**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TPHCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

**------oOo-----**

A logo of hands holding a book

Description automatically generated

**ĐỀ TÀI BÁO CÁO**

**MÔN NHẬP MÔN KHOA HỌC DỮ LIỆU**

**PHÂN TÍCH**

**KHẢ NĂNG KHÁCH HÀNG   
CÓ THỂ TỪ BỎ DỊCH VỤ  
 ĐANG SỬ DỤNG**

**GVHD: *Ths. LÊ MINH TÂN***

**Lớp HP**: [***INDS331085\_23\_2\_02CLC***](https://fhqx.hcmute.edu.vn/course/view.php?id=57937)

**Học kì**:***II***

**Năm Học**: ***2023 – 2024***

***Thành phố Hồ Chí Minh, tháng 5 năm*** ***2024***

**NHẬN XÉT, ĐÁNH GIÁ CỦA GIÁO VIÊN**

**VÀ DANH SÁCH SINH VIÊN THỰC HIỆN**

Đề tài**:** ***Phân tích khả năng khách hàng có thể từ bỏ dịch vụ đang sử dụng.***

|  |  |
| --- | --- |
| **Nhận xét của giáo viên** | **Điểm số** |
|  |  |

***Danh sách thành viên***

|  |  |  |  |
| --- | --- | --- | --- |
| **MSSV** | **Họ và tên** | **Mức độ hoàn thành** | **Nhiệm vụ** |
| 21110234 | Nguyễn Thành Lợi | 100% | Giới thiệu, phân tích Dataset “Internet Service Provider Customer Churn” |
| 21110836 | Mai Anh Khoa | 100% | Giới thiệu, phân tích Dataset “Credit Card Churn Prediction” |
| 21110168 | Huỳnh Minh Đức | 100% | Giới thiệu, phân tích Dataset “Telecom Company Churn Rate, Call Center Data” |

*Nhóm trưởng: Mai Anh Khoa*

**Tp. Hồ Chí Minh, tháng 5 năm 2024**

**MỤC LỤC**

[**CHƯƠNG 1: TỔNG QUAN VỀ ĐỀ TÀI** 1](#_Toc165745763)

[**1.1** **LÝ DO CHỌN ĐỀ TÀI** 1](#_Toc165745764)

[**1.2** **MÔ TẢ ĐỀ TÀI** 1](#_Toc165745765)

[**1.3** **DANH SÁCH DATASETS** 2](#_Toc165745766)

[**CHƯƠNG 2: GIỚI THIỆU VỀ DATASETS** 3](#_Toc165745767)

[**2.1 DATASET *“Internet Service Provider Customer Churn”*** 3](#_Toc165745768)

[**2.1.1 Giới thiệu chung về bối cảnh, khó khăn trong việc giữ chân khách hàng của dịch vụ Internet** 3](#_Toc165745769)

[**2.1.2 Giới thiệu về dataset** 4](#_Toc165745770)

[**2.1.3 Tổng quan dữ liệu** 5](#_Toc165745771)

[**2.1.4 Phân bố giá trị trong các thuộc tính** 6](#_Toc165745772)

[**2.2 DATASET *“Credit Card Churn Prediction”*** 10](#_Toc165745773)

[**2.2.1** **Giới thiệu về dataset** 10](#_Toc165745774)

[**2.2.2 Các thuộc tính của dữ liệu** 11](#_Toc165745775)

[**2.2.3 Phân bố giá trị trong các thuộc tính** 12](#_Toc165745776)

[**2.3 DATASET *“Telecom Company Churn Rate, Call Center Data”*** 25](#_Toc165745777)

[**2.3.1 Giới thiệu về dataset** 25](#_Toc165745778)

[**2.3.2 Phân bố giá trị trong các thuộc tính** 26](#_Toc165745779)

[**CHƯƠNG 3: PHÂN TÍCH DATASETS** 33](#_Toc165745780)

[**3.1 DATASET *“Internet Service Provider Customer Churn”*** 33](#_Toc165745781)

[**3.1.1 Phân tích thuộc tính “churn”** 33](#_Toc165745782)

[**3.1.2 Phân tích yếu tố ảnh hưởng “churn”** 33](#_Toc165745783)

[**3.1.3 Kết luận thu được sau khi thực hiện việc phân tích dữ liệu** 40](#_Toc165745784)

[**3.1.4 Ưu điểm và nhược điểm của dataset** 40](#_Toc165745785)

[**3.2 DATASET *“Credit Card Churn Prediction”*** 41](#_Toc165745786)

[**3.2.1. Tỉ lệ churn** 41](#_Toc165745787)

[**3.2.2. Mối quan hệ giữa tỉ lệ churn với các thuộc tính phân loại (category column)** 42](#_Toc165745788)

[**3.2.3. Mối quan hệ giữa tỉ lệ churn với các thuộc tính định lượng (numerical column)** 45](#_Toc165745789)

[**3.3.4 Bản đồ nhiệt cho sự tương quan giữa các thuộc tính** 46](#_Toc165745790)

[**3.3.5 Mối tương quan giữa tỉ lệ rời bỏ và mỗi giá trị trong các cột phân loại (category columns)** 47](#_Toc165745791)

[**3.3.6 Kết luận thu được sau khi thực hiện việc phân tích dữ liệu** 48](#_Toc165745792)

[**3.3.7 Ưu điểm và nhược điểm của dataset** 48](#_Toc165745793)

[**3.3 DATASET *“Telecom Company Churn Rate, Call Center Data”*** 49](#_Toc165745794)

[**3.3.1 Phân tích thuộc tính “churn”** 49](#_Toc165745795)

[**3.3.2 Phân tích yếu tố ảnh hưởng “churn”** 49](#_Toc165745796)

[**3.3.3 Bản đồ nhiệt cho sự tương quan giữa các thuộc tính và mối tương quan giữa tỉ lệ rời bỏ và mỗi giá trị trong các cột** 53](#_Toc165745797)

[**3.3.4 Ưu điểm và nhược điểm của dataset** 54](#_Toc165745798)

[**3.4 NHẬN XÉT CHUNG** 55](#_Toc165745799)

[**TÀI LIỆU THAM KHẢO** 56](#_Toc165745800)

# **CHƯƠNG 1: TỔNG QUAN VỀ ĐỀ TÀI**

* 1. **LÝ DO CHỌN ĐỀ TÀI**

Lựa chọn đề tài ***“Phân tích và dự đoán khả năng khách hàng có thể từ bỏ dịch vụ đang sử dụng”*** mang lại nhiều lợi ích quan trọng trong lĩnh vực kinh doanh và quản lý khách hàng. Dưới đây là một số lý do nhóm chúng em chọn đề tài này.

*Nắm bắt yếu tố ảnh hưởng và cải thiện dịch vụ:* Phân tích dữ liệu từ khách hàng giúp xác định vấn đề chung và tìm ra giải pháp cải thiện dịch vụ, tăng cường sự hài lòng của khách hàng.

*Giảm tỷ lệ từ bỏ dịch vụ:* Dự đoán khả năng từ bỏ giúp xác định nhóm khách hàng có nguy cơ cao từ bỏ và triển khai biện pháp giữ chân, giảm thiểu tỷ lệ này.

*Cải thiện trải nghiệm khách hàng:* Phân tích dữ liệu giúp nhận ra điểm yếu trong trải nghiệm khách hàng và cải thiện chúng, tạo ra trải nghiệm tốt hơn và tăng cường lòng trung thành của khách hàng.

*Tối ưu hóa chi phí và tài nguyên****:***Tập trung vào duy trì và giữ chân khách hàng hiện tại giúp giảm chi phí tiếp thị và quảng cáo, tăng hiệu quả trong việc duy trì mạng lưới khách hàng.

*Đồng bộ hóa chiến lược phát triển:* Phân tích và dự đoán từ bỏ dịch vụ giúp nhận biết các khía cạnh không hài lòng của khách hàng và thay đổi sản phẩm, dịch vụ để đáp ứng tốt hơn nhu cầu của họ.

*Áp dụng kiến thức vào thực tế:* Thực hiện dự án giúp sinh viên áp dụng kiến thức vào thực tế thông qua các bước phân tích dữ liệu thực tế, hiểu rõ hơn về quá trình làm việc với dữ liệu và áp dụng các phương pháp trong lĩnh vực khoa học dữ liệu.

* 1. **MÔ TẢ ĐỀ TÀI**

Đề tàitập trung vào việc phân tích và dự đoán khả năng khách hàng từ bỏ dịch vụ mà họ đang sử dụng. Nghiên cứu sẽ tập trung vào việc hiểu rõ các yếu tố nào ảnh hưởng đến quyết định này và làm thế nào để chúng ta có thể dự đoán khả năng từ bỏ dịch vụ của khách hàng từ đó đưa ra được các chiến lược giúp giảm thiểu hành vi này trong tương lai.

