**TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**BÁO CÁO MÔN HỌC**

**ĐỀ TÀI: PHÂN TÍCH, DỰ ĐOÁN TỈ LỆ RỜI BỎ CỦA KHÁCH HÀNG**

**Môn học: Phân tích dữ liệu**

**GVHD: ThS. Hồ Hướng Thiên**

**LỚP: DH22IM02**

**SINH VIÊN THỰC HIỆN:**

**Nguyễn Minh Châu - 2254052007**

**Võ Thị Hiền Thanh - 2254052074**

**Thành phố Hồ Chí Minh, 2024**

**Mục lục**

[**CHƯƠNG I. TỔNG QUAN ĐỀ TÀI** 2](#_Toc186146116)

[1. Sơ lược về đề tài 2](#_Toc186146117)

[2. Mục tiêu nghiên cứu 2](#_Toc186146118)

[3. Tài nguyên sử dụng 2](#_Toc186146119)

[**CHƯƠNG II. TỔNG QUAN VỀ BỘ DỮ LIỆU** 2](#_Toc186146120)

[1. Sơ lược về bộ dữ liệu 2](#_Toc186146121)

[2. Mô tả thuộc tính của dữ liệu 2](#_Toc186146122)

[**CHƯƠNG III. KIỂM TRA VÀ LÀM SẠCH DỮ LIỆU** 5](#_Toc186146123)

[1. Kiểm tra về kiểu dữ liệu 5](#_Toc186146124)

[2. Kiểm tra vấn đề Missing Data 6](#_Toc186146125)

[3. Kiểm tra giá trị trùng lặp 7](#_Toc186146126)

[4. Kiểm tra giá trị outlier 7](#_Toc186146127)

[**CHƯƠNG IV. TRỰC QUAN HÓA DỮ LIỆU** 9](#_Toc186146128)

[1. Trực quan theo từng thuộc tính 9](#_Toc186146129)

[2. Trực quan theo biến mục tiêu 16](#_Toc186146130)

[**CHƯƠNG V. XÂY DỰNG MÔ HÌNH DỰ ĐOÁN** 26](#_Toc186146131)

[1. Mã hóa và chuẩn hóa dữ liệu 26](#_Toc186146132)

[2. Xây dựng mô hình dự đoán 27](#_Toc186146133)

[3. Lý do rời rạc và chuẩn hoá 28](#_Toc186146134)

[**CHƯƠNG VI. KẾT LUẬN** 29](#_Toc186146135)

# **CHƯƠNG I. TỔNG QUAN ĐỀ TÀI**

1. Sơ lược về đề tài

- Tỉ lệ khách hàng rời bỏ là tỉ lệ khách hàng đã ngừng sử dụng sản phẩm hoặc dịch vụ của một công ty trong một khoảng thời gian nhất định.

- “Tỉ lệ khách hàng rời bỏ sau khi sử dụng dịch vụ” là con số luôn được quan tâm và cũng là vấn đề quan trọng đối với bất kì công ty nào.

- Giữ chân khách hàng cũ dễ hơn tìm kiếm khách hàng mới.

1. Mục tiêu nghiên cứu

Dự đoán khả năng rời bỏ của khách hàng mới dựa trên thông tin của họ giúp cho doanh nghiệp chủ động tiếp cận và giữ chân khách hàng có nguy cơ cao.

1. Tài nguyên sử dụng

Bộ dữ liệu: “Telco-Customer-Churn.csv”

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

# **CHƯƠNG II. TỔNG QUAN VỀ BỘ DỮ LIỆU**

1. Sơ lược về bộ dữ liệu

- Tên bộ dữ liệu: “Telco-Customer-Churn.csv”.

- Nguồn cung cấp: Công ty IBM (International Business Machines Corporation).

- Tập dữ liệu bao gồm 7043 khách hàng mỗi khách hàng có 21 thuộc tính của một công ty cung cấp dịch vụ viễn thông.Công ty IBM (International Business Machines Corporati

1. Mô tả thuộc tính của dữ liệu

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Tên thuộc tích | Loại dữ liệu | Mô tả thuộc tính |
| 1 | customerID | object | Mã khách hàng |
| 2 | gender | object | Giới tính của khách hàng ("Male", "Female") |
| 3 | SeniorCitizen | Int64 | Có phải người cao tuổi không  (0: không phải, 1: người cao tuổi) |
| 4 | Partner | object | Khách hàng có người bạn đời không ( "Yes", "No") |
| 5 | Dependents | object | Khách hàng có người phụ thuộc không ("Yes", "No") |
| 6 | tenure | Int64 | Thời gian khách hàng đã sử dụng dịch vụ, tính bằng tháng |
| 7 | PhoneService | Object | Có sử dụng dịch vụ điện thoại hay không ("Yes", "No") |
| 8 | MultipleLines | object | Có sử dụng nhiều đường dây điện thoại không ("No phone service", "No", "Yes") |
| 9 | InternetService | object | Loại dịch vụ Internet khách hàng đang sử dụng ("DSL", "Fiber optic", "No") |
| 10 | OnlineSecurity | object | Có sử dụng dịch vụ bảo mật trực tuyến hay không ("Yes", "No", "No internet service") |
| 11 | OnlineBackup | object | Có đăng ký dịch vụ sao lưu trực tuyến không ("Yes", "No", "No internet service") |
| 12 | DeviceProtection | object | Có đăng ký dịch vụ bảo vệ thiết bị không ("Yes", "No", "No internet service") |
| 13 | TechSupport | object | Có đăng ký dịch vụ hỗ trợ kỹ thuật không ("Yes", "No", "No internet service") |
| 14 | StreamingTV | object | Có sử dụng dịch vụ truyền hình trực tiếp không ("Yes", "No", "No internet service") |
| 15 | StreamingMovies | object | Có sử dụng dịch vụ xem phim trực tuyến không ("Yes", "No", "No internet service) |
| 16 | Contract | object | Loại hợp đồng ("Month-to-month", "One year", "Two year") |
| 17 | PaperlessBilling | object | Có sử dụng hóa đơn không điện tử không ("Yes", "No") |
| 18 | PaymentMethod | object | Phương thức thanh toán ("Electronic check", "Mailed check", "Bank transfer (automatic)", "Credit card (automatic)") |
| 19 | MonthlyCharges | float64 | Số tiền mà khách hàng phải trả hàng tháng |
| 20 | TotalCharges | object | Tổng số tiền khách hàng đã thanh toán |
| 21 | Churn | object | Trạng thái khách hàng có rời bỏ dịch vụ hay không ("Yes", "No") |

# **CHƯƠNG III. KIỂM TRA VÀ LÀM SẠCH DỮ LIỆU**

1. Kiểm tra về kiểu dữ liệu

customerID object

gender object

SeniorCitizen int64

Partner object

Dependents object

tenure int64

PhoneService object

MultipleLines object

InternetService object

OnlineSecurity object

OnlineBackup object

DeviceProtection object

TechSupport object

StreamingTV object

StreamingMovies object

Contract object

PaperlessBilling object

PaymentMethod object

MonthlyCharges float64

TotalCharges object

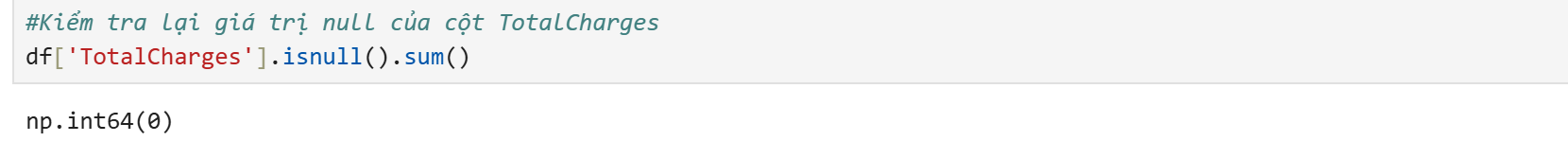
Churn object

- Chuyển cột TotalCharges từ dạng object sang float.

=> Kết quả: Xuất hiện 11 giá trị NULL trong cột TotalCharges sau khi chuyển đổi.

=> Hướng giải quyết: Điền các giá trị trung bình của cột TotalCharges vào các vị trí bị NULL

- Sau đó kiểm tra lại không có giá trị nào rỗng.



