KẾT QUẢ NGHIÊN CỨU VÀ KIẾN NGHỊ

1.1. Dữ liệu và phân tích khám phá dữ liệu

1.1.1. Dữ liệu

Bởi vì quy định pháp luật về việc bảo vệ thông tin cá nhân, vậy nên các ngân hàng thương mại sẽ không được quyền tiết lộ thông tin về tài khoản, gian dịch của khách hàng. Vì vậy trong bài nghiên cứu này sẽ thực nghiệm trên bộ dữ liệu "Bank_Churners.csv" được tải xuống từ kho lưu trữ dữ liệu mở Zenodo của tổ chức CERN vào ngày 24/02/2024. Sau quá trình xử lí, bộ dữ liệu bao gồm 10127 quan sát với 20 biến, bao gồm 19 biến độc lập và 1 biến phụ thuộc. Trong số quan sát này, có 83.9% khách hàng không rời bỏ tương ứng với 8500 người; 16.1% tương ứng với 1627 khách hàng rời bỏ dịch vụ.

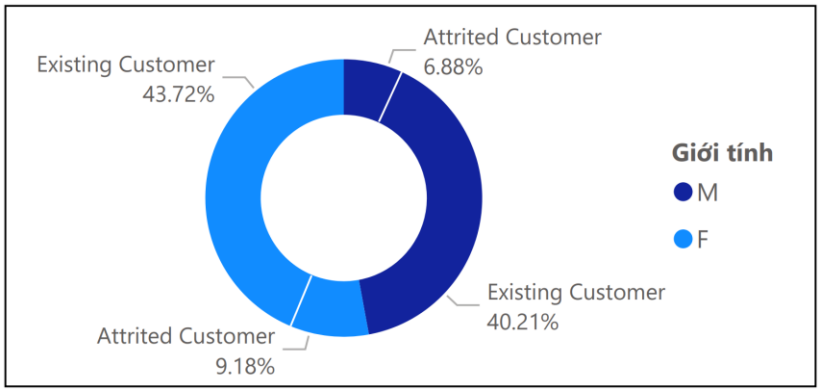
Bảng 3.1: Bảng mô tả bộ dữ liệu

Nhóm biến nhân khẩu học (6 biến):	
Customer_age	Tuổi của khách hàng
Gender	Giới tính của khách hàng
Dependent_Count	Số người phụ thuộc trong gia đình của khách hàng
Education_Level	Trình độ học vấn của khách hàng
Marital_Status	Tình trạng hôn nhân của khách hàng
Income_Category	Phân loại thu nhập của khách hàng tính bằng đô la
Nhóm biến về mối quan hệ giữa khách hàng và ngân hàng (3 biến)	
Months_on_book	Số tháng mà khách hàng đã sử dụng dịch vụ của ngân hàng
Total_Relationship_Count	Tổng số sản phẩm mà khách hàng sử dụng
Contacts_Count_12_mon	Số lần liên hệ giữa khách hàng và ngân hàng trong 12 tháng gần đây
Nhóm biến về lịch sử giao dịch thẻ tín dụng của khách hàng (10 biến)	
Card_category	Loại thẻ tín dụng mà khách hàng đang sử dụng

Credit_Limit	Hạn mức tín dụng của thẻ
Total_Revolving_Bal	Tổng tín dụng quay vòng
Avg_Open_To_Buy	Trung bình số dư khả dụng của thẻ tín dụng trong 12 tháng qua
Total_Trans_Amt	Tổng mức chi tiêu thẻ tín dụng trong 12 tháng qua
Avg_Utilization_Ratio	Tỷ lệ chi tiêu thẻ trung bình
Total_Amt_Chng_Q4_Q1	Thay đổi tổng số lần giao dịch từ quý 4 đến quý 1
Total_Trans_Ct	Tổng số lượng giao dịch trong 12 tháng qua
Total_Ct_Chng_Q4_Q1	Thay đổi tổng số lượng giao dịch quý 4 so với quý 1
Months_Inactive_12_mon	Số tháng không sử dụng thẻ trong 12 tháng qua
Biến mục tiêu (1 biến)	
Attrition_ Flag	Cho biết khách hàng đã rời bỏ việc sử dụng thẻ tín dụng hay chưa

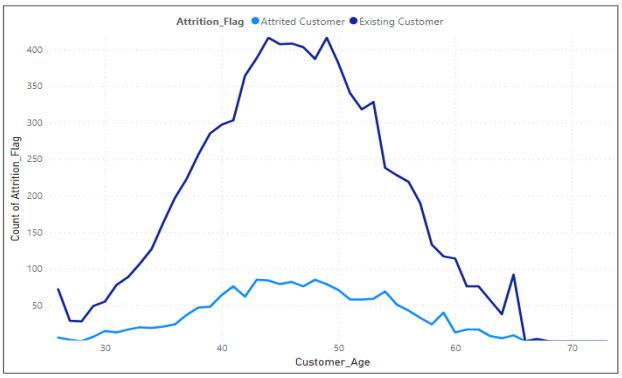
1.1.2. Mô tả một số biến nhân khẩu học

Trong bộ dữ liệu, tỷ trọng về giới tính khá đồng đều. Trong số 10127 khách hàng, có 47.09% tương ứng với 4769 khách hàng nam, 5358 khách hàng nữ, tương ứng với 52.91%. Theo đó, tính trên tổng khách hàng, số lượng khách hàng rời bỏ sử dụng thẻ tín dụng lần lượt là 6.88% đối với nam giới và 9.18% đối với nữ giới. Điều này có nghĩa là không có sự khác biệt quá nhiều giữa 2 giới trong bộ dữ liệu và trong việc rời bỏ sử dụng thẻ tín dụng ngân hàng.



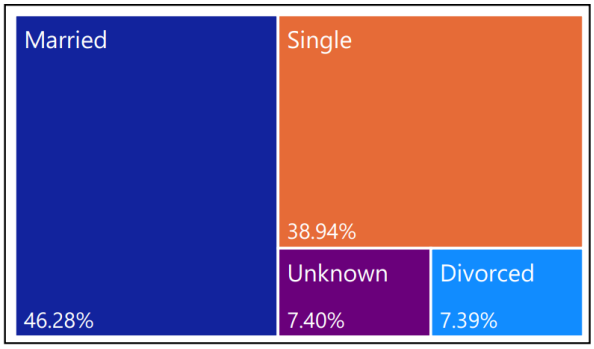
Hình 3.1: Tỷ trọng khách hàng theo giới tính và trạng thái sử dụng thẻ tín dụng

Về độ tuổi, người trẻ tuổi nhất là 26 tuổi, lớn nhất là 73. Từ đồ thị, ta thấy tuổi của khách hàng gần như phân phối chuẩn. Độ tuổi trung bình là 46.33 tuổi, trong đó những người đang 46 tuổi chiếm số lượng lớn nhất.



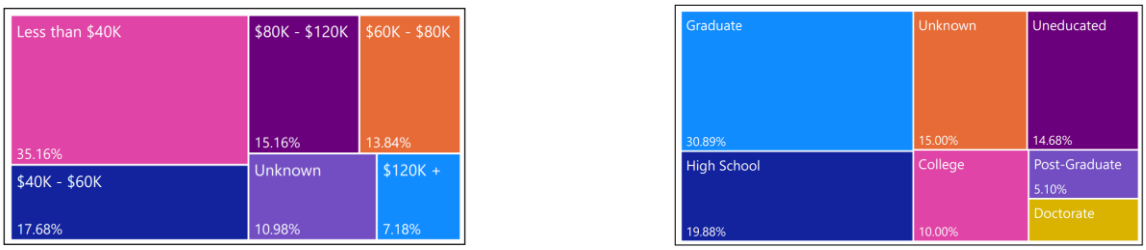
Hình 3.2: Độ tuổi của khách hàng chia theo trạng thái sử dụng

Trong bộ dữ liệu nghiên cứu, số khách hàng đã kết hôn chiếm phần lớn, 46.28% tương ứng 4687 người, tiếp theo là khách hàng độc thân với tỷ trọng là 38.94%, tương ứng với 3943 người. Nhóm khách hàng không tiết lộ tình trạng hôn nhân và đã li di thì xấp xỉ nhau, khoảng hơn 700 người với tỉ trọng khoảng 7.4%. Theo số lượng người phụ thuộc, có hơn 91% khách hàng có ít nhất một người phụ thuộc vào họ để được hỗ trợ và phúc lợi. Trong khi đó, có hơn 50% số khách hàng có hai hoặc ba người.