Đồ án này sẽ được thực hiện trong khoảng thời gian học kỳ 2 của năm 3 (từ tháng 1/2024 đến tháng 5/2024). Phạm vi nghiên cứu của đề tài sẽ không chỉ tập trung vào các khách hàng đã từ bỏ dịch vụ, mà đồng thời sẽ mở rộng để xem xét các yếu tố nền tảng khác có thể kể đến như thị trường, xu hướng ngành và các yếu tố về kinh tế.

Đối tượng nghiên cứu của đề tài sẽ tập trung vào các khách hàng đang sử dụng dịch vụ và những đối tượng liên quan trong hệ sinh thái dịch vụ đó. Cụ thể hơn, nghiên cứu sẽ xem xét đến các yếu tố quan trọng có thể đo lường và quan sát được, có thể kể đến các yếu tố như hành vi sử dụng dịch vụ, hậu quả kinh tế, hậu quả dịch bệnh và các đánh giá về chất lượng dịch vụ. Dựa vào đó, ta có thể quan sát và đo lường trên một tập đối tượng để có thể đưa ra được những kết luận, áp dụng cho đối tượng mục tiêu lớn hơn.

Nghiên cứu cũng sẽ xem xét đến các yếu tố có thể ảnh hưởng đến độ hữu ích của dữ liệu. Có thể kể đến như tác động của thời gian, khi mà dữ liệu được nghiên cứu có khả năng không phản ánh đúng tình hình hiện tại hoặc các yếu tố địa lý như văn hóa, dân số, môi trường, và pháp luật cũng cần phải được xem xét kỹ lưỡng do các yếu tố trên cũng ảnh hưởng đến quyết định từ bỏ dịch vụ của khách hàng.

Ngoài ra, đối với đối tượng mẫu, đồ án cũng sẽ xác định những yếu tố phổ biến quan trọng như chi phí, chất lượng dịch vụ, sự hài lòng và các yếu tố tâm lý khác. Thông qua đó, ta có thể nghiên cứu được những yếu tố ảnh hưởng đến quyết địch từ bỏ dịch vụ, bao gồm cả yếu tố nội tại (như sự hài lòng của khách hàng) và các yếu tố ngoại vi (như thị trường cạnh tranh).

Về phương thức thu thập dữ liệu, nghiên cứu sẽ tập trung chủ yếu vào các dữ liệu có sẵn từ các nền tảng trực tuyến do số lượng và chất lượng của các dữ liệu đó đa phần đều đã được kiểm chứng.

* 1. **DANH SÁCH DATASETS**
* “Internet Service Provider Customer Churn”,

<https://www.kaggle.com/datasets/mehmetsabrikunt/internet-service-churn>

* “Credit Card Churn Prediction”,

<https://www.kaggle.com/datasets/anwarsan/credit-card-bank-churn>

* “Telecom Company Churn Rate, Call Center Data”,

<https://www.kaggle.com/datasets/datazng/telecom-company-churn-rate-call-center-data/data>

# **CHƯƠNG 2: GIỚI THIỆU VỀ DATASETS**

## **2.1 DATASET *“Internet Service Provider Customer Churn”***

### **2.1.1 Giới thiệu chung về bối cảnh, khó khăn trong việc giữ chân khách hàng của dịch vụ Internet**

Dịch vụ Internet đã trở thành một phần không thể thiếu trong cuộc sống hàng ngày của chúng ta. Với sự phát triển nhanh chóng của công nghệ và sự gia tăng của các ứng dụng trực tuyến, người dùng ngày càng có nhu cầu cao về kết nối Internet ổn định và tốc độ nhanh.Tuy nhiên, ngành công nghiệp Internet Service Provider (ISP) cũng đối mặt với những thách thức lớn trong việc giữ chân khách hàng. Việc giảm tỷ lệ churn, tức là việc ngăn chặn khách hàng hủy dịch vụ và chuyển sang nhà cung cấp khác, đang trở thành vấn đề quan trọng.

Một trong những khó khăn chính là sự cạnh tranh gay gắt giữa các nhà cung cấp dịch vụ Internet. Có nhiều công ty cung cấp dịch vụ tương tự và khách hàng có nhiều lựa chọn để chuyển đổi nếu họ không hài lòng với dịch vụ hiện tại. Điều này tạo ra một môi trường cạnh tranh mạnh mẽ, khiến việc giữ chân khách hàng trở nên khó khăn hơn.

Một thách thức khác là sự không ổn định trong chất lượng dịch vụ Internet. Khách hàng yêu cầu một kết nối ổn định và tốc độ cao để có thể truy cập các ứng dụng trực tuyến mà họ cần. Nếu mạng Internet của một nhà cung cấp không đáp ứng được yêu cầu này, khách hàng có thể sẽ tìm kiếm các lựa chọn khác.

Thành công trong việc giữ chân khách hàng của ISP cũng phụ thuộc vào mối quan hệ tốt với khách hàng. Sự hài lòng của khách hàng với dịch vụ, hỗ trợ và tương tác với nhân viên đóng vai trò quan trọng trong việc quyết định khách hàng có tiếp tục sử dụng dịch vụ hay không. Các công ty cũng phải đối mặt với việc xử lý các yếu tố giá cả và sản phẩm. Khách hàng có thể chuyển đổi sang nhà cung cấp khác nếu họ tìm thấy một gói sản phẩm tốt hơn hoặc giá cả cạnh tranh hơn.

Để giảm tỷ lệ churn, các nhà cung cấp dịch vụ Internet cần tìm hiểu và đáp ứng nhu cầu của khách hàng một cách tốt nhất. Điều này bao gồm cải thiện chất lượng dịch vụ, cung cấp giá cả cạnh tranh, tạo mối quan hệ tốt với khách hàng và tăng cường hoạt động tiếp thị để tạo niềm tin và sự hài lòng từ phía khách hàng.

Việc giữ chân khách hàng là một thách thức quan trọng đối với các nhà cung cấp dịch vụ Internet. Để thành công trong việc giảm tỷ lệ churn, họ cần tập trung vào cải thiện chất lượng dịch vụ, xây dựng mối quan hệ tốt với khách hàng và đáp ứng các yêu cầu và mong muốn của khách hàng một cách tốt nhất.

### **2.1.2 Giới thiệu về dataset**

- Bộ dữ liệu *“Internet Service Provider Customer Churn”* cung cấp thông tin về việc liệu khách hàng có từ bỏ nhà cung cấp dịch vụ Internet hay không, đồng thời chứa các thuộc tính có thể ảnh hưởng đến quyết định này.

- Bộ dữ liệu được thu thập vào năm 2022 chứa 72275 hàng và 11 cột. Các thuộc tính trong bộ dữ liệu bao gồm:

+ id : Mã định danh khách hàng

+ is\_tv\_subscriber: Khách hàng có đăng ký tv hay không, giá trị bao gồm 0 là không đăng ký và 1 là đăng ký

+ is\_movie\_package\_subscriber: Khách hàng có đăng ký gói xem phim điện ảnh hay không, giá trị bao gồm 0 là không đăng ký và 1 là đăng ký

+ subscription\_age: Thời gian khách hàng sử dụng dịch vụ của nhà cung cấp, giá trị đo lường theo năm

+ bill\_avg: hóa đơn trung bình 3 tháng qua - tính từ thời điểm thu thập dữ liệu

+ reamining\_contract: Thời hạn còn lại của hợp đồng với khách hàng, giá trị đo lường theo năm.

* Nếu null; khách hàng không có hợp đồng. Khách hàng có thời gian hợp đồng phải sử dụng dịch vụ của họ cho đến khi kết thúc hợp đồng. Nếu họ hủy dịch vụ trước khi kết thúc thời gian hợp đồng, họ sẽ phải trả một khoản phí phạt.
* Có hai cách mà khách hàng có thể sử dụng dịch vụ. Thông qua hợp đồng có thời hạn với chi phí thấp hơn hoặc thông qua hợp đồng đăng ký hàng tháng thông thường rõ ràng là có chi phí cao hơn. Vì vậy, ngay cả khi một người không ký hợp đồng, họ vẫn là người dùng trả phí hàng tháng.

+ service\_failure\_count: số cuộc gọi đến trung tâm vì lỗi dịch vụ trong 3 tháng qua

+ download\_avg: sử dụng internet 3 tháng qua - tính từ thời điểm thu thập dữ liệu, giá trị đo lường theo GB

+ upload\_avg: trung bình tải lên 3 tháng qua - tính từ thời điểm thu thập dữ liệu, giá trị đo lường theo GB

+ download\_over\_limit: giới hạn vượt quá số lượng trong 9 tháng qua. Hầu hết khách hàng đều có giới hạn tải xuống. Nếu họ đạt đến giới hạn này, họ phải trả tiền cho việc này.

