

Thông thường, biến phụ thuộc là rời rạc có thể nhận hai hoặc nhiều giá trị có thể (ví dụ 1: Nam, 0: Nữ). Mô hình hồi quy logistic là mô hình hồi quy được sử dụng thường xuyên nhất để phân tích các dữ liệu này (Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X., 2013). Hồi quy Logistic cũng được sử dụng trong nhiều lĩnh vực như y tế: dự đoán bệnh tim mạch (Adeshina, A. F., & Adeyemo, O. O., 2017), giáo dục: xác định yếu tố dẫn đến sinh viên năm nhất bỏ học (Gyeongmi Lee & 9 Jihyun Kim, 2020), du lịch: dự đoán khách hàng sẽ đặt phòng hay không (Hieu Nguyen, Anh Nguyen & Thinh Nguyen, 2020),...

Training Set và Testing Set trong mô hình hồi quy Logistic

- Training Set (Tập huấn luyện)

Tập huấn luyện (training set) là tập dữ liệu được sử dụng để huấn luyện mô hình. Các thuật toán học máy sẽ học các mô hình từ tập huấn luyện này. Việc học sẽ khác nhau tùy thuộc vào thuật toán và mô hình sử dụng. Ví dụ, khi sử dụng môt hình Hồi quy tuyến tính (Linear Regression), các điểm trong tập huấn luyện được sử dụng để tìm ra hàm số hay đường phù hợp nhất mô tả quan hệ giữa đầu vào và đầu ra của tập dữ liệu huấn luyện bằng cách sử dụng một số phương pháp tối ưu hóa như công thức nghiệm ở bài trước hoặc các thuật toán tối ưu gần đúng như gradient descent hay stochastic gradient descent. Trong thực tế, tập dữ liệu huấn luyện thường bao gồm các cặp vectơ đầu vào và vectơ đầu ra tương ứng, trong đó vector đầu ra thường được gọi là nhãn (label hoặc target). Các thuật toán nói chung sẽ tìm cách tối ưu sai số dự đoán trên tập huấn luyện này đến mức đủ tốt. Trong trường hợp overfitting sai số dự đoán của mô hình trên tập huấn luyện có thể rất thấp, thậm chí = 0%.

- Testing Set (Tập kiểm thử)

Mục tiêu của machine learning là tạo ra những mô hình có khả năng tổng quát hóa để dự đoán tốt trên cả dữ liệu chưa thấy bao giờ (nằm ngoài tập huấn luyện), do đó, để biết một thuật toán hay mô hình có tốt hay không thì sau khi được huấn luyện, mô hình cần được đánh giá hiệu quả thông qua bộ dữ liệu kiểm thử (testing set). Bộ dữ liệu này được sử dụng để tính độ chính xác hoặc sai số của mô hình dự đoán đã được huấn luyện. Chúng ta biết nhãn thực của mọi điểm trong tập hợp dữ liệu kiểm thử này, nhưng chúng ta sẽ tạm thời giả vờ như không biết và đưa các giá trị đầu vào của tập vào mô hình dự đoán để nhận kết quả dự đoán đầu ra. Sau đó chúng ta có thể nhìn vào các nhãn thực và so sánh nó với kết quả dự đoán của các đầu vào tương ứng này và xem liệu mô hình có dự đoán đúng hay không. Việc tính tổng trung bình của toàn bộ các lỗi này chúng ta có thể tính toán được lỗi dự đoán trên tập kiểm thử.

Các lỗi dự đoán này được đánh giá thông qua rất nhiều chỉ số khác nhau như độ chính xác (precision), độ hồi tưởng (recall), F1-Score, RMSE, MAE,…Các chỉ số này sẽ được trình bày chi tiết hơn trong bài tiếp theo. Một lưu ý là các chỉ số đo mức độ hiệu quả của mô hình này trên tập kiểm thử có thể khác với các lossfunction hay objective function sử dụng để tối ưu hóa mô hình trên tập huấn luyện. Nghĩa là quá trình kiểm thử và quá trình huấn luyện là hoàn toàn độc lập với nhau, cả về bộ dữ liệu lẫn cách thức so sánh chỉ số.
Tập dữ liệu kiểm thử tốt là một tập dữ liệu độc lập với tập dữ liệu huấn luyện (để ngoài và không được tham gia vào quá trình huấn luyện), nhưng tuân theo cùng một phân phối xác suất như tập dữ liệu huấn luyện. Điều này giúp cho việc đánh giá không bị thiên vị. Nếu một mô hình phù hợp với trainining set nhưng lại sai khác trên testing set, thì việc rất có khả năng nó bị overfitting. Ngược lại, sai số không quá nhiều thì thường chúng là một mô hình phù hợp.

3.2. Dữ liệu

3.2.1 Mô tả dữ liệu

Dữ liệu bao gồm các thông tin của 10000 khách hàng đã và đang sử dụng thẻ tín dụng (Credit Card Customers) từ một ngân hàng tại Mỹ. Tập dữ liệu bao gồm 21 biến như tuổi, giới tính, học vấn, thu nhập, trạng thái hôn nhân, loại thẻ tín dụng và các chỉ số liên quan đến hoạt động sử dụng thẻ của khách hàng.
Đầu tiên, tập dữ liệu đã được chia thành các biến phân loại và biến liên tục làm biến độc lập và một biến phụ thuộc (khách hàng rời bỏ). Tiếp theo, tôi phân tích các biến bằng các số liệu thống kê khác nhau bao gồm giá trị tối thiểu, giá trị tối đa, phương sai, độ lệch chuẩn, kiểm định chi bình phương (cho biến phân loại) và phân tích tương quan (cho biến liên tục). Phân tích ban đầu đã cho thấy mối quan hệ tuyến tính giữa các biến không tồn tại; do đó, một mô hình phi tuyến được áp dụng để phát triển mô hình dự đoán khách hàng rời bỏ.

3.2.2 Đề xuất mô hình nghiên cứu

Các biến và kỳ vọng về dấu của các biến độc lập

Biến phụ thuộc:

Attrition_Flag là biến đại diện cho việc rời bỏ thẻ tín dụng của khách hàng.
Attrition_Flag được quy ước bằng 1 nếu khách hàng là ‘Attrited Customer’ và bằng 0 nếu Existing Customer.

Biến giải thích định tính:

- Giới tính (Gender): Đặc điểm về giới tính có ảnh hưởng đến khả năng thiết lập kế hoạch chỉ tiêu tài chính và các mặt hàng tiêu dùng, do đó có ảnh hưởng đến xu hướng sở hữu và hành vi sử dụng thẻ tín dụng trong tiêu dùng. Kết quả của nghiên cứu của Judith L. Zaichkowsky và Yunchuan Liu cho thấy rằng giới tính có ảnh

hưởng đáng kể đến quyết định sử dụng thẻ tín dụng và nam giới có xu hướng sử dụng thẻ tín dụng nhiều hơn nữ giới.