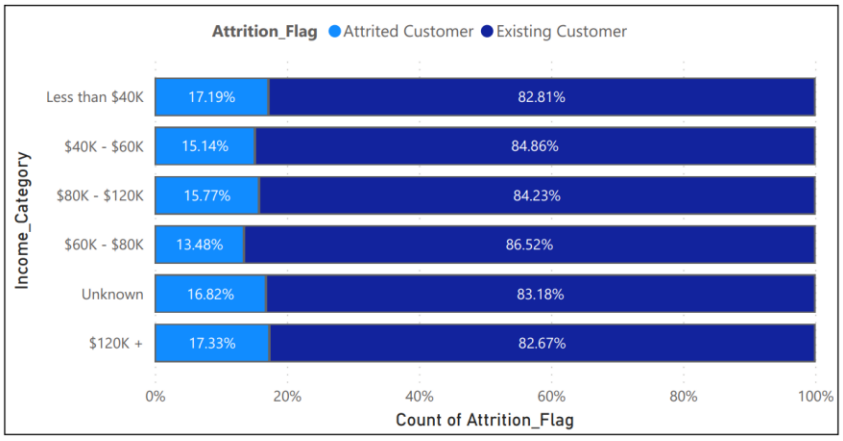
Hình 3.3: Tỷ trọng tình trạng hôn nhân của khách hàng

Về trình độ học vấn, số khách hàng tốt nghiệp đại học chiếm đa số (30.89%), tiếp theo là tốt nghiệp trung học phổ thông với 19.88%. Trình độ sau đại học và tiến sĩ chỉ chiếm khoảng 4%-5%.



Hình 3.4, 3.5: Khung thu nhập và trình độ học vấn của khách hàng

Xét theo thu nhập, có hơn 50% số khách hàng có thu nhập lớn hơn 40 nghìn đô la, tuy nhiên, tỉ trọng khách hàng có thu nhập nhỏ hơn 40 đô lại chiếm phần lớn so với các nhóm còn lại, chiếm 35.16%. Khung thu nhập phổ biến nhất của tập khách hàng này là 40 nghìn đô.

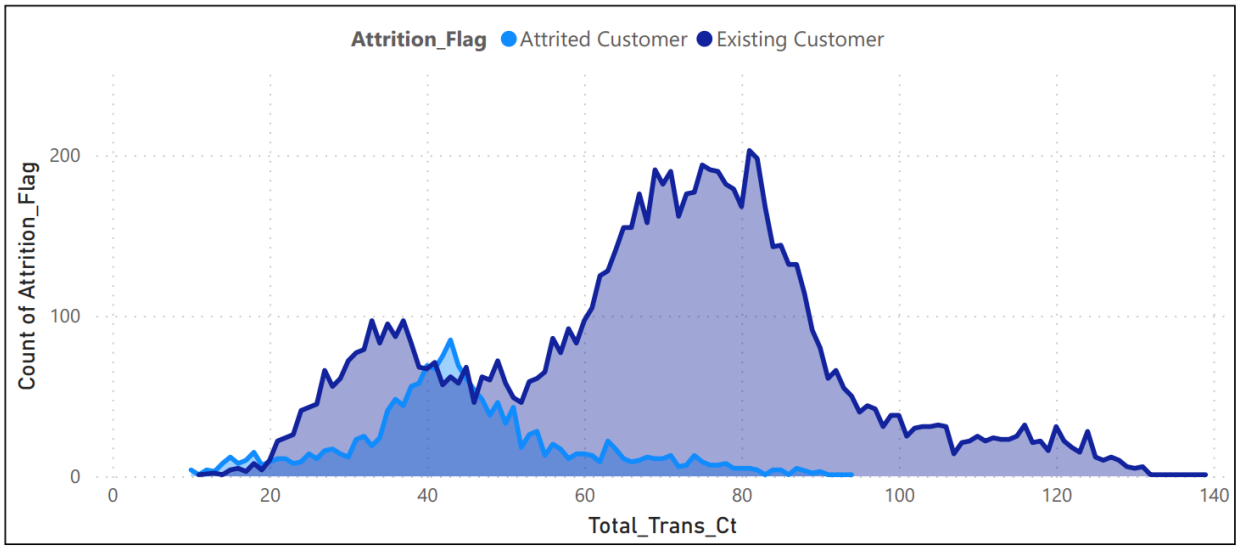


Hình 3.6: Tỷ lệ rời bỏ của khách hàng chia theo khung thu nhập

Một điều đáng chú ý ở đây là, nhóm thu nhập cao nhất (hơn 120 nghìn đô) thì lại có tỷ lệ % rời bỏ dịch vụ cao nhất so với các nhóm thu nhập khác, mặc dù không nhiều. Tỷ lệ rời bỏ cao tiếp theo là nhóm thu nhập thấp nhất trong bộ dữ liệu. Do đó, có thể nhận định rằng nhóm khách hàng có thu nhập trung bình có tỷ lệ rời bỏ dịch vụ thẻ tín dụng thấp hơn các nhóm khác.

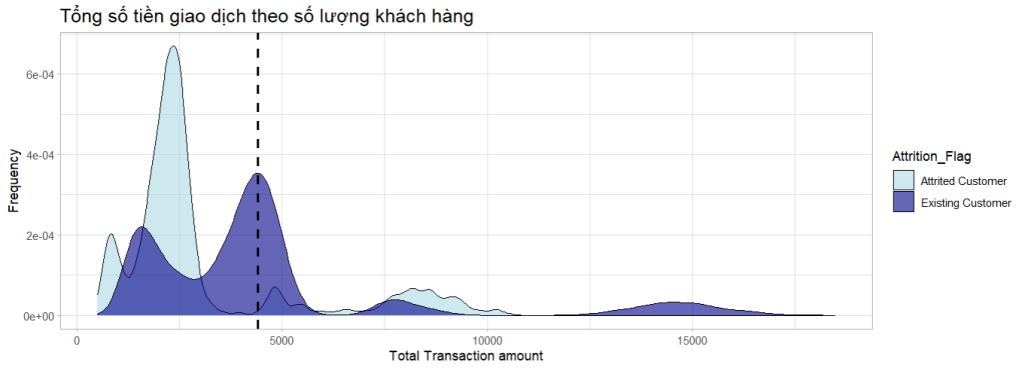
1.1.3. Mô tả một số biến liên quan đến hành vi chi tiêu

Trong vòng 12 tháng, số lượng giao dịch của nhóm khách hàng trung thành nhiều hơn đáng kể so với nhóm khách hàng rời bỏ. 8500 khách hàng không rời bỏ dịch vụ đã thực hiện 656824 giao dịch, trong khi những khách hàng rời bỏ chỉ thực hiện 73107 giao dịch,