+ churn: đây là thuộc tính mục đích – là 1 nếu khách hàng có từ bỏ nhà cung cấp dịch vụ Internet, ngược lại là 0

### **2.1.3 Tổng quan dữ liệu**

- Giá trị null chiếm nhiều sẽ ảnh hưởng quá trình phân tích dữ liệu, việc phân tích và xử lý giá trị null giúp cho giúp cho việc phân tích dữ liệu trở nên chính xác và đáng tin cậy hơn.

A screenshot of a computer code

Description automatically generated

- Xử lý dữ liệu null cột reamining\_contract

A close-up of a computer screen

Description automatically generated

- Xử lý dữ liệu null cột download\_avg và upload\_avg

A close up of text

Description automatically generated

### **2.1.4 Phân bố giá trị trong các thuộc tính**

|  |  |
| --- | --- |
| **Thuộc tính** | **Trực quan phân bố giá trị** |
| is\_tv\_subscriber |  |
| **Nhận xét:**  Số lượng khách hàng có đăng ký tv chiếm phần lớn (58922 người) và số lượng khách hàng không có đăng ký chiếm phần ít hơn có đăng ký (13352 người).. |
| is\_movie\_package\_subscriber |  |
| **Nhận xét:**  Số lượng khách hàng có đăng ký gói xem phim điện ảnh (24185 người) ít hơn khoảng một nửa số người không đăng ký (48089 người). |
| **Nhận xét chung:** is\_tv\_subscriber và is\_movie\_package\_subscribe đều có khách hàng có đăng ký chiếm đa số. Vì thế 2 dịch vụ này chính là 2 dịch vụ thu hút khách hàng, trong đó is\_tv\_subscriber có vẻ thu hút được nhiều khách hàng hơn. | |
| subscription\_age |  |
| **Nhận xét:**   * Khách hàng đăng ký có thời gian sử dụng lâu nhất là 12.8 năm * Số lượng khách hàng tỷ lệ nghịch với thời gian sử dụng dịch vụ của nhà cung cấp, khách hàng đăng ký có thời gian sử dụng khoảng 0 đến 2 năm với số lượng khá cao (trên 12000 người) – cao nhất là trên 14000 người, và số lượng giảm dần theo thời gian sử dụng tăng. * Tỉ lệ nghịch giữa số lượng khách hàng và thời gian sử dụng dịch vụ có thể có nhiều nguyên nhân:   + Chuyển đổi và thử nghiệm: Một số khách hàng mới có thể đăng ký dịch vụ TV để thử nghiệm hoặc do một chương trình khuyến mãi. Tuy nhiên, sau khi thử nghiệm, họ có thể quyết định không tiếp tục sử dụng dịch vụ nếu họ không cảm thấy hài lòng hoặc không cần thiết. Điều này có thể dẫn đến một số lượng lớn khách hàng mới đăng ký nhưng sau đó giảm dần theo thời gian.   + Sự đa dạng hóa của nhu cầu: Khi thời gian sử dụng tăng lên, nhu cầu của khách hàng có thể thay đổi. Có thể họ tìm kiếm các dịch vụ hoặc sản phẩm mới phù hợp hơn với nhu cầu hiện tại của họ hoặc đáp ứng các yêu cầu cụ thể hơn.   + Sự cạnh tranh và phát triển: Thị trường thường xuyên thay đổi với sự xuất hiện của các đối thủ mới và các lựa chọn thay thế. Khi có sự cạnh tranh mạnh mẽ hoặc các dịch vụ mới xuất hiện, một số khách hàng có thể chuyển sang sử dụng các lựa chọn mới này thay vì tiếp tục sử dụng dịch vụ hiện tại. |
| bill\_avg |  |
| **Nhận xét:**   * Hóa đơn trung bình trong 3 tháng nhiều nhất là 406 đô la * Hóa đơn trung bình trong 3 tháng qua nằm trong khoảng nhỏ hơn 50 đô la là chủ yếu * Số hóa đơn tỷ lệ nghịch số lượng khách hàng,có nghĩa là phần lớn khách hàng thực hiện ít hóa đơn hoặc là sử dụng các dịch vụ ít tốn tiền. |
| reamining\_contract |  |
| **Nhận xét:**   * Thời hạn còn lại của hợp đồng cao nhất là 2.92 năm. * Phần lớn khách hàng không có hợp đồng hoặc hết hợp đồng (thời hạn còn lại bằng 0) |
| service\_failure\_count |  |
| **Nhận xét:**  - Số cuộc gọi đến trung tâm vì lỗi dịch vụ trong 3 tháng qua đa số nhỏ hơn 2 cuộc (68072 người).  - Phần khách hàng còn lại gọi vì lỗi chiếm khá ít (dưới 5000 người) và số cuộc gọi nhiều nhất là 19 cuộc. |
| download\_avg |  |
| **Nhận xét:**   * Đa số khách hàng download trung bình 3 tháng qua cao nhất là 4415.2GB và đa số nhỏ hơn 252 GB (trên 70000 người). * Theo bài viết [Average Internet Data Usage](https://headphonesaddict.com/average-internet-data-usage/), người dùng internet trung bình ở Mỹ đã tiêu thụ 154.6 GB dữ liệu cố định và di động mỗi tháng vào năm 2021, có vẻ như có sự khác biệt lớn giữa lượng dữ liệu download trung bình của người dùng ở Mỹ và khách hàng trong dataset. Điều này cho thấy nhu cầu download của khách hàng trong dataset không quá cao. |
| upload\_avg |  |
| **Nhận xét:**   * Đa số khách hàng upload trung bình 3 tháng qua cao nhất là 453.3GB và đa số nhỏ hơn 36 GB (trên 70000 người). * Tương tự cũng trong bài viết [Average Internet Data Usage](https://headphonesaddict.com/average-internet-data-usage/), người upload internet trung bình ở Mỹ trên 30GB dữ liệu mỗi tháng vào năm 2021. Cũng như download, nhu cầu upload của khách hàng trong dataset không quá cao. |
| download\_over\_limit |  |
| **Nhận xét:**   * Đa số khách hàng không download quá giới hạn trong 9 tháng qua (68373 người) * Giá trị download quá giới hạn cao nhất khoảng là 7 GB * Download quá giới hạn thông thường sẽ được thông báo trước khi khách hàng chạm mức và quá giới hạn sẽ bị tính phí thêm nên đa phần khách hàng sẽ không sử dụng quá giới hạn quá nhiều |

## **2.2 DATASET *“Credit Card Churn Prediction”***

### **Giới thiệu về dataset**

Bộ dữ liệu “*Credit Card Churn Prediction*” được tạo ra nhằm dùng để phân tích dữ liệu và tìm ra nguyên nhân đằng sau vấn đề suy giảm khách hàng của một ngân hàng thẻ tín dụng tiêu dùng và tận dụng để dự đoán những khác hàng có xu hướng sẽ rời bỏ việc sử dụng dịch vụ của ngân hàng.

Bộ dữ liệu bao gồm 23 cột và 10127 dòng, không có cột nào chưa giá trị null.

Đối với dataset này ta sẽ lấy toàn bộ dataset làm dữ liệu mẫu. Điều này là do trong bộ dữ liệu không có sự chênh lệch quá nhiều về một phía (giả sử như giới tính được chia với tỉ lệ gần 50:50, tỉ lệ hôn nhân hay học vấn cũng được chia đều), đồng thời, theo mô tả trên trang *kaggle.com*, tỉ lệ valid của toàn bộ thuộc tính là 100%, mismatched là 0% và missing là 0% thế nên kết luận là lấy toàn bộ dataset làm bộ dữ liệu mẫu là hợp lý.

Tuy nhiên, ta sẽ bỏ đi dữ liệu của hai cột cuối là “Naive\_Bayes\_Classifier\_Attrition\_Flag\_Card\_Category\_Contacts\_Count\_12\_mon\_Dependent\_count\_Education\_Level\_Months\_Inactive\_12\_mon\_1” và “Naive\_Bayes\_Classifier\_Attrition\_Flag\_Card\_Category\_Contacts\_Count\_12\_mon\_Dependent\_count\_Education\_Level\_Months\_Inactive\_12\_mon\_2” do tác giả khuyến nghị.

### **2.2.2 Các thuộc tính của dữ liệu**

**# CLIENTNUM:** mã khách hàng – là mã định danh độc nhất cho mỗi khách hàng đang nắm giữ tài khoản

**# Attrition\_Flag:** là một biến sự kiến nội bộ thể hiện hoạt động của khác hàng. Gồm hai giá tri: bằng 1 nếu tài khoản đã đóng và bằng 0 nếu ngược lại.