- Trình độ học vấn (Education_Level): có ảnh hưởng đến thái độ và cách đánh giá đối với thẻ tín dụng và các tiện ích dịch vụ trong thanh thẻ tín dụng. Do đó, có ảnh hưởng đến khả năng sử dụng thẻ tín dụng để thanh toán trong tiêu dùng. Theo Kim et al. (2015), khách hàng có trình độ học vấn cao có xu hướng sử dụng thẻ tín dụng nhiều hơn so với những khách hàng có trình độ học vấn thấp.
- Tình trạng hôn nhân (Marital_Status):có ảnh hưởng đến kế hoạch chi tiêu trong tiêu dùng, đặc điểm các loại sản phẩm tiêu dùng, do đó gián tiếp ảnh hưởng đến việc sở hữu thẻ tín dụng, số lượng thẻ và thói quen sử dụng thẻ tín dụng trong thanh toán. Yang et al. (2020) cũng chỉ ra rằngkhách hàng đã kết hôn có xu hướng sử dụng thẻ tín dụng nhiều hơn so với những khách hàng độc thân.
- Thu nhập (Income_Category):có ảnh hưởng đến kế hoạch và thói quen chi tiêu trong tiêu dùng của khách hàng, do dó có ảnh hưởng đến việc sở hữu và thoái quan sử dụng thẻ tín dụng trong thanh toán. Chia-Lin Chang (2017) cũng cho thấy rằng những người có thu nhập cao hơn có xu hướng sử dụng thẻ tín dụng nhiều hơn và có khả năng trả nợ thẻ tốt hơn

Biến giải thích định lượng:

- Customer_Age (tuổi của khách hàng): có ảnh hưởng đến thái độ của khách hàng đối với việc sử dụng thể tin dụng với các vấn đề như: việc vay nợ, tính an toàn và tính dễ sử dụng.
- Total_Relationship_Count (số lượng sản phẩm tài chính sử dụng bởi khách hàng) Tổng số sản phẩm tài chính (bao gồm tài khoản tiết kiệm, tài khoản vãng lai và tài khoản đầu tư) mà khách hàng có tại ngân hàng. Nghiên cứu của Học viện khoa học Nga (2016) cho thấy rằng khách hàng có nhiều sản phẩm tài chính tại một ngân hàng có xu hướng ít có khả năng rời đi hơn.
- Total_Amt_Chng_Q4_Q1 (thay đổi về số tiền chi tiêu của khách hàng từ quý 1 đến quý 4)
- Total_Trans_Ct (tổng số giao dịch của khách hàng) : Berkeley (2012) cho thấy rằng khách hàng sử dụng thẻ tín dụng nhiều hơn có xu hướng ít có khả năng rời đi hơn so với những người sử dụng ít hơn. Tuy nhiên, điều này còn phụ thuộc vào độ tuổi, giới tính và thu nhập của khách hàng.
- Total_Ct_Chng_Q4_Q1 (số lần giao dịch từ quý 1 đến quý 4) Theo Liu và cộng sự (2018) biến này có tác động tích cựclên biến phụ thuộc, cụ thể khi số lần thực hiện giao dịch trong tài khoản tín dụng của khách hàng cao khiến cho khả năng khách hàng vẫn tiếp tục sử dụng thẻ tín dụng cao hơn. Ngoài ra, nếu một khách hàng có xu hướng tiêu dùng thẻ tín dụng một cách không có kế hoạch và số tiền sử dụng

nhiều hơn, thì khách hàng này sẽ có khả năng từ bỏ thẻ tín dụng cao hơn (Nguyen Van Tien Nguyen & Tran Thi Hong Tran, 2021) cũng chỉ ra rằng số lần sử dụng thẻ tín dụng có ảnh hưởng đáng kể đến tình trạng từ bỏ thẻ tín dụng của khách hàng. Cụ thể, khách hàng sử dụng thẻ tín dụng ít hơn một lần mỗi tháng có xu hướng từ bỏ thẻ tín dụng nhiều hơn so với những khách hàng sử dụng thẻ tín dụng nhiều hơn một lần mỗi tháng.

	Biến	Kỳ vọng dấu	Paper ủng hộ
Biến phụ thuộc	Attrition_ Flag		
Biến độc lập	Customer _Age		Keramati et al. (2016)
	Gender	(+/-)	Keramati et al.(2016)
	Education_Level	(-)	Keramati et al.(2016)
	Marital_Status	(+)	G. L. Nie, W. Rowe, L. L. Zhang, Y. J. Tian, and Y. Shi
	Income_Category	(+)	Agyapong et al.
	Credit_Limit	(+)	(Chongqi Wu, Lan Wang, 2022
	Total_Amt_Chng_Q4_Q1	(+)	Xinyu Miao và Haoran Wang (2022)
	Total_Trans_Ct	(+)	Xinyu Miao và Haoran Wang (2022)
	Total_Ct_Chng_Q4_Q1	(-)	(Liu, H., Zhou, N., & Wang, T. , 2018)

IV. Kết quả nghiên cứu

4.1 Thống kê mô tả

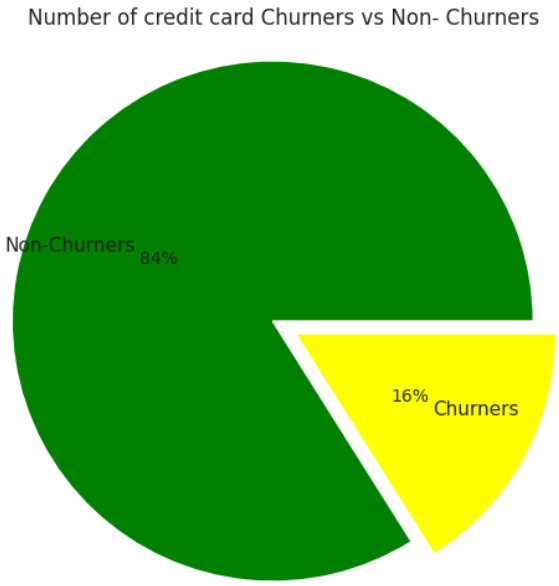
	CLIENTNUM	Customer_Age	Credit_Limit	Total_Amt_Chng_Q4_Q1	Total_Trans_Amt	Total_Trans_Ct	Total_Ct_Chng_Q4_Q1
count	1.012700e+04	10127.000000	10127.000000	10127.000000	10127.000000	10127.000000	10127.000000
mean	7.391776e+08	46.325960	8631.953698	0.759941	4404.086304	64.858695	0.712222
std	3.690378e+07	8.016814	9088.776650	0.219207	3397.129254	23.472570	0.238086
min	7.080821e+08	26.000000	1438.300000	0.000000	510.000000	10.000000	0.000000
25%	7.130368e+08	41.000000	2555.000000	0.631000	2155.500000	45.000000	0.582000
50%	7.179264e+08	46.000000	4549.000000	0.736000	3899.000000	67.000000	0.702000
75%	7.731435e+08	52.000000	11067.500000	0.859000	4741.000000	81.000000	0.818000
max	8.283431e+08	73.000000	34516.000000	3.397000	18484.000000	139.000000	3.714000