1. Kiểm tra vấn đề Missing Data

Kết quả kiểm tra Missing data

customerID False

gender False

SeniorCitizen False

Partner False

Dependents False

tenure False

PhoneService False

MultipleLines False

InternetService False

OnlineSecurity False

OnlineBackup False

DeviceProtection False

TechSupport False

StreamingTV False

StreamingMovies False

Contract False

PaperlessBilling False

PaymentMethod False

MonthlyCharges False

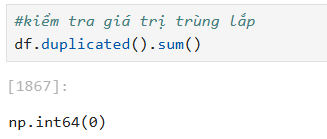
TotalCharges False

Churn False

dtype: bool

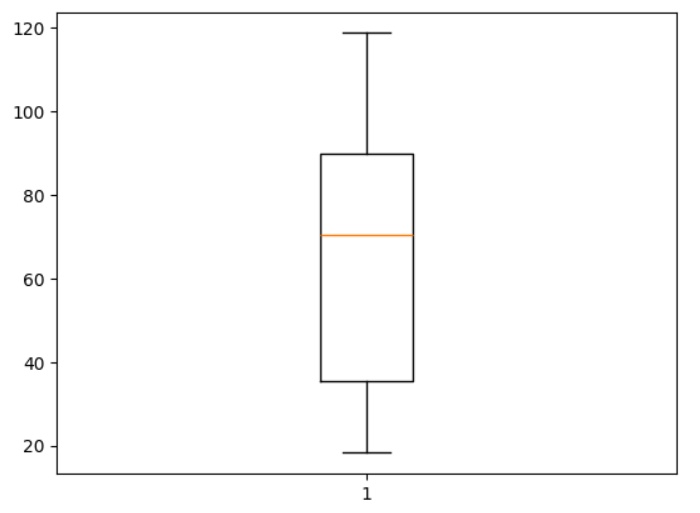
-> Không có giá trị thiếu ở bất kỳ cột nào.

1. Kiểm tra giá trị trùng lặp

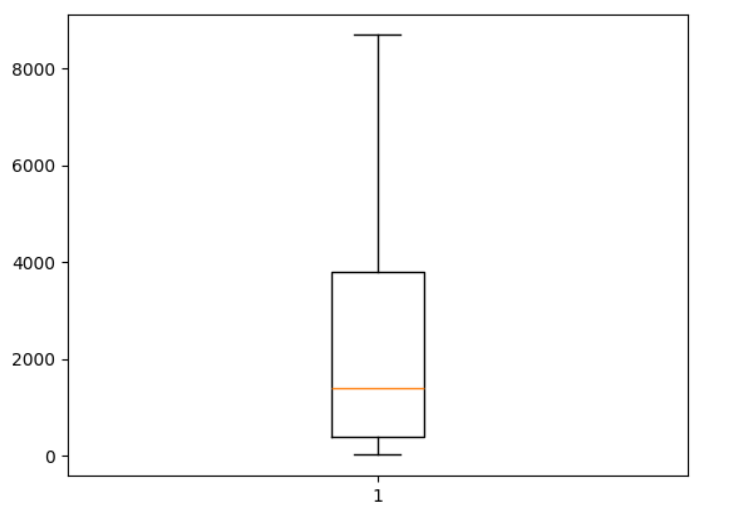


=> Không có dòng dữ liệu nào bị trùng lặp

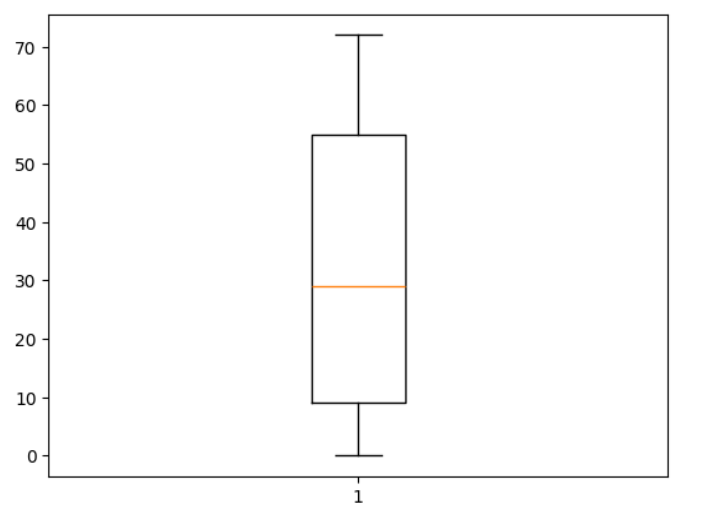
1. Kiểm tra giá trị outlier



Thuộc tính MonthlyCharges



Thuộc tính TotalCharges



Thuộc tính tenure

=> Cả ba thuộc tính đều không xuất hiện giá trị ngoại lệ (outlier).

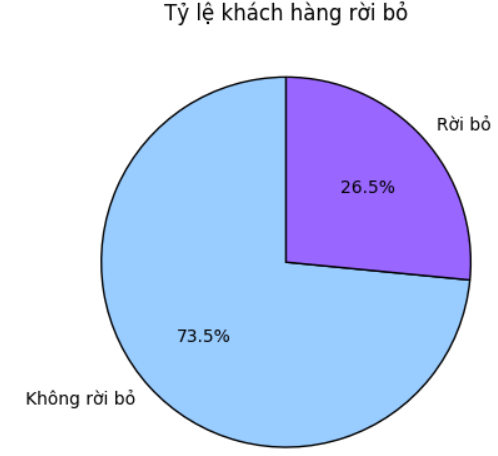
# **CHƯƠNG IV. TRỰC QUAN HÓA DỮ LIỆU**

1. Trực quan theo từng thuộc tính

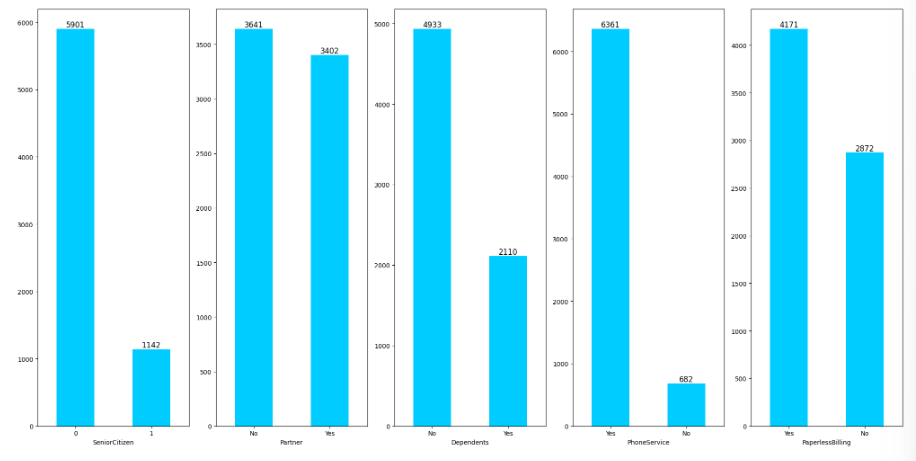
A blue rectangular bars with white text

Description automatically generated

Số lượng khách hàng nam (3555) và nữ (3488) không có sự chênh lệch lớn.



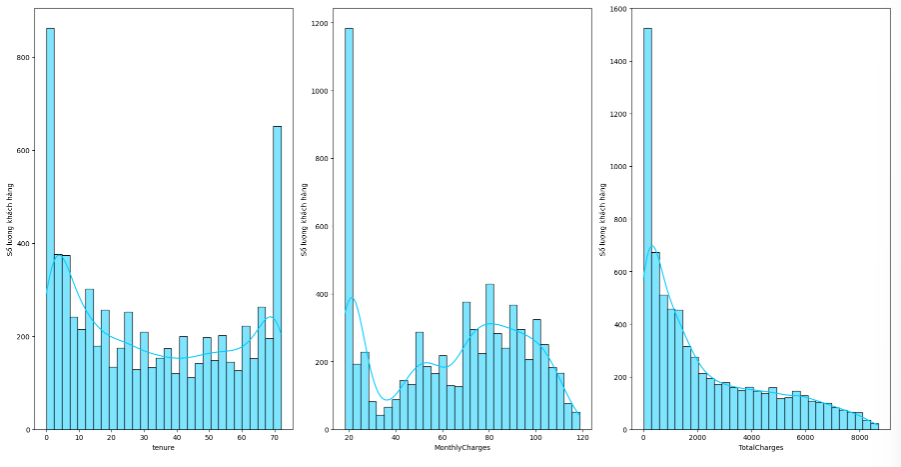
**Nhận xét**: Biểu đồ thể hiện tỷ lệ phần trăm của khách hàng rời bỏ và không rời bỏ dịch vụ. Cho thấy tỷ lệ khách hàng rời bỏ (26.5%) và không rời bỏ (73.5%).



