# **ỨNG DỤNG CÁC KỸ THUẬT KHAI PHÁ DỮ LIỆU ĐỂ PHÂN TÍCH ĐẶC ĐIỂM KHÁCH HÀNG**

## **Thiết lập mục tiêu:**

* Tìm ra đặc điểm của khách hàng rời bỏ và ở lại.
* Làm rõ những yếu tố ảnh hưởng đến quyết định rời bỏ của khách hàng/ những dịch vụ được khách hàng quyết định tiếp tục sử dụng.
* Từ những yếu tố ảnh hưởng đến sự rời bỏ của khách hàng mà ta có thể đưa ra dự báo những giải pháp, những sự thay đổi của các dịch vụ để giữ chân khách hàng.

## **Chọn dữ liệu:**

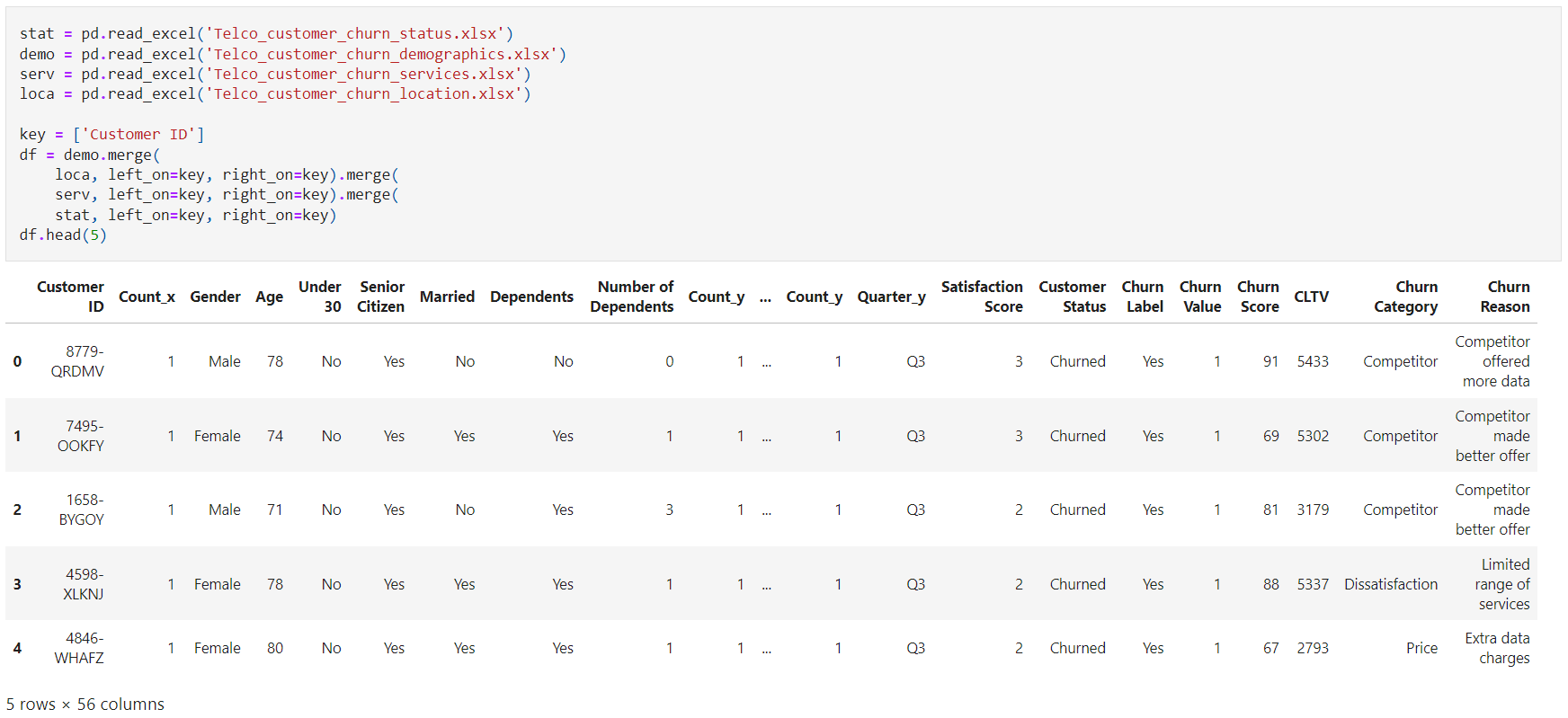
* Với mục tiêu như trên tụi em quyết định chọn bộ dữ liệu *“Telco Customer Churn”.*
* Để có thể tìm ra chân dung khách hàng, chúng em đã chọn 2 bảng dữ liệu: *Location* và *Population* (Telco\_customer\_churn\_location.xlsx và Telco\_customer\_churn\_population.xlsx)
* Để có thể phân tích sự rời bỏ của khách hàng trong quá trình sử dụng dịch vụ của công ty truyền thông chúng em chọn 2 bảng dữ liệu: *Services* và *Status* (Telco\_customer\_churn\_services.xlsx và Telco\_customer\_churn\_status.xlsx)
* Link dữ liệu trên kaggle:

<https://www.kaggle.com/datasets/ylchang/telco-customer-churn-1113>

## **Phân tích cụ thể:**

### ***Tiền xử lý dữ liệu:***

#### *Tích hợp dữ liệu:*



* Trong quá trình tìm kiếm dữ liệu, chúng em đã tìm được bộ dữ liệu gồm 5 file:
  + Telco\_customer\_churn\_location.xlsx
  + Telco\_customer\_churn\_population.xlsx
  + Telco\_customer\_churn\_services.xlsx
  + Telco\_customer\_churn\_status.xlsx
  + Telco\_customer\_churn\_demographics.xlsx
* Để phù hợp với mục tiêu đã được đề ra chúng em đã tích hợp các file:
  + Telco\_customer\_churn\_location.xlsx
  + Telco\_customer\_churn\_population.xlsx
  + Telco\_customer\_churn\_services.xlsx
  + Telco\_customer\_churn\_status.xlsx

lại với nhau để có thể sử dụng một cách tốt nhất, đầy đủ nhất với tất cả các thông tin bộ dữ liệu này mang lại.

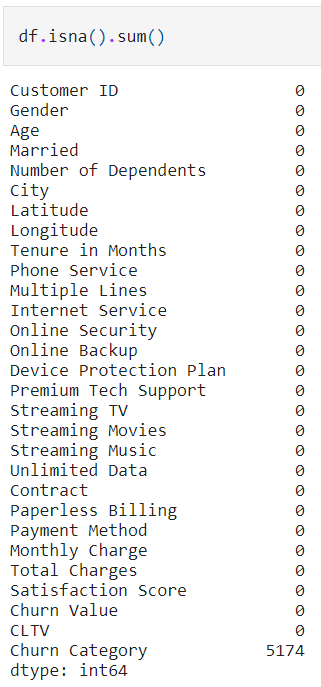
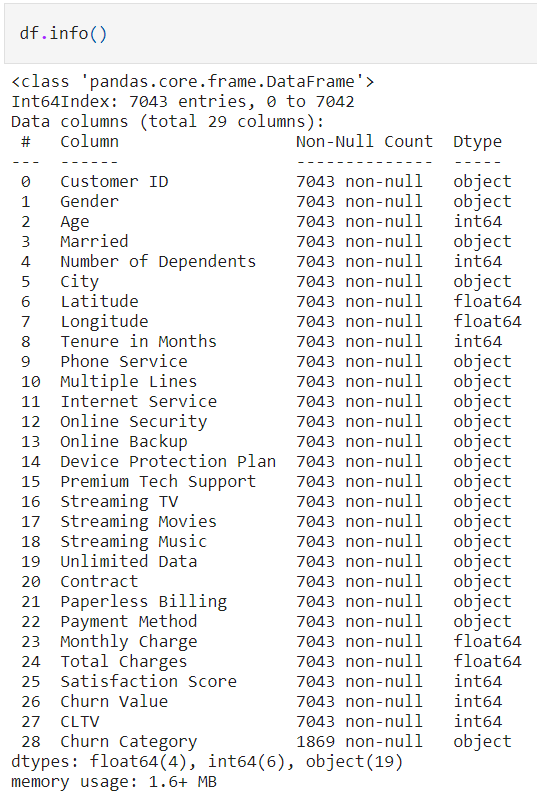
* Dữ liệu gồm 7043 dòng 56 cột.

#### *Lựa chọn dữ liệu:*



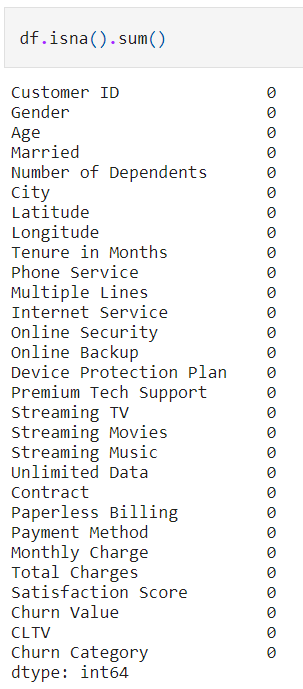
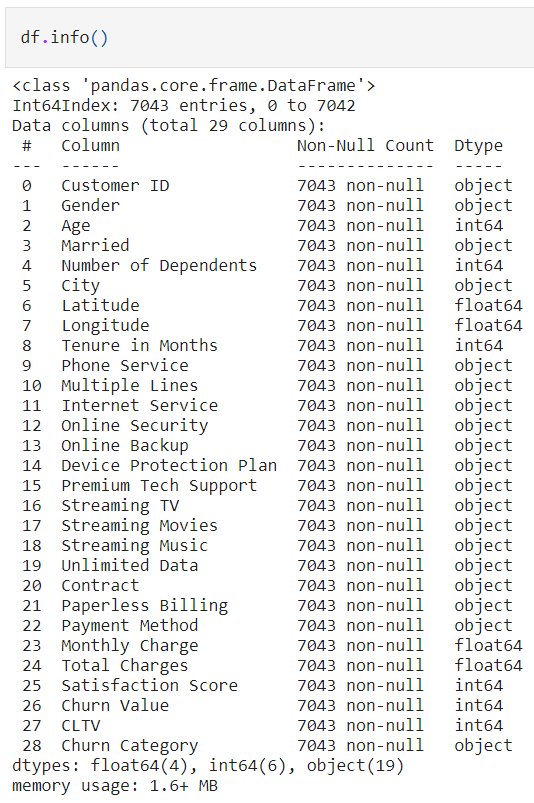
* Trong những file dữ liệu đã được tích hợp, dựa vào mục tiêu đã được đề ra, chúng em đã chọn lựa ra những cột dữ liệu cần thiết nhất có để sử dụng để làm rõ được những ý muốn của nhóm. File dữ liệu sử dụng được xuất ra gồm 29 cột: Customer ID, Gender, Age, Married, Number of Dependents, City, Latitude, Longitude, Tenure in Months, Phone Service, Multiple Lines, Internet Service, Online Security, Online Backup, Device Protection Plan, Premium Tech Support, Streaming TV, Streaming Movies, Streaming Music, Unlimited Data, Contract, Paperless Billing, Payment Method, Monthly Charge, Total Charges, Satisfaction Score, Churn Value, CLTV, Churn Category.

#### *Làm sạch dữ liệu:*

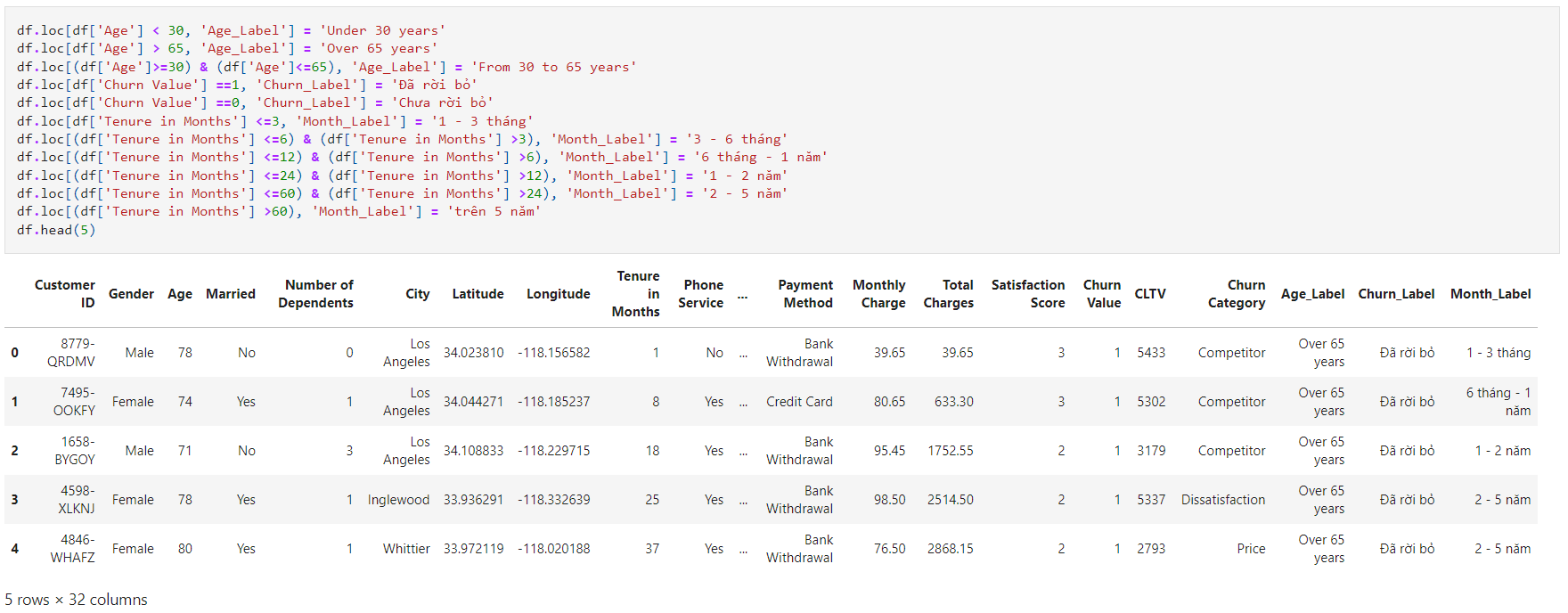


* Dữ liệu được tích hợp vẫn có những thiếu sót hoặc thiếu thông tin làm khó có thể sử dụng được trong quá trình phân tích. Do vậy, trước tiên chúng em đã lựa chọn làm sạch dữ liệu trước.
* Cột Churn Category so với các cột khác đã bị để trống dữ liệu (chỉ có 1869 dòng, so với các cột khác là 7043 dòng) và những điểm dữ liệu đó là do những khách hàng đó không rời bỏ nên không có dữ liệu về lý do rời bỏ. Nên tụi em đã điền vào những đoạn trống từ “Not Churn” để có thể khai thác cột dữ liệu này để phân tích.