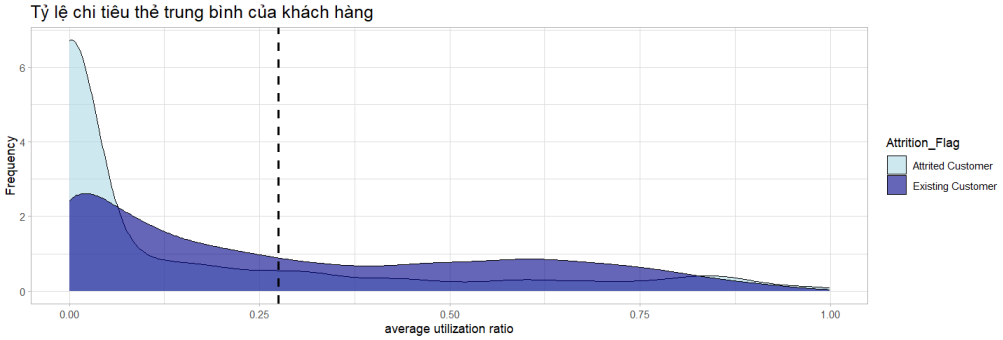
Hình 3.7: Số lượng giao dịch của khách hàng trong vòng 12 tháng theo trạng thái sử dụng

Từ biểu đồ, ta thấy những khách hàng có số lượng giao dịch trong vòng 12 tháng qua là 94 giao dịch trở lên thì sẽ không từ bỏ sử dụng thẻ tín dụng của ngân hàng này. Trong đó, nhóm khách hàng trung thành thường thực hiện 81 giao dịch 1 năm, nhóm khách hàng rời bỏ thường thực hiện 43 giao dịch và cũng có 50% số lượng khách hàng rời dịch vụ có số giao dịch trong 12 tháng qua nhỏ hơn hoặc bằng 43.



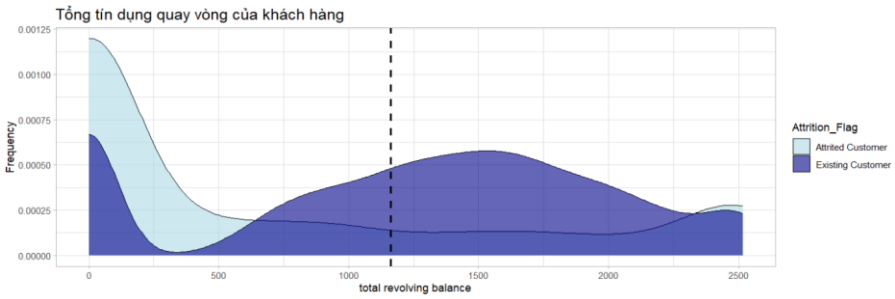
Hình 3.8: Tổng số tiền giao dịch của khách hàng theo trạng thái sử dụng thẻ

Về khối lượng giao dịch của khách hàng trong vòng 1 năm qua, tổng số tiền giao dịch đang được phân bố không đồng đều với cả 2 loại khách hàng. Mặc dù đều xuất hiện phân phối nhị thức nhưng nhóm khách hàng ngưng sử dụng dịch vụ thẻ cũng có phân phối chuẩn. Đáng chú ý là, trung vị của nhóm khách hàng rời bỏ nhỏ hơn nhóm khách hàng không rời bỏ, do đó, nhóm khách hàng trung thành có xu hướng thực hiện giao dịch có khối lượng lớn hơn so với nhóm còn lại.



Hình 3.9: Tỷ lệ chi tiêu thẻ trung bình của khách hàng theo trạng thái sử dụng

Tỷ lệ chi tiêu thẻ trung bình của 2 nhóm khách hàng cũng có sự khác biệt. Nhóm khách hàng trung thành có xu hướng có tỷ lệ này trung bình cao hơn so với nhóm rời bỏ, xấp xỉ 0.25; và giá trị này ở nhóm khách hàng còn lại là gần 0.

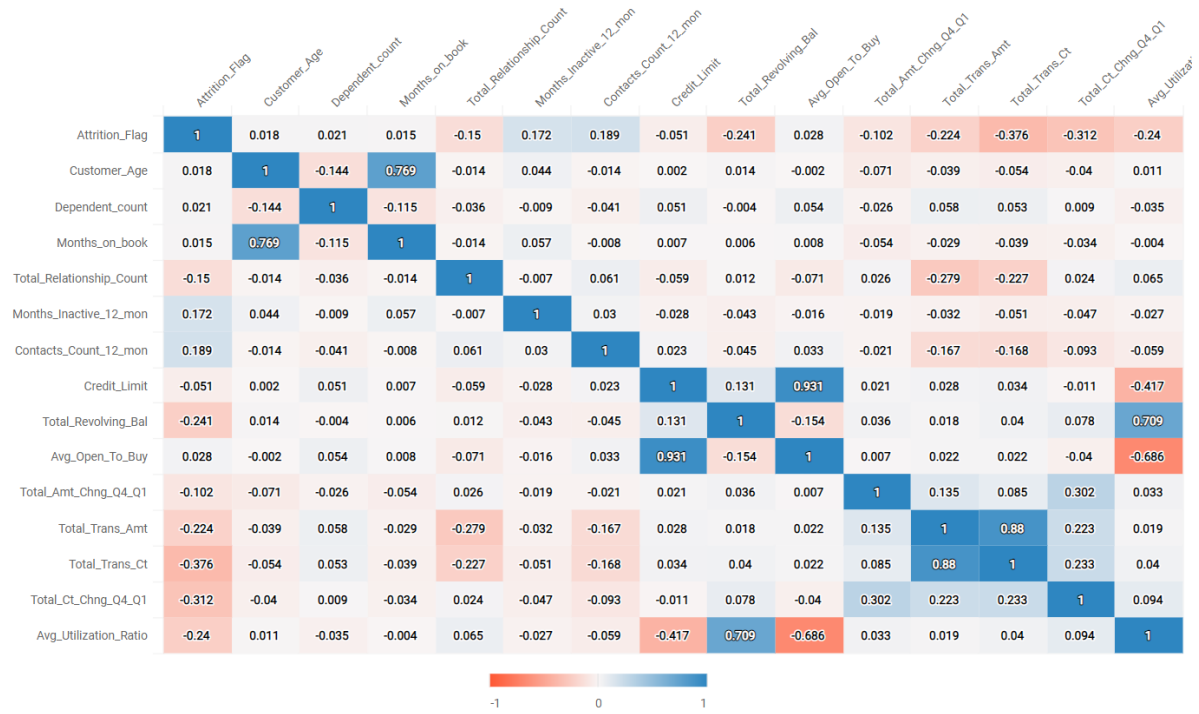


Hình 3.10: Tổng tín dụng quay vòng của khách hàng theo trạng thái sử dụng thẻ

Đối với biến tổng số dư quay vòng, ta thấy có sự phân phối rất khác biệt giữa hai nhóm khách hàng. Những khách hàng rời bỏ dịch vụ có tổng tín dụng quay vòng trung bình khoảng 600 đô, nhưng trung vị lại là 0. Trong khi đó, nhóm khách trung thành có tổng số vòng quay trung bình cao hơn nhiều, gần 1300 và trung vị là 1000.

1.1.4. Đánh giá tương quan giữa các biến

Trước khi đưa dữ liệu vào để huấn luyện mô hình, các biến cần được đánh giá tương quan. Chuyên đề sẽ loại bỏ các biến có liên quan mật thiết với nhau, các biến có độ tương quan cao, có thể dẫn đến vấn đề đa cộng tuyến trong mô hình hồi quy và các biến dư thừa hoặc không có ý nghĩa thống kê để cải thiện hiệu quả mô hình.