**Nhận xét:**

* **SeniorCitizen (Người cao tuổi):** Số lượng người không phải là người cao tuổi cao hơn đáng kể so với số lượng người là người cao tuổi.Cụ thể, số lượng người không phải người cao tuổi là 5901, trong khi số lượng người cao tuổi chỉ là 1142. Điều này cho thấy tập khách hàng chủ yếu là người trẻ và trung niên.
* **Partner (Bạn đời):** Số lượng người có bạn đời gần bằng với số lượng người không có bạn đời. Số liệu cụ thể là 3402 người có bạn đời và 3641 người không có bạn đời.
* **Dependents (Người phụ thuộc):** Số lượng người không có người phụ thuộc cao hơn đáng kể so với số lượng người có người phụ thuộc. Cụ thể, có 4833 người không có người phụ thuộc và chỉ có 2110 người có người phụ thuộc.
* **PhoneService (Dịch vụ điện thoại):** Số lượng người sử dụng dịch vụ điện thoại rất cao (6361) so với số lượng người không sử dụng (chỉ 682). Điều này cho thấy dịch vụ điện thoại rất phổ biến trong tập khách hàng này.
* **PaperlessBilling (Hóa đơn điện tử):** Số lượng người sử dụng hóa đơn điện tử cũng cao hơn đáng kể so với số lượng người không sử dụng.Cụ thể, có 4171 người sử dụng hóa đơn điện tử và 2872 người không sử dụng. Xu hướng sử dụng hóa đơn điện tử khá phổ biến.

=> Biểu đồ cho thấy tập khách hàng chủ yếu là người trẻ và trung niên, không có người phụ thuộc, sử dụng dịch vụ điện thoại và có xu hướng sử dụng hóa đơn điện tử. Số lượng người cao tuổi và người có người phụ thuộc chiếm tỷ lệ ít hơn. Số lượng người có và không có bạn đời gần như tương đương.



**Nhận xét:**

* **Biểu đồ Histogram của tenure (Thời gian gắn bó):** Phân bố: Biểu đồ cho thấy một phân bố đa đỉnh (multimodal), với hai đỉnh rõ rệt. Một đỉnh nằm ở gần 0 (cho thấy một lượng lớn khách hàng mới), và một đỉnh khác nằm ở khoảng 70 (cho thấy một lượng lớn khách hàng gắn bó lâu dài).

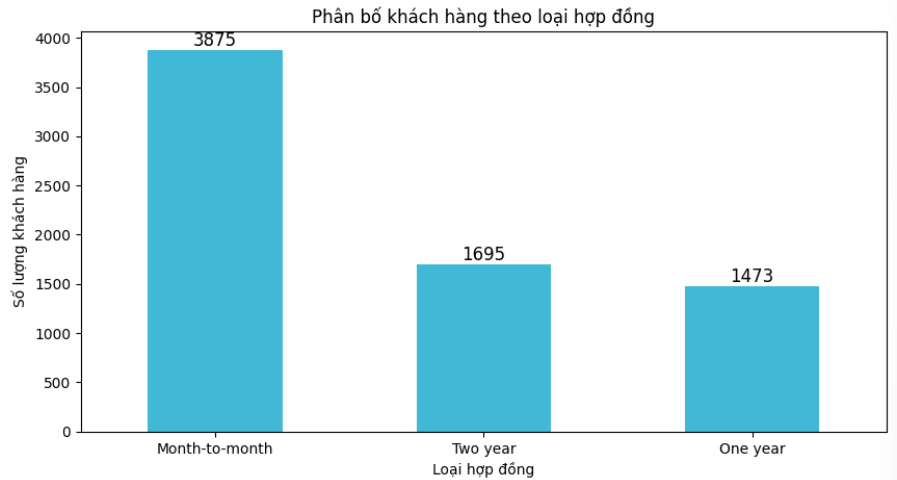
Ý nghĩa: Điều này cho thấy có hai nhóm khách hàng chính: khách hàng mới đăng ký dịch vụ và khách hàng đã sử dụng dịch vụ trong một thời gian dài. Có thể có một số yếu tố ảnh hưởng đến việc khách hàng tiếp tục sử dụng dịch vụ sau một khoảng thời gian nhất định (ví dụ: hết hợp đồng, thay đổi nhu cầu).

* **Biểu đồ Histogram của MonthlyCharges (Phí hàng tháng):** Phân bố: Biểu đồ cho thấy một phân bố lệch phải (right-skewed), với một đỉnh lớn ở khoảng 20 và một phân bố kéo dài sang bên phải.

Ý nghĩa: Điều này cho thấy phần lớn khách hàng trả phí hàng tháng thấp, trong khi một số ít khách hàng trả phí cao hơn đáng kể. Có thể có nhiều gói dịch vụ khác nhau với mức giá khác nhau, hoặc một số khách hàng sử dụng các dịch vụ bổ sung tốn kém.

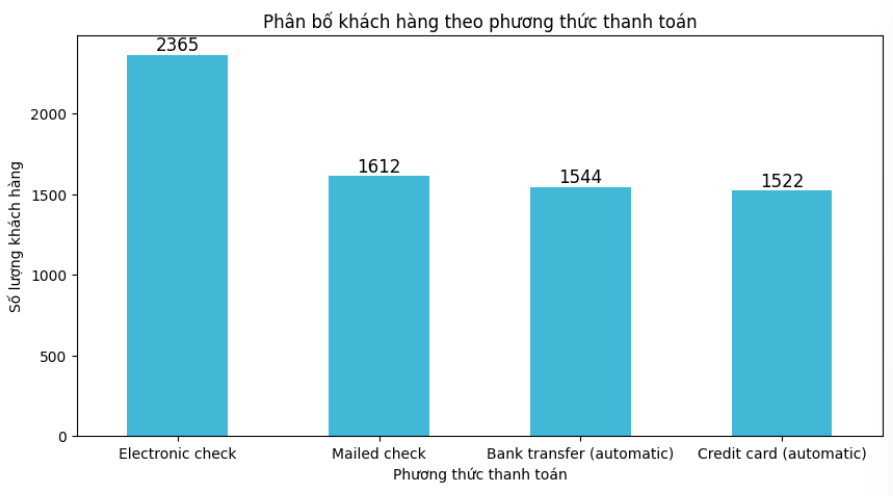
* **Biểu đồ Histogram của TotalCharges (Tổng phí):** Phân bố: Biểu đồ cho thấy một phân bố lệch phải (right-skewed), tương tự như MonthlyCharges, nhưng mức độ lệch rõ rệt hơn.

Ý nghĩa: Điều này là dễ hiểu, vì TotalCharges được tính bằng cách tích lũy MonthlyCharges theo thời gian (tenure). Khách hàng có thời gian gắn bó lâu hơn và/hoặc trả phí hàng tháng cao hơn sẽ có TotalCharges cao hơn. Mối quan hệ với tenure và MonthlyCharges: Phân bố này củng cố mối quan hệ giữa TotalCharges, tenure và MonthlyCharges.