#### *Biến đổi dữ liệu:*



* Tạo thêm 3 cột là Age\_Label, Churn\_Label và Month\_Label, những cột này được tạo nên dựa vào 3 cột là Age, Churn Value, Tenure in Months
* Cột Age\_Label dựa trên cột Age có các giá trị:
* “Under 30 years” (Age < 30)
* “From 30 to 65 years” (30<=Age<=65)
* “Over 65 years”(Age>65)
* Cột Churn\_Label dựa trên với cột Churn Value có các giá trị:
* “Chưa rời bỏ” (Churn Value=0)
* “Đã rời bỏ” (Churn Value=1)
* Cột Month\_Label dựa trên cột Tenure in Months có các giá trị:
* “Từ 1-3 tháng” (Tenure in Months<=3)
* “Từ 1-6 tháng”(3<Tenure in Months<=6)
* “1 năm”(6<Tenure in Months<=12)
* “2 năm”(12<Tenure in Months<=24)
* “5 năm”(24<Tenure in Months<=60)
* “Trên 5 năm”(Tenure in Months>60)
* Thay đổi giá trị của cột *Number of Dependent* thành 2 giá trị Yes và No.

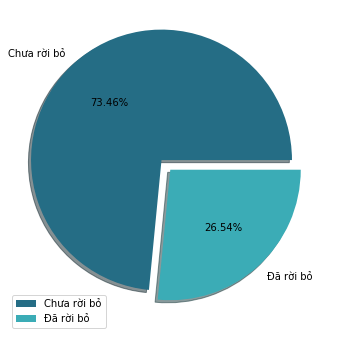
### ***Mô tả cột:***

Dữ liệu gồm 7043 dòng 32 cột.

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Tên cột | Ý nghĩa cột | Giá trị cột |
| 1 | Customer ID | Mã khách hàng |  |
| 2 | Gender | Giới tính khách hàng | “Female”: nữ  “Male”: nam |
| 3 | Age | Tuổi của khách hàng |  |
| 4 | Married | Cho biết nếu khách hàng đã kết hôn | Yes: đã kết hôn  No: chưa kết hôn |
| 5 | Number of Dependents | Cho biết số người phụ thuộc với khách hàng | Yes: Có người phụ thuộc  No: Không có người phụ thuộc |
| 6 | City | Thành phố cư trú của khách hàng |  |
| 7 | Latitude | Vĩ độ của thành phố cư trú của khách hàng |  |
| 8 | Longitude | Kinh độ của thành phố cư trú của khách hàng |  |
| 9 | Tenure in Months | Cho biết tổng số tháng mà khách hàng đã ở với công ty vào cuối quý |  |
| 10 | Phone Service | Khách hàng sử dụng dịch vụ điện thoại | No: Không sử dụng  Yes: Sử dụng |
| 11 | Multiple Lines | Khách hàng có nhiều đường dây hay không | Yes: có nhiều đường dây  No: không có nhiều đường dây |
| 12 | Internet Service | Nhà cung cấp dịch vụ mạng Internet cho khách hàng | Yes: Có dùng dịch vụ mạng Internet  No: Không có dùng dịch vụ mạng Internet |
| 13 | Online Security | Bảo mật trực tuyến | Yes: có bảo mật  No: không có bảo mật |
| 14 | Online Backup | Sao lưu trực tuyến | Yes: có sao lưu trực tuyến  No: không có sao lưu trực tuyến |
| 15 | Device Protection Plan | Bảo vệ thiết bị | Yes: có bảo vệ thiết bị  No: không có bảo vệ thiết bị |
| 16 | Premium Tech Support | Hỗ trợ kỹ thuật | Yes: có Hỗ trợ kỹ thuật  No: không có Hỗ trợ kỹ thuật |
| 17 | Streaming TV | Truyền hình trực tuyến | Yes: có xem truyền hình trực tuyến  No: không có xem truyền hình trực tuyến |
| 18 | Streaming Movies | Xem phim trực tuyến | Yes: có xem phim trực tuyến  No: không xem phim trực tuyến |
| 19 | Streaming Music | Phát nhạc trực tuyến | Yes: có phát nhạc trực tuyến  No: không phát nhạc trực tuyến |
| 20 | Unlimited Data | Dữ liệu không giới hạn | Yes: có trả phí  No: không trả phí |
| 21 | Contract | Hợp đồng theo thời gian | Month-to-month: theo tháng  One year: theo 1 năm  Two year: theo 2 năm |
| 22 | Paperless Billing | Thanh toán không giấy tờ | Yes: thanh toán không giấy tờ  No: không thanh toán không giấy tờ |
| 23 | Payment Method | Phương thức thanh toán | Electronic check: Séc điện tử  Mailed check: Séc được gửi qua thư  Bank transfer (automatic): Chuyển khoản ngân hàng (tự động)  Credit card (automatic): Thẻ tín dụng (tự động) |
| 24 | Monthly Charges | Chi phí theo tháng |  |
| 25 | Total Charges | Tổng chi phí dịch vụ |  |
| 26 | Satisfaction Score | Mức độ hài lòng của khách hàng | Từ 1 (Rất không hài lòng) đến 5 (Rất hài lòng). |
| 27 | Churn Value |  | 1 = khách hàng rời công ty  0 = khách hàng vẫn ở lại với công ty |
| 28 | CLTV | Giá trị trọn đời của khách hàng |  |
| 29 | Churn Category | Danh mục lý do khách hàng rời bỏ | Attitude: thái độ  Competitor: đối thủ cạnh tranh  Dissatisfaction: không hài lòng  Price: giá cả  Other: khác |
| 30 | Age\_Label | Nhóm tuổi | Under 30 years: Tuổi dưới 30  From 30 to 65 years: Tuổi từ 30 đến 65  Over 65 years: Tuổi trên 65 |
| 31 | Month\_Label | Mốc thời gian | Từ 1 - 3 tháng: Đã sử dụng dưới 3 tháng  Từ 1 - 6 tháng: Đã sử dụng từ 3 đến 6 tháng  1 năm: Đã sử dụng 1 năm  2 năm: Đã sử dụng 2 năm  5 năm: Đã sử dụng 5 năm  Trên 5 năm: Đã sử dụng trên 5 năm |
| 32 | Churn\_Label | Trạng thái khách hàng | Chưa rời bỏ: Khách hàng chưa rời bỏ công ty  Đã rời bỏ: Khách hàng đã rời bỏ công ty |

### ***Khai phá dữ liệu:***

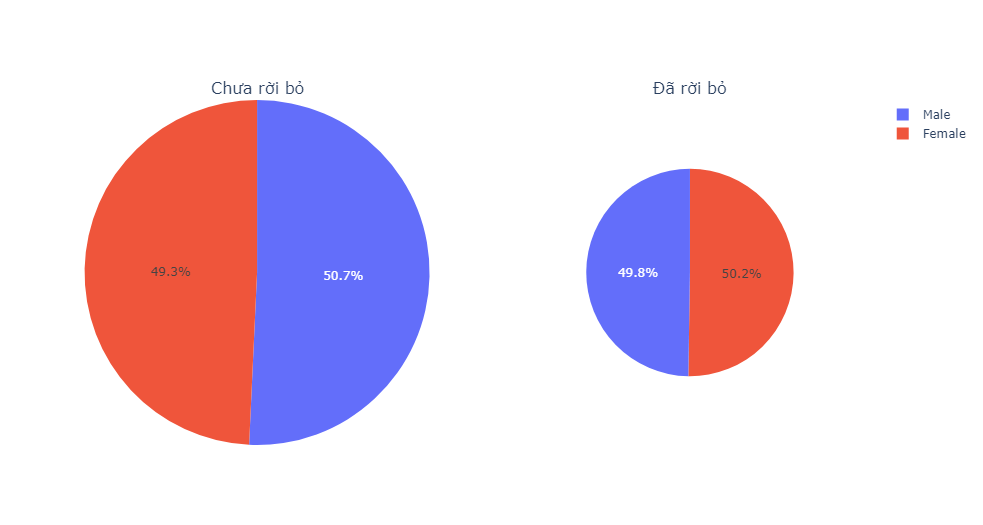
#### *Phân tích các đặc điểm của khách hàng rời bỏ và ở lại:*



* *Số lượng khách hàng rời bỏ chiếm 26,54%, khách hàng chưa rời bỏ chiếm 73,46%. Số lượng khách hàng rời bỏ chiếm gần ¼ tổng số lượng khách hàng.*
* Câu hỏi đặt ra là *“Nhóm khách hàng rời bỏ có đặc điểm gì?”* từ đây đưa ra những phương pháp giúp giữ chân những khách hàng đang có những đặc điểm tương tự.

##### *Nhân khẩu học:*

* Giới tính:

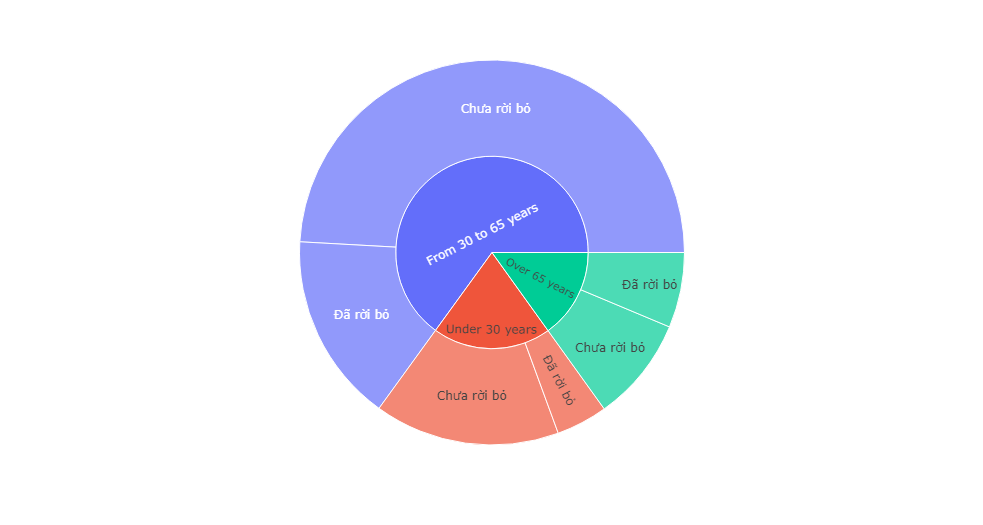


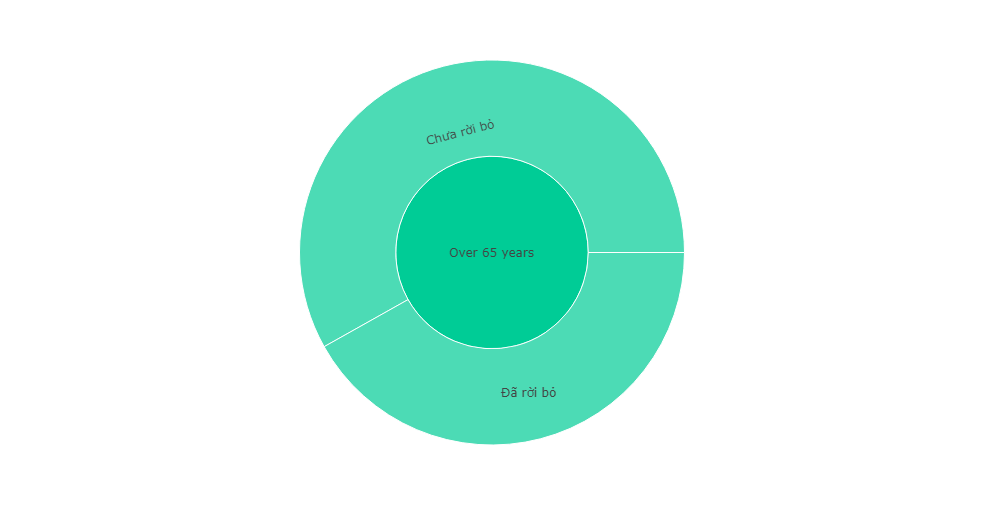
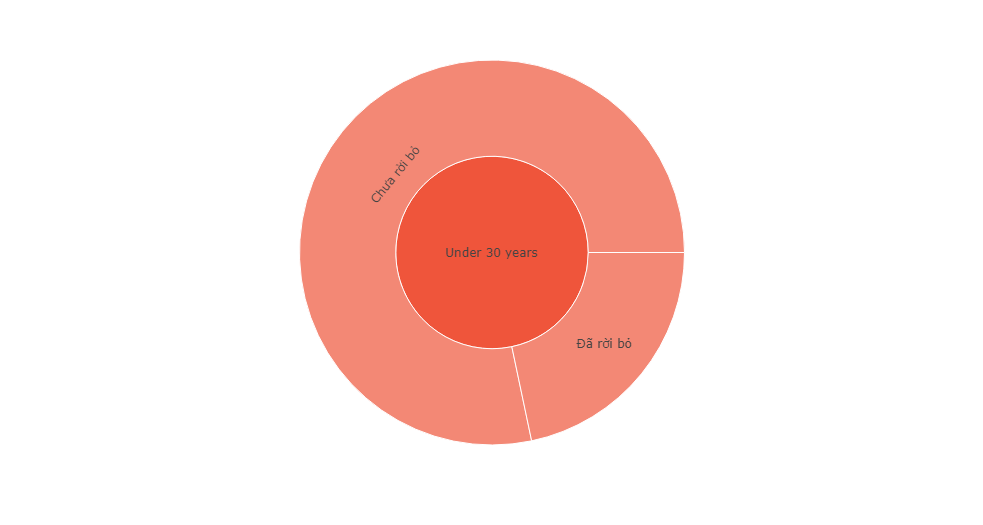
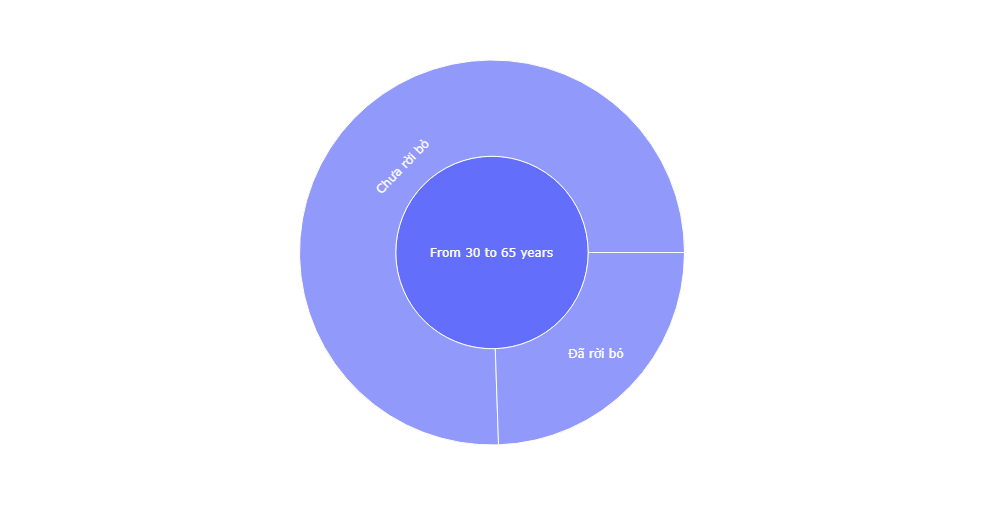
*Theo 2 biểu đồ trên*

* *Nhóm khách hàng “chưa rời bỏ” có số lượng nam nữ chênh lệch 1,5%, trong đó số lượng nam nhiều hơn nữ.*
* *Nhóm khách hàng “đã rời bỏ” có sự chênh lệch không nhiều (chỉ chiếm 0.4%), trong đó số lượng nữ nhiều hơn nam.*

*→ Có thể thấy, giới tính gần như không ảnh hưởng nhiều đến sự rời bỏ của khách hàng.*

* Tuổi:





*Trong cả 3 nhóm tuổi thì khách hàng có độ tuổi từ 30 đến 65 là nhiều nhất. Trong đó:*

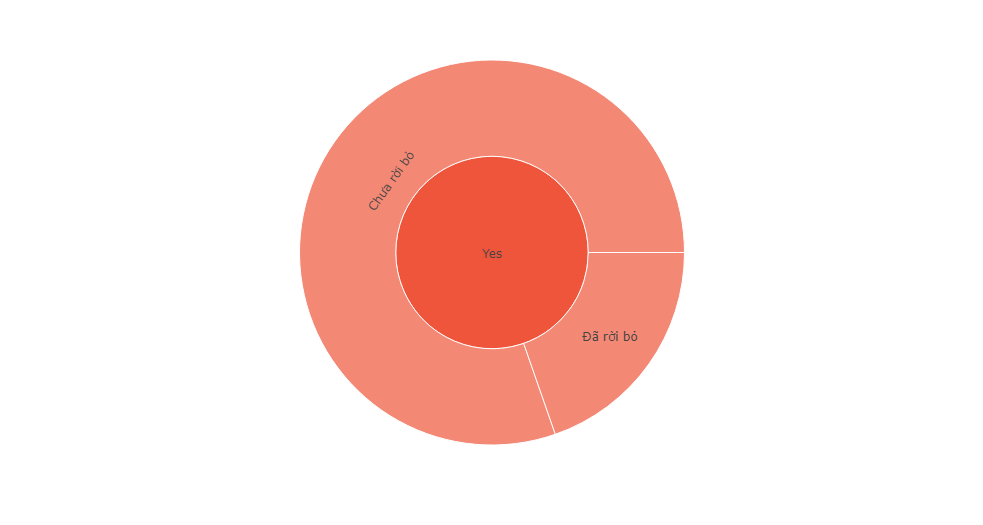
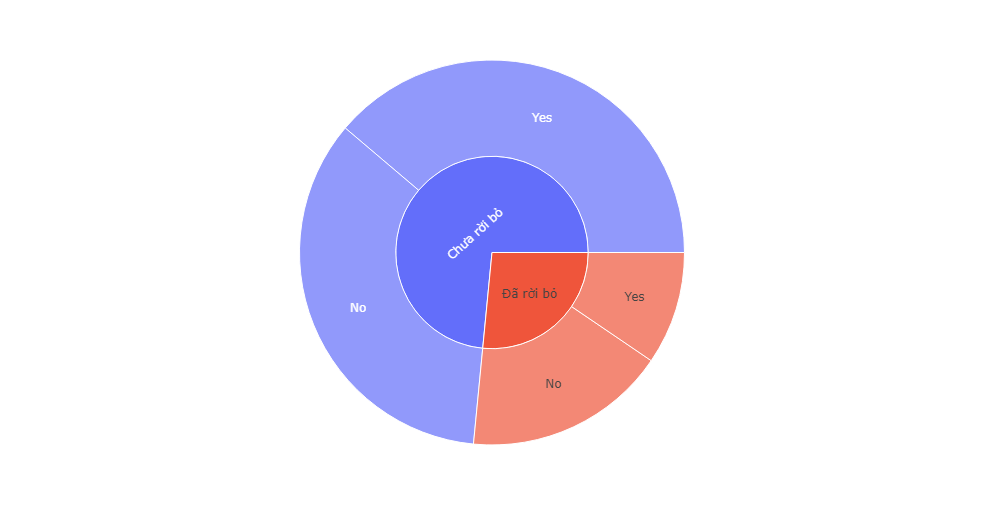
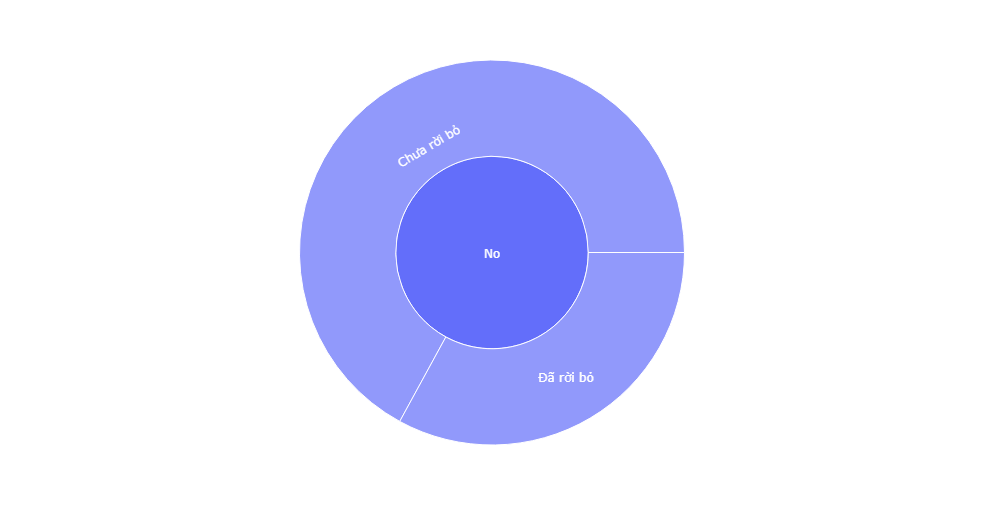
* *Nhóm khách hàng Trên 65 tuổi có tỷ lệ rời bỏ dịch vụ của công ty nhiều hơn so với các nhóm khác . Tỷ lệ này gần bằng 45% tổng số khách hàng ở độ tuổi này.*

*→ Những sản phẩm của công ty là những dịch vụ liên quan đến internet và người lớn tuổi thường có ít nhu cầu sử dụng những dịch vụ này nên có khả năng rời bỏ cao.*

* *Trong khi đó nhóm khách hàng Dưới 30 tuổi và Từ 30 đến 65 tuổi có tỷ lệ khách hàng chưa rời bỏ nhiều hơn đã rời bỏ, nhóm khách hàng đã rời bỏ gần bằng ⅓ số khách hàng ở lại của cả 2 độ tuổi.*

*→ Những người trẻ tuổi thường theo kịp xu hướng, nên việc sử dụng các dịch vụ liên quan đến internet sẽ nhiều hơn.*

* Tình hình hôn nhân:

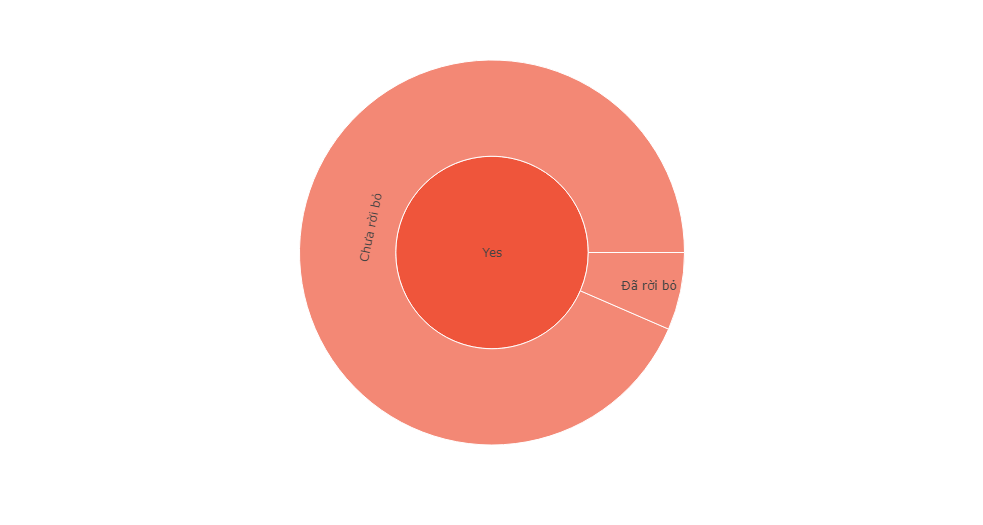
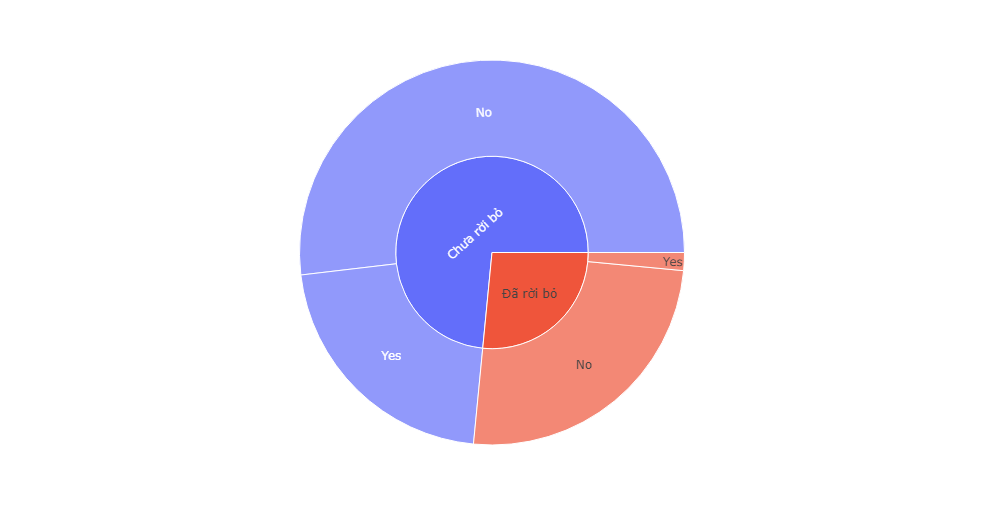
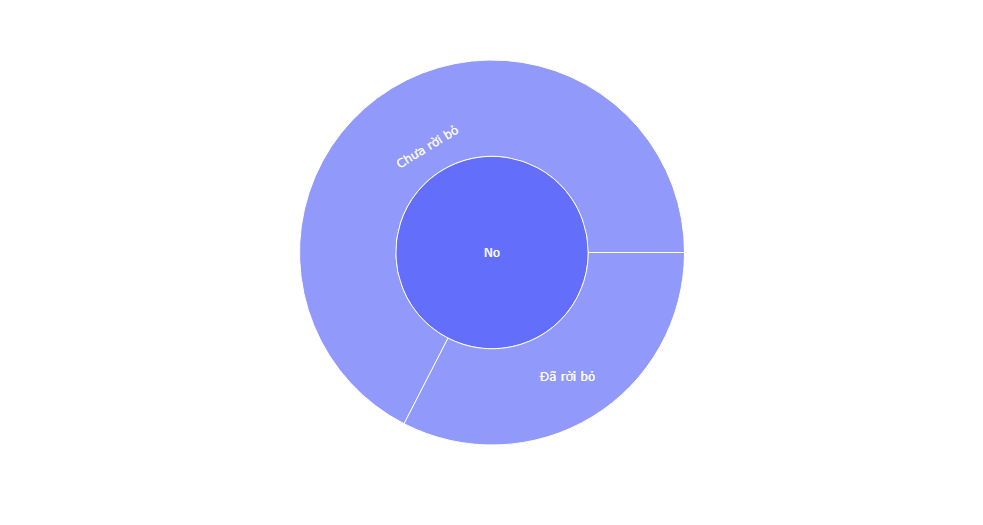