Hình 3.11: Ma trận tương quan giữa các biến định lượng

Từ biểu đồ trên, ta có thể thấy hệ số tương quan giữa các biến:

- Months_on_book và Customer_Age là 0.769
- Avg_Open_To_Buy và Credit_Limit là 0.931
- Total_Trans_Ct và Total_Trans_Amt là 0.88
- Avg_Utilization_Ratio và Total_Revolving_Bal là 0.709

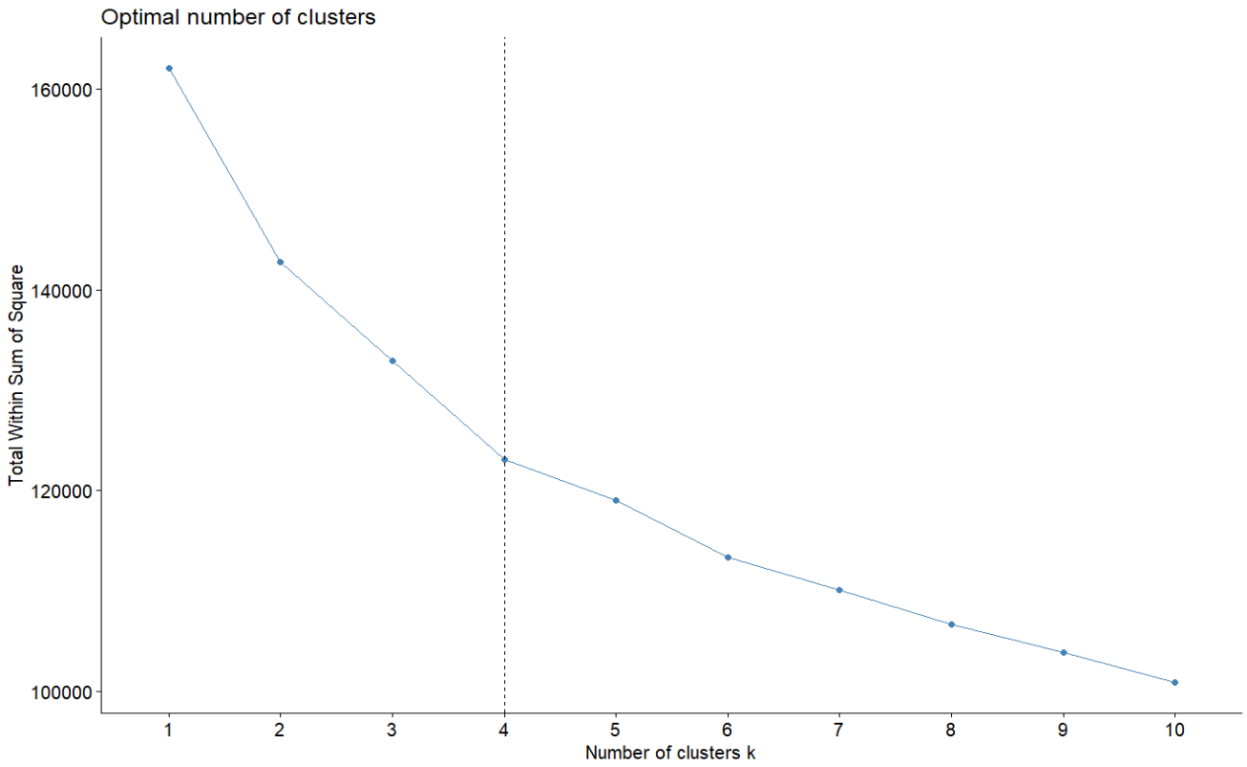
Giữa các biến còn lại, hệ số tương quan đều nhỏ hơn 0.7. Vì vậy, để đưa vào huấn luyện mô hình, chuyên đề sẽ không lựa chọn các biến: Total_Revolving_Bal, Total_Trans_Ct, Avg_Open_To_Buy, và Customer_Age.

1.2. Phân cụm khách hàng và xử lí bất cân bằng dữ liệu

1.2.1. Kết quả phân cụm khách hàng

Việc phân chia khách hàng thành các nhóm có đặc điểm tương đồng giúp mô hình tập trung vào từng nhóm cụ thể, từ đó nâng cao hiệu quả dự đoán và phân loại, phần nào cải thiện được các chỉ số trong quá trình dự đoán. Bên cạnh đó, trước khi đưa dữ liệu vào huấn luyện, phân cụm khách hàng giúp tránh tình trạng mô hình bị ảnh hưởng bởi các nhóm khách hàng có đặc điểm quá khác biệt, dẫn đến kết quả dự đoán không chính xác.

Chuyên đề sử dụng thuật toán phân cụm K-Means để phân loại các khách hàng. Với giới hạn giá trị k là [2,10] , bộ số liệu sẽ được chia ra thành từ 2 cho đến 10 cụm. Tuy nhiên việc chia ra quá nhiều cụm dẫn đến việc sẽ có nhiều cụm khách hàng. Việc phân tích nhiều cụm như vậy sẽ dẫn đến việc lãng phí tài nguyên nghiên cứu cũng như ngân hàng cần mất nhiều thời gian, chi phí để xây dựng chính sách giữ chân khách hàng cho từng cụm đối tượng. Để chọn giá trị k tối ưu, chuyên đề sử dụng phương pháp khuỷu tay, dựa trên nguyên lí tối thiểu hóa tổng bình phương khoảng cách trong các cụm.



Hình 3.12: Đồ thị xác định k tối ưu cho thuật toán phân cụm

Dựa vào đồ thị trên, giá trị của k thích hợp là 4. Việc chia tập khách hàng ngân hàng thành 3 cụm cũng là vừa đủ để thực hiện nghiên cứu. Trong đó, kích thước của 4 cụm sẽ lần lượt là 3588 quan sát; 4965 quan sát; 922 quan sát, 652 quan sát tương ứng với tỷ lệ là 35.43%, 49.03%, 9.1%, 6.44%. Sau khi bộ dữ liệu ban đầu được phân được thành 4 cụm, chuyên đề sẽ sử dụng 4 cụm này lưu thành 4 bộ dữ liệu con riêng biệt để tiến hành các bước huấn luyện tiếp theo.

1.2.2. Xử lí bất cân bằng cho từng cụm

SMOTE là kĩ thuật lấy mẫu quá mức tổng hợp, được sử dụng để giải quyết vấn đề mất cân bằng dữ liệu. Mỗi bộ dữ liệu trong 5 tập dữ liệu trong chuyên đề này sẽ được chia thành 2 tập con là tập huấn luyện chiếm 80% và tập kiểm tra chiếm 20%. Sau khi chia ra thành các tập con, những tập dùng để huấn luyện sẽ được cân bằng lại do có hiện tượng bất cân bằng dữ liệu, sao cho 2 lớp chiếm tỷ trọng xấp xỉ nhau.

Bảng 3.2: Tỷ lệ khách hàng trong các tập dữ liệu trước khi sử dụng SMOTE

Tập dữ liệu	Rời bỏ	Không rời bỏ	Tổng	Tỷ lệ
Bộ dữ liệu gốc	1298	6804	8102	1:5.24
Cụm 1	468	2403	2871	1:5.13
Cụm 2	706	3266	3972	1:4.63
Cụm 3	39	699	738	1:17.92
Cụm 4	87	435	522	1:5

Việc áp dụng kĩ thuật Borderline-SMOTE không làm thay đổi số lượng lớp đa số mà chỉ sinh thêm số lượng lớp thiểu số. Các mẫu mới sinh ra không chỉ là bản sao của của các trường hợp thiểu số hiện có; thay vào đó, thuật thuật toán lấy mẫu không gian đặc trưng cho từng lớp mục tiêu và các lớp lân cận gần nhất của nó, đồng thời tạo ra các mẫu mới kết hợp các đặc điểm của trường hợp mục tiêu với các đặc điểm của các lớp lân cận. Cách tiếp cận này làm tăng các thuộc tính có sẵn cho mỗi lớp và làm cho các mẫu tổng quát hơn.

Bảng 3.3: Tỷ lệ khách hàng trong các tập dữ liệu sau khi sử dụng SMOTE

Tập dữ liệu	Rời bỏ	Không rời bỏ	Tổng	Tỷ lệ
Bộ dữ liệu gốc	6490	6804	13294	1:1.04
Cụm 1	2340	2403	4770	1:1.03
Cụm 2	2824	3266	6090	1:1.16
Cụm 3	663	699	1362	1:1.05
Cụm 4	435	435	870	1:1

Sau khi áp dụng kĩ thuật SMOTE, cả 5 bộ dữ liệu để cho các mô hình học đã gần như cân bằng, tỷ lệ giữa 2 lớp cho biến mục tiêu đều xấp xỉ 1:1.