**Nhận xét:**

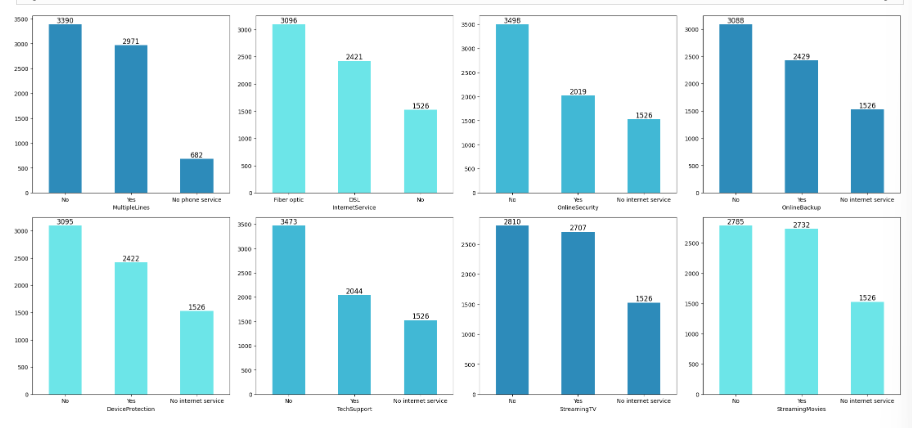
* **Loại hợp đồng "Month-to-month":** là loại hợp đồng phổ biến nhất với số lượng khách hàng là 3875, chiếm phần lớn so với hai loại hợp đồng còn lại. Điều này cho thấy khách hàng có xu hướng thích sự linh hoạt và không muốn bị ràng buộc bởi hợp đồng dài hạn.
* **Loại hợp đồng "Two year":** có số lượng khách hàng là 1695. Mặc dù ít hơn so với hợp đồng hàng tháng, nhưng vẫn cao hơn so với hợp đồng một năm. Điều này cho thấy một bộ phận khách hàng vẫn sẵn sàng ký hợp đồng dài hạn để có thể nhận được các ưu đãi hoặc giá tốt hơn.
* **Loại hợp đồng "One year":** là loại hợp đồng ít phổ biến nhất với số lượng khách hàng là 1473. Tóm lại: Khách hàng ưa chuộng sự linh hoạt với hợp đồng hàng tháng. Một lượng khách hàng đáng kể chọn hợp đồng hai năm, có thể do các ưu đãi đi kèm. Hợp đồng một năm ít được khách hàng lựa chọn nhất.



**Nhận xét:**

* **Electronic check (Chuyển khoản điện tử):** là hình thức thanh toán được sử dụng nhiều nhất với 2365 khách hàng. Điều này cho thấy khách hàng ưa chuộng sự tiện lợi và nhanh chóng của hình thức thanh toán trực tuyến.
* **Mailed check (Gửi séc qua bưu điện):** được sử dụng bởi 1612 khách hàng, đứng thứ hai. Số lượng này thấp hơn đáng kể so với chuyển khoản điện tử, cho thấy xu hướng giảm dần của việc sử dụng các hình thức thanh toán truyền thống.
* **Bank transfer (automatic) (Chuyển khoản ngân hàng tự động):** có 1544 khách hàng sử dụng hình thức này. Số lượng này gần tương đương với hình thức thanh toán bằng thẻ tín dụng tự động.
* **Credit card (automatic) (Thẻ tín dụng tự động):**là hình thức thanh toán ít được sử dụng nhất với 1522 khách hàng.

**=>** Chuyển khoản điện tử là hình thức thanh toán phổ biến nhất, cho thấy xu hướng số hóa trong thanh toán. Các hình thức thanh toán tự động (chuyển khoản ngân hàng và thẻ tín dụng) có số lượng khách hàng tương đương nhau và thấp hơn so với chuyển khoản điện tử. Gửi séc qua bưu điện vẫn được một bộ phận khách hàng sử dụng, nhưng số lượng đang giảm dần.

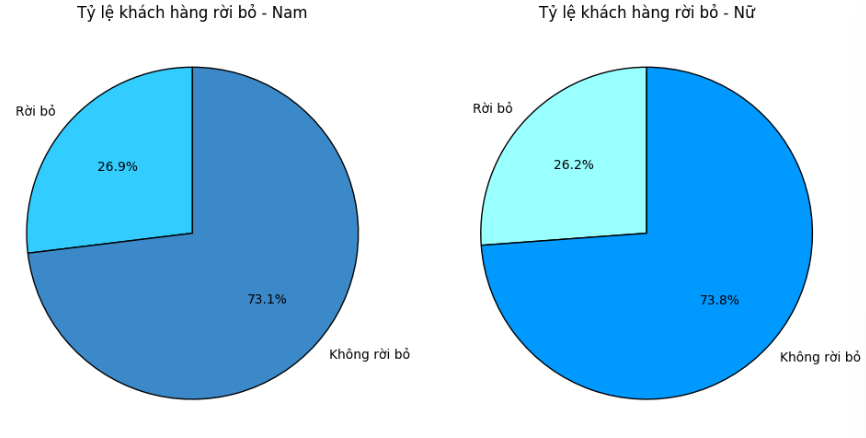


**Nhận xét:**

* **MultipleLines (Nhiều đường dây điện thoại):** Số lượng khách hàng sử dụng nhiều đường dây điện thoại ít hơn đáng kể so với số lượng khách hàng không sử dụng. Có một số lượng nhỏ khách hàng không có dịch vụ điện thoại.
* **InternetService (Dịch vụ Internet):** Fiber optic là loại dịch vụ Internet phổ biến nhất, tiếp theo là DSL. Có một số lượng đáng kể khách hàng không sử dụng dịch vụ Internet.
* **OnlineSecurity (Bảo mật trực tuyến):** Số lượng khách hàng không sử dụng dịch vụ bảo mật trực tuyến nhiều hơn đáng kể so với số lượng khách hàng sử dụng. Có một số lượng đáng kể khách hàng không có dịch vụ Internet.
* **OnlineBackup (Sao lưu trực tuyến)**: Tương tự như bảo mật trực tuyến, số lượng khách hàng không sử dụng dịch vụ sao lưu trực tuyến nhiều hơn đáng kể so với số lượng khách hàng sử dụng. Cũng có một số lượng đáng kể khách hàng không có dịch vụ Internet.
* **DeviceProtection (Bảo vệ thiết bị):** Tương tự như hai dịch vụ trên, số lượng khách hàng không sử dụng dịch vụ bảo vệ thiết bị nhiều hơn đáng kể so với số lượng khách hàng sử dụng. Có một số lượng đáng kể khách hàng không có dịch vụ Internet.
* **TechSupport (Hỗ trợ kỹ thuật):** Số lượng khách hàng không sử dụng dịch vụ hỗ trợ kỹ thuật nhiều hơn đáng kể so với số lượng khách hàng sử dụng. Có một số lượng đáng kể khách hàng không có dịch vụ Internet.
* **StreamingTV (Xem TV trực tuyến):** Số lượng khách hàng sử dụng dịch vụ xem TV trực tuyến và không sử dụng gần như tương đương. Có một số lượng đáng kể khách hàng không có dịch vụ Internet.
* **StreamingMovies (Xem phim trực tuyến):** Tương tự như xem TV trực tuyến, số lượng khách hàng sử dụng dịch vụ xem phim trực tuyến và không sử dụng gần như tương đương. Có một số lượng đáng kể khách hàng không có dịch vụ Internet.

=> Số lượng khách hàng không có dịch vụ Internet là khá lớn trong các dịch vụ như bảo mật trực tuyến, sao lưu trực tuyến, bảo vệ thiết bị và hỗ trợ kỹ thuật. Điều này cho thấy có thể có một phân khúc khách hàng không quan tâm đến các dịch vụ này hoặc không có khả năng tiếp cận Internet. Việc số lượng khách hàng không sử dụng các dịch vụ bảo mật, sao lưu và bảo vệ thiết bị nhiều hơn số lượng khách hàng sử dụng có thể là một vấn đề cần được quan tâm. Có thể cần có các chiến dịch giáo dục hoặc khuyến mãi để khuyến khích khách hàng sử dụng các dịch vụ này. Việc số lượng khách hàng sử dụng và không sử dụng dịch vụ xem TV và phim trực tuyến gần như tương đương cho thấy thị trường cho các dịch vụ này khá cân bằng.