*Tỷ lệ khách hàng đã kết hôn và chưa kết hôn gần như ngang nhau. Trong đó:*

* *Tỷ lệ khách hàng rời bỏ chiếm gần ⅓ trong nhóm khách hàng chưa kết hôn.*
* *Tỷ lệ khách hàng rời bỏ chiếm gần ⅕ trong nhóm khách hàng đã kết hôn.*

*→ Khách hàng đã kết hôn ít rời bỏ công ty hơn vì nhóm này thường thích ổn định nên sẽ ít thay đổi những dịch vụ đang sử dụng.*

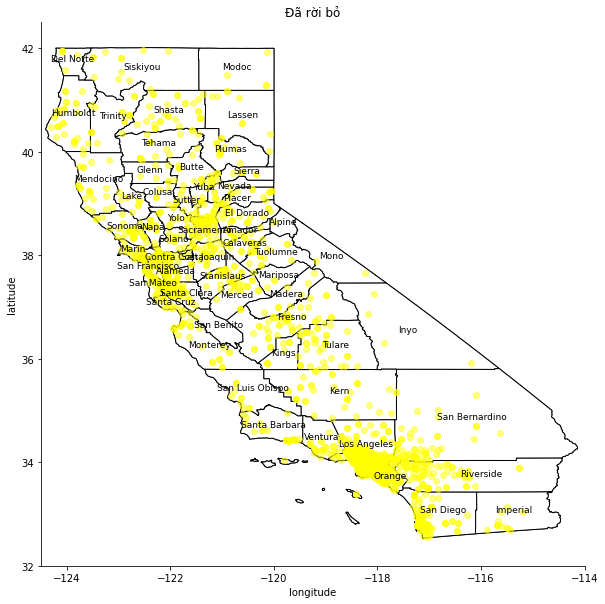
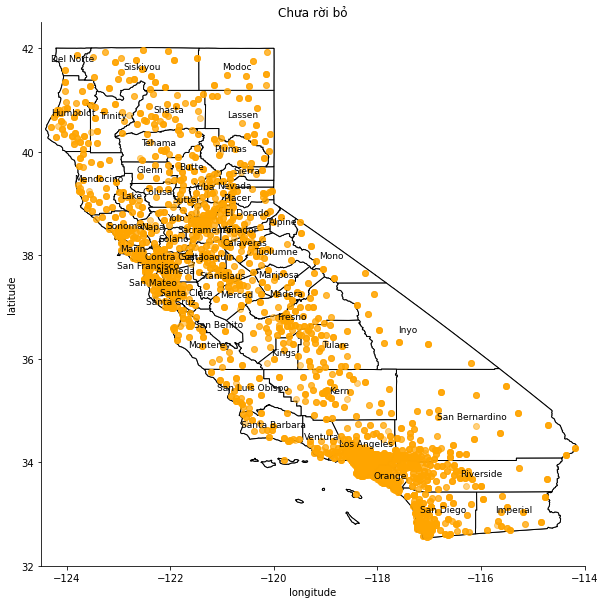
* Số người phụ thuộc:



* *Trong nhóm khách hàng chưa rời bỏ, nhóm khách hàng không có người phụ thuộc chiếm 70.6% trong khi đó, ở nhóm khách hàng đã rời bỏ, nhóm khách hàng không có người phụ thuộc chiếm 94.33%.*
* *Trong nhóm khách hàng không có người phụ thuộc, nhóm khách hàng chưa rời bỏ chiếm 67,45% trong khi đó, ở nhóm khách hàng có người phụ thuộc, nhóm khách hàng chưa rời bỏ chiếm 93.48%.*

*→ Nhóm khách hàng chưa rời bỏ dịch vụ hầu hết là khách hàng có người phụ thuộc.*

* Nơi ở của khách hàng:



* *Ở cả hai nhóm, khách hàng của công ty đều trải dài khắp tất cả các khu vực của California, tập trung nhiều nhất là ở Orange, San Diego, Los Angeles,...*

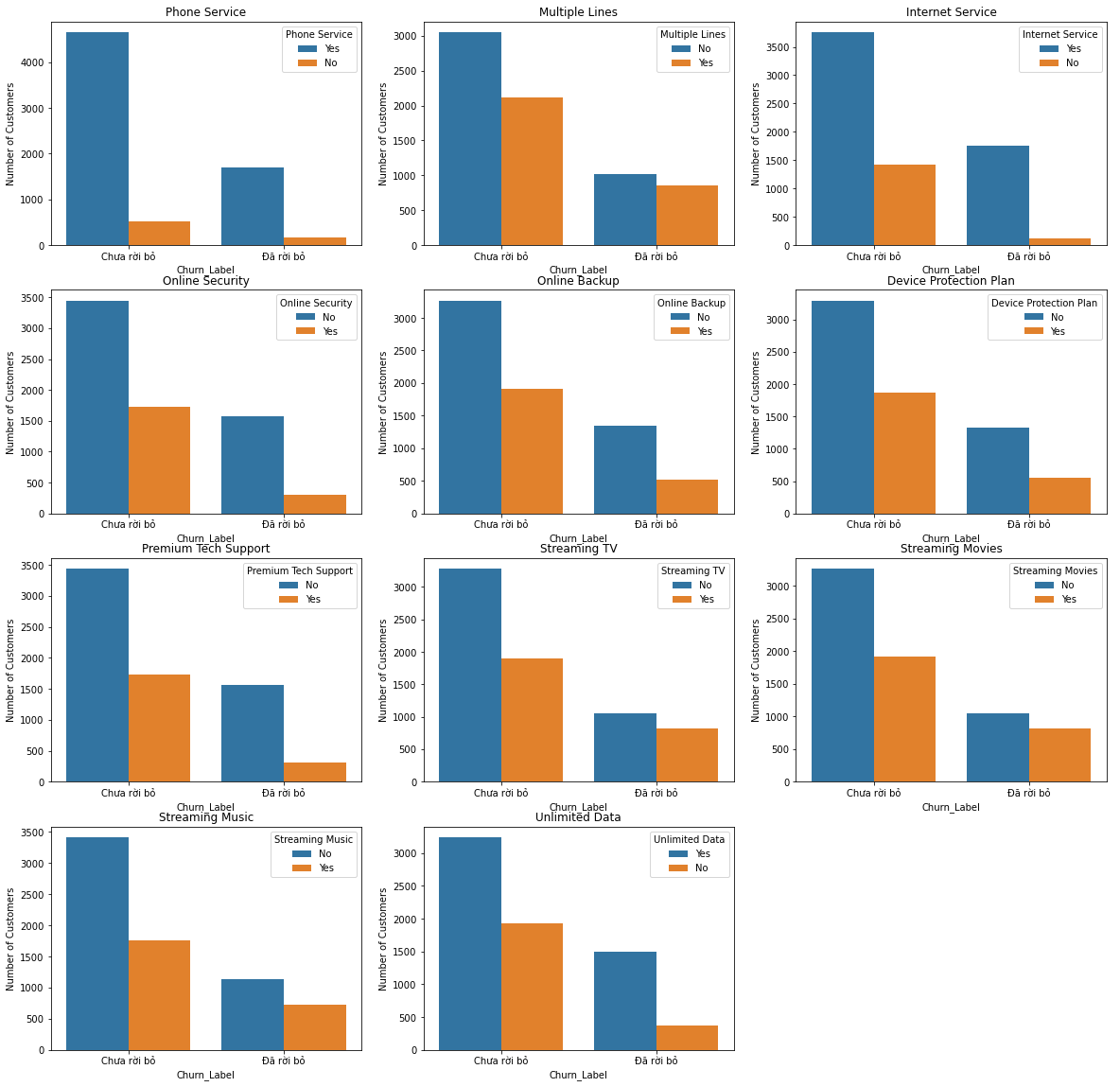
*→ Khách hàng ở khu vực nào càng nhiều thì số lượng khách hàng rời bỏ ở các khu vực đó cũng nhiều hơn so với các nơi khác → không có gì được xem là bất thường.*

*→ Yếu tố nơi ở không ảnh hưởng đến quyết định rời bỏ của khách hàng.*

* **Kết luận:** Trong các yếu tố phân tích về nhân khẩu học thì giới tính, nơi ở không ảnh hưởng đến quyết định rời bỏ của khách hàng. Còn lại là các yếu tố khác như tuổi, tình trạng hôn nhân và số người phụ thuộc:
* Nhóm khách hàng đã rời bỏ thì thường là những khách hàng trên 65 tuổi.
* Nhóm khách hàng chưa rời bỏ thì sẽ có độ tuổi dưới 65, những khách hàng đã kết hôn và có người phụ thuộc.

##### *Dịch vụ:*

* Những dịch vụ của công ty:



* *Nhóm khách hàng chưa rời bỏ:*
* *Khách hàng đăng ký sử dụng dịch vụ Phone Service nhiều nhất (hơn 4500 khách hàng).*
* *Tiếp theo là các dịch vụ: Internet Service, Online Security, Premium Tech Support với số khách hàng đăng ký sử dụng là hơn 3500 khách hàng.*
* *Và cuối cùng là các dịch vụ: Multiple Lines, Online Backup, Device Protection, Streaming TV, Streaming Movies, Streaming Music, Unlimited Data với số khách hàng là hơn 3000 khách hàng.*

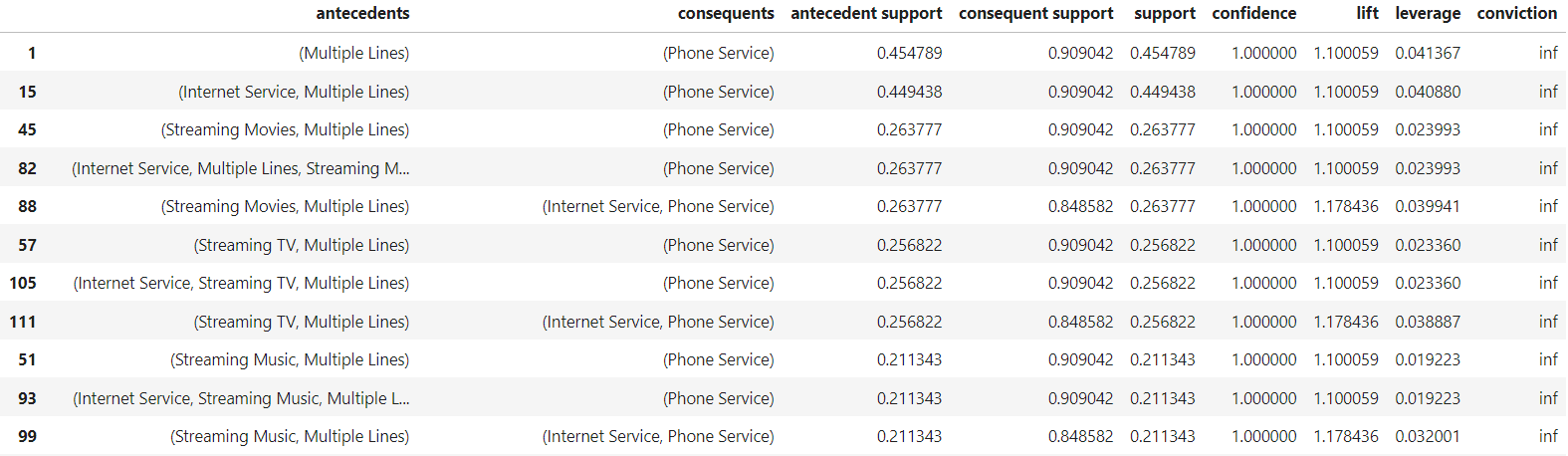
*→ Phone Service được sử dụng nhiều nhất vì độ tiện lợi của nó ở hiện tại nên số khách hàng không sử dụng dịch vụ này thường rất ít, các dịch vụ khác có số khách hàng không sử dụng dịch vụ gần như bằng nhau (dao động từ 1500 - 2000 khách hàng).*