1.3. Kết quả huấn luyện các mô hình

1.3.1. Mô hình phân loại Naive Bayes

Sau khi thu được các tập dữ liệu huấn luyện đã được xử lí bất cân xứng, chuyên đề sẽ đưa chúng vào mô hình Naive Bayes đơn giản để huấn luyện. Kết quả huấn luyện được thể hiện bảng sau:

Bảng 3.4 : Hiệu suất mô hình phân loại Naive Bayes (%)

Bộ dữ liệu	Accuracy	Sensitivity	Specificity	Precision	NPR	F score
Bộ dữ liệu gốc	73.68	72.70	78.72	94.63	35.87	82.22
Cụm 1	72.38	71.08	80.00	95.39	32.18	81.46
Cụm 2	77.34	75.78	84.04	95.31	44.76	84.42
Cụm 3	85.33	85.38	84.62	98.65	30.56	91.53
Cụm 4	76.15	80.73	52.38	89.80	34.38	85.02

Từ bảng kết quả, ta thấy mô hình phân loại Naive Bayes hoạt động ở mức tương đối với độ chính xác trong khoảng 73.68%-85.33%. Nhìn chung, việc phân cụm dữ liệu và áp dụng kĩ thuật xử lí bất cân bằng đã giúp cải thiện hiệu suất chung của mô hình. Hầu như các chỉ số đánh giá được tính toán từ ma trận tương quan đều được cải thiện. Trong đó, giá trị Sensitivity và Preision khá cao, cho thấy mô hình này phân loại lớp khách hàng không rời bỏ khá tốt.

Tuy nhiên, giá trị NPR rất thấp, điều cho thấy mô hình này hoạt động không tốt khi phân loại nhóm khách hàng rời bỏ. Cụ thể, ví dụ ở bộ dữ liệu gốc, nếu mô hình này dự đoán có 100 khách hàng rời bỏ dịch vụ thì chỉ đúng được khoảng 35 người, tương tự với các cụm khách hàng khác. Việc dự đoán thừa quá nhiều khách hàng không có ý định rời bỏ thành những khách hàng có khả năng rời bỏ sẽ khiên ngân hàng tốn thêm rất nhiều nguồn lực không cần thiết để giữ chân khách hàng. Điều này xảy ra có thể do các giả thuyết của mô hình phân loại này quá "ngây thơ", việc các thuộc tính của bộ dữ liệu hoàn toàn độc lập với nhau là điều khó xảy ra trong thực tế

1.3.2. Hồi quy Logistic

Bảng 3.5: Hiệu suất mô hình hồi quy Logistic (%)

Bộ dữ liệu	Accuracy	Sensitivity	Specificity	Precision	NPR	F score
Bộ dữ liệu gốc	79.41	79.89	76.90	94.69	42.59	86.66
Cụm 1	76.97	78.34	70.25	92.72	39.91	84.59
Cụm 2	82.36	82.47	81.87	95.65	49.12	88.57
Cụm 3	94.41	95.32	75.00	98.79	42.86	97.02
Cụm 4	66.91	71.30	50.00	84.62	31.11	77.38

Tương tự với mô hình phân loại Naive Bayes, mô hình hồi quy Logistic cũng cho hiệu quả chung khá cao, tỷ lệ chính xác dao động từ 66% - 94% mặc dù có sự khác biệt rõ rệt giữa các cụm và bộ dữ liệu gốc. Cụm 3 là cụm có hiệu suất mô hình cao nhất với mọi tỷ số đều cao hơn tất cả các cụm khác. Đặc biệt là hệ số Accuracy lên tới 94.41% và hệ số Fscore lên tới 97.02%, cho thấy khả năng dự đoán đúng lớp khách hàng không rời bỏ là rất cao. Tuy nhiên, trong bài toán giữ chân khách hàng, việc dự đoán đúng tập khách hàng có ý định rời bỏ mới là điều quan trọng. Đối với mô hình này, tỷ lệ dự đoán khách hàng rời bỏ trên tổng số khách hàng rời bỏ thực tế ở mức khá thấp. Bên cạnh đó, hệ số NPR cũng khá thấp, tương tự đối với mô hình Naive Bayes, điều này sẽ khiến cho ngân hàng tốn chi phí và thời gian để tăng cường mối quan hệ với cả những khách hàng không có ý định rời bỏ.

Sau khi huấn luyện mô hình hồi quy Logistics cho cả 5 tập dữ liệu, ta sẽ thu được kết quả hồi quy ứng với từng tập. Chuyên đề này chỉ nhận xét kết quả hồi quy khi chạy mô hình đối với tập dữ liệu gốc (Bảng 3.6). Các kết quả hồi quy cho từng cụm nhỏ sẽ được thể hiện ở phụ lục C.

Bảng 3.6 : Kết quả hồi quy Logistic cho tập huấn luyện của bộ dữ liệu gốc

Biến	$\boldsymbol{\beta_j}$	**Std.Error**	**z value**	**Pr(>|z|) (Sig)**
(Intercept)	2.9990	0.2274	13.1860	0.0000
Dependent_count	0.0320	0.0186	1.7180	**0.0857**
Months_on_book	-0.0144	0.0030	-4.7500	0.0000

Total_Relationship_Count	-0.4084	0.0162	-25.2870	0.0000
Months_Inactive_12_mon	0.4924	0.0248	19.8790	0.0000
Contacts_Count_12_mon	0.5883	0.0230	25.5360	0.0000
Credit_Limit	0.0000	0.0000	-8.6730	0.0000
Total_Amt_Chng_Q4_Q1	0.1117	0.1208	0.9240	**0.3553**
Total_Trans_Amt	-0.0002	0.0000	-17.6450	0.0000
Total_Ct_Chng_Q4_Q1	-3.7690	0.1310	-28.7640	0.0000
Avg_Utilization_Ratio	-2.5870	0.0966	-26.7740	0.0000
Gender_1 (Male)	-0.5426	0.0916	-5.9250	0.0000
Education_Level_1 (Uneducated)	0.0151	0.0868	0.1740	**0.8619**
Education_Level_2 (High School)	-0.0522	0.0805	-0.6480	**0.5170**
Education_Level_3 (College)	-0.2474	0.0973	-2.5410	0.0110
Education_Level_4 (Graduate)	-0.1853	0.0750	-2.4720	0.0134
Education_Level_5 (Post-Graduate)	0.3052	0.1191	2.5630	0.0104
Education_Level_6 (Doctorate)	0.2415	0.1211	1.9930	0.0462
Marital_Status_1 (Divorced)	-0.1095	0.1213	-0.9020	**0.3668**
Marital_Status_2 (Single)	-0.0530	0.0935	-0.5670	**0.5707**
Marital_Status_3 (Married)	-0.3222	0.0928	-3.4720	0.0005
Income_Category_1 (Less than $40K)	0.4049	0.0833	4.8620	0.0000
Income_Category_2 ($40K - $60K)	0.1998	0.0990	2.0180	0.0435

Income_Category_3 ($60K - $80K)	0.2644	0.1308	2.0210	0.0433
Income_Category_4 ($80K - $120K)	0.6837	0.1295	5.2810	0.0000
Income_Category_5 ($120K +)	0.7444	0.1465	5.0800	0.0000
Card_Category_1 (Silver)	0.4343	0.1205	3.6050	0.0003
Card_Category_2 (Gold)	1.0730	0.2204	4.8670	0.0000
Card_Category_3 (Platinum)	1.0710	0.5085	2.1060	0.0352

Kết qủa hồi quy cho thấy 2 biến: số người phụ thuộc, tổng khối lượng giao dịch của quý 1 thay đổi so với quý 4 không thực sự có ý nghĩa thống kê ở mức ý nghĩa 10%. Một vài biến giả trong 2 nhóm biến giả trình độ học vấn và tình trạng hôn nhân có giá trị sig lớn hơn 0.05 nhưng không phải tất cả nên chúng ta có thể kết luận rằng 2 biến này vẫn có sự tác động đến biến phân loại.