1. Trực quan theo biến mục tiêu



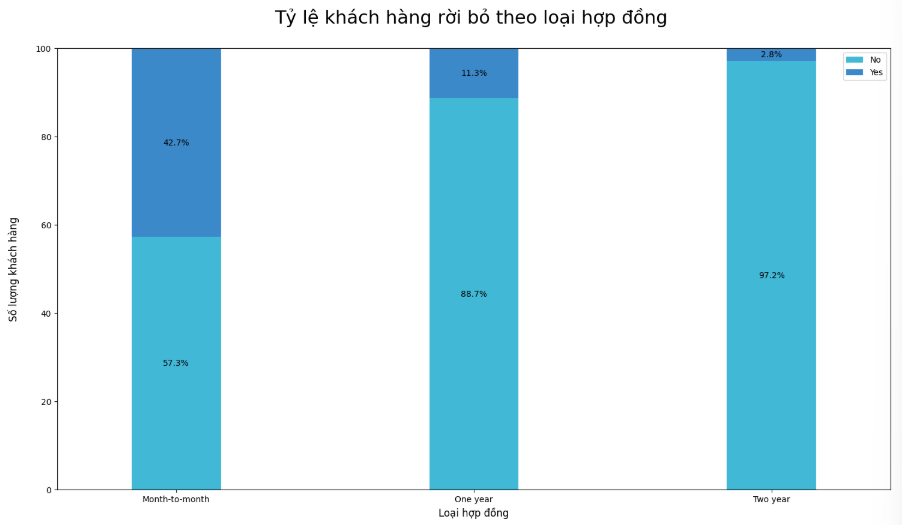
|  |  |  |  |
| --- | --- | --- | --- |
| Giới tính | Số lượng | Số lượng rời bỏ | Tỷ lệ |
| Male | 3555 | 930 | 26.2% |
| Female | 3488 | 939 | 26.9% |

**Nhận xét:**

* **Tỷ lệ rời bỏ chung**: Cả ở nam và nữ, tỷ lệ khách hàng không rời bỏ dịch vụ chiếm phần lớn (hơn 73%). Điều này cho thấy phần lớn khách hàng vẫn tiếp tục sử dụng dịch vụ.
* **So sánh theo giới tính:**

**+**  **Nam**: Tỷ lệ khách hàng nam rời bỏ dịch vụ là 26.9%, trong khi tỷ lệ không rời bỏ là 73.1%.

**+ Nữ**: Tỷ lệ khách hàng nữ rời bỏ dịch vụ là 26.2%, trong khi tỷ lệ không rời bỏ là 73.8%. Sự khác biệt giữa nam và nữ: Có một sự khác biệt rất nhỏ về tỷ lệ rời bỏ giữa nam và nữ. Tỷ lệ rời bỏ ở nam cao hơn một chút (0.7%) so với nữ. Sự khác biệt này không đáng kể và có thể được coi là không có sự khác biệt đáng kể về tỷ lệ rời bỏ giữa hai giới tính. Kết luận: Nhìn chung, biểu đồ cho thấy tỷ lệ khách hàng tiếp tục sử dụng dịch vụ là rất cao, không có sự khác biệt lớn về tỷ lệ rời bỏ giữa khách hàng nam và nữ.

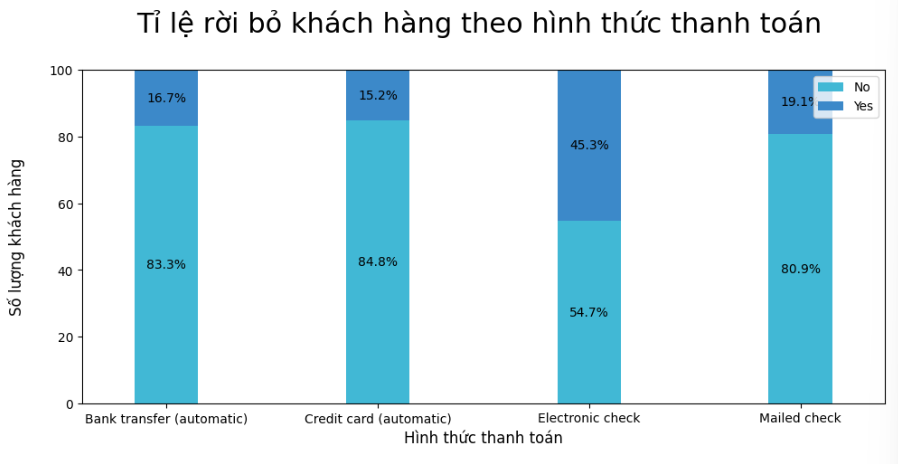


|  |  |  |  |
| --- | --- | --- | --- |
| Loại | Số lượng | Số lượng rời bỏ | Tỷ lệ |
| Month-to-month | 3875 | 1655 | 42.7% |
| One year | 1473 | 166 | 11.3% |
| Two year | 1695 | 48 | 2.8% |

**Nhận xét:**

* **Month-to-month (Hàng tháng):** Đây là loại hợp đồng có tỷ lệ khách hàng rời bỏ cao nhất (42.7%). Điều này cho thấy khách hàng sử dụng hợp đồng hàng tháng có xu hướng rời bỏ dịch vụ cao hơn so với các loại hợp đồng khác.
* **One year (Một năm):** Loại hợp đồng này có tỷ lệ khách hàng rời bỏ thấp hơn đáng kể so với hợp đồng hàng tháng, chỉ 11.3%.
* **Two year (Hai năm):** Đây là loại hợp đồng có tỷ lệ khách hàng rời bỏ thấp nhất, chỉ 2.8%. Điều này cho thấy khách hàng ký hợp đồng dài hạn (hai năm) có xu hướng gắn bó với dịch vụ hơn.

=> Biểu đồ cho thấy một mối tương quan rõ ràng giữa loại hợp đồng và tỷ lệ khách hàng rời bỏ. Hợp đồng ngắn hạn (hàng tháng) có tỷ lệ rời bỏ cao nhất, trong khi hợp đồng dài hạn (hai năm) có tỷ lệ rời bỏ thấp nhất. Điều này gợi ý rằng việc khuyến khích khách hàng ký hợp đồng dài hạn có thể là một chiến lược hiệu quả để giảm tỷ lệ khách hàng rời bỏ.

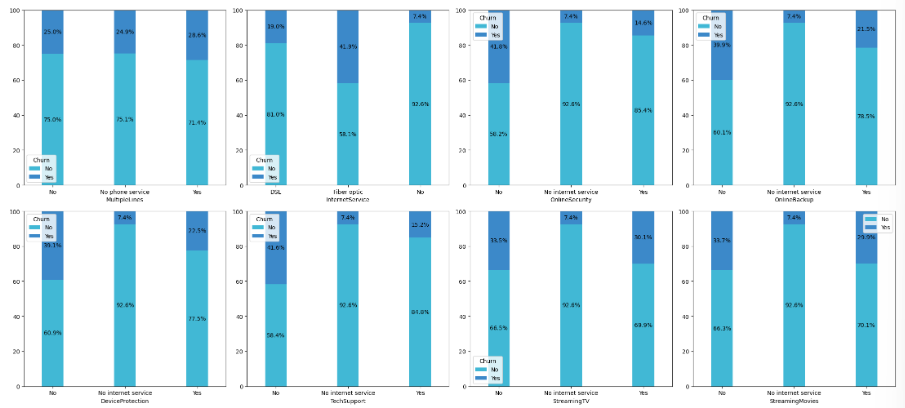


|  |  |  |  |
| --- | --- | --- | --- |
| Loại | Số lượng | Số lượng rời bỏ | Tỷ lệ |
| Bank transfer | 1544 | 258 | 16.7% |
| Credit card | 1522 | 232 | 15.2% |
| Electronicheck | 2365 | 1071 | 45.3% |
| Mailecheck | 1612 | 308 | 19.1% |

**Nhận xét**

* **Electronic check (Séc điện tử)**: Đây là phương thức thanh toán có tỷ lệ khách hàng rời bỏ cao nhất (45.3%). Điều này cho thấy khách hàng sử dụng séc điện tử có xu hướng rời bỏ dịch vụ cao hơn đáng kể so với các phương thức thanh toán khác. Hơn nữa, đây là phương thức duy nhất mà tỷ lệ rời bỏ gần bằng hoặc vượt quá 1/2.
* **Bank transfer (automatic) (Chuyển khoản ngân hàng tự động) và Credit card (automatic) (Thẻ tín dụng tự động):** Hai phương thức này có tỷ lệ khách hàng rời bỏ tương đối thấp và gần nhau, lần lượt là 16.7% và 15.2%. Điều này cho thấy việc sử dụng thanh toán tự động (bằng chuyển khoản ngân hàng hoặc thẻ tín dụng) có thể giúp giảm thiểu tỷ lệ khách hàng rời bỏ.
* **Mailed check (Séc gửi qua bưu điện):** Phương thức này có tỷ lệ khách hàng rời bỏ ở mức trung bình (19.1%), cao hơn một chút so với hai phương thức thanh toán tự động nhưng thấp hơn nhiều so với séc điện tử.