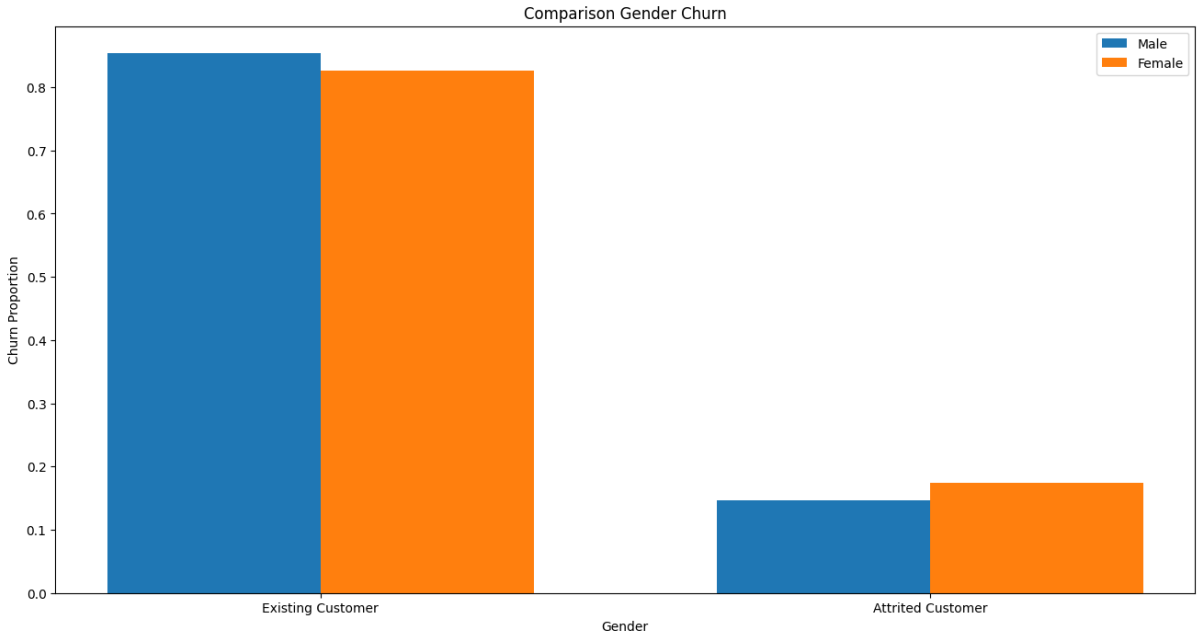
Kết quả thống kê mô tả này bao gồm các giá trị thống kê cơ bản cho 6 biến khác
nhau trong một tập dữ liệu gồm 10.127 quan sát:

- Độ tuổi trung bình của khách hàng là 46.33 và độ tuổi nhỏ nhất và lớn nhất của khách hàng trong tập dữ liệu lần lượt là 26 và 73 với độ lệch chuẩn là 8.02
- Trung bình hạn mức tín dụng là 8631.95 với độ lệch chuẩn là 9088.776. Độ lệch chuẩn của

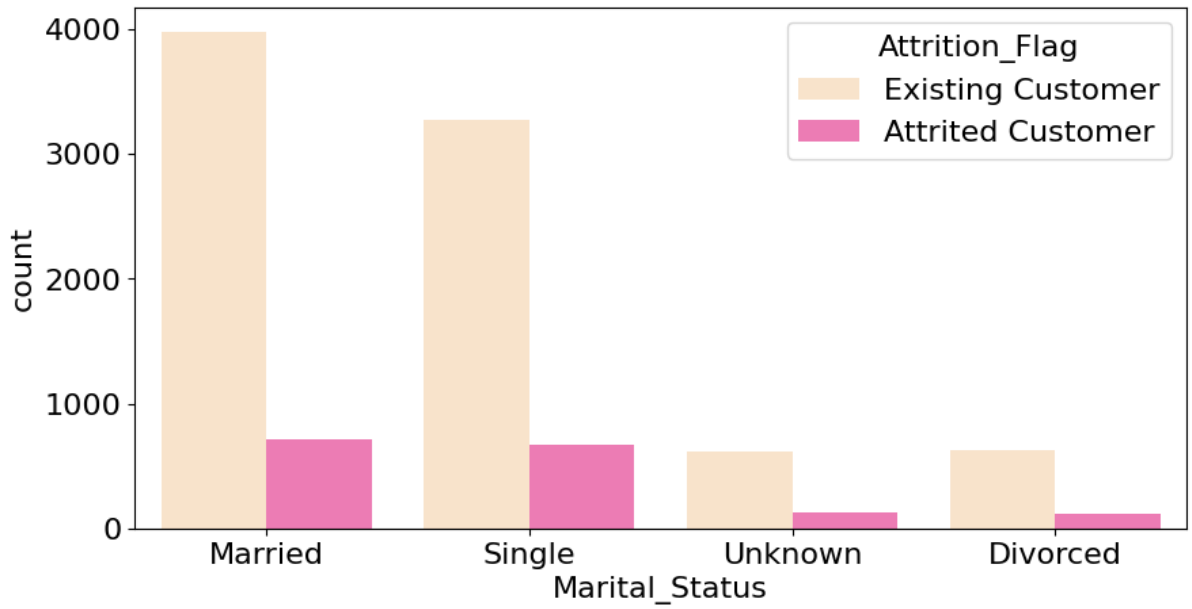
giới hạn tín dụng (Credit_Limit) trong tập dữ liệu này khá cao, cho thấy rằng giới hạn tín dụng của các khách hàng khác nhau rất nhiều và giá trị trung vị của giới hạn tín dụng (Credit_Limit) là 4549 đô la Mỹ, cho thấy rằng nửa số khách hàng có giới hạn tín dụng thấp hơn giá trị này và nửa số khác có giới hạn tín dụng cao hơn giá trị này.

- Trung bình thay đổi về tổng số giao dịch của khách hàng trong 12 tháng gần nhất là 4404.08 lần với độ lệch chuẩn là 3397.12
- Trung bình phần trăm thay đổi số tiền giao dịch giữa quý 4 và quý 1 là 0.76 với độ lệch chuẩn là 0.219
- Trung bình số giao dịch của khách hàng trong vòng 12 tháng gần đây là 64.86 với độ lệch chuẩn là 23.472
- Trung bình phần trăm thay đổi trong số lượng giao dịch của khách hàng giữa quý 4 và quý 1 là 0.71 với độ lệch chuẩn là 0.24