Giá trị ước lương B_j cho thấy chiều hướng tác động của các biến tới biến rời bỏ dịch vụ thẻ của khách hàng. Cụ thể, các biến về khung thu nhập, loại thẻ, số tháng không hoạt động, số lần liên hệ trong vòng 12 tháng, tổng khối lượng giao dịch, số người phụ thuộc càng tăng thì có thể: khả năng khách hàng ngưng sử dụng thẻ tín dụng của ngân hàng càng cao. Ngược lại, các biến còn lại có hệ số tương quan âm, nghĩa là khi chúng càng tăng thì khả năng khách hàng rời bỏ dịch vụ càng thấp.

1.3.3. Máy vector hỗ trợ

Chuyên đề sẽ tiếp tục áp dụng mô hình Support vector machine cho cả 5 bộ dữ liệu. Với tham số Kernel = radial, còn gọi là Radial Basis Function, hàm này đo lường mức độ tương đồng giữa các điểm dữ liệu dựa trên khoảng cách Euclidean. Tiếp theo các điểm dữ liệu sẽ được ánh xạ vào không gian đặc trưng có chiều cao hơn, nơi việc phân tách các lớp dữ liệu trở nên dễ dàng hơn. Sau khi ánh xạ, SVM được áp dụng trong không gian đặc trưng này để tìm ra siêu phẳng phân chia tốt nhất.

Để tìm ra siêu phẳng tốt nhất, mô hình SVM phi tuyến này đã tạo ra 4827 support vector trong tập huấn luyện của bộ dữ liệu gốc, hơn 2 nghìn vector cho cụm 1 hoặc 2 và hơn 200 vector cho cụm 3 hoặc 4.

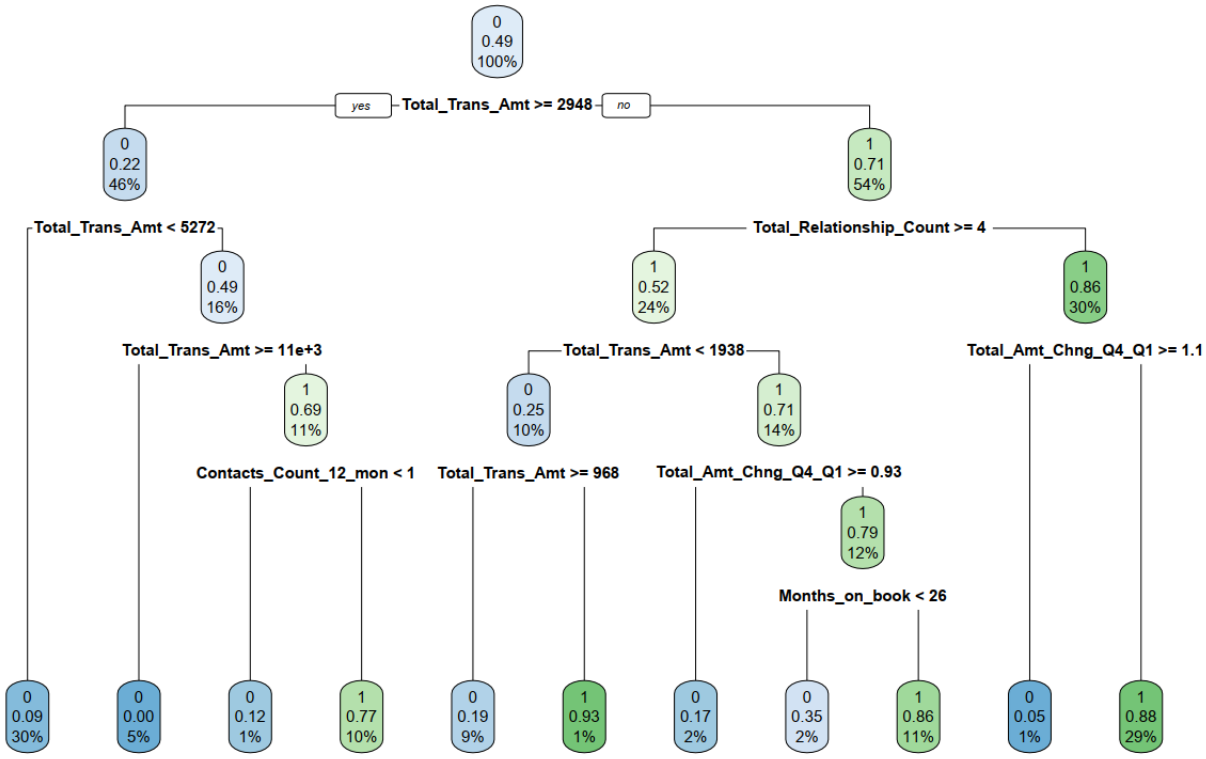
Bảng 3.8: Hiệu suất mô hình máy vector hỗ trợ (%)

Bộ dữ liệu	Accuracy	Sensitivity	Specificity	Precision	NPR	F score
Bộ dữ liệu gốc	89.78	92.04	78.12	95.59	65.56	93.78
Cụm 1	88.15	89.22	81.9	96.64	56.58	92.78
Cụm 2	88.72	90.06	82.98	95.77	66.10	92.83
Cụm 3	95.11	97.66	61.54	97.09	66.67	97.37
Cụm 4	78.46	84.49	52.38	90.10	37.93	86.67

Nhìn chung, hiệu suất của mô hình SVM cao hơn mô hình hồi quy Logistic, độ chính xác đều ở mức trên 78%. Giá trị hệ số Sensitivity và Precision rất cao, điều này cũng giống với 2 mô hình trên, cho thấy khả năng phân loại nhóm khách hàng trung thành là có hiệu quả hơn. Tuy nhiên, hệ số đánh giá mang tính quyết định trong chuyên đề này là NPR và Speciticity lại khá thấp, mặc dù nhìn chung cao hơn hai mô hình trước. Điều đáng chú ý là đối với cụm 4, mọi kết quả phân loại đều ở mức thấp nhất, trong đó hệ số quan trọng nhất của chuyên đề là NPR và Specificity thì lại thấp nhất. Khả năng phân loại khách hàng rời bỏ của thuật toán SVM đối với cụm 4 là cực kì thấp. Điều này cho thấy nếu chỉ áp dụng riêng mô hình này vào nhóm khách hàng cụm 4 thì ngân hàng sẽ bỏ sót rất nhiều khách hàng có ý định rời bỏ sử dụng thẻ tín dụng.

1.3.4. Cây quyết định

Mô hình cây quyết định dựa trên các thuộc tính đầu vào ở tập dữ liệu gốc đã tạo ra 10 nút quyết định, trong đó có 1 nút gốc dựa vào biến Total_Trans_Amt (Tổng khối lượng giao dịch của khách hàng) và 11 nút cuối biểu hiện phần trăm số khách hàng và xác suất dự báo.