* *Nhóm khách hàng đã rời bỏ:*
* *Số lượng khách hàng sử dụng các dịch vụ chênh lệch không nhiều.*
* *Số lượng khách hàng không sử dụng các dịch vụ luôn dưới 800 khách hàng.*

*→ Phone Service, Internet Service vẫn là những dịch vụ được ưa chuộng, số lượng khách hàng không sử dụng dịch vụ ở nhóm này vẫn là ít nhất.*

***⇒*** *Việc phân tích riêng lẻ các dịch vụ không thể tìm ra những đặc điểm nổi bật giữa 2 nhóm khách hàng. Vì vậy nhóm đã quyết định sử dụng Luật kết hợp để khai thác những quy luật về hành vi lựa chọn dịch vụ của khách hàng.*

* Sử dụng phương pháp luật kết hợp Association Rules:
* *Đối với khách hàng đã rời bỏ:*

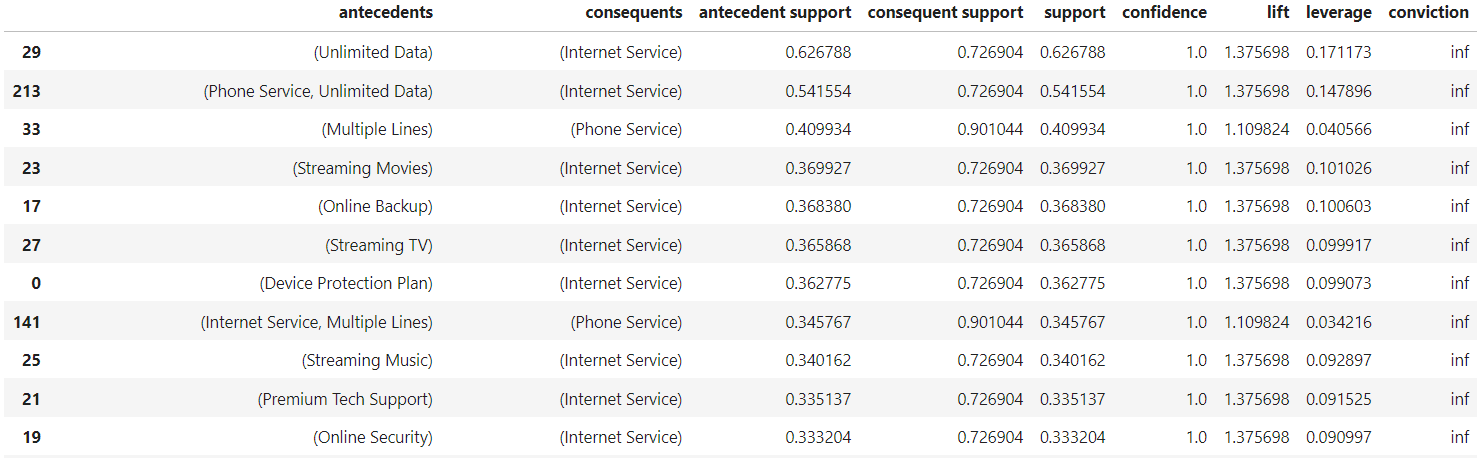
**

*Ta có thể thấy được sự kết hợp của các dịch vụ trong nhóm khách hàng đã rời bỏ thông qua các chỉ số như support, confidence, lift.*

* *Dịch vụ Multiple Lines và Phone Service có tần suất xuất hiện cùng nhau là hơn 45.47% trong tổng các giao dịch của khách hàng. Trong khi đó, khi khách hàng sử dụng các dịch vụ Multiple Lines thì 100% sẽ sử dụng thêm dịch vụ Phone Service bởi vì chỉ số confidence là 1.0. Với chỉ số lift là 1.1(>1.0) thì ta có thể xác định 2 loại dịch vụ này đi cùng với nhau theo một quy luật chứ không phải chỉ là sự trùng hợp ngẫu nhiên.*
* *Khi sử dụng các dịch vụ như Internet Service, Multiple Lines thì sẽ sử dụng dịch vụ Phone Service là 100%, trong khi đó tần suất xuất hiện chung của các dịch vụ này là 44.94%. Với chỉ số lift là 1.1(>1.0) thì ta có thể xác định được các loại dịch vụ này đi cùng với nhau theo một quy luật chứ không phải chỉ là sự trùng hợp ngẫu nhiên.*
* *Khi sử dụng các dịch vụ như Streaming TV, Multiple Lines thì sẽ sử dụng dịch vụ Internet Service, Phone Service là 100%, trong khi đó tần suất xuất hiện chung của các dịch vụ này là 25.68%. Với chỉ số lift là 1.178(>1.0) thì ta có thể xác định được các loại dịch vụ này đi cùng với nhau theo một quy luật chứ không phải chỉ là sự trùng hợp ngẫu nhiên.*

*→ Với những chỉ số như trên ta có thể thấy được một vài quy luật cụ thể của nhóm khách hàng Đã rời bỏ khi lựa chọn sử dụng các dịch vụ của công ty.*

* *Khi khách hàng sử dụng các dịch vụ Multiple Lines thì kèm theo đó khách hàng sẽ sử dụng dịch vụ Phone Service của công ty.*
* *Khách hàng sử dụng dịch vụ như Internet Service, Multiple Lines thì sẽ sử dụng luôn dịch vụ Phone Service vì nó cũng có sự hỗ trợ nhau trong quá trình sử dụng.*
* *Hay là các loại dịch vụ như Streaming TV, Multiple Lines thì sẽ luôn có Internet Service, Phone Service đi kèm. Bởi vì muốn sử dụng những loại dịch vụ trên thì ta cần phải có Internet, hay các đường dây điện thoại nên việc chúng đi cùng nhau chính là một quy luật.*
* *Đối với khách hàng chưa rời bỏ:*

**

*Ta có thể thấy được sự kết hợp của các dịch vụ trong nhóm khách hàng chưa rời bỏ thông qua các chỉ số như support, confidence, lift.*

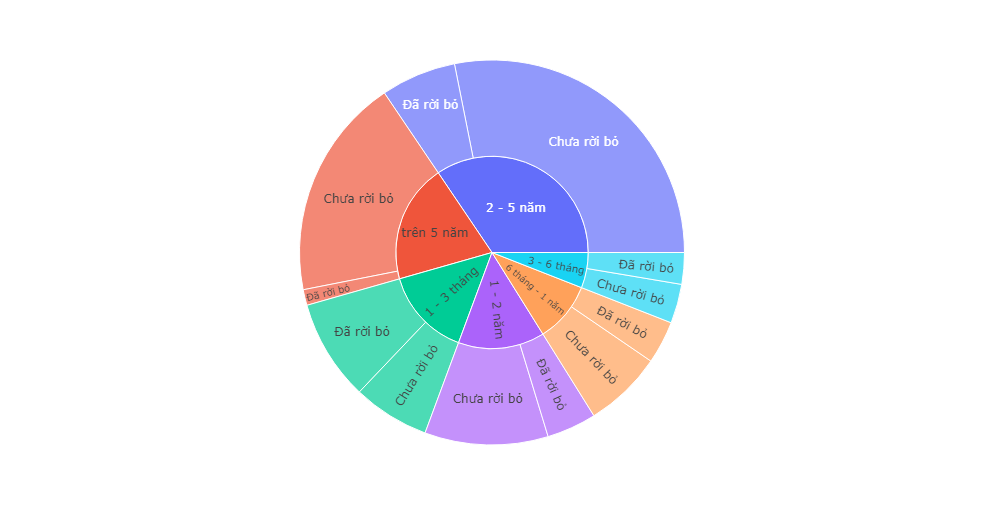
* *Dịch vụ Unlimited Data và Internet Service có tần suất xuất hiện cùng nhau là 62.7% trong tổng các giao dịch của khách hàng. Trong khi đó, khi khách hàng sử dụng dịch vụ Unlimited Data thì 100% sẽ sử dụng thêm dịch vụ Internet Service.Với chỉ số lift là 1.37 thì ta có thể xác định 2 loại dịch vụ này đi cùng với nhau theo một quy luật chứ không phải chỉ là sự trùng hợp ngẫu nhiên.*
* *Dịch vụ Phone Service, Unlimited Data và Internet Service có tần suất xuất hiện cùng nhau là gần 54.2% trong tổng các giao dịch của khách hàng. Trong khi đó, khi khách hàng sử dụng dịch vụ Phone Service, Unlimited Data thì 100% sẽ sử dụng thêm dịch vụ Internet Service. Với chỉ số lift là 1.37 thì ta có thể xác định 2 loại dịch vụ này đi cùng với nhau theo một quy luật chứ không phải chỉ là sự trùng hợp ngẫu nhiên.*
* *Dịch vụ Multiple Line và Phone Service có tần suất xuất hiện cùng nhau là gần 41% trong tổng các giao dịch của khách hàng. Trong khi đó, khi khách hàng sử dụng dịch vụ Multiple Line thì 100% sẽ sử dụng thêm dịch vụ Phone Service. Với chỉ số lift là 1.11 thì ta có thể xác định 2 loại dịch vụ này đi cùng với nhau theo một quy luật chứ không phải chỉ là sự trùng hợp ngẫu nhiên.*

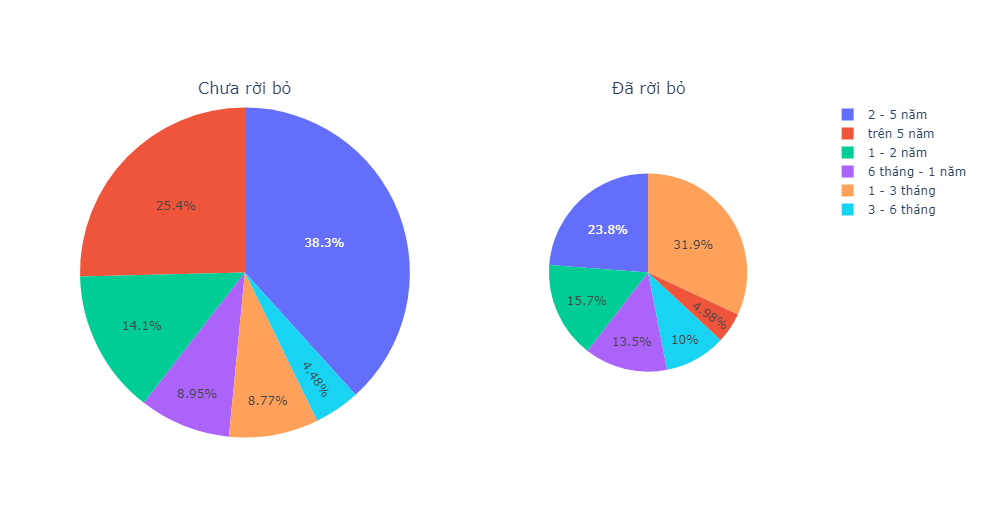
*→ Với những chỉ số như trên ta có thể thấy được một vài quy luật cụ thể khi khách hàng Chưa rời bỏ lựa chọn sử dụng dịch vụ của công ty.*

* *Khách hàng sử dụng Unlimited Data thì sẽ sử dụng Internet Service.*
* *Khách hàng sử dụng Phone Service, Unlimited Data thì sẽ sử dụng Internet Service.*
* *Khách hàng sử dụng Multiple Line thì sẽ sử dụng Phone Service.*

*⇒ Khách hàng rời bỏ thường có xu hướng sử dụng Multiple Lines hết hợp với một số dịch vụ khác như Phone Service, Internet Service…Khách hàng chưa rời bỏ thường có xu hướng sử dụng Unlimited Data kết hợp với một số dịch vụ khác như Phone Service và Internet Service, …*

* Số tháng sử dụng:

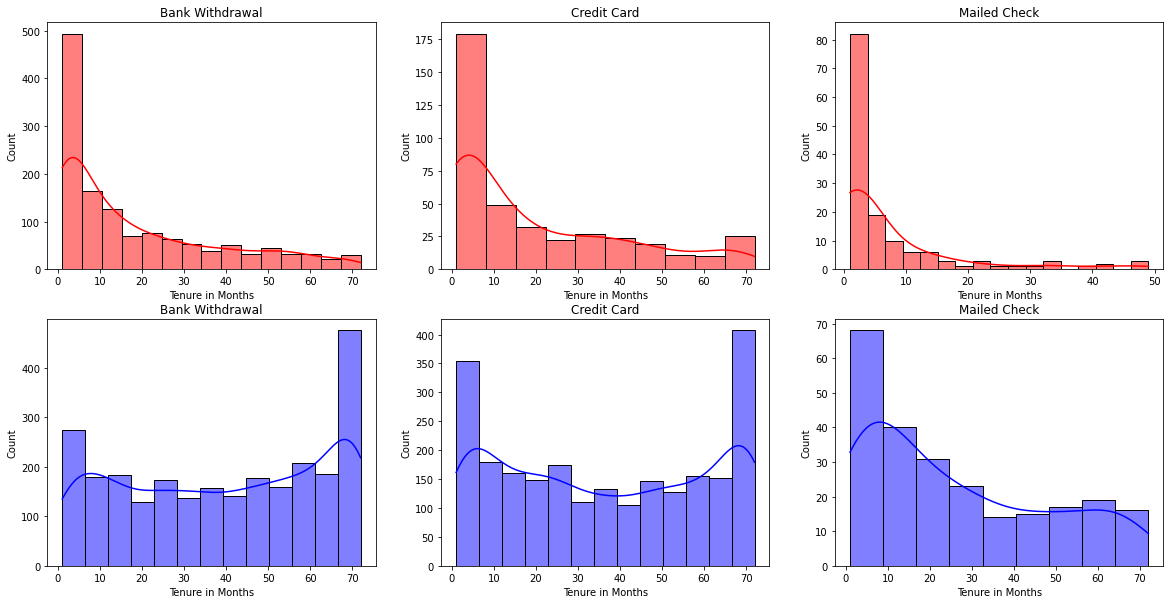




* *Nhóm khách hàng “chưa rời bỏ”:*
* *Nhóm khách hàng gắn bó từ 2-5 năm chiếm tỷ lệ lớn nhất tiếp đến là khách hàng gắn bó trên 5 năm với tỷ lệ 25.4%.*
* *Nhóm khách hàng gắn bó từ 1-2 năm chiếm 14,1%.*
* *Nhóm khách hàng mới sử dụng dịch vụ của công ty từ 3 - 6 tháng chiếm tỷ lệ thấp nhất.*
* *Nhóm khách hàng “đã rời bỏ”:*
* *Nhóm khách hàng gắn bó từ 1 - 3 tháng chiếm tỷ lệ nhiều nhất với 31,9%.*
* *Nhóm khách hàng gắn bó trên 5 năm chiếm tỷ lệ ít nhất với 4,98%.*

*→ Nhận thấy khách hàng gắn bó lâu dài với công ty có khả năng ở lại rất cao. Đồng thời, khách hàng có thời gian sử dụng dịch vụ càng ngắn thì khả năng rời bỏ càng cao.*

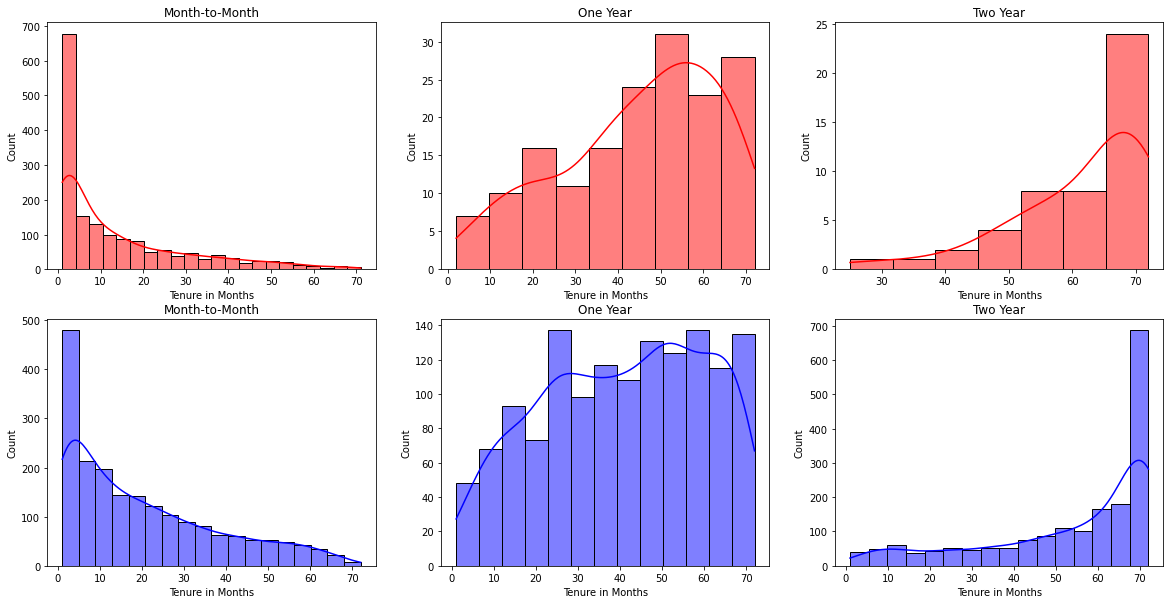
* Phương thức thanh toán:



* *Nhóm khách hàng chưa rời bỏ:*
* *Số lượng khách hàng sử dụng các phương thức thanh toán không biến động nhiều qua mỗi tháng sử dụng dịch vụ.*
* *Số lượng khách hàng sử dụng phương thức thanh toán là chuyển khoản ngân hàng hoặc trả bằng thẻ tín dụng cao hơn nhiều so với Mailed Check. Chiếm số lượng nhiều nhất khi khách hàng sử dụng dịch vụ trong 10 tháng đầu và từ tháng 70.*
* *Số lượng khách hàng sử dụng dịch vụ Mailed Check ít hơn rất nhiều, cao nhất chỉ ở mức 70 khách hàng trong 10 tháng đầu.*
* *Nhóm khách hàng đã rời bỏ:*
* *Nhìn chung, số lượng khách hàng sử dụng cả 3 phương thức thanh toán trong 3 tháng đầu rất cao, sau đó giảm mạnh vào tháng tiếp theo. Sau 10 tháng, số lượng khách hàng biến động theo xu hướng giảm dần khi số tháng sử dụng tăng lên.*
* *Số lượng khách hàng sử dụng phương thức thanh toán chuyển khoản ngân hàng cao hơn nhiều so với chuyển khoản bằng thẻ và Mailed Check.*

*→ Những khách hàng sử dụng phương thức thanh toán tự động (Chuyển khoản ngân hàng, Thẻ tín dụng) có niềm tin và sự tin tưởng nhiều hơn vào công ty, có xu hướng sử dụng dịch vụ lâu hơn so với khách hàng sử dụng hình thức thanh toán khác.*

* Thời hạn ký hợp đồng.



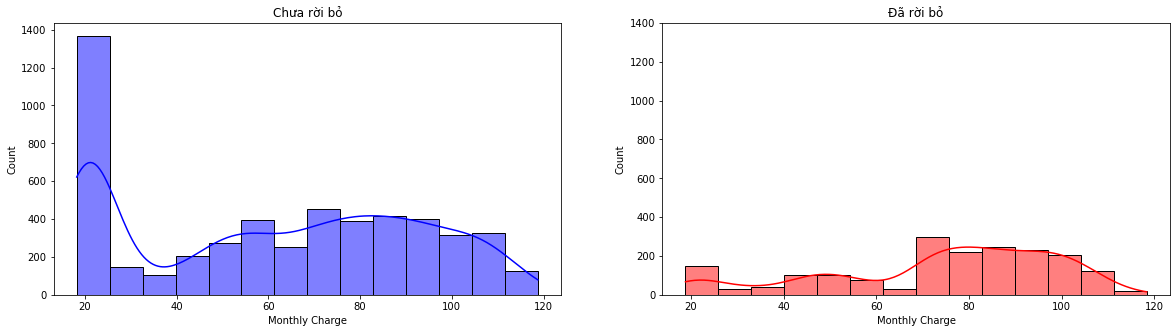
* *Nhóm khách hàng rời bỏ:*
* *Nhìn chung, số lượng khách hàng rời bỏ ký hợp đồng Month-to-Month cao hơn rất nhiều so với hai loại hợp đồng dài hạn khác.*
* *Đối với hợp đồng Month-to-Month, số khách hàng sử dụng lên đến 700 người, sau đó giảm mạnh cho đến sau 3 tháng sử dụng rồi biến động theo xu hướng giảm khi thời gian sử dụng tăng lên.*
* *Số lượng khách hàng khi ký hợp đồng One Year và Two Year biến động rất nhỏ khi thời gian sử dụng tăng lên.*

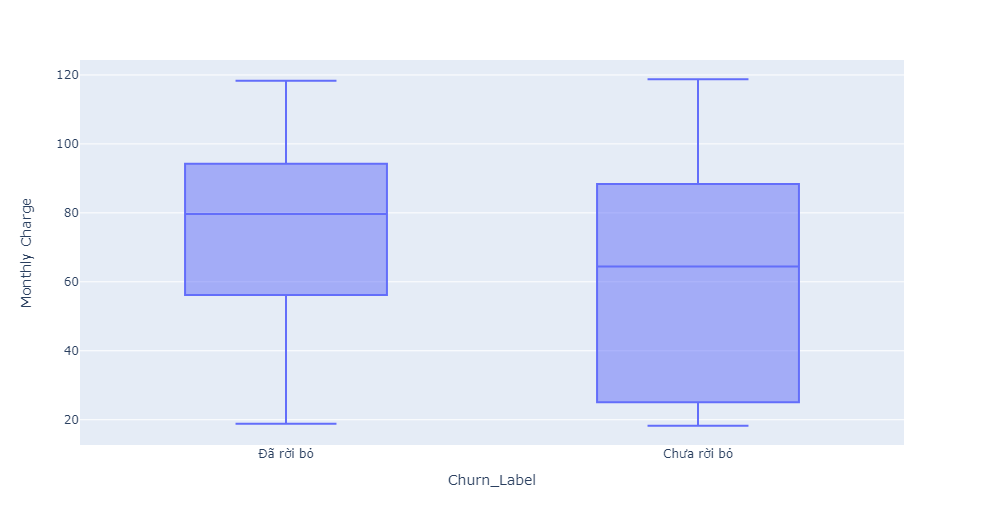
*→ Khách hàng mới sử dụng dịch vụ thường có xu hướng ký hợp đồng Month-to-Month. Khách hàng ký hợp đồng Month-to-Month có khả năng rời bỏ cao.*

* *Nhóm khách hàng chưa rời bỏ:*
* *Với loại hợp đồng Month-to-Month thì có số khách hàng giảm dần khi số tháng sử dụng dịch vụ tăng lên.*
* *Nhóm khách hàng ký hợp đồng One Year có số khách hàng không nhiều bằng 2 nhóm còn lại, thường không có sự chênh lệch rõ rệt. Từ trên 10 tháng có số khách hàng dường như gần bằng nhau.*
* *Nhóm khách hàng ký hợp đồng Two Year thường sẽ có thời gian sử dụng dịch vụ rất lâu, đa số là trên 5 năm.*

*→ Biểu đồ cho thấy những khách hàng ký hợp đồng dài hạn có khả năng sử dụng dịch vụ trong thời gian dài hơn và ít rời bỏ hơn.*

* *Chi phí từng tháng:*

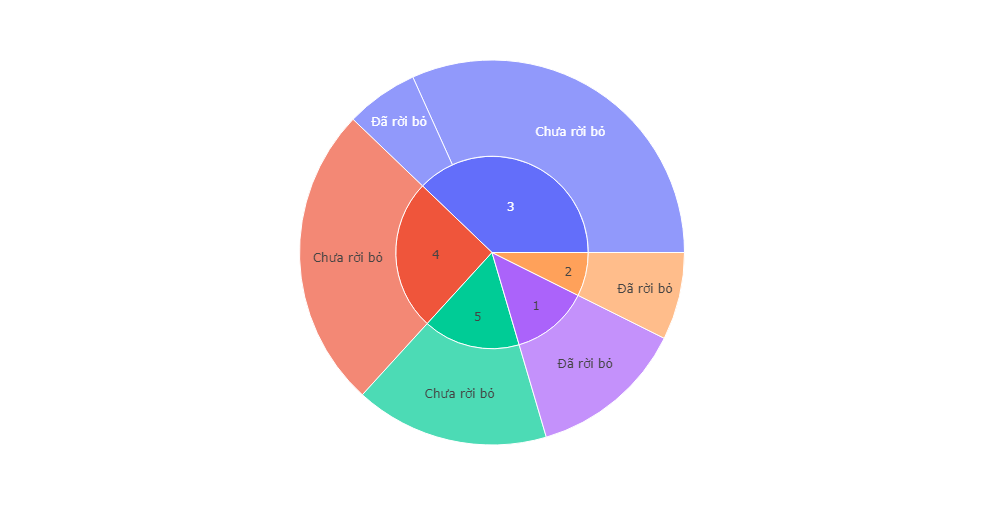




* *Khách hàng đã rời bỏ:*
* *Mức chi trả hàng tháng dao động từ 18,85 - 118,35$, giá trị trung vị chi trả hàng tháng rơi vào khoảng 79,65$.*
* *Chi phí thường dao động trong khoảng 56,15 - 94,21$ và có xu hướng lệch về bên dưới giá trị trung vị.*
* *Số lượng khách hàng thấp (khoảng 100) và dao động nhỏ ở mức 20 - 60$.*
* *Khách hàng chưa rời bỏ:*
* *Mức chi trả hàng tháng dao động từ 18,25 - 118,75$, giá trị trung vị chi trả hàng tháng rơi vào khoảng 64,425$.*
* *Chi phí thường giao động trong khoảng 25,1 - 88,4$ và có xu hướng lệch nhiều về bên dưới giá trị trung vị.*
* *Số lượng khách hàng chi trả ở mức 20 cao đột biến (gần 1400) và không biến động nhiều ở mức 50 - 110$ (300 - 400).*