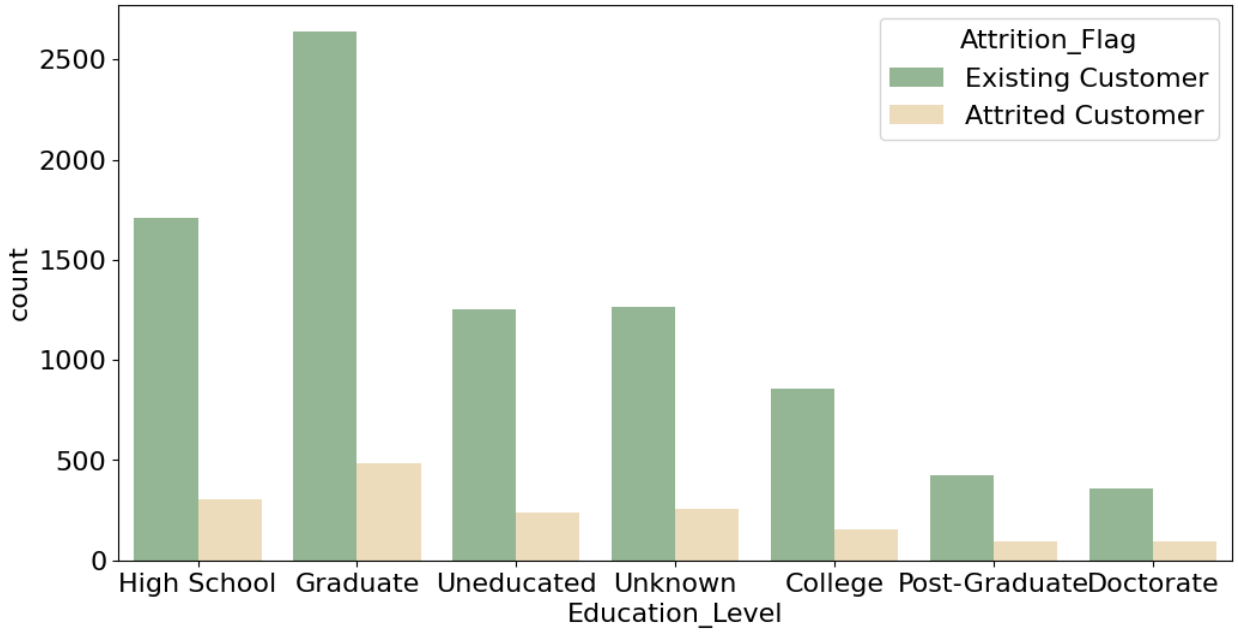




Theo dữ liệu, 16% khách hàng đã rời ngân hàng trong khi 86% ở lại. Đa số khách hàng là độc thân và tỷ lệ nữ cao hơn nam khoảng 3%. Mức thu nhập chung thường thấp hơn 40000$. Hơn 30% người dùng thẻ tín dụng có bằng cử nhân. Gần một nửa số khách hàng đã kết hôn, và số lượng khách hàng độc thân lên đến 40%.



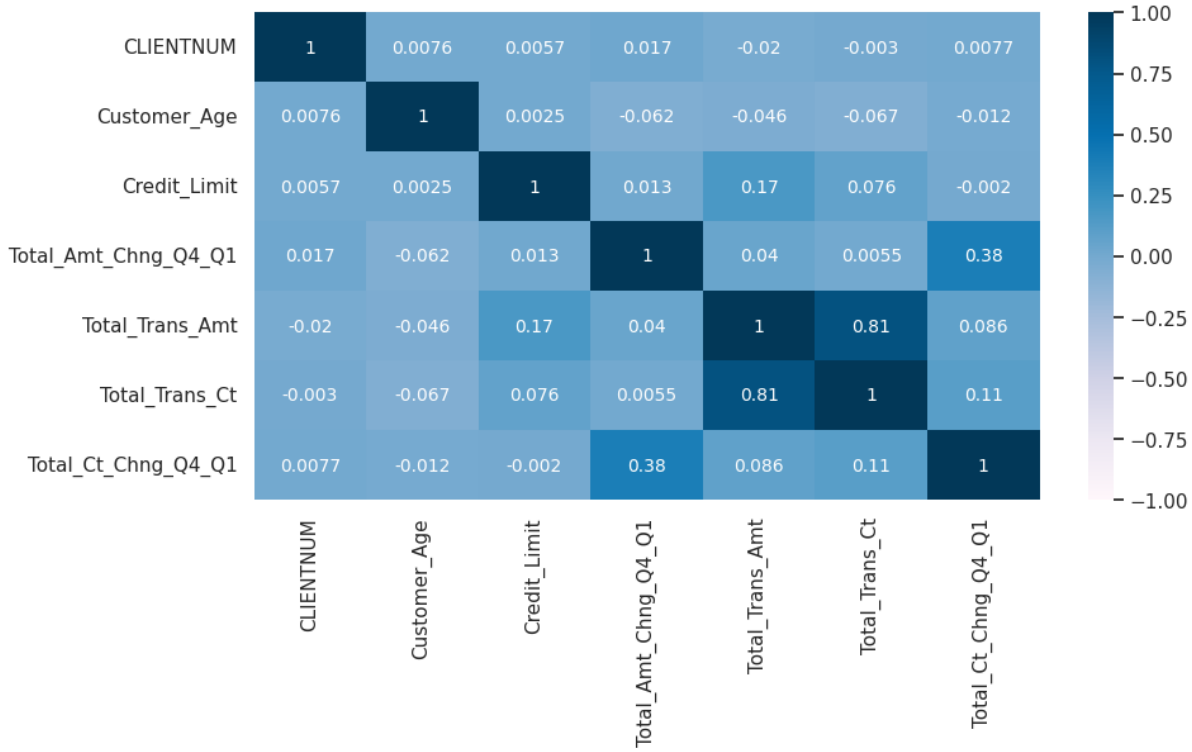
Số lượng khách hàng rời bỏ trong nhóm đã ly dị là thấp nhất, trong khi số lượng khách hàng hiện tại trong nhóm đã kết hôn là cao nhất. Số lượng khách hàng rời bỏ trong nhóm độc thân và số lượng khách hàng hiện tại trong nhóm đã kết hôn là tương đối giống nhau. Số lượng duy trì thẻ tín dụng cao nhất là khi kết hôn do nhu cầu quản lý tiền bạc tăng lên của các cặp vợ chồng.



Số lượng khách hàng Attrited Customer trong mỗi nhóm trình độ học vấn khác nhau khá đồng đều, trong khoảng từ 92 đến 487. Tuy nhiên, số lượng khách hàng Existing

Customer có sự chênh lệch lớn, từ 2641 (trình độ Graduate) xuống còn 356 (trình độ Doctorate). Trình độ Graduate có số lượng khách hàng rời đi cao nhất (487), trong khi đó trình độ Postgraduate có số lượng khách hàng rời đi thấp nhất (92). Tuy nhiên, nếu tính tỉ lệ khách hàng rời đi trong từng nhóm trình độ học vấn thì kết quả có thể khác. Nhóm Unknown là nhóm có số lượng khách hàng rời đi cao thứ hai (256), trong khi đó số lượng khách hàng hiện có lại khá tương đối (1263). Điều này có thể chỉ ra rằng, các khách hàng trong nhóm này có thể không cung cấp đầy đủ thông tin về trình độ học vấn của mình. Từ kết quả này, có thể thấy rằng trình độ học vấn có thể ảnh hưởng đến tình trạng khách hàng. Cần phân tích thêm để hiểu rõ hơn về mối quan hệ giữa các yếu tố này và đưa ra các biện pháp cải thiện tình trạng khách hàng.