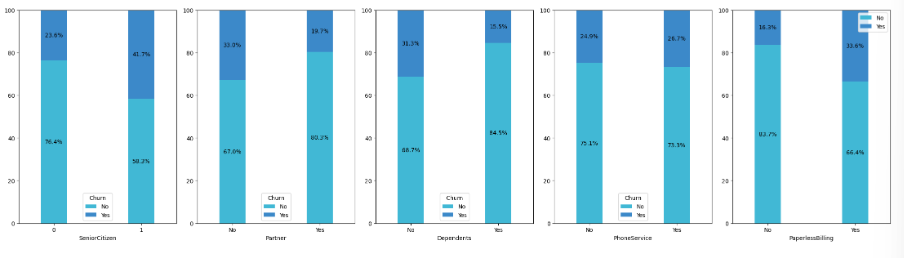
=> Biểu đồ cho thấy một mối tương quan rõ ràng giữa phương thức thanh toán và tỷ lệ khách hàng rời bỏ. Việc sử dụng séc điện tử có liên quan đến tỷ lệ rời bỏ cao nhất, trong khi các phương thức thanh toán tự động (chuyển khoản ngân hàng và thẻ tín dụng) có liên quan đến tỷ lệ rời bỏ thấp nhất. Điều này cho thấy việc khuyến khích khách hàng chuyển sang các phương thức thanh toán tự động có thể là một chiến lược hiệu quả để giảm tỷ lệ khách hàng rời bỏ.



**Nhận xét:**

* **MultipleLines (Nhiều đường dây điện thoại):** Khách hàng có nhiều đường dây điện thoại có tỷ lệ rời bỏ cao hơn một chút (khoảng 25%) so với khách hàng không có ("No") (khoảng 23%).Khách hàng không có dịch vụ điện thoại có tỷ lệ rời bỏ thấp nhất (khoảng 7%). Có thể nhóm này ít tương tác với các dịch vụ khác, dẫn đến ít khả năng rời bỏ hơn.
* **InternetService (Dịch vụ Internet):** Khách hàng sử dụng dịch vụ Fiber optic (Cáp quang) có tỷ lệ rời bỏ cao nhất (khoảng 42%). Điều này khá bất ngờ và cần được điều tra thêm. Có thể do cạnh tranh từ các nhà cung cấp cáp quang khác, hoặc có thể do vấn đề về chất lượng dịch vụ cáp quang. Khách hàng sử dụng DSL (Đường dây thuê bao số) có tỷ lệ rời bỏ thấp hơn (khoảng 19%). Khách hàng không sử dụng dịch vụ Internet có tỷ lệ rời bỏ thấp nhất (khoảng 7%).
* **OnlineSecurity (Bảo mật trực tuyến):** Khách hàng không sử dụng dịch vụ bảo mật trực tuyến có tỷ lệ rời bỏ cao nhất (khoảng 42%). Điều này cho thấy bảo mật trực tuyến có thể là một yếu tố quan trọng giữ chân khách hàng. Khách hàng có sử dụng dịch vụ bảo mật trực tuyến có tỷ lệ rời bỏ thấp hơn đáng kể (khoảng 14%).Khách hàng không có dịch vụ Internet có tỷ lệ rời bỏ thấp nhất (khoảng 7%).
* **OnlineBackup (Sao lưu trực tuyến):** Xu hướng tương tự như Bảo mật trực tuyến, khách hàng không sử dụng dịch vụ sao lưu trực tuyến có tỷ lệ rời bỏ cao nhất (khoảng 40%). Khách hàng có sử dụng dịch vụ sao lưu trực tuyến có tỷ lệ rời bỏ thấp hơn (khoảng 22%). Khách hàng không có dịch vụ Internet có tỷ lệ rời bỏ thấp nhất (khoảng 7%).
* **DeviceProtection (Bảo vệ thiết bị):** Khách hàng không sử dụng dịch vụ bảo vệ thiết bị có tỷ lệ rời bỏ cao nhất (khoảng 39%). Khách hàng có sử dụng dịch vụ bảo vệ thiết bị có tỷ lệ rời bỏ thấp hơn (khoảng 22%). Khách hàng không có dịch vụ Internet có tỷ lệ rời bỏ thấp nhất (khoảng 7%).
* **TechSupport (Hỗ trợ kỹ thuật):** Khách hàng không sử dụng dịch vụ hỗ trợ kỹ thuật có tỷ lệ rời bỏ cao nhất (khoảng 41%). Khách hàng có sử dụng dịch vụ hỗ trợ kỹ thuật có tỷ lệ rời bỏ thấp hơn (khoảng 15%). Khách hàng không có dịch vụ Internet có tỷ lệ rời bỏ thấp nhất (khoảng 7%).
* **StreamingTV (Xem TV trực tuyến) và StreamingMovies (Xem phim trực tuyến):** Xu hướng tương tự nhau. Khách hàng không sử dụng các dịch vụ xem TV/phim trực tuyến có tỷ lệ rời bỏ cao hơn (khoảng 34% cho TV và 33% cho Phim) so với khách hàng có sử dụng (khoảng 30% cho cả hai).Khách hàng không có dịch vụ Internet có tỷ lệ rời bỏ thấp nhất (khoảng 7%).

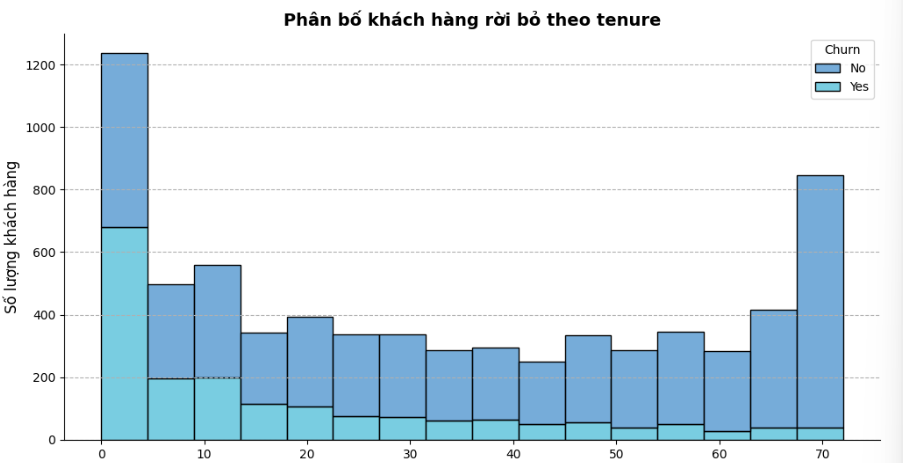
=> Khách hàng không sử dụng các dịch vụ bổ sung (bảo mật, sao lưu, bảo vệ thiết bị, hỗ trợ kỹ thuật) có xu hướng rời bỏ cao hơn. Điều này cho thấy các dịch vụ bổ sung này đóng vai trò quan trọng trong việc giữ chân khách hàng. Khách hàng sử dụng dịch vụ Fiber optic có tỷ lệ rời bỏ cao bất thường và cần được điều tra thêm. Khách hàng không có dịch vụ Internet có tỷ lệ rời bỏ thấp nhất trong tất cả các trường hợp.