*→ Khách hàng rời bỏ có giá trị trung vị chi trả hàng tháng và mức chi trả dao động cao hơn khách hàng chưa rời bỏ.*

* Mức độ hài lòng dịch vụ:



*Mức độ hài lòng của khách hàng được tính theo thang đo Likert 5 mức độ:*

*1: Rất không hài lòng*

*2: Không hài lòng*

*3: Bình thường*

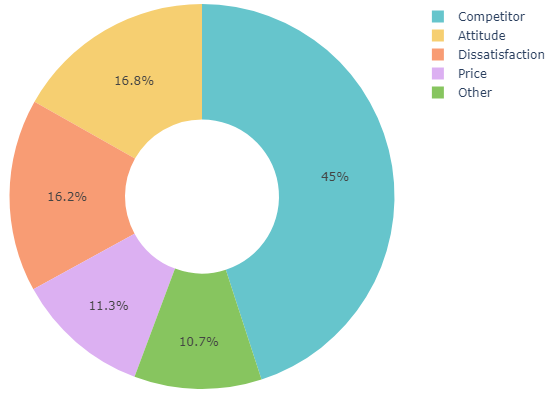
*4: Hài lòng*

*5: Rất hài lòng*

* *Nhóm khách hàng chưa rời bỏ có mức độ hài lòng là 3,4,5.*
* *Nhóm khách hàng đã rời bỏ có mức độ hài lòng là 1,2,3.*
* *Ở mức hài lòng bình thường, số lượng khách hàng rời bỏ (429) chiếm khoảng ⅙ số lượng khách hàng chưa rời bỏ (2236).*

*→ Nhóm khách hàng rời bỏ thường có mức độ hài lòng về dịch vụ dưới mức trung bình. Ngược lại, nhóm khách hàng chưa rời bỏ hầu hết có mức độ hài lòng về dịch vụ ở trên mức trung bình.*

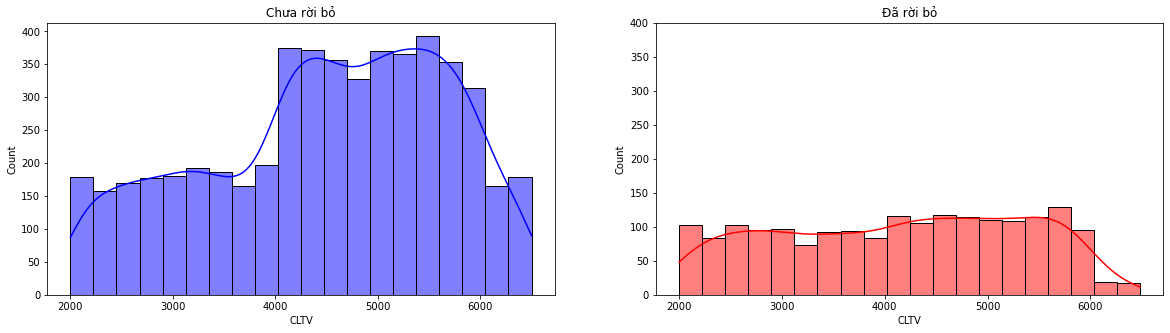
* Lý do rời bỏ:



*Trong các lý do rời bỏ:*

* *Đối thủ cạnh tranh (chiếm 45%) là nguyên nhân lớn nhất ảnh hưởng đến sự rời bỏ của khách hàng. Khách hàng cho rằng dịch vụ của đối thủ cạnh tranh tốt hơn bao gồm các điều khoản, ưu đãi, thiết bị và tốc độ download .*
* *Hơn 15% khách hàng không hài lòng về thái độ phục vụ từ phía nhà cung cấp, người hỗ trợ.*
* *Hơn 15% khách hàng không hài lòng về dịch vụ chẳng hạn như phạm vi hạn chế dịch vụ, hỗ trợ trực tuyến kém hiệu quả, độ tin cậy của mạng, ...*
* *Yếu tố giá chiếm tỷ lệ thấp nhất với 11,3%, khách hàng cho rằng giá quá cao, cước đường dài, phí dữ liệu bổ sung.*
* **Kết luận:** 
  + Khách hàng chưa rời bỏ:
* Có xu hướng sử dụng Unlimited Data kết hợp với một số dịch vụ khách như Phone Service và Internet Service.
* Khách hàng có thời gian sử dụng lâu trên 2 năm.
* Đa phần ký hợp đồng One Year và Two Year.
* Chi phí hàng tháng thường dao động trong khoảng 56,15 - 94,21$, trung vị 79,65$.
* Có mức độ hài lòng về dịch vụ trên mức trung bình.
  + Khách hàng đã rời bỏ:
* Đa phần đều sử dụng dịch vụ Multiple Lines kết hợp với một số dịch vụ khách như Internet Service.
* Đa phần có thời gian sử dụng dịch vụ ngắn dưới 1 năm.
* Thường là khách hàng mới sử dụng dịch vụ (thời gian sử dụng dưới 10 tháng) và ký hợp đồng Month-to-Month.
* Chi phí thường giao động trong khoảng 25,1 - 88,4$, trung vị 64,425$.
* Có mức độ hài lòng dưới mức trung bình.
* Nguyên nhân rời bỏ:
* Đối thủ cạnh tranh (45%): Tốt hơn về điều khoản, ưu đãi, thiết bị và tốc độ download.
* Thái độ phục vụ (hơn 15%): Thái độ phục vụ không tốt bên phía nhà cung cấp và người hỗ trợ.
* Không hài lòng với dịch vụ (hơn 15%): chẳng hạn như phạm vi hạn chế dịch vụ, hỗ trợ trực tuyến kém hiệu quả, độ tin cậy của mạng, ...
* Giá (hơn 10%): giá quá cao, cước đường dài, phí dữ liệu bổ sung.
* Giá trị trọn đời:

*Thể hiện lợi nhuận ròng mà bạn dự đoán sẽ có thể ghi nhận cho một khách hàng cụ thể trong khoảng thời gian giữ mối quan hệ với họ.*



* *Nhóm khách hàng đã rời bỏ dường như không chênh lệch nhiều ở từng mức giá trị, tuy nhiên ở mức giá trị trọn đời là 6000 trở lên thì ít đột ngột*
* *Nhóm khách hàng chưa rời bỏ có số khách hàng ở từng mức giá trị nhiều hơn ở nhóm đã rời bỏ. Từ mức giá trị từ 4000 - 6000 lại cao hơn rất nhiều so với các mức giá trị khác.*

*→ Nhóm khách hàng chưa rời bỏ có giá trị trọn đời cao hơn nhóm đã rời bỏ.*

*→ Những khách hàng có giá trị trọn đời cao là những khách hàng trung thành của công ty, có thời gian sử dụng dịch vụ lâu dài. Cũng chính nhóm khách hàng này sẽ góp phần đem lại lợi nhuận lâu dài và bền vững cho công ty.*

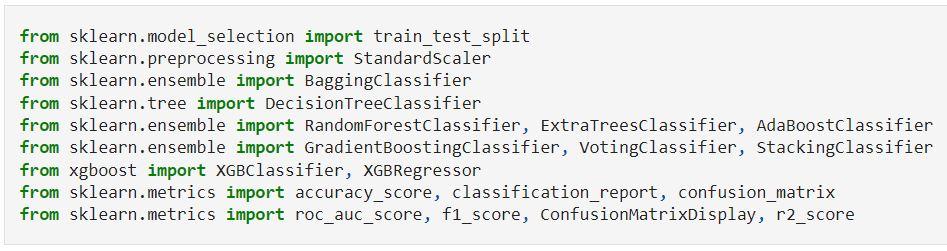
##### *Giải pháp của doanh nghiệp:*

* Với mục đích tối đa hóa lợi nhuận đồng nghĩa với việc công ty phải sở hữu một tập khách hàng lớn. Trong khi đó, quá trình thu hút khách hàng mới là một quá trình lâu dài và gian khổ. Từ tiếp xúc đầu tiên đến phễu tiếp thị và sau đó chuyển sang phễu bán hàng cho đến giao dịch, yêu cầu đến sự tham gia của nhiều bộ phận chưa kể đến công sức và thời gian để thuyết phục khách hàng. Do đó, giữ chân khách hàng là giải pháp tối ưu và yêu cầu ít nguồn lực hơn.
* Việc cần làm của doanh nghiệp là xem xét những đặc điểm nổi bật của 2 nhóm khách hàng Đã rời bỏ và Chưa rời bỏ được tìm ra ở trên, sau đó đưa ra những giải pháp hợp lý để có thể giữ chân những khách hàng đang có dấu hiệu rời bỏ.
* Giải pháp giữ chân nhóm khách hàng:
* Về dịch vụ:
* Xem xét cải tiến chất lượng dịch vụ Multiple Lines.
* Cải tiến nâng cấp thiết bị, các yếu tố như tốc độ truy cập và tốc độ tải, giới hạn dịch vụ.
* Cải thiện các vấn đề bảo mật.
* Đào tạo một đội ngũ chuyên xử lý các vấn đề lỗi kỹ thuật, lỗi mạng khi nhận được phản hồi từ khách hàng một cách nhanh chóng, kịp thời và hiệu quả. Đảm bảo dịch vụ của công ty luôn ở trạng thái tốt nhất.
* Về các chương trình quảng cáo:
* Đưa ra các gói Combo khuyến mãi tích hợp các dịch vụ:
* Combo 1: *Unlimited Data, Internet Service*
* Combo 2: *Phone Service, Unlimited Data, Internet Service*
* Combo 3: *Multiple Line, Phone Service*
* Khuyến khích khách hàng ký hợp đồng One Year, Two Year và thanh toán số bằng cách đưa ra các ưu đãi thú vị như khuyến mãi, giảm giá, chiết khấu tặng kèm.
  + Tổ chức các sự kiện tri ân và tặng quà cho khách hàng gắn bó lâu năm.
* Hỗ trợ khách hàng các loại phí như cước đường dài hay phí dữ liệu bổ sung bằng cách đưa ra các gói tháng bổ sung dữ liệu và gói tháng cho các cuộc gọi đường dài.
* Nhân viên:
* Mở các khóa học, các lớp training cho nhân viên về kỹ năng, hiểu biết, những bài học về thái độ đối với khách hàng nhằm mục đích nâng cao sự chuyên nghiệp và thái độ phục vụ của nhân viên.
* Lắng nghe phàn nàn của khách hàng để phản hồi với công ty giúp cải thiện dịch vụ theo tiêu chí làm hài lòng khách hàng.
  + Phân chia khách hàng theo giá trị trọn đời(CLTV):

Đối với Nhóm khách hàng có giá trị trọn đời cao: Cần phải gia tăng lòng trung thành của nhóm này bằng cách tổ chức các chiến dịch đánh mạnh vào đối tượng này, đưa ra những ưu đãi đặc biệt dành riêng để nâng cao trải nghiệm và sự gắn bó của khách hàng với công ty.