Hệ số tương quan



Kiểm định đa cộng tuyến

	VIF	features
0	13.170799	Customer_Age
1	1.900076	Credit_Limit
2	13.119733	Total_Amt_Chng_Q4_Q1
3	7.636121	Total_Trans_Ct
4	11.491021	Total_Ct_Chng_Q4_Q1

Sau khi loại bỏ những biến có vif lớn hơn 10 ta được:

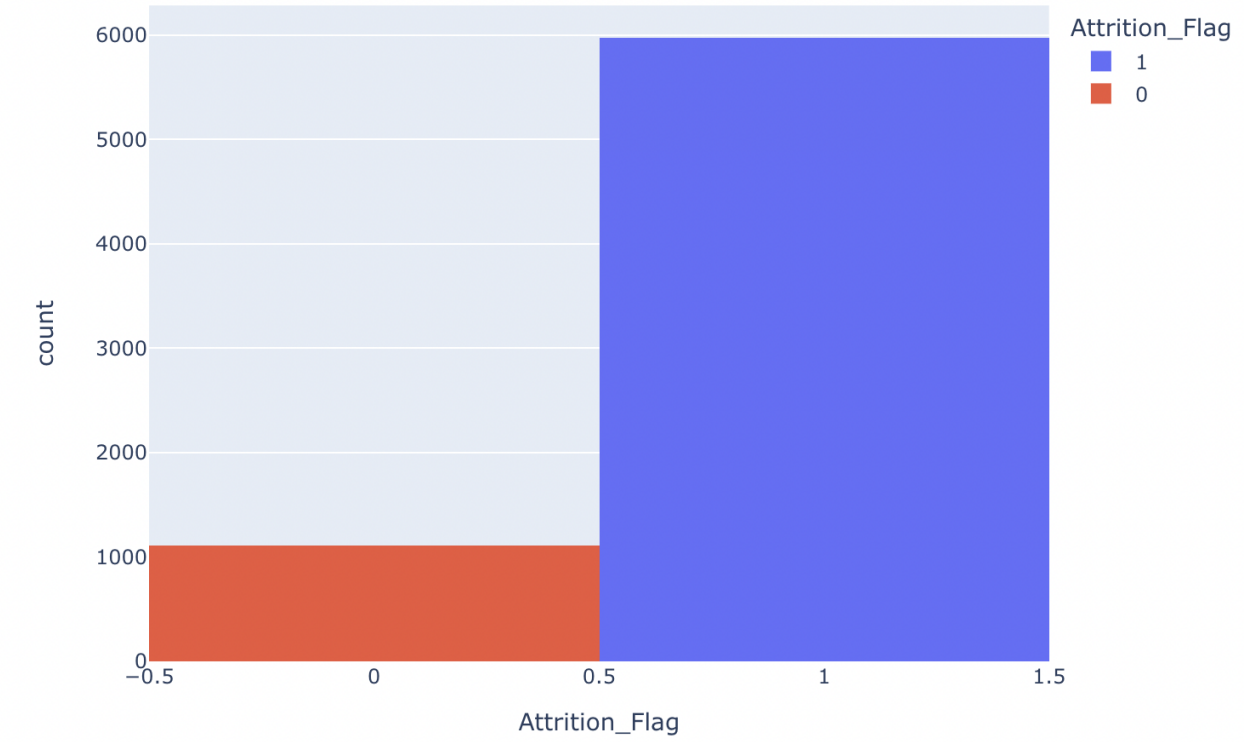
	VIF	features
0	1.798428	Credit_Limit
1	1.798428	Total_Trans_Ct

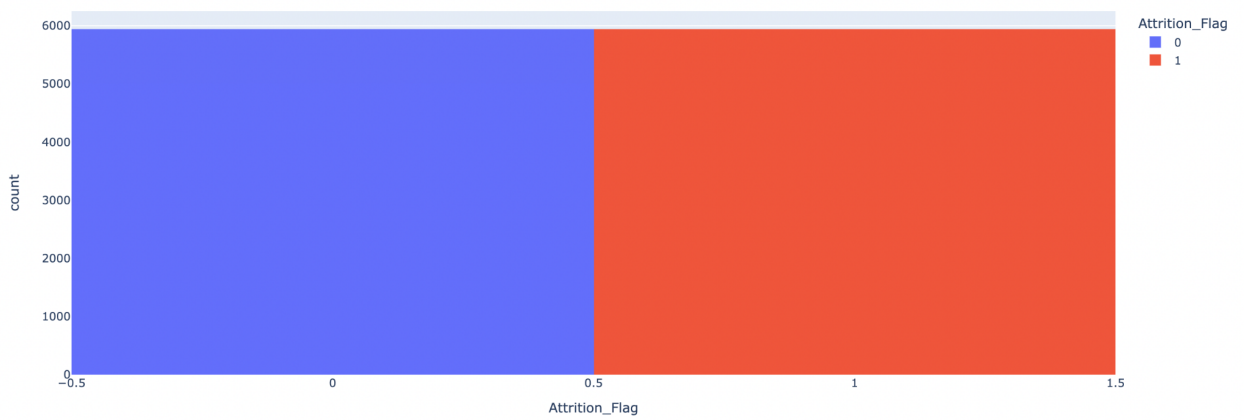
Sau khi loại bỏ các biến không phù hợp thì mô hình không còn biến nào có VIF>10. Do đó, mô hình đã không còn bị đa cộng tuyến.

Kiểm tra tính cân bằng của dữ liệu

Trong mô hình logit, kiểm tra cân bằng dữ liệu (data balance check) được sử dụng để đảm bảo rằng tỷ lệ các trường hợp trong mẫu của bạn được phân chia đúng đắn giữa các nhóm được dự báo và không được dự báo. Việc kiểm tra cân bằng dữ liệu chỉ là một trong các bước cần thiết để xây dựng mô hình logit chính xác. Nếu tỷ lệ của các trường hợp không được phân 14 chia đúng đắn, có thể sử dụng các kỹ thuật điều chỉnh dữ liệu như oversampling hoặc undersampling để cân bằng lại dữ liệu.

Từ biểu đồ ta thấy sự mất cân bằng dữ liệu giữa hai lớp là rất lớn, với số lượng mẫu gắn nhãn "1" gần năm lần lớn hơn số lượng mẫu gắn nhãn "0". Điều này có thể dẫn đến các vấn đề trong việc huấn luyện mô hình và đánh giá hiệu suất của nó. Do đó, việc xử lý mất cân bằng dữ liệu này có thể cần thiết để đạt được hiệu suất tốt hơn khi phân loại các mẫu trong tập dữ liệu.





Tổng số dữ liệu đã tăng lên đáng kể từ kích thước ban đầu và số lượng mẫu của cả hai lớp giờ đây đều bằng nhau, tỉ lệ 50:50. Điều này sẽ giúp cho mô hình dự đoán đạt được độ chính xác tốt hơn trên cả hai lớp và tránh tình trạng mô hình chỉ dự đoán tốt trên lớp đa số và không tốt trên lớp thiểu số.

```
length of oversampled data is  11876
Number of 0s in oversampled data 5938
Number of 1s 5938
Proportion of 0s data in oversampled data is  0.5
Proportion of 1s data in oversampled data is  0.5
```

4.2 Mô hình hồi quy Logistic

4.2.1 Quá trình training và testing mô hình logit

Để dự báo bằng mô hình Logit, ta cần xác định và chia tập dữ liệu thành 2 phần.