**# Customer\_Age:** Độ tuổi của khách hàng.

**# Gender:** giới tính. Gồm hai giá trị: M=Male và F=Female.

**# Dependent\_count:** biểu thị số người phụ thuộc (ví dụ: con cái, vợ/chồng) mà thẻ tín dụng này đang hỗ trợ

**# Education\_Level:** Trình độ học vấn của chủ tài khoản (ví dụ: tốt nghiệp cấp 3, cao đẳng, v.v)

**# Marital\_Status:** Tình trạng hôn nhân, gồm các giá trị: Married (đã kết hôn), Single (độc thân), Disvorced (đã ly hôn), và Unknown (chưa biết).

**# Income\_Category:** thu nhập hàng năm của chủ tài khoản.

**# Card\_Category:** loại thẻ (xanh, bạc, vàng, bạch kim)

**# Months\_on\_book:** khoảngthời gian làm việc với ngân hàng

**# Total\_Relationship\_Count:** tổng số sản phẩm mà khách hàng nắm giữ

**# Months\_Inactive\_12\_mon:** khoảng thời gian không hoạt động trong vòng 12 tháng gần nhất.

**# Contacts\_Count\_12\_mon:** số lượng thực hiện liên lạc trong 12 tháng gần nhất.

**# Credit\_Limit:** giới hạn tài khoản của thẻ tín dụng.

**# Total\_Revolving\_Bal:** số tiền đã chi tiêu bằng thẻ tín dụng mà không được thanh toán vào cuối chu kỳ thanh toán

**# Avg\_Open\_To\_Buy:** số tiền tín dụng người dùng có thể sử dụng để không vượt quá giới hạn tín dụng đã được xác định (trung bình trong 12 tháng gần nhất).

**# Total\_Amt\_Chng\_Q4\_Q1:** thay đổi trong tổng số tiền giao dịch (từ quý 4 đến quý 1)

**# Total\_Trans\_Amt:** tổng số tiền đã giao dịch trong khoảng 12 tháng gần nhất

**# Total\_Trans\_Ct:** tổng số lượng đã giao dịch trong khoảng 12 tháng gần nhất

**# Total\_Ct\_Chng\_Q4\_Q1:** thay đổi trong số lượng giao dịch từ quý 4 đến quý 1

**# Avg\_Utilization\_Ratio:** tỉ lệ sử dụng thẻ trung bình

### **2.2.3 Phân bố giá trị trong các thuộc tính**

|  |  |  |
| --- | --- | --- |
| **Thuộc tính** | **Trực quan phân bố giá trị** | **Thống kê** |
| CLIENTNUM | **A graph of blue squares  Description automatically generated** | Max: 828343083  Min: 708082083  Mean: 717926358.0  Std: 36903783.5 |
| Attrition\_Flag | **A blue and red pie chart  Description automatically generated** | Tỉ lệ Existing Customer (khách hàng vẫn còn tài khoản): 83.9%  Tỉ lệ Attrited Customer (khách hàng không còn tài khoản): 16.1%  Tỉ lệ phổ biến: tỉ lệ Existing Customer |
| Customer\_Age |  | |
| **Nhận xét:** Ta thấy rằng giá trị của thuộc tính customer\_age tuân theo phân phối chuẩn và những khách hàng có độ tuổi lớn hơn 70 tuổi được đánh nhãn là những ngoại lệ của thuộc tính. Tuy nhiên, ở rất nhiều quốc gia, những khách hàng ở độ tuổi này sỡ hữu và sử dụng thẻ tín dụng là bình thường. Vậy nên trong bài phân tích này những giá trị bất thường về tuổi tác này có thể không được xem là một ngoại lệ. | |
| **Phân tích ngoại lệ:** | |
| **Nhận xét**: Không thấy có sự bất thường trong hành vi của những khách hàng lớn tuổi. | |
| Gender | **A blue and red circle with black text  Description automatically generated** | Giá trị độc nhất: 2 (Male và Female)  Tỉ lệ Male (nam): 47.1%  Tỉ lệ Female (nữ): 52.9%  Tỉ lệ phổ biến: tỉ lệ Female |
| Dependent\_count |  | |
| **Nhận xét**: Theo kết quả trên, ta nhận thấy:   * Đồ thị phân bố có xu hướng lệch dương, điều này cho thấy hầu hết các tài khoản có sự phụ thuộc nhiều hơn các tài khoản không có lượng phụ thuộc. * Hầu hết các khách hàng đều có số người phụ thuộc vài tài khoản ấy vào khoảng 2-3 người. * Không có các giá trị ngoại lệ. | |
| Education\_Level | **A pie chart with numbers and a black background  Description automatically generated** | Giá trị độc nhất: 7  Tỉ lệ Graduate (tốt nghiệp): 30.9%  Tỉ lệ High School (trung học): 19.9%  Tỉ lệ others (các loại trình độ khác): 49.2%  Tỉ lệ phổ biến: tỉ lệ Graduate |
| Marital\_Status | **A pie chart with numbers and a black background  Description automatically generated** | Giá trị độc nhất: 4  Tỉ lệ Married (đã kết hôn): 46.3%  Tỉ lệ Single (độc thân): 38.9%  Tỉ lệ others: 14.8%  Tỉ lệ phổ biến: tỉ lệ Married |
| Income\_Category | **A pie chart with numbers and a black background  Description automatically generated** | Giá trị độc nhất: 6  Tỉ lệ Less than $40k : 35.2%  Tỉ lệ 40k-60k: 17.7%  Tỉ lệ others: 47.2%  Tỉ lệ phổ biến: tỉ lệ Less than $40k |
| Card\_Category | **A blue circle with red and black text  Description automatically generated** | Giá trị độc nhất: 4  Tỉ lệ Blue: 93.2%  Tỉ lệ others: 6.8%  Tỉ lệ phổ biến: tỉ lệ Blue |
| Months\_on\_book |  | |
| **Nhận xét**: Dựa trên phần trực quan hóa dữ liệu trên, ta có một số nhận xét như sau:   * Quan sát biểu đồ historgram, ta nhận thấy dữ liệu dường như có phân phối chuẩn. Trong khi đó ở biểu đồ boxplot, ta nhận thấy ràng các khách hàng gắn bó với ngân hàng từ 0-18 tháng và những khách háng có trên 55 tháng gắn bó với ngân hàng được xem là những ngoại lệ. * Tuy nhiên, trên thực tế, ngân hàng có thể có những khách hàng mới hoặc những khách hàng gắn bó rất lâu, thế nên, những dữ liệu này sẽ không được xem là những ngoại lệ. Việc biểu đồ boxplot phân loại các dữ liệu này là những dữ liệu ngoại lệ là do những khách hàng có số tháng gắn bó với ngân hàng là 36 tháng có số lượng nhiều hơn rất nhiều so với những điểm dữ liệu khác. * Như đã nói trước đó, dữ liệu về months\_on\_book dường như tuân theo phân phối chuẩn do các giá trị bên trái và bên phải của trung vị dường như bằng nhau. Tuy nhiên, số liệu này có thể bị ảnh hưởng do có đến tận 3500 khách hàng có 36 tháng gắn bó với ngân hàng. Điều này nảy sinh một số nghi ngờ đối với sự phân phối của dữ liệu. | |
| **Phân tích ngoại lệ:** | |
| **Nhận xét**: Nhìn chung ta thấy sự phân bổ của hai tệp khách hàng này tương đối giống nhau và chỉ một phần của những tệp khách hàng này đã rời bỏ ngân hàng. | |
| Total\_Relationship\_Count |  | |
| **Nhận xét**:   * Biểu đồ phân bố có thiên hướng lệch âm, tức là những giá đầu sẽ có số lượng lớn hơn những giá trị sau. * Không có giá trị ngoại lệ. * Hầu hết các khách hàng sở hữu từ 3-4 sản phẩm từ ngân hàng | |
| Months\_Inactive\_12\_mon |  | |
| **Nhận xét**:   * Đồ thị phân bố có thiên hướng lệch trái. * Các giá trị ngoại lệ được phân loại là những khách hàng có 0, 5, 6 tháng không không thực hiện hoạt động nào trong vòng 12 tháng. * Với các khách hàng cố số tháng không hoạt động là 0, tức là trong 12 tháng, những khách hàng này luôn hoạt động. Ngoài ra còn có những khách hàng không hoạt động trong vòng nửa năm. Những giá trị này tương đối bình thường và đều là những trường hợp có thể xảy ra trong thực tế nên chúng ta sẽ không phân loại chúng là những ngoại lệ. Chúng ta sẽ thực hiện phân tích hành vi của những giá trị này ngay sau đây. | |
| **Phân tích ngoại lệ:** | |
| **Nhận xét**: Dựa vào các biểu đồ trên, ta đưa ra được một số nhận xét như sau:   * Với các biểu đồ bên trái (số tháng là 0), ta thấy mặc dù các khách hàng có hoạt động trong vòng 12 nhưng tỉ lệ từ bỏ của họ lại rất cao. Hầu hết những khách hàng của thuộc tính này là phụ nữ và sỡ hữu thẻ xanh (blue card). * Với các biểu đồ bên phải, tỉ lệ từ bỏ là tương đối thấp (hầu như giống với các thuộc tính trước đó). Ngoài ra, tệp khách hàng này sỡ hữu nhiều thẻ vàng (gold card) hơn, nhiều khách hàng trình độ Graduate hơn so với dữ liệu về khách hàng có 0 tháng không hoạt động. | |
| Contacts\_Count\_12\_mon |  | |
| **Nhận xét**:   * Đồ thị phân bố có thiên hướng lệch trái. * Những giá trị 0, 5, 6 được đánh giá là những giá trị ngoại lệ. Tuy nhiên, những giá trị này sẽ không được coi là bất thường vì những khách hàng có thể 0 liên lạc, hoặc liên lạc 5, 6 lần với ngân hàng là điều bình thường. | |
| **Phân tích ngoại lệ:** | |
| **Nhận xét**:   * Hầu hết tất cả các khách hàng có 0 lần liên lạc với ngân hàng thì đều vẫn còn là khách hàng của ngân hàng, và nữ chiếm đa số. * Những khách hàng có 5, 6 lần liên lạc với khách hàng thì có tỉ lệ khách hàng rời bỏ ngân hàng cao hơn so với những khách hàng có 0 lần liên lạc, và hầu hết là những khách hàng nam. * Điều này có thể chỉ ra cho chúng ta thấy ràng có sự liên hệ giữa thuộc tính này với tỉ lệ rời bỏ của khách hàng. | |
| Credit\_Limit |  | |
| **Nhận xét**:   * Đồ thị lệch về phía bên trái. * Thuộc tính này có những ngoại lệ mang giá trị cao ở cuối (higher end outliner). Tức là có một tệp khách hàng sỡ hữu mức tín dụng cao và được xem là những ngoại lệ. | |
| **Phân tích ngoại lệ:** | |
| **Nhận xét**:   * Ta thấy được ràng, những tệp khách hàng của nhóm ngoại lệ này hầu hết là những khách hàng có mức thu nhập cao. * Điều đặc biệt là, thuộc tính này có tệp khách hàng ở trong mọi phân loại thẻ (Card Category) và có khoảng 90% khách hàng sỡ hữu thẻ xanh và thẻ bạc, hơn 80% khách hàng là những khách hàng chưa từ bỏ ngân hàng. * Điều này khá mới mẻ, vì đây là lần đầu mà tất cả các loại thẻ xuất hiện với một giá trị đáng xem xét. | |
| Total\_Revolving\_Bal | **Nhận xét**:   * Hầu hết khách hàng có total revolving balance là 0. Điều này có thể là do họ không sử dụng thẻ tín dụng. * Không có giá trị ngoại lệ. | |
| Avg\_Open\_To\_Buy |  | |
| **Nhận xét**:   * Đồ thị phân bố bị lệch trái như credit limit * Thuộc tính này cũng có những ngoại lệ mang giá trị cao ở cuối như credit limit. Điều này có thể chỉ ra ràng một vài khách hàng sử dụng rất ít lượng hạn mức tín dụng của họ | |
| **Phân tích ngoại lê:** | |
| **Nhận xét**:   * Ta nhận thấy ràng hành vi của những giá trị ngoại lệ này rất giống với những giá trị ngoại lệ của credit limit. Điều này có thể chỉ ra ràng có mối tương quan tương đối mạnh với 2 thuộc tính credit limit và average open to buy. * Những ngoại lệ này cũng có những khách hàng có nguồn thu nhập cao và hầu hết cũng là nam giới. * Cũng có tất cả các thuộc tính thẻ, với hơn 90% là thẻ xanh và thẻ bạc. * Hầu hết các khách hàng chưa từ bỏ ngân hàng. | |
| Total\_Amt\_Chng\_Q4\_Q1 |  | |
| **Nhận xét**:   * Biểu đồ phân phối cũng cho thấy hình dạng của phân phối chuẩn, nhưng tương tự như thuộc tính Months\_on\_book, ta cũng sẽ kiểm định cho giả thuyết này. * Thuộc tính có các ngoại lệ kết thúc giá trị cao và giá trị thấp (higher and lower end outliners). | |
| **Phân tích ngoại lệ:** | |
| **Nhận xét**:   * Hầu hết những khách hàng có mức thay đổi nhỏ trong giao dịch sẽ có số lượng người rời bỏ cao hơn. * Hầu hết các khách hàng có mức thay đổi nhỏ là khách hàng nữ. * Hầu hết các khách hàng có mức thay đổi cao là những khách hàng vẫn chưa từ bỏ ngân hàng. | |
| Total\_Trans\_Amt |  | |
| **Nhận xét**:   * Ta nhận thấy ràng có sự phân phối đa điểm phức tạp trong phân bố của dữ liệu. Điều này cho thấy, dữ liệu trên không phải là một phân phối đơn giản, cụ thể là nó có thể chứa các nhóm dữ liệu riêng biệt hoặc các cụm dữ liệu riêng biệt. Để hiểu rõ hơn, ta sẽ dùng phân tích cụm giữa thuộc tính này với các thuộc tính định lượng khác ở phần sau. * Đồng thời, dữ liệu cũng có những giá trị ngoại lệ cao. | |
| **Phân tích ngoại lệ:**    **Nhận xét**:   * Những khách hàng có tổng số lượng giao dịch cao có lượng khách hàng ở lại cao. * Mặc dù có tổng lượng giao dịch cao hơn, tuy nhiên, hầu hết khách hàng vẫn có mức thu nhập dưới $40k | |
| Total\_Trans\_Ct |  | |
| **Nhận xét**:   * Thuộc tính nay cũng cho thấy sự đa điểm trong phân bố dữ liệu như total transaction amount. * Có ít ngoại lệ cao. | |
| Total\_Ct\_Chng\_Q4\_Q1 |  | |
| **Nhận xét**: Có những ngoại lệ thấp và cao | |
| Avg\_Utilization\_Ratio |  | |
| **Nhận xét**: Hầu hết các khách hàng không sử dụng thẻ tín dụng | |