#### *Xây dựng mô hình dự báo:*

* Mục đích: Xây dựng mô hình dự báo để xác định với đặc điểm về nhân khẩu học và cách khách hàng sử dụng dịch vụ để dự đoán khả năng khách hàng rời bỏ và số tháng khách hàng gắn bó với công ty.
* Dữ liệu để xây dựng hình: Age, Tenure in Months, Monthly Charge, Churn Value, Gender, Married, Number of Dependents, Phone Service, Multiple Lines, Internet Service, Online Security, Online Backup, Device Protection Plan, Premium Tech Support, Streaming TV, Streaming Movies, Streaming Music, Unlimited Data, Contract, Paperless Billing, Payment Method





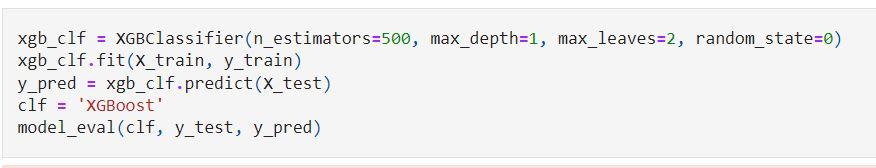
##### *So sánh các mô hình dự báo:*

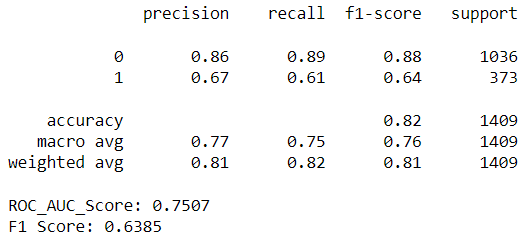
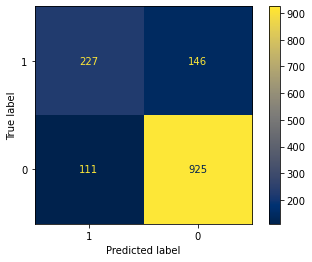
|  |  |
| --- | --- |
| Tên mô hình | Kết quả mô hình |
| XGBoost | https://lh3.googleusercontent.com/fpZKzQcbbrXCAy0rE2WgTKniRB-RnEmjNxktBM8Ww9T9WnRoJEEyH7A20Y7kJTgKKgfFJdwYemLF1xKr0LYkgBrskfReZlsQ2YpwG9GD4prX5S5JLWDkKkmIvLEivLNmVz1MRWeYqGo88UGwYVK38doNXLqdZY2hzSEWp-HlR3SVyPFThqc2PRsUC6-OWghttps://lh5.googleusercontent.com/7y6PPMFspazVCK3X87aPqZMQ6whZR0b8drFKN655eHWvKKGwheNgNvRnVtos50Gr5T_QK8lNoQWI7rRiX_q6XleaCmGj6EE3DAKFjDIiSimdSd9BEDTsOnJOgOXxhXBpfw1eWcrH4k5od8_GJN-k7iR6f_DDQSku5U2yHRtNmXURz298GyiAg-FoGZ6jWQ |
| Logistic  Regression | https://lh3.googleusercontent.com/bid3WVGFq74CVUfJCzb1E0Q7XBw3D7HTQ4orgg2O0OpdY3kaOVUKLXWo_VHqsKBfafFuCn7jJZK-BYEbJZijIC02JbbPC-23kyHkMHhObu5orZsXG0MvnCAqdZtKyTOo3sgGN6ztP6KC7nW0rNh3AGAiYIFUeTzczORwpGU1vxkmQDvTU1Y_Ru-GQmSQEghttps://lh6.googleusercontent.com/Fc_iDqecEOyggRvuFNLMIiMgKPwf2bGPJX4fhYw6Ko409AP9zZ9u-SnKyKcTc3JYe3FwEEoCO0Xn_KpuRa3DSuzCv9DT7y1QQSynUMNMApWulplnqNj-k757ZrHhhGd2bTr22nY3GQvV5RK9pG2ski2AHR9HHsJ_55y6s_YOJyF90tRSQcoos4Xlbh52lg |
| Support Vector Machines | https://lh5.googleusercontent.com/CT6uueNYP4J1uHZKg_i-bBpjEYCopv5u6EypP3HxHygEP-St3QAyDYgQEzy0EPpyX6msFvaCiL587jrIq86FhtuVeJ8l_qMdQ0zGRtO08mrzaGNskMLUV0pWfPVof7slLn8CLWGBaOrJH4KNXaxuAyEjuoN2Uvv8DmTvVWDBEoKDEOhC8aQiB4wE3nbfLQhttps://lh5.googleusercontent.com/bne0UnxTeilSEv82f0ikBCSRJ4Jw6iWRN5H8K10bFa9QzUDaJ1ELYkojXJSiAoDPPQsYaXYFe3Uy8boiFBQs6cxF-OuD-Rb-XHncinZrm6Qez3A49RA9ldXYc9r0Q5UrFxJhN-7fMcMXe_MOF4frv3addnat_S50Iu1lEmkafOcV4tFjMo7ZRJD605RzOw |
| Decision tree classifier | https://lh5.googleusercontent.com/FgqaqU-by6oa-nUjITyOlqJNRJ0Ge8rW_-aZ3DKzulY-TCaK_6OZ2qvDuQwIb31zmtqGbqC6KXw7DcNqbxREHnKZSF7wK1WlmGoOKdU9uwQgLWQgCGY4_WnhzYawTXRLr39U48dO8IR7g7P9KS1b9-VnXevYhHzyCC6TY4B5YZ5dXS9fa9qaP4lHG66aOQhttps://lh3.googleusercontent.com/wPKZMPEYBUZq99VqqJJgcA6evXV58ypRmi5HFLtxfmX6ypwGJeFSF55QYOmmBdRgmR8A2DKo14aSwijFdW8kUyE63X1HQBD4XXk7OoV7XFuXJhRDh6lX9nxwYEfqn6_dsBjtaCw0dDj2kXu2vng9F-jF-CNHhTrFc0ljiRMg3PDN9n_20mj2jgtKjVkjPw |
| AdaBoost  Classifier | https://lh5.googleusercontent.com/8jey6cIMRt-eTVUNPMFNbi00PFZ-U_x65DTgnTRU0XuLzBI-OPpmvwo_Zyk0GBe-mvE2wXfrnzgUlkQOgT2GHYHSycFvqGMRP5AY6SRg9IOFxEtBsQ5zVSD3znTfB0wjASDSYYdisRRs-j_zNP4DtafXmHUykoAUly3hw1PBQT6FFXtpA2XLTCW-nGLnUwhttps://lh4.googleusercontent.com/xYa4xdAaapIqif83dMF8gNwDJZuCxmVr_kAGIUNEXaJqP9nyfjZsKxHFnUqMNdWDdl5-JtB9ZqGqvKLiZJYGiC-Adnb5L8NCHb_iIEJZ9tW8m9KVKgICmqYxoSiZSlYJwclt-HRCkzsm9eGRJ6hlnDqgDaV1RSpFsWHonuM-3qc40YZ1FEVkmvLyuBNpaw |
| Random Forest Classifier | https://lh3.googleusercontent.com/LFHASEApqXnf_HZJ7EoRACdXCNuN1vysp3bOU9K_hjoTd1a2w5Ithfy5oUrheOD3PknnT7y-b7hmVBiLdAiw5muUtm3KKDXZ1quxcBZdmU8Icd29RjRfTtrkzyWJXD7qqA-Py6pJjML5hbzxMDKHS3mhHBFLM2Ms7RcsuUyNsA7HIXlp7UcejF0a9ao41ghttps://lh3.googleusercontent.com/1uRYslYOk2OVZzkkIWArqsNsf49JVOpD4250nbGcEHxXhQ4T9zv5AVULbi4K-Ovk2p_uxMWblUKp-CIJWspi-oOCgYwWOvcEuxaJvVp0hPGz26viLm6gWv5D_Xm1n0Wclj8hsYGpYot5oUjEO98PqDYCpTByQmzTWUfzzovZIPUdXDbr7eoEIKYPeHBbIw |

* Nhận xét: Xét kết quả đánh giá của từng mô hình, ta nhận thấy các chỉ số (Precision, Recall, F1-Score, Accuracy và ROC\_AUC\_Score) của mô hình XGBoost đều cao hơn các mô hình dự báo khác. Do đó, ta có thể sử dụng mô hình XGBoost để xây dựng và tối ưu mô hình dự báo.
* Giới thiệu mô hình phân lớp XGBoost:
* Ưu điểm của mô hình: XGBoost là thuật toán hiện đại nhất để giải quyết các bài toán học có giám sát có tốc độ xử lý nhanh và cho ra kết quả với độ chính xác cao.
* Ý tưởng thuật toán: XGBoost đơn giản là giải thuật Tree (Cây), áp dụng các kỹ thuật để kết hợp các Tree, làm mịn Training Loss (Sai số khi huấn luyện) và Regularization (Chuẩn hóa sai số, hệ số và số biến)

##### *Xây dựng mô hình dự đoán khả năng khách hàng rời bỏ:*

* Sử dụng mô hình XGBClassifier của thư viện XGBoost
* Input: Age, Monthly Charge, Gender, Married, Number of Dependents, Phone Service, Multiple Lines, Internet Service, Online Security, Online Backup, Device Protection Plan, Premium Tech Support, Streaming TV, Streaming Movies, Streaming Music, Unlimited Data, Contract, Paperless Billing, Payment Method
* Output: Churn Value
* Chia tập dữ liệu thành 2 tập training và testing với training:testing = 8:2.

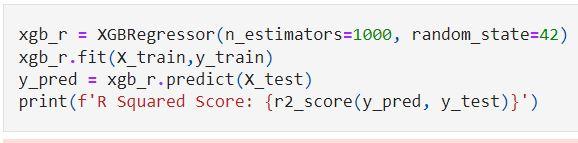




* *Các chỉ số:*
* *Hiệu suất của mô hình: AUC = 75,07%*
* *Độ chính xác: Accuracy = 82%*
* *Theo như ma trận nhầm lẫn:*
* *có 227 khách hàng rời bỏ được dự đoán rời bỏ.*
* *có 146 khách hàng rời bỏ được dự đoán không rời bỏ.*
* *có 111 khách hàng không rời bỏ được dự đoán rời bỏ.*
* *có 925 khách hàng không rời bỏ được dự đoán không rời bỏ.*
* *Precision = 81% thể hiện sự phát hiện các điểm dữ liệu khách hàng rời bỏ khá chuẩn.*
* *Recall = 82% thể hiện sự bỏ sót các điểm dữ liệu khách hàng rời bỏ khá thấp.*
* *F1 Score = 63,85% thể hiện sự dung hòa của 2 chỉ số Precision và Recall khá cao, từ đó kết luận rằng mô hình đang được đánh giá khá tốt.*

##### *Xây dựng mô hình dự đoán số tháng khách hàng gắn bó với công ty:*

* Sử dụng mô hình XGBRegressor của thư viện XGBoost
* Input: Age, Monthly Charge, Gender, Married, Number of Dependents, Phone Service, Multiple Lines, Internet Service, Online Security, Online Backup, Device Protection Plan, Premium Tech Support, Streaming TV, Streaming Movies, Streaming Music, Unlimited Data, Contract, Paperless Billing, Payment Method
* Output: Tenure in Months
* Chia tập dữ liệu thành 2 tập training và testing với training:testing = 8:2.





* *Sau khi xây dựng mô hình thì ta có chỉ số* ***R Squared Score là 0,996*** *(gần tiến về 1). Chỉ số R Squared cho thấy được mức độ phù hợp của dữ liệu với mô hình. → Với những yếu tố Input của mô hình được xây dựng thì có thể giải thích được 99.6% Output, con số này rất cao → có thể dự đoán ra số tháng gần đúng với thực tế.*

##### *Dự đoán trường hợp cụ thể:*

|  |  |  |  |
| --- | --- | --- | --- |
|  | User Input |  | User Input |
| Age | 78 | Online Backup | No |
| Monthly Charge | 39 | Device Protection Plan | No |
| Total Charges | 165 | Premium Tech Support | No |
| Gender | Male | Streaming TV | No |
| Married | No | Streaming Movies | No |
| Number of Dependents | No | Streaming Music | No |
| Phone Service | No | Unlimited Data | No |
| Multiple Lines | No | Contract | Month-to-Month |
| Internet Service | Yes | Paperless Billing | No |
| Online Security | No | Payment Method | Credit Card |

*Khi lấy các thông tin trên đưa vào mô hình thì ta có kết quả là*

*Số tháng khách hàng sẽ ở lại với cty: 3.8653066*

*KH có rời bỏ không? Yes*

*xác suất khách hàng ở lại công ty là: 41.24824404716492*

|  |  |  |  |
| --- | --- | --- | --- |
|  | User Input |  | User Input |
| Age | 30 | Online Backup | No |
| Monthly Charge | 39 | Device Protection Plan | No |
| Total Charges | 3640 | Premium Tech Support | No |
| Gender | Male | Streaming TV | No |
| Married | Yes | Streaming Movies | No |
| Number of Dependents | Yes | Streaming Music | No |
| Phone Service | No | Unlimited Data | No |
| Multiple Lines | No | Contract | Month-to-Month |
| Internet Service | Yes | Paperless Billing | No |
| Online Security | No | Payment Method | Credit Card |

*Khi lấy các thông tin trên đưa vào mô hình thì ta có kết quả là*

*Số tháng khách hàng sẽ ở lại với cty: 70.81738*

*KH có rời bỏ không? No*

*xác suất khách hàng ở lại công ty là: 95.4227976500988*

# **KẾT LUẬN**

* Việc khai thác đặc điểm khách hàng doanh nghiệp là một quy trình quan trọng, kết quả phân tích giúp các nhà quản lý, các cấp lãnh đạo nắm rõ tình hình doanh nghiệp cụ thể là về khách hàng chẳng hạn như đặc điểm nhân khẩu học, tâm lý khách hàng, hành vi khách hàng, các vấn đề còn tồn đọng trong sản phẩm, dịch vụ của công ty... Dựa trên các căn cứ này, ban lãnh đạo công ty có thể đưa ra những quyết định đổi mới đúng hướng nhằm tạo mối quan hệ bền chặt hơn với khách hàng.
* Qua bài báo cáo này, nhóm đã hiểu rõ hơn về cách triển khai phân tích một bài toán thực tế trong doanh nghiệp bằng cách sử dụng các phương pháp, mô hình, công cụ trong Khai phá dữ liệu. Cụ thể như các bước đặt vấn đề, trình tự tiến hành và quá trình thực hiện triển khai các ý tưởng. Ngoài ra, dự án nhỏ này còn giúp các thành viên trong nhóm hiểu hơn về tầm quan trọng của khách hàng, các giải pháp của doanh nghiệp và những điều thú vị về đặc điểm tập khách hàng của doanh nghiệp.