Hình 3.13: Đồ thị cây quyết định với tập dữ liệu chung

Cây quyết định trong tập dữ liệu gốc (sau khi xử lí bất cân xứng) chia tập khách hàng thành 2 nhánh lớn: những người có tổng khối lượng giao dịch nhỏ hơn 2948 đô và những người có tổng khối lượng giao dịch lớn hơn 2948 đô. Từ các nút tiếp theo, bộ dữ liệu được phân chia theo các thuộc tính: số lượng liên hệ được thực hiện trong vòng 12 tháng, tổng khối lượng giao dịch thay đổi của quý 1 so với quý 4, số lượng sản phẩm mà khách hàng đang nắm giữ và số tháng khách hàng đã gắn bó với ngân hàng. Từ cây quyết định của tập dữ liệu gốc, chuyên đề rút ra được một số quy luật khi khách hàng có ý định rời bỏ dịch vụ thẻ tín dụng như:

- Khách hàng có số lần liên hệ với ngân hàng lớn hơn 1 lần, tổng khối lượng giao dịch trong vòng 1 năm nhỏ hơn 11 nghìn đô

- Khách hàng có tổng khối lượng giao dịch thấp hơn 968 đô trong 1 năm, số sản phẩm nắm giữ lớn hơn 4
- Khách hàng có thời gian gắn bó với ngân hàng từ 26 tháng trở lên, tổng khối lượng giao dịch của quý 1 thay đổi so với quý 4 nhỏ hơn 0.93, tổng khối lượng giao dịch nhỏ hơn 1939 đô và số lượng sản phẩm nắm giữ lớn hơn 4
- Khách hàng có tổng khối lượng giao dịch của quý 1 thay đổi so với quý 4 nhỏ hơn 1.11 nhưng số lượng sản phẩm nắm giữ nhỏ hơn 4

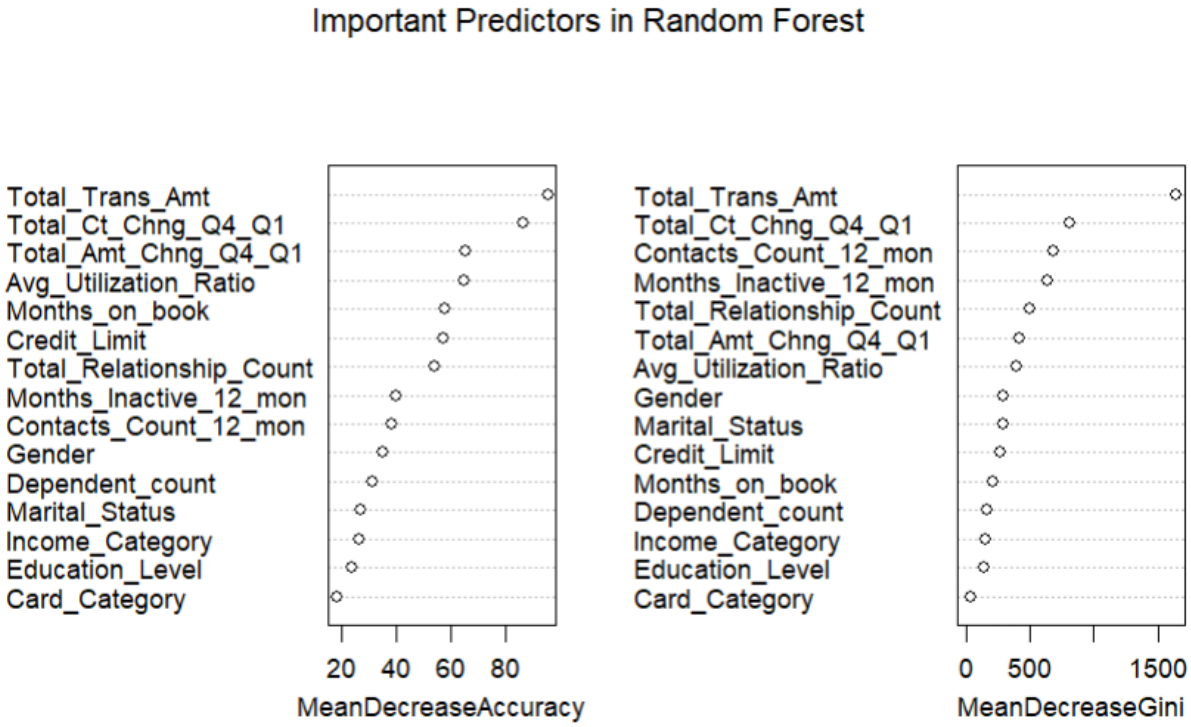
Bảng 3.9: Hiệu suất mô hình cây quyết định (%)

Bộ dữ liệu	Accuracy	Precision	Sensitivity	Specificity	NPR	F score
Bộ dữ liệu gốc	90.12	96.57	91.45	83.28	65.39	93.94
Cụm 1	89.81	96.55	91.34	80.95	61.59	93.87
Cụm 2	91.13	96.86	92.04	87.23	71.93	94.39
Cụm 3	93.47	98.77	94.15	84.61	52.38	96.41
Cụm 4	77.70	89.22	83.48	47.61	35.71	86.25

Dựa vào kết quả được tính toán dựa vào confusion matrix, các thước đo cho thấy thuật toán Decision Tree hoạt động tốt ở tập dữ liệu gốc, cụm 1, 2, và 3. Nhìn chung tỷ lệ chính xác, tỷ lệ dự đoán đúng lớp khách hàng không rời bỏ và hệ số Fscore đều cao, nhưng NPR lại khá thấp, thuật toán này đã bỏ lỡ hơn ít nhất 29% số lượng khách hàng rời bỏ. Đặc biệt đối với cụm 4, giá trị này rất thấp, điều này cho thấy decision tree không phù hợp với bộ dữ liệu có những đặc điểm giống cụm 4.

1.3.5. Rừng ngẫu nhiên

Tiếp tục áp dụng mô hình Random Forest vào 5 tập dữ liệu đã được cân bằng, trước tiên thu được kết quả các biến quan trọng nhất trong khi huấn luyện tập dữ liệu chung như sau:



Hình 3.14: Mức độ quan trọng của các biến trong tập dữ liệu chung

Biểu đồ trên thể hiện tầm quan trọng của các biến khi theo 2 cách tiếp cận là mức độ giảm trung bình độ chính xác (Accuracy) và mức độ giảm trung bình trong Gini. Biến tổng khối lượng giao dịch của khách hàng "Total_Trans_Amt'' là biến có mức độ quan trọng nhất, về cả accuracy và chỉ số Gini đều cao hơn hẳn các biến còn lại. Điều này cũng khá dễ hiểu, các khách hàng có khả năng cao sẽ rời bỏ dịch vụ thẻ thì sẽ giảm dần mức chi tiêu của họ. Mức độ quan trọng thứ 2 là biến tổng số lượng thay đổi giữa quý 1 so với quý 4 ''Total_Ct_Chng_Q4_Q1''. 3 biến có mức độ quan trọng cao tiếp theo trong top 5 đều là biến liên quan đến hành vi chi tiêu thẻ tín dụng của khách hàng.

Các biến ít có mức độ quan trọng đến biến phân loại là loại thẻ- "Card_Category", trình độ học vấn - "Education_Level", khung thu nhập - "Income_Category", tình trạng hôn nhân - "Marital_Status ", số người phụ thuộc vào chủ thẻ - "Dependent_count" , giới tính chủ thẻ -"Gender". Đây đều là các nhóm biến liên quan đến đặc điểm nhân khẩu của chủ thẻ.

Kết hợp với bảng kết quả hồi quy bên trên, ta có thể nhận định rằng: khi tổng số giao dịch thay đổi của quý 1 so với quý 4, tỷ lệ chi tiêu tín dụng trung bình, số lượng sản phẩm đang nắm giữ càng cao hoặc càng tăng thì khả năng khách hàng đó rời bỏ dịch vụ

thẻ tín dụng sẽ càng thấp. Ngược lại, khi số lần liên hệ được thực hiện trong vòng 12 tháng và số tháng không hoạt động càng tăng thì khả năng khách hàng đó rời bỏ dịch vụ thẻ tín dụng sẽ càng cao.

Bảng 3.10: Hiệu suất mô hình Random Forest (%)

Bộ dữ liệu	Accuracy	Precision	Sensitivity	Specificity	NPR	F score
Bộ dữ liệu gốc	94.17	95.77	97.34	77.82	85.05	96.55
Cụm 1	92.89	95.17	96.57	71.43	78.13	95.86
Cụm 2	94.06	96.04	96.65	82.97	85.25	96.35
Cụm 3	95.65	96.57	98.83	53.84	77.78	97.68
Cụm 4	88.46	90.52	96.33	47.62	71.42	93.33

Mô hình Random Forest có chính xác chung (accuracy) cao nhất trong 5 mô hình phân loại mà chuyên đề này sử dụng, dao động từ 88.46% đến 95.65% tùy từng tập dữ liệu. Các hệ số đánh giá khả năng phân loại đúng lớp đa số (khách hàng không rời bỏ) của mô hình này ở mức rất cao, đều hơn 90%. Đối với dữ liệu của cụm khách hàng 3, tỷ lệ dự đoán đúng khách hàng trung thành – Sensitivity lên tới 98.83%, điều này có nghĩa, trong 100 khách hàng trung thành thì mô hình này đã đoán đúng được 98 người.