## **2.3 DATASET *“Telecom Company Churn Rate, Call Center Data”***

### **2.3.1 Giới thiệu về dataset**

Bộ dữ liệu *“Telecom Company Churn Rate, Call Center Data”* cung cấp thông tin về việc liệu khách hàng có từ bỏ dịch vụ viễn thông hay không trong , đồng thời chứa các thuộc tính có thể ảnh hưởng đến quyết định này.

Bộ dữ liệu được thu thập trong giai đoạn tháng 1 năm 2021- tháng 10 năm 2022 tại Switzerland gồm 2 tập dữ liệu: Telecom Churn Rate Dataset(tập dữ liệu chính) và Telecom Company Call-Center-Dataset. Telecom Churn Rate Dataset là tập dữ liệu cung cấp thông tin về việc liệu khách hàng có từ bỏ dịch vụ viễn thông hay không, Telecom Company Call-Center-Dataset cung cấp thông tin về các cuộc gọi ở trung tâm cuộc gọi các công ty viễn thông. Chúng ta sẽ tập trung vào tập dữ liệu Telecom Churn Rate Dataset.

Các thuộc tính trong tập dữ liệu Telecom Churn Rate Dataset bao gồm :

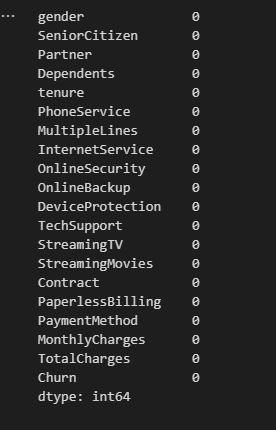
* customerID: ID khách hàng
* gender: giới tính (nữ, nam)
* SeniorCitizen: Khách hàng có phải là người cao tuổi hay không (1, 0)
* Partner: khách hàng có đối tác hay không (Có, Không)
* Dependents: Khách hàng có người phụ thuộc hay không (Có, Không)
* tenure: Số tháng khách hàng đã gắn bó với công ty
* PhoneService: Khách hàng có dịch vụ điện thoại hay không (Có, Không)
* MultipleLines: Khách hàng có dịch vụ đa dòng hay không (Có, Không, Không có dịch vụ điện thoại)
* InternetService: Nhà cung cấp dịch vụ internet của khách hàng (DSL, Fiber Optic, No)
* OnlineSecurity: Khách hàng có bảo mật trực tuyến hay không (Có, Không, Không có dịch vụ internet)
* OnlineBackup: Khách hàng có sao lưu trực tuyến hay không (Có, Không, Không có dịch vụ internet)
* DeviceProtection: Khách hàng có bảo vệ thiết bị hay không (Có, Không, Không có dịch vụ internet)
* TechSupport: Khách hàng có được hỗ trợ kỹ thuật hay không (Có, Không, Không có dịch vụ internet)
* StreamingTV: Khách hàng có truyền hình trực tuyến hay không (Có, Không, Không có dịch vụ internet)
* StreamingMovies: Khách hàng có streaming phim hay không (Có, Không, Không có dịch vụ internet)
* Contract: Thời hạn hợp đồng của khách hàng (Theo tháng, Một năm, Hai năm)
* PaperlessBilling: Khách hàng có thanh toán không cần giấy tờ hay không (Có, Không)
* PaymentMethod: Phương thức thanh toán của khách hàng (

Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))

* MonthlyCharges: Số tiền phải trả mà khách hàng hàng tháng
* TotalCharges: Tổng số tiền phải trả mà khách hàng
* Churn: Khách hàng có rời bỏ hay không (Có hoặc Không)

### **2.3.2 Phân bố giá trị trong các thuộc tính**

Các thuộc tính trong Telecom Churn Rate Dataset không có giá trị null.



|  |  |
| --- | --- |
| **Thuộc tính** | **Trực quan phân bố giá trị** |
| Gender |  |
| **Nhận xét:** Khách hàng nam (3555 người) lớn hơn khách hàng nữ (3488 người). |
| SeniorCitizen |  |
| **Nhận xét:** Khách hàng là người cao tuổi(1142 người) chiếm phần nhỏ hơn so khách hàng không phải là người cao tuổi (5901 người). |
| Partner |  |
| **Nhận xét:** Khách hàng có người sống cùng (3402 người) nhỏ hơn khách hàng một mình (3641 người). |
| Dependents |  |
|  | **Nhận xét:** Khách hàng sống phụ thuộc (2110 người) nhỏ hơn nhiều khách hàng tự lập (4933 người). |
| tenure | **A graph with numbers and lines  Description automatically generated** |
| **Nhận xét:** Thời gian khách hàng đã sử dụng dịch vụ của công ty, trong đó sử dụng dịch vụ từ hơn 7 tháng trở xuống có 1612 người trong khi có 444 người sử dụng từ 36 đến hơn 42 tháng. |
| PhoneService |  |
| **Nhận xét:** Khách hàng dùng dịch vụ(6361 người) điện thoại cao hơn khách hàng không dùng(682 người). |
| MultipleLines |  |
| **Nhận xét:** Từ dữ liệu trong thuộc tính PhoneService ,ta có 2971 khách hàng dùng dịch vụ đa dòng và 3390 khách hàng không dùng dịch vụ này. |
| InternetService |  |
| **Nhận xét:** Có 2421 khách hàng dùng dịch vụ DSL, 3096 khách hàng dùng dịch vụ Fiber optic và 1526 không dùng dịch vụ Internet. |
| OnlineSecurity |  |
| **Nhận xét:** Từ dữ liệu trong thuộc tính InternetService,ta có 2019 khách hàng dùng dịch vụ bảo vệ trực tuyến và 3498 khách hàng không dùng dịch vụ này. |
| OnlineBackup |  |
| **Nhận xét:** Từ dữ liệu trong thuộc tính InternetService,ta có 2429 khách hàng dùng dịch vụ sao lưu trực tuyến và 3088 khách hàng không dùng dịch vụ này. |
| DeviceProtection |  |
| **Nhận xét:** Từ dữ liệu trong thuộc tính InternetService,ta có 2422 khách hàng dùng dịch vụ bảo vệ thiết bị và 3095 khách hàng không dùng dịch vụ này. |
| TechSupport |  |
| **Nhận xét:** Từ dữ liệu trong thuộc tính InternetService,ta có 2044 khách hàng dùng dịch vụ bảo vệ thiết bị và 3473 khách hàng không dùng dịch vụ này. |
| StreamingTV |  |
| **Nhận xét:** Từ dữ liệu trong thuộc tính InternetService,ta có 2707 khách hàng dùng dịch vụ truyền hình trực tuyến và 2810 khách hàng không dùng dịch vụ này |
| StreamingMovies |  |
| **Nhận xét:** Từ dữ liệu trong thuộc tính InternetService,ta có 2732 khách hàng dùng dịch vụ xem phim trực tuyến và 2785 khách hàng không dùng dịch vụ này |
| Contract |  |
| **Nhận xét:** Có 3875 khách hàng có hợp đồng theo một tháng, 1473 khách hàng có hợp đồng theo một năm và 1695 khách hàng có hợp đồng theo hai năm. |
| PaperlessBilling |  |
| **Nhận xét:** Khách hàng thanh toán không cần giấy tờ (4171 người) nhiều hơn khách hàng thanh toán cần giấy tờ(2872 người) |
| PaymentMethod |  |
| **Nhận xét:**  - Số lượng khách hàng thanh toán bằng Séc điện tử là nhiều nhất(2365 người)  - 1612 người thanh toán bằng Séc được gửi qua thư.  - 1544 người thanh toán bằng chuyển khoản ngân hàng (tự động), 1522 người thanh toán bằng thẻ tín dụng (tự động) |
| MonthlyCharges | **A graph of a graph with numbers and a bar chart  Description automatically generated with medium confidence** |
| **Nhận xét:** Có lượng lớn khách hàng(1606 người) phải trả từ 18.25 đến 28.25 $.  Một ít khách hàng(191 người) trả từ 28.25 đến 38.25 $. |
| TotalCharges | **A graph with numbers and a bar  Description automatically generated** |
| **Nhận xét:** Có đến 1788 khách hàng phải trả tổng số tiền cho dịch vụ là dưới 410$, trong khi chỉ 3 khách hàng trả tổng tiền trong khoảng từ 8610 đến 9020$ |
| Churn |  |
| **Nhận xét:** Số lượng khách hàng từ bỏ dịch vụ(1869 người) ít hơn khách hàng còn sử dụng dịch vụ(5174 người) |

# **CHƯƠNG 3: PHÂN TÍCH DATASETS**

## **3.1 DATASET *“Internet Service Provider Customer Churn”***

### **3.1.1 Phân tích thuộc tính “churn”**

A blue and orange pie chart

Description automatically generated- Churn là thuộc tính mục đích quan trọng trong dataset này nên việc xem xét và đánh giá dữ liệu trên thuộc tính này cũng quan trọng có thể quyết định có thể quyết định khả năng dự đoán của dữ liệu tương lai.

- Từ dữ liệu trực quan phân bố giá trị của “churn”, ta có thể tính số lượng khách hàng có khả năng churn trong dataset này như sau:

+ churn = 72274 \* 55.41% = 40050 người

+ not\_churn = 72274 \* 44.59% = 32224 người

- Tỷ lệ churn cho ta thấy số lượng khách hàng có khả năng từ bỏ nhà cung cấp dịch vụ Internet chiếm 55.41% tổng số. Tuy nhiên, sự khác biệt giữa tỷ lệ churn và không churn không đáng kể (chênh lệch khoảng 10.82%). Điều này cho thấy rằng bộ dữ liệu không bị thiên vị và phân bố dữ liệu trong thuộc tính mục tiêu này khá cân bằng.