Gồm 1 tập Train và 1 tập Test. Tập Train sẽ được sử dụng để huấn luyện mô hình và tập Test sẽ được sử dụng để đánh giá mô hình.

Với bộ dữ liệu hơn 10.000 khách hàng hiện có của bài tiểu luận, bao gồm các biến độc lập và phụ thuộc đã được chọn lần lượt là Gender, Customer_ Age, Education_Level,Marital_Status,Income_Category,Total_Amt_Chng_Q4_Q1,Total_Trans_Ct, Total_Ct_Chng_Q4_Q1 Credit_Limit.

Ta sẽ sử dụng phần mềm python để chạy vòng lặp cho tất cả tỷ lệ tập Train - Test để tìm ra tỷ lệ tập Train - Test cho kết quả dự báo tốt nhất. Sau khi lựa chọn được tỷ lệ tập Train - Test cho kết quả dự báo hiệu quả nhất, ta sẽ tiến hành chạy mô hình Logistic từ dữ liệu tỷ lệ tập Train - Test vừa xác định được ở phần mềm Python.

Cách thức thực hiện được tiến hành như sau:

Bước 1: Truy cập phần mềm Python và đăng nhập các thư viện cần thiết để tạo mô hình hồi quy Logistic.

Bước 2: Nhập dữ liệu .

Bước 3: Xóa tất cả cột dữ liệu không sử dụng trong đề tài chỉ giữ lại 9 cột dữ liệu gồm Gender, Customer_Age,Education_Level,Marital_Status,Income_Category,Total_Amt_Chng_Q4_Q1,Total_Trans_Ct, Total_Ct_Chng_Q4_Q1 Credit_Limit.

Bước 4: Tiến hành hồi quy Logistic.

- Gán biến giả cho các biến định tính gồm Attrition_Flag và Gender, Education_Level, Marital_Status,Income_Category
- Kiểm tra tính cân bằng trong dữ liệu và sử dụng SMOKE để tìm tỷ lệ lớp mẫu phù hợp cho dự báo.
- Kiểm tra khả năng dự báo của các biến độc lập được chọn.
- Triển khai mô hình hồi quy Logistic và kiểm tra mức độ dự báo chính xác của tậ Train Test vừa xác định được.
- Đánh giá mô hình hồi quy Logistic (giữa dữ liệu dự báo được và dữ liệu gốc).

4.2.2 Kết quả mô hình logit

- Triển khai mô hình hồi quy Logistic và kiểm tra mức độ dự báo chính xác của tập Train - Test vừa xác định được.

Training Model

```
# fit our train inputs
# that is basically the whole training part of the machine learning
t= targets.values.ravel()
reg.fit(scaled_inputs,t)
```

```
▾          LogisticRegression
LogisticRegression(solver='liblinear')
```

```
# assess the train accuracy of the model
Logistic_Model_Score=reg.score(scaled_inputs,t)*100
Logistic_Model_Score
```

```
88.4719176926969
```

Testing Model

Bước 1: Chuẩn bị dữ liệu

Trong quá trình testing, ta cần chuẩn bị dữ liệu giống như trong quá trình training.

Bước 2: Sử dụng mô hình đã được huấn luyện để dự đoán xác suất của sự kiện nhị phân.

Sau khi đã có mô hình, ta sử dụng nó để dự đoán xác suất của sự kiện nhị phân trên tập dữ liệu kiểm tra.

Tính các giá trị FPR và TPR tại các ngưỡng xác suất khác nhau

- False Positive Rate (FPR): Tỷ lệ các trường hợp được dự đoán là positive (positive predictions) mà thực tế lại là negative (negative ground truth).

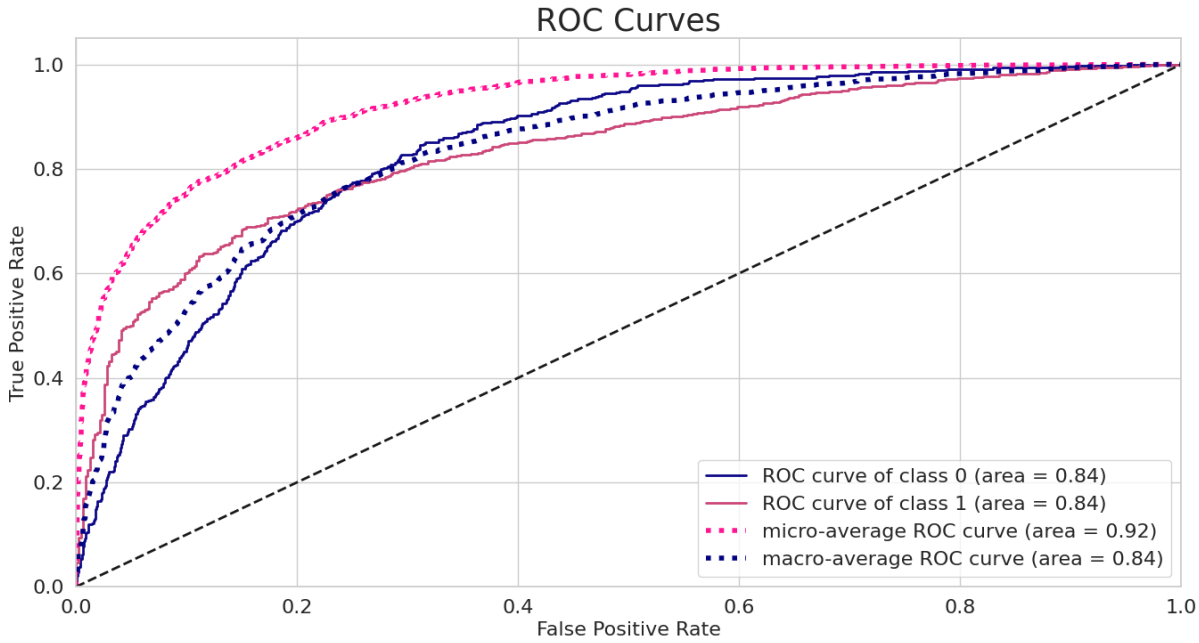
- True Positive Rate (TPR): Tỷ lệ các trường hợp được dự đoán là positive và thực tế cũng là positive.

- Threshold là ngưỡng xác suất mà mô hình sử dụng để quyết định liệu một quan sát có được phân loại là positive hay negative.