**Nhận xét:**

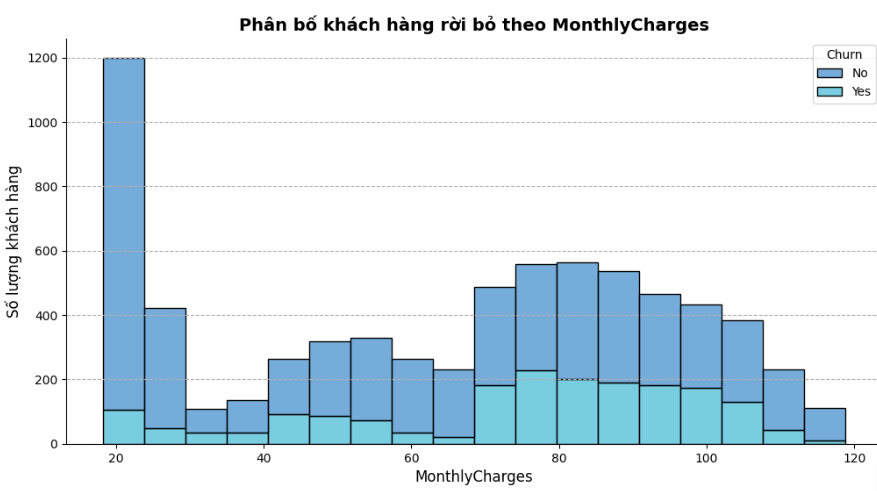
* **SeniorCitizen (Người cao tuổi):** Khách hàng là người cao tuổi có tỷ lệ rời bỏ cao hơn đáng kể (khoảng 41.7%) so với khách hàng không phải người cao tuổi(khoảng 23.6%). Điều này cho thấy người cao tuổi có thể gặp khó khăn với dịch vụ hoặc có nhu cầu khác biệt.
* **Partner (Có vợ/chồng):** Khách hàng không có vợ/chồng có tỷ lệ rời bỏ cao hơn (khoảng 33.0%) so với khách hàng có vợ/chồng (khoảng 19.2%). Có thể những người sống độc thân có xu hướng thay đổi dịch vụ thường xuyên hơn.
* **Dependents (Có người phụ thuộc):** Khách hàng không có người phụ thuộc có tỷ lệ rời bỏ cao hơn (khoảng 30.3%) so với khách hàng có người phụ thuộc (khoảng 15.5%). Tương tự như trường hợp có vợ/chồng, có thể việc có người phụ thuộc tạo sự ổn định hơn trong việc sử dụng dịch vụ.
* **PhoneService (Dịch vụ điện thoại):** Khách hàng có dịch vụ điện thoại có tỷ lệ rời bỏ (khoảng 24.9%) cao hơn một chút so với khách hàng không có dịch vụ điện thoại (khoảng 25.1%). Sự khác biệt này là không đáng kể. Điều này cho thấy dịch vụ điện thoại không ảnh hưởng nhiều đến quyết định rời bỏ.
* **PaperlessBilling (Hóa đơn điện tử):** Khách hàng sử dụng hóa đơn điện tử có tỷ lệ rời bỏ cao hơn (khoảng 33.6%) so với khách hàng không sử dụng hóa đơn điện tử (khoảng 16.3%). Điều này có thể do khách hàng gặp khó khăn với việc quản lý hóa đơn điện tử hoặc thích hình thức hóa đơn truyền thống hơn.

=> Người cao tuổi, người sống độc thân và người không có người phụ thuộc có xu hướng rời bỏ dịch vụ cao hơn. Đây là những nhóm khách hàng cần được quan tâm đặc biệt. Việc sử dụng hóa đơn điện tử cũng liên quan đến tỷ lệ rời bỏ cao hơn. Cần xem xét lại quy trình hoặc cung cấp thêm hỗ trợ cho khách hàng sử dụng hình thức này. Dịch vụ điện thoại dường như không ảnh hưởng nhiều đến quyết định rời bỏ.



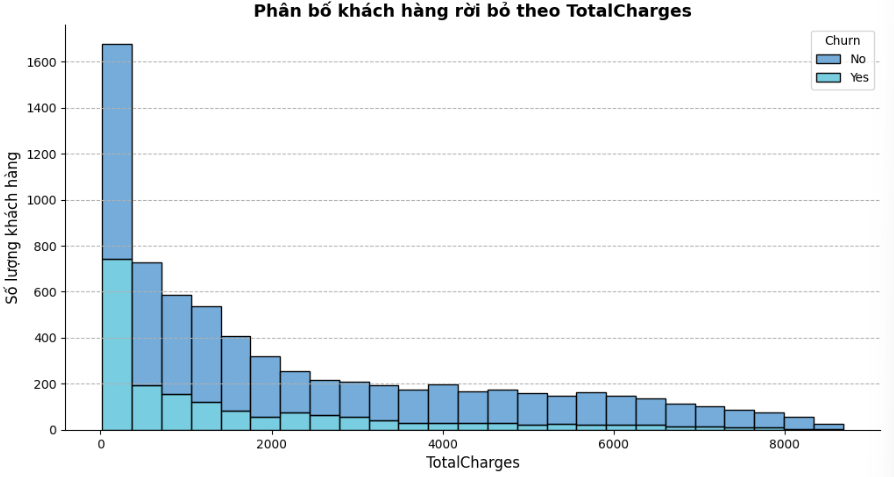
**Nhận xét:**

* **Khách hàng mới (tenure gần 0):** Tỷ lệ khách hàng rời bỏ dịch vụ *rất cao*. Có một lượng lớn khách hàng đăng ký dịch vụ nhưng rời đi ngay trong tháng đầu tiên.
* **Tenure từ 1 đến khoảng 20:** Tỷ lệ rời bỏ vẫn còn khá cao, nhưng bắt đầu giảm dần theo thời gian.
* **Tenure từ khoảng 20 trở lên**: Tỷ lệ rời bỏ dịch vụ khá thấp và tương đối ổn định. Điều này cho thấy khách hàng đã sử dụng dịch vụ được một thời gian thường có xu hướng tiếp tục sử dụng. Đặc biệt, ở khoảng tenure cao (từ 60 trở lên), số lượng khách hàng không rời bỏ dịch vụ tăng lên đáng kể.



**Nhận xét:**

* **Phí hàng tháng thấp (khoảng 20):** Điều này cho thấy khách hàng sử dụng các gói dịch vụ giá rẻ có xu hướng rời bỏ dịch vụ nhiều hơn.
* **Phí trung bình (khoảng 40-70):** Tỷ lệ khách hàng rời bỏ dịch vụ thấp hơn so với nhóm phí thấp.
* **Phí cao (khoảng 80-100):** Tỷ lệ khách hàng rời bỏ dịch vụ lại tăng lên, đặc biệt ở khoảng 80-90. Điều này cho thấy ngay cả khách hàng sử dụng các gói dịch vụ đắt tiền cũng có xu hướng rời bỏ.