### **3.1.2 Phân tích yếu tố ảnh hưởng “churn”**

|  |  |
| --- | --- |
| **Thuộc tính** | **Trực quan** |
| is\_tv\_subscriber |  |
| **Nhận xét:**   * Đối với khách hàng đăng ký TV, số lượng not\_churn cao hơn số churn.Điều này có thể cho thấy rằng dịch vụ TV có thể đóng vai trò quan trọng trong việc giữ chân khách hàng. * Đối với khách hàng không đăng ký TV, số lượng churn lại nhiều hơn so với not\_churn.Điều này có thể chỉ ra rằng việc không đăng ký TV có thể là một trong những yếu tố góp phần vào việc khách hàng chọn rời bỏ dịch vụ |
| is\_movie\_package\_subscriber |  |
| **Nhận xét:**   * Đối với khách hàng đăng ký movie, số khách hàng not\_churn cao hơn so với churn. Điều này có thể cho thấy rằng việc đăng ký dịch vụ movie có thể giúp giữ chân khách hàng. * Đối với khách hàng không đăng ký movie, số khách hàng not\_churn cao hơn so với churn. Điều này cho thấy rằng việc không đăng ký dịch vụ movie không phải lúc nào cũng là nguyên nhân chính gây ra churn. |
| **Nhận xét chung**: Có sự biến động đáng kể về số lượng churn và not\_churn giữa các nhóm khách hàng có đăng ký và không đăng ký TV (đối với is\_tv\_subscriber) hay nhóm khách hàng có đăng ký và không đăng ký movie (đối với is\_movie\_package\_subscribe). Điều này cho thấy rằng 2 thuộc tính trên có thể có ảnh hưởng đáng kể đến việc xác định khả năng churn của khách hàng. | |
| subscription\_age |  |
| **Nhận xét:**  Chúng ta có thể thấy rằng hầu hết các khách hàng không có khả năng từ bỏ dịch vụ đã có thời gian đăng ký từ 1 đến 4 năm, và có một số trường hợp có thời gian đăng ký cao mà vẫn chưa từ bỏ dịch vụ. Tuy nhiên, đối với những khách hàng đã từ bỏ, thời gian đăng ký thường nằm trong khoảng từ 1.5 đến 3 năm, và có những trường hợp ngoại lệ với khách hàng có thời gian đăng ký cao. Theo như biểu đồ thì hầu hết các khách hàng có xu hướng muốn rời khỏi dịch vụ trước khi đạt đến 3 năm sử dụng. |
|  |
| **Nhận xét:**  Tuổi đăng ký thường nằm trong khoảng từ 0 đến 5 năm thì sự phân bố của giá trị churn cao hơn not\_churn nhưng not\_churn lại cao hơn churn từ 5 năm trở về sau, và sự phân bố dường như bằng nhau và tiến sát về 0 kể từ 10 năm trở về sau. Ta thấy rằng sự phân bố các giá trị churn thay đổi qua dần qua số tuổi đăng ký nên thuộc tính subscription\_age thể hệt sự tác động rõ rệt lên churn. |
| bill\_avg |  |
| **Nhận xét:**  Trung bình hóa đơn chủ yếu nằm trong khoảng từ 10 đến 40 đô la cho cả khách hàng từ bỏ và không từ bỏ dịch vụ. Tuy nhiên, có những trường hợp ngoại lệ với các khách hàng có trung bình hóa đơn cao là những người không từ bỏ và cũng có các khách hàng có trung bình hóa đơn cao là những người từ bỏ dịch vụ. |
| reamining\_contract |  |
| **Nhận xét:**  Ta nhận thấy rằng khách hàng không có hợp đồng thường hay đã hết thời hạn hợp đồng (giá trị bằng 0) là những người có khả năng từ bỏ dịch vụ, bởi vì nếu họ hủy trong khi còn trong hợp đồng, họ phải trả phí phạt. Tuy nhiên, cũng có những trường hợp đặc biệt với khách hàng đã ký hợp đồng nhưng vẫn từ bỏ dịch vụ. Do đó, hợp đồng còn là một yếu tố quan trọng ảnh hưởng đến khả năng từ bỏ dịch vụ (churn). |
| service\_failure\_count |  |
| **Nhận xét:**  Những khách hàng có số cuộc gọi về lỗi dịch vụ cao thường từ bỏ dịch vụ, nhưng ta cũng có thể thấy rằng những khách hàng có số lượng cuộc gọi về lỗi dịch vụ cao mà không từ bỏ dịch vụ. Mặc dù service\_failure\_count có vẻ không tương quan với các giá trị churn nhưng trong thực tế thì thuộc tính này có ảnh hưởng nhất định đến quyết định của khách hàng. |
| download\_avg |  | |
| upload\_avg |  | |
| download\_over\_limit |  | |
| **Nhận xét:**   * Ta thấy rằng sự giao động về phân bố không nhiều giữa churn và not\_churn ở các thuộc tính churn, đồng nghĩa với việc khả năng từ bỏ dịch vụ của khách hàng không quá bị ảnh từ các thuộc tính trên. * download, upload tùy vào nhu cầu sử dụng của khách hàng (ví dụ streaming video, trò chơi trực tuyến hoặc tải xuống tập tin lớn thì download\_avg sẽ cao) chứ không ảnh hưởng đến việc từ bỏ dịch vụ của khách hàng. * Về download\_over\_limit, trong thực tế, khách hàng thường rất cẩn thận về việc vượt quá giới hạn tải xuống vì họ thường nhận được cảnh báo khi họ đạt đến giới hạn đó. Ngoài ra, từ biểu đồ, chúng ta không thể thấy một mối quan hệ giữa việc vượt quá giới hạn tải xuống và sự churn. | | |

A graph with blue squares

Description automatically generated

Theo sự phân tích về sự ảnh hưởng của các thuộc tính đối với churn cũng như hình ảnh thể hiện sự tương quan của các thuộc tính đối với churn, chúng ta có thể thấy rằng:

* reamining\_contract tác động đến khả năng từ bỏ dịch vụ của khách hàng nhiều nhất.
* bill\_avg, service\_failure\_count ít tác động đến khả năng từ bỏ dịch vụ của khách hàng nhất. Tuy nhiên, service\_failure\_count thực tế vẫn có tác động nhất định đến khả năng churn, chỉ là trong trường hợp dataset này thì thuộc tính này ít tác động.
* Các thuộc tính còn lại có sự tác động nhất định đến churn, tuy nhiên các thuộc tính không ảnh hưởng phân tích ở trên có thể tác động phần nào ở bộ dữ liệu này, nhưng khi chuyển qua bộ dữ liệu khác thì không chắc sẽ đúng.

A screenshot of a computer screen

Description automatically generated

Biểu đồ nhiệt thể hiện sự tương quan giữa các cặp thuộc tính

Dựa trên biểu đồ nhiệt, mối tương quan giữa các yếu tố liên quan đến khách hàng dịch vụ Internet trong bộ dữ liệu này được thể hiện như sau:

* **Mối Tương Quan Mạnh**: Có vẻ như có một mối tương quan mạnh mẽ giữa subscription\_age và id, với giá trị là 0.6, cho thấy khách hàng lâu năm có xu hướng có ID thấp hơn (có thể là khách hàng đầu tiên).
* **Tương Quan Tiêu Cực**: churn có mối tương quan tiêu cực với remaining\_contract (-0.68), có nghĩa là khách hàng càng ít thời gian hợp đồng còn lại thì càng có khả năng chấm dứt dịch vụ.
* **Không Tương Quan**: Các yếu tố như bill\_avg, service\_failure\_count, download\_avg, và upload\_avg có mối tương quan gần như không đáng kể với churn, cho thấy những yếu tố này không ảnh hưởng nhiều đến quyết định chấm dứt dịch vụ của khách hàng.
* **Tương Quan Dương**: is\_tv\_subscriber và movie\_package\_subscriber có mối tương quan dương với nhau (0.34), có thể hiểu rằng khách hàng đăng ký dịch vụ truyền hình có xu hướng đăng ký gói dịch vụ phim.

### **3.1.3 Kết luận thu được sau khi thực hiện việc phân tích dữ liệu**

* Các thuộc tính không có mối quan hệ lớn với tỉ lệ churn bao gồm:
  + bill\_avg
  + download\_avg
  + upload\_avg
* service\_failure\_count cũng có thể không có quan hệ với tỉ lệ churn trong bộ dữ liệu này nhưng xét thực tế thì thuộc tính này có khả năng gây ảnh đáng kể đến khả năng từ bỏ dịch vụ của khách hàng.
* Thuộc tính reamining\_contract ảnh hưởng đến khả năng từ bỏ dịch vụ của khách hàng nhiều nhất. Với khách hàng có ít thời gian hợp đồng thì có khả năng chấm dứt hợp đồng cao.
* Khách hàng đăng ký truyền hình thường có xu thế sẽ đăng ký luôn dịch vụ phim và những khách hàng có tuổi đăng ký lâu thường là những khách hàng đầu tiên.

### **3.1.4 Ưu điểm và nhược điểm của dataset**

* + - 1. **Ưu điểm**
* *Dữ liệu đa dạng:* Dataset chứa thông tin đa dạng về các khách hàng của nhà cung cấp dịch vụ Internet, bao gồm các thuộc tính như hợp đồng, dịch vụ đăng ký, tuổi đăng ký, hóa đơn trung bình, ... . Điều này cung cấp cơ sở dữ liệu phong phú cho việc phân tích và dự đoán sự churn của khách hàng.
* *Thông tin chi tiết:* Dữ liệu cung cấp thông tin chi tiết về sự từ bỏ dịch vụ của khách hàng (churn), bao gồm cả số lượng khách hàng từ bỏ và không từ bỏ, giúp hiểu rõ hơn về tình trạng churn của công ty.
* *Dữ liệu cân bằng:* Tỷ lệ churn và không churn không quá chênh lệch, chỉ khác biệt khoảng 10.82%, điều này cho thấy dữ liệu không bị thiên vị và phân bố khá cân bằng.
* *Có sự kết hợp giữa các loại dịch vụ:* Dataset bao gồm thông tin về việc đăng ký các loại dịch vụ như TV và gói phim, cho phép phân tích mối quan hệ giữa các loại dịch vụ và sự churn của khách hàng.
  + - 1. **Nhược điểm**
* *Thiếu một số thông tin quan trọng:* Có thể thiếu một số thông tin quan trọng như đánh giá từ khách hàng, dịch vụ cụ thể mà họ sử dụng, hoặc thông tin về hỗ trợ khách hàng.
* *Có thể thiếu thông tin về nguyên nhân churn:* Dataset không cung cấp thông tin chi tiết về nguyên nhân gây ra sự churn của khách hàng, điều này có thể làm hạn chế khả năng phân tích và dự đoán.
* *Dữ liệu có thể không phản ánh đầy đủ thực tế:* Dữ liệu có thể bị hạn chế về một số khía cạnh của sự churn hoặc không thể phản ánh đầy đủ thực tế của tình hình sự churn trong môi trường thực tế.
* *Không có thông tin về phân phối địa lý:* Dataset không cung cấp thông tin về phân phối địa lý của khách hàng, điều này có thể làm giảm khả năng phân tích tác động của vị trí địa lý lên sự churn.