	False Positive Rate	True Positive Rate	Threshold
0	0.000000	0.000000	1.999674
1	0.000000	0.000390	0.999674
2	0.000000	0.097228	0.985442
3	0.002092	0.097228	0.985427
4	0.002092	0.191332	0.970700
...	...	...	...
467	0.939331	0.999219	0.037580
468	0.939331	0.999610	0.037443
469	0.974895	0.999610	0.019985
470	0.974895	1.000000	0.019793
471	1.000000	1.000000	0.008127

Kiểm tra đường cong roc_auc (ROC-AUC curve)

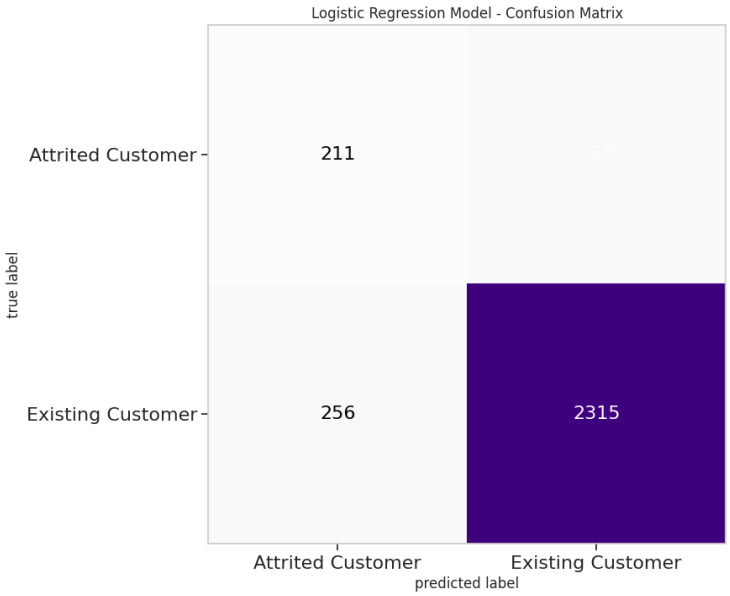
Đường ROC (Receiver Operating Characteristic) là một trong những đánh giá phổ biến nhất cho mô hình phân loại, nó đo lường hiệu suất của mô hình dự đoán trên cả hai biến phụ thuộc và không phụ thuộc. Đường ROC có thể được đánh giá bằng diện tích dưới đường cong ROC (AUC - Area Under the ROC Curve), giá trị AUC càng lớn thì mô hình càng tốt. Khi đường ROC tiến dần về 1, điều này cho thấy mô hình có khả năng phân loại tốt và tốt hơn so với mô hình có đường ROC tiến dần về 0.5 hoặc dưới đó. Điều này có nghĩa là mô hình có khả năng phân loại chính xác giữa hai nhóm dữ liệu, đó là những khách hàng sẽ rời đi và những khách hàng sẽ ở lại.



Dựa vào đường cong ROC, có thể thấy đường cong đi dọc theo biên trái và rồi đi dọc theo biên phía trên của không gian ROC, chứng tỏ kết quả dự đoán của mô hình chính xác với AUC là 0.899, cho thấy mô hình có khả năng phân loại tốt trên tập dữ liệu đánh giá. Kết quả AUC là chỉ số đo lường hiệu suất của mô hình phân loại, được tính bằng diện tích dưới đường cong ROC (Receiver Operating Characteristic).

Đánh giá độ chính xác của mô hình

Ma trận confusion của mô hình Logistic Regression



Nhận xét:

- Có 211 trường hợp mô hình dự đoán là "Attrited Customer" và thực tế cũng là "Attrited Customer".
- Có 257 trường hợp mô hình dự đoán là "Attrited Customer" nhưng thực tế lại là "Existing Customer".
- Có 256 trường hợp mô hình dự đoán là "Existing Customer" nhưng thực tế lại là "Attrited Customer".
- Có 2315 trường hợp mô hình dự đoán là "Existing Customer" và thực tế cũng là "Existing Customer".

```
from colorama import Fore
print(Fore.GREEN + "Accuracy of Logistic Regression is
```

```
Accuracy of Logistic Regression is :  84.63%
```

Nhận xét: mô hình Logistic Regression đã được đánh giá với độ chính xác (accuracy) là 84.63%➔Mô hình đã hoạt động tốt trong việc dự đoán kết quả và có khả năng áp dụng trên dữ liệu mới với độ chính xác tương tự.

Tính toán các thông số đánh giá mô hình như precision, recall và f1-score

	precision	recall	f1-score	support
0	0.51	0.50	0.50	477
1	0.91	0.91	0.91	2562
accuracy			0.85	3039
macro avg	0.71	0.70	0.71	3039
weighted avg	0.84	0.85	0.85	3039

- Precision: Chỉ số này đo lường tỷ lệ các trường hợp dự đoán là True (1) mà thực sự là True (1) so với tổng số các trường hợp dự đoán là True (1). Trong trường hợp này, precision của lớp 0 là 0.51 và của lớp 1 là 0.91.
- Recall: Chỉ số này đo lường tỷ lệ các trường hợp dự đoán là True (1) mà thực sự là True (1) so với tổng số các trường hợp thực sự là True (1). Trong trường hợp này, recall của lớp 0 là 0.5 và của lớp 1 là 0.91.
- F1-score: Đây là trung bình giữa precision và recall. F1-score càng cao thì mô hình càng tốt. Trong trường hợp này, F1-score của lớp 0 là 0.5 và của lớp 1 là 0.91.
- Accuracy: Chỉ số này đo lường tỷ lệ số dự đoán chính xác so với tổng số lượng điểm dữ liệu. Trong trường hợp này, accuracy là 0.85.
- Macro avg: Đây là trung bình của các chỉ số precision, recall và F1-score của các lớp. Trong trường hợp này, macro avg của precision, recall và F1-score là 0.71, 0.870 và 0.71.
- Weighted avg: Đây là trung bình có trọng số của các chỉ số precision, recall và F1-score của các lớp dựa trên số lượng mẫu trong mỗi lớp. Trong trường hợp này, weighted avg của precision, recall và F1-score là 0.84, 0.85 và 0.85.
Kết quả cho thấy mô hình có tương đối tốt với accuracy đạt 0.85. Tuy nhiên, kết quả cần được đánh giá kỹ hơn bằng cách xem xét các chỉ số precision, recall và F1- score của từng lớp. Mô hình đạt được kết quả tốt đối với lớp 1 với các chỉ số precision, recall và F1-score đều đạt 0.91. Tuy nhiên, mô hình chưa thể đạt được kết quả tốt đối với lớp 0, với chỉ số precision của lớp 0 chỉ đạt 0.51 và recall của lớp 0 chỉ đạt 0.5, cho thấy mô hình có xu hướng dự đoán sai nhiều điểm dữ liệu của lớp 0. Mặc dù vậy, weighted avg của precision, recall và F1- score đạt 0.87, 0.85 và 0.86, cho thấy mô hình có hiệu quả trung bình tốt trên toàn bộ tập dữ liệu.
Đánh giá mô hình hồi quy Logistic.

```
                                Results: Logit
================================================================================
Model:                    Logit                 Pseudo R-squared:      0.509
Dependent Variable:       Attrition_Flag        AIC:                   8128.7714
Date:                     2023-05-04 15:36      BIC:                   8305.9460
No. Observations:         11876                 Log-Likelihood:        -4040.4
Df Model:                 23                    LL-Null:               -8231.8
Df Residuals:             11852                 LLR p-value:           0.0000
Converged:                1.0000                Scale:                 1.0000
No. Iterations:           7.0000
--------------------------------------------------------------------------------
                              Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
--------------------------------------------------------------------------------
Customer_Age                 -0.1322   0.0029 -44.8447 0.0000 -0.1379 -0.1264
Credit_Limit                 -0.0000   0.0000  -4.7502 0.0000 -0.0000 -0.0000
Total_Amt_Chng_Q4_Q1         -1.6125   0.1393 -11.5793 0.0000 -1.8855 -1.3396
Total_Trans_Amt              -0.0003   0.0000 -20.0853 0.0000 -0.0004 -0.0003
Total_Trans_Ct                0.0752   0.0025  29.7129 0.0000  0.0702  0.0801
Total_Ct_Chng_Q4_Q1           1.9926   0.1450  13.7453 0.0000  1.7085  2.2768
Gender_M                      1.0206   0.0935  10.9199 0.0000  0.8374  1.2038
Education_Level_Doctorate     1.9730   0.1665  11.8503 0.0000  1.6466  2.2993
Education_Level_Graduate      1.7542   0.0785  22.3499 0.0000  1.6004  1.9081
Education_Level_High School   2.0411   0.0943  21.6420 0.0000  1.8562  2.2259
Education_Level_Post-Graduate 1.9500   0.1686  11.5691 0.0000  1.6197  2.2804
Education_Level_Uneducated    1.8663   0.1013  18.4285 0.0000  1.6678  2.0648
Education_Level_Unknown       1.9467   0.1021  19.0571 0.0000  1.7465  2.1469
Marital_Status_Married        1.3651   0.0780  17.5024 0.0000  1.2123  1.5180
Marital_Status_Single         0.8857   0.0807  10.9714 0.0000  0.7275  1.0439
Marital_Status_Unknown        1.1681   0.1465   7.9749 0.0000  0.8810  1.4551
Income_Category_$40K - $60K   1.6377   0.1062  15.4234 0.0000  1.4296  1.8459
Income_Category_$60K - $80K   1.1739   0.1184   9.9182 0.0000  0.9420  1.4059
Income_Category_$80K - $120K  1.2838   0.1108  11.5891 0.0000  1.0667  1.5010
Income_Category_Less than $40K 1.2766  0.0920  13.8785 0.0000  1.0963  1.4569
Income_Category_Unknown       1.8131   0.1300  13.9483 0.0000  1.5583  2.0679
```