Tuy nhiên, chỉ số đánh giá quan trọng hơn trong bài toán giữ chân khách hàng lại là Specificity và NPR. Nhìn chung 2 hệ số này đều cao hơn 4 mô hình trước đó. Chỉ số NPR ở mô hình này đã cao hơn đáng kể, thấp nhất là 71.42% và cao nhất là 85.25%. Điều này có nghĩa trong 100 khách hàng có ý định rời bỏ thẻ tín dụng thì mô hình này đã đoán đúng được ít nhất 71 người. Mặc dù đây không phải con số quá cao nhưng lại là một tín hiệu đáng mừng trong việc phát hiện sớm những chủ thẻ muốn ngưng sử dụng dịch vụ thẻ.

1.4. Đánh giá chung kết quả huấn luyện các mô hình

Chuyên đề này đã sử dụng 5 mô hình học máy để dự đoán khả năng khách hàng rời bỏ dịch vụ thẻ tín dụng. Xét hiệu suất dự đoán tổng quát, cả 5 mô hình được áp dụng trên 5 tập dữ liệu đều hoạt động tương đối tốt, Accuracy dao động từ 67.69% (khi áp dụng mô hình hồi quy logistic cho cụm 4) đến 96.2% (khi áp dụng mô hình hồi quy logistic cho cụm 3). Đa số độ chính xác của các mô hình đều rơi vào khoảng 80% đến 90%. Bên cạnh đó,

khi hai chỉ số Sensitivity và Precision có mức độ cao thấp ngược nhau, ta xét đến thước đo Fscore. Giá trị này ở các mô hình cũng khá cao, chủ yếu nằm trong khoảng 85% đến 95%.

Bảng 3.11: Accuracy và Negative Predictive Rate của các mô hình

Mô hình	Accuracy	Negative Predictive Rate (NPR)
Naïve Bayes	72.38 – 85.33	30.56 – 44.76
Hồi quy Logistics	66.91 – 94.41	31.11 – 49.12
Máy vector hỗ trợ	78.46 – 95.11	37.93 – 66.67
Cây quyết định	77.70 – 93.47	35.71 – 71.93
Rừng ngẫu nhiên	88.46 – 95.65	71.42 – 85.25

Xét về khả năng dự đoán lớp tối thiểu – những khách hàng rời bỏ thẻ tín dụng, 5 mô hình này hoạt động tạm ổn. Khả năng đoán đúng chủ yếu ở mức 60% - 70%. Hai mô hình phân loại Naive Bayes và hồi quy Logistic cho kết quả thang đo NPR thấp nhất trong 4 mô hình, tiếp theo là 2 mô hình máy vector hỗ trợ và cây quyết định. Việc áp dụng riêng các mô hình này vào thực tế sẽ khiến ngân hàng dự đoán chưa được chuẩn các khách hàng muốn rời bỏ. Thang đo này có giá trị lớn nhất khi áp dụng thuật toán rừng ngẫu nhiên, với khả năng dự đoán đúng trong khoảng 71.42% - 85.25%.

Một điều đáng chú ý khi áp dụng các thuật toán này vào các bộ dữ liệu là tất cả các thang đo đánh giá hiệu suất của mô hình đối với cụm 4 đều thấp hơn 3 cụm khác và tập dữ liệu chung. Điều này có thể lí giải là do tập dữ liệu ở cụm 4 có quá nhiều nhiễu hoặc ngoại lệ, điều này dẫn đến việc các thuật toán nhạy cảm với nhiễu sẽ học sai các thuộc tính từ bộ dữ liệu. Ngoài ra, điều này còn do dữ liệu ở cụm 4 chất lượng thấp hơn so với các cụm khác hoặc sự phân bố dữ liệu khác biệt so với các cụm khác, khiến các mô hình được đào tạo trên các cụm 1, 2, 3 không phù hợp với cụm này.

1.5. Kiến nghị và đề xuất

Dựa trên kết quả nghiên cứu, chuyên đề có thể đưa ra một số gợi ý về chính sách giữ chân khách hàng cho các nhà quản lí ngân hàng như sau:

Thứ nhất, với khả năng dự đoán chính xác cao nhất cả lớp đa số và lớp thiểu số, các ngân hàng nên áp dụng thuật toán rừng ngẫu nhiên để phân loại và dự báo những khách hàng có khả năng rời bỏ sử dụng dịch vụ thẻ tín dụng. Tiếp theo, các nhà quản lí nên thiết

kế các chính sách chăm sóc khách hàng riêng biệt cho nhóm khách hàng này thay vì chăm sóc toàn bộ tập khách hàng của ngân hàng. Điều này tiết kiệm được rất nhiều nguồn lực cũng như tối ưu được chi phí.

Thứ hai, từ những thuộc tính có tác động mạnh đến khả năng dự báo khách hàng rời bỏ thẻ tín dụng, các nhà quản lý cũng nên quan tâm về những yếu tố mà có thể tác động để giữ chân nhóm khách hàng có nguy cơ rời bỏ. Cụ thể, các chính sách này nhằm không khuyến khích khách hàng giảm chi tiêu thẻ tín dụng hay giảm số lượng giao dịch thẻ tín dụng, nên tăng hạn mức thẻ và tăng số lượng sản phẩm sử dụng với ngân hàng. Dành riêng cho nhóm khách hàng này, ngân hàng có thể đưa ra những quy định ưu đãi về phí thường niên, khuyến mãi đặc biệt khi thanh toán bằng thẻ tín dụng hoặc tăng tỷ lệ hoàn tiền hay các sản phẩm đặc thù khác. Đồng thời, cần đẩy mạnh nhiều sản phẩm ngân hàng khác để thu hút sự quan tâm, tham gia của nhóm khách hàng này, từ đó làm tăng số lượng sản phẩm mà khách hàng nắm giữ.

Thứ ba, các nghiên cứu trong tương lai của ngân hàng có thể cân nhắc đến việc phân cụm chỉ dựa vào nhóm biến hành vi sử dụng thẻ tín dụng thay vì phân cụm khách hàng dựa trên tất cả các biến như trong chuyên đề. Mặc dù phát hiện của Xiahou và Harada (2022) đã chứng minh rằng mỗi chỉ số đo hiệu suất dự đoán được cải thiện đáng kể sau khi phân khúc khách hàng nhưng kết quả của chuyên đề chỉ ra rằng việc phân cụm khách hàng vẫn chưa thực sự nâng cao khả năng dự đoán sự rời bỏ của khách hàng. Kết quả từ nghiên cứu này chỉ dừng lại ở mức thử nghiệm, đối chiếu và quyết định mô hình học máy nào là tốt nhất để phân loại khách hàng rời bỏ dưới sự phân cụm các khách hàng trong ngành ngân hàng.

Thứ tư, các ghiên cứu trong tương lai nên xem xét việc áp dụng các mô hình học sâu, học kết hợp, không dừng lại ở việc chạy các mô hình truyền thống và cơ bản như trong chuyên đề, nhằm cải thiện được các chỉ số quan trọng trong bài toán giữ chân khách hàng.

Cuối cùng, tập dữ liệu nghiên cứu có thể cân nhắc được mở rộng về cả số quan sát và số thuộc tính, hay mở rộng quy mô phân tích ở một quy mô lớn hơn như nhiều ngân hàng thay vì chỉ tập trung vào một ngân hàng như trong chuyên đề.