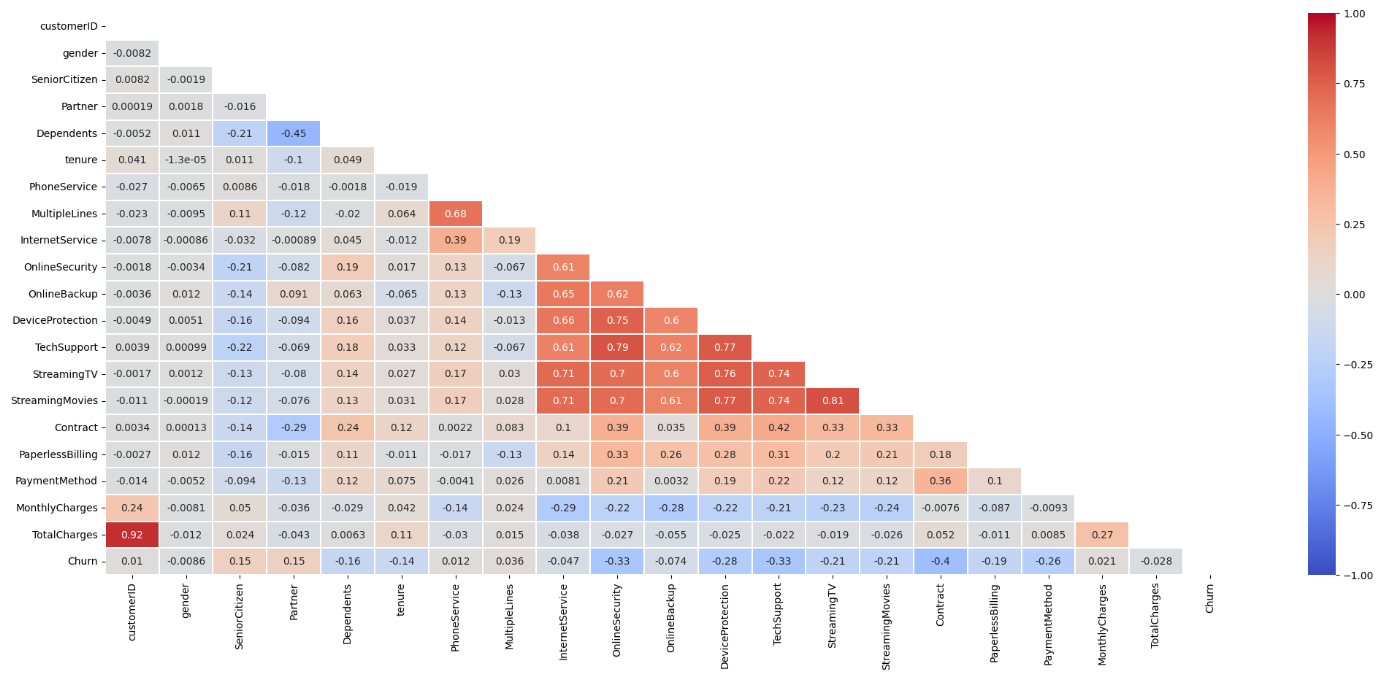


**Nhận xét:**

* **Phần lớn khách hàng có tổng phí thấp**: Một lượng lớn khách hàng có TotalCharges gần 0, và số lượng khách hàng giảm dần khi TotalCharges tăng lên.
* **Khách hàng có tổng phí thấp có xu hướng rời bỏ cao hơn:** Tỷ lệ khách hàng rời bỏ (màu xanh nhạt) đặc biệt cao ở những khách hàng có TotalCharges thấp (gần 0 đến khoảng 1000). Điều này cho thấy những khách hàng mới, chưa chi nhiều tiền cho dịch vụ, có xu hướng rời bỏ dịch vụ nhiều hơn.
* Khi TotalCharges tăng lên (từ khoảng 1000 trở lên), tỷ lệ khách hàng rời bỏ dịch vụ giảm đáng kể. Điều này cho thấy những khách hàng đã gắn bó với dịch vụ và chi nhiều tiền hơn có xu hướng tiếp tục sử dụng dịch vụ.

**Ma trận hệ số tương quan**

Ma trận hệ số tương quan đánh giá mức độ tương quan giữa các thuộc tính trong bộ dữ liệu.

****

**Kiểm tra mối tương quan của cột TotalCharges và tenure\*MonthlyCharges**



=> Có mối quan hệ tuyến tính gần như hoàn hảo.

# **CHƯƠNG V. XÂY DỰNG MÔ HÌNH DỰ ĐOÁN**

1. Mã hóa và chuẩn hóa dữ liệu
   1. Mã hóa

Dùng Label Encoding để mã hóa các thuộc tính với các giá trị cố định. Cụ thể

Partner, Dependents, PhoneService, PaperlessBilling, Churn, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport , StreamingTV, StreamingMovies.

* 1. Chuẩn hóa dữ liệu
  2. Rời rạc hóa dữ liệu bằng phương pháp rời rạc bằng khoảng

Column áp dụng: Tenure

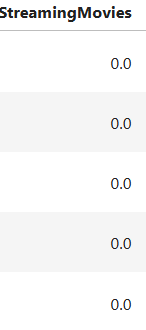
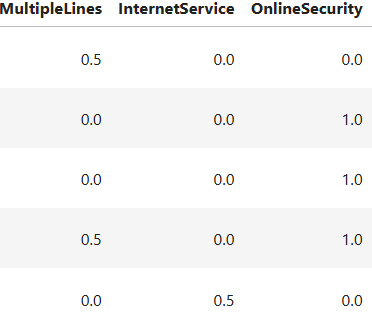
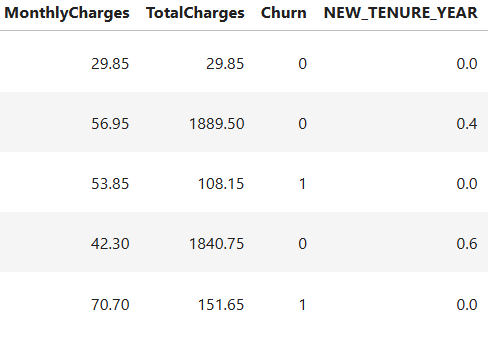
|  |  |
| --- | --- |
| Khoảng dữ liệu | NEW\_TENURE\_YEAR |
| 0-12 | 0-1 Year |
| 13-24 | 1-2 Year |
| 25-36 | 2-3 Year |
| 37-48 | 3-4 Year |
| 49-60 | 4-5 Year |
| 61-72 | 5-6 Year |

Ý nghĩa: giúp cải thiện hiệu suất mô hình, tăng tính dễ hiểu.

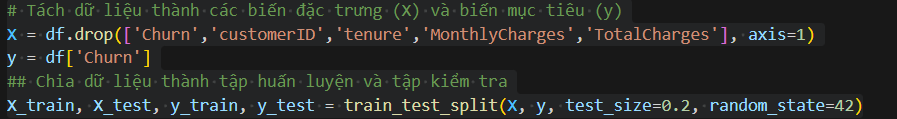
Sau khi rời rạc xong thì mã hóa dữ liệu bằng Label Encoding.

* 1. Chuẩn hóa dữ liệu bằng Min-Max Scaler

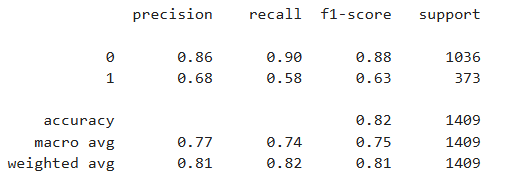
Áp dụng vào các thuộc tính MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport , StreamingTV, StreamingMovies, MonthlyCharges, TotalCharges,NEW\_TENURE\_YEAR.

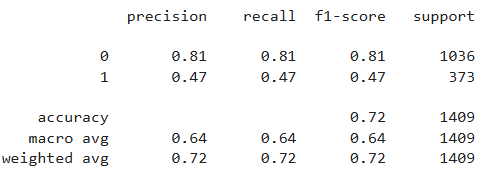
1. Xây dựng mô hình dự đoán

Chia dữ liệu thành 2 phần với 20% là test và 80% là train dữ liệu cho mô hình học máy:

1. Logistic Regression



1. Decision Tree



## 3. Lý do rời rạc và chuẩn hoá

Ta sẽ xem qua kết quả chạy của 2 mô hình:

A screenshot of a computer

Description automatically generated

Kết quả sau khi rời rạc hoá và chuẩn hoá

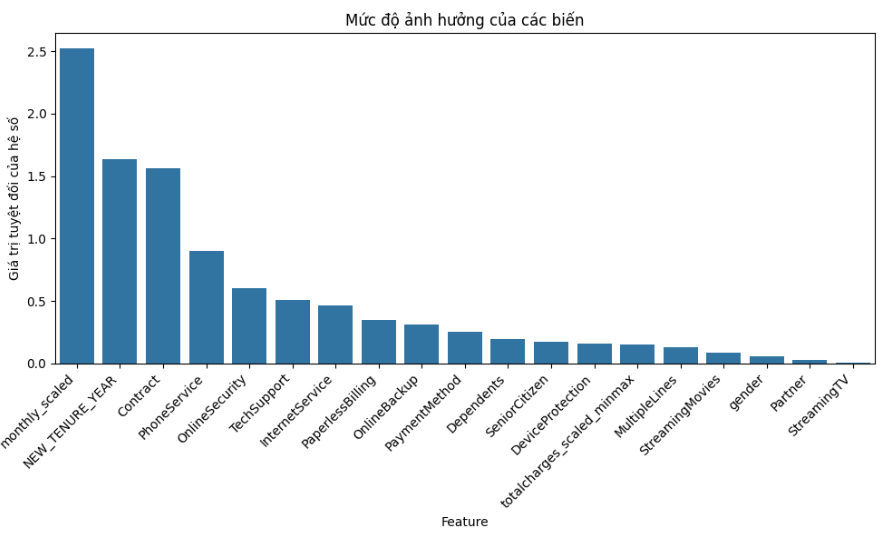
A screenshot of a computer

Description automatically generated

Nhận xét: Việc rời rạc hoá cột tenure và chuẩn hoá 3 cột MonthlyCharges, TotalCharges, NEW\_TENURE\_YEAR để đưa vào mô hình học máy đã làm cho độ chính xác tăng thêm.

# **CHƯƠNG VI. KẾT LUẬN**

Dựa trên cả bốn độ đo, Logistic Regression thể hiện hiệu suất tốt hơn so với Decision Tree trong bài toán phân loại này. Nó có độ chính xác tổng thể cao hơn, dự đoán chính xác hơn các mẫu positive và ít bỏ sót các mẫu positive hơn.



**Nhận xét:** Biểu đồ trên cho thấy mức độ ảnh hưởng của các thuộc tính khi huấn luyện bằng mô hình Logistic Regression. Ta có thể thấy rõ biến monthly\_scaled là ảnh hưởng nhiều nhất và có thể hiểu rằng việc khách hàng rời bỏ hay ở lại với dịch vụ phụ thuộc khá lớn vào việc số chi phí họ phải chi trả hàng tháng.