Bảng kết quả chạy hồi quy Logistic cho thấy tất cả các giá trị P > |z| đều nhỏ hơn 0.05 điều này chứng tỏ tất cả các biến độc lập đều có ảnh hưởng đến biến phụ thuộc

V. Kết luận

Mô hình đã đạt được một độ chính xác tương đối cao trên tập train và tập test. Tuy nhiên, kết quả của mô hình cần được đánh giá kỹ hơn bằng cách xem xét các chỉ số precision, recall và F1-score của từng lớp. Kết quả cho thấy mô hình có hiệu quả trung bình tốt trên toàn bộ tập dữ liệu.

Mô hình có khả năng phân loại positive tốt (TPR cao) khi ngưỡng xác suất nhỏ hơn 0.1, tuy nhiên, khi ngưỡng tăng lên, FPR cũng tăng lên rất nhanh, cho thấy mô hình có xu hướng dự đoán sai nhiều điểm dữ liệu của lớp 0. Việc đánh giá kết quả mô hình trên từng lớp cũng cho thấy rằng mô hình có hiệu quả tốt đối với lớp 1, nhưng chưa thể đạt được kết quả tốt đối với lớp 0.

Tóm lại, mô hình logit đã đạt được một số kết quả khả quan, nhưng cần được cải thiện đối với việc dự đoán lớp 0 và việc tối ưu hóa ngưỡng phân loại để giảm thiểu FPR.
Ngoài ra, đối với tổng số tiền giao dịch và số lần giao dịch, chúng rất tương tự nhau. Cả hai đều có thể phản ánh tình trạng sử dụng của một khách hàng, vì hóa đơn có thể là một số chi tiêu lớn hoặc chi trả thường xuyên với số tiền nhỏ. Nó rất dễ hiểu rằng, một khách hàng sử dụng thẻ tín dụng càng nhiều, thì khả năng anh ta rời bỏ dịch vụ ngân hàng càng ít. Qua quá trình sử dụng, khách hàng có thể trở nên phụ thuộc hơn vào thẻ tín dụng hoặc hài lòng hơn với các dịch vụ và sản phẩm, do đó, rõ ràng là họ sẽ tiếp tục sử dụng thẻ.

Tài liệu tham khảo

[1] Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. Expert Systems with Applications, 38(12), 15273-15285.
[2] Dalvi, P. K., Khandge, S. K., Deomore, A., Bankar, A., & Kanade, V. A. (2016, March). Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. In 2016 symposium on colossal data analysis and networking (CDAN) (pp. 1-4). IEEE.
[3] Lee, T. S., Chiu, C. C., Chou, Y. C., & Lu, C. J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. Computational Statistics & Data Analysis, 50(4), 1113-1130.
[4] Westgaard, S., & Van der Wijst, N. (2001). Default probabilities in a corporate bank portfolio: A logistic model approach. European journal of operational research, 135(2), 338-349.
[5] Nazari, M., & Alidadi, M. (2013). Measuring credit risk of bank customers using artificial neural network. Journal of Management Research, 5(2), 17.
[6] Carter, T. C., Pinto, M. B., & Kahle, L. R. (2006). Gender differences in credit card behaviors and attitudes among college students. Journal of Consumer Affairs, 295-316.
[7] Dana AL-Najjar, Nadia Al-Rousan & Hazem AL-Najjar. (2022). Machine Learning to Develop Credit Card Customer Churn Prediction. Journal of Theoretical and Applied Electronic Commerce Research, 1529-1542.
[8] Guangli Nie, Wei Rowe, Lingling Zhang, Yingjie Tian & Yong Shi. (2011). Credit card churn forecasting by logistic regression and decision tree. Expert Systems with Applications, 15273-15285. Gyeongmi Lee & Jihyun Kim. (2020).
[9] Castanedo, F. (2014). Using Deep Learning to Predict Customer Churn in a Mobile Telecommunication Network, 1–8.
[10] Guo-en, X. I. A., & Wei-dong, J. I. N. (2008). Model of Customer Churn Prediction on Support Vector Machine. Systems Engineering - Theory & Practice, 28(1),71–77. https://doi.org/10.1016/S1874- 8651(09)60003-X
[11] Lu, N., Lin, H., Lu, J., & Zhang, G. (2011). A Customer Churn Prediction Model in Telecom Industry Using Boosting, (c), 1–7.
[12] Mitkees, I. M. M., Ibrahim, A., & Elseddawy, B. (2017). Customer Churn Prediction Model using Data Mining techniques, 262–268.