## **3.2 DATASET *“Credit Card Churn Prediction”***

### **3.2.1. Tỉ lệ churn**

Tỉ lệ Churn (từ bỏ dịch vụ): ở dataset “*Credit Card Churn Prediction*” tỉ lệ churn tương ứng với tỉ lệ Atrrited Customer trong thuộc tính Attrition\_Flag. Tỉ lệ này trong bộ dữ liệu trên có giá trị là 16%, tức là có 16% khách hàng của dịch vụ thẻ tín dụng này đã đóng tài khoản, từ bỏ sử dụng dịch vụ.

A blue and orange pie chart

Description automatically generated

### **3.2.2. Mối quan hệ giữa tỉ lệ churn với các thuộc tính phân loại (category column)**

|  |  |
| --- | --- |
| **Thuộc tính được xét** | **Biểu diễn phụ thuộc** |
| Gender |  |
| **Nhận xét:** Tập dữ liệu có tỉ lệ Nữ nhiều hơn Nam, tuy nhiên không có sự khách biệt quá lớn giữa hai giới trong tỉ lệ churn. |
| Education\_Level |  |
| **Nhận xét**:   * Có 15% giá trị unknown biểu thị cho việc 15% khách hàng không cung cấp thông tin về trình độ học vấn. * Trình độ học vấn phổ thông chiếm khoảng 50%, trong khi các trình độ cao hơn chiếm khoảng 20%. * Nhìn chung, không có sự khách biệt nhau rõ ràng giữa các giá trị đối với tỉ lệ churn, từ đó có thể nhận định thuộc tính này không có sự ảnh hưởng quá lớn đối với tỉ lệ rời bỏ của khách hàng. |
| Marital\_Status |  |
| **Nhận xét**:   * Hầu hết tình trạng hôn nhân của khách hàng là Married và Single. * Sự khác nhau giữa Churn và không Churn giữa các thuộc tính tình trạng hôn nhân là không rõ ràng. |
| Income\_Category |  |
| **Nhận xét**:   * Phần lớn khách hàng có thu nhập bình quân hàng năm nằm ở mức trung bình. * Có gần 11% giá trị Unknown. Có thể là do khách hàng không cung cấp thông tin về thu nhập cho phía ngân hàng. * Chỉ có một số ít khách hàng có thu nhập ở mức cao. * Những khách hàng có mức thu nhập thấp thì chiếm số lượng rời bỏ cao hơn. * Sự tương quan giữa thu nhập và tỉ lệ churn cũng không được rõ ràng như những thuộc tính phân loại trước. |
| Card\_Category |  |
| **Nhận xét**:   * Ta thấy rằng hầu hết khách hàng sử dụng thẻ xanh (khoảng 90%). * Những loại thẻ cao cấp hơn có tỉ lệ rất thấp (tổng dưới 10%). * Tương tự các thuộc tính trước, không có sự rõ ràng về mối quan hệ giữa phân loại thẻ và tỉ lệ churn. |

**Nhận xét chung:**

* Tất cả các đồ thị của mối quan hệ giữa các thuộc tính phân loại (category) và thuộc tính Attrition\_Flag dường như lệch về bên phải. Tức là các phân loại có tỉ lệ cao hơn sẽ có tệp Existing Customer và Attrited Customer cao hơn, và với mỗi phân loại thì tỉ lệ Existing Customer sẽ cao hơn và khoảng chênh lệch giữa hai customer này ở mỗi phân loại của từng thuộc tính là gần như không đổi
* Từ điều đó, ta có thể thấy dường như các thuộc tính phân loại sẽ không có hoặc ít ảnh hưởng hơn tới tỉ lệ từ bỏ dịch vụ

### **3.2.3. Mối quan hệ giữa tỉ lệ churn với các thuộc tính định lượng (numerical column)**

A diagram of a graph

Description automatically generated with medium confidence

**Nhận xét**: Dựa trên đồ thị pairplot ở trên, ta nhận ra một số đặc điểm như sau:

* Có sự tương quan tuyến tính rõ ràng giữa thuộc tính credit\_limit và avg\_open\_to\_buy
* Các giá trị thấp của thuộc tính total\_trans\_amt, total\_amt\_chg\_q4\_q1, total\_trans\_ct có số lượng khách hàng rời bỏ cao hơn nhiều so với những giá trị cao.
* Ta có thể thấy được có một số cụm được hình thành ở các thuộc tính total\_trans\_amt, total\_trans\_ct, total-ct\_chng\_q4\_q1.

### **3.3.4 Bản đồ nhiệt cho sự tương quan giữa các thuộc tính**

A screenshot of a computer screen

Description automatically generated

**Nhận xét**: Quan sát biểu đồ nhiệt trên, ta thấy được một vài đặc điểm như sau:

* Xét về thuộc tính attrtion\_flag, ta thấy ràng ít có mối tương quan mạnh với thuộc tính này. Một vài mối tương quan tiêu biểu có thể kể đến như: với total\_trans\_ct(-0.37), total\_ct\_chng\_q4\_q1(-0.29), months\_inactive\_12\_mon(0.15), contacts\_count\_12\_mon(0.2).
* Đối với mối quan tương quan giữa các thuộc tính khách, ta thấy credit\_limit và avg\_open\_to\_buy có mức tương quan rất cao(=1), months\_on\_book và customer\_age (0.79) và , total\_revolving\_balvà avg\_utilization\_ratio(0.62). Những mối tương quan này tương đối dễ hiểu, khách hàng có hạn mức càng cao thì khả năng sẵn sàng chi trả của họ càng cao, khách hàng có độ tuổi càng cao thì số tháng gắn bó với ngân hàng của họ cũng cao hơn, hay khách hàng số vốn xoay vòng cao thì trung bình họ sử dụng thẻ tín dụng để chi trả cũng cao hơn rất nhiều.

### **3.3.5 Mối tương quan giữa tỉ lệ rời bỏ và mỗi giá trị trong các cột phân loại (category columns)**

Thay vì ở phần trước, ta chi xét mối tương quan giữa attrition\_flag với các thuộc tính đinh lượng, ở phần này ta sẽ xét mối tương quan của thuộc tính với cả thuộc tính định lượng và phân loại để có cái nhìn chắc chắc hơn về mối quan hệ của các thuộc tính. Để làm được điều đó, ta cần tạo ra một dữ liệu dummies, thay thế các giá trị phân loại thành các giá trị định lượng, và sử dụng biểu đồ cột để so sáng mối tương quan.

A screen shot of a computer

Description automatically generated

**Nhận xét**: Dựa vào hình trên, ta có thể đưa ra những nhận định rõ ràng hơn về mối tương quan giữa các thuộc tính với tỉ lệ churn như sau:

* Đối với mối tương quan dương, ta có những thuộc tính có mối tương quan ở mức cao với tỉ lệ churn bao gôm contacts\_ct\_12\_mon và months\_inactive\_12\_mon.
* Đối với mối tương quan âm, ta có những thuộc tính nổi bật như: total\_trans\_ct, total\_ct\_chng\_q4\_q1, và total\_revolving\_bal.
* Những nhận định trên tương đối thực tế, ví dụ như, khách hàng có số lần giao dịch cao thì có ít nguy cơ từ bỏ dịch vụ hơn, hay khách hàng có số tháng không hoạt động cao thì nguy cơ từ bỏ của họ cao hơn mức bình thường.
* Một điểm thú vị là những khách hàng có số lần liên lạc với ngân hàng nhiều có nguy cơ rời bỏ cao nhất so với những thuộc tính khác. Điều này xảy ra có thể là do việc khách hàng liên lạc nhiều lần có thể là do họ phàn nàn về dịch vụ và sau nhiều lần phàn nàn không thành công thì họ hết kiên nhẫn và rời bỏ dịch vụ đó.

### **3.3.6 Kết luận thu được sau khi thực hiện việc phân tích dữ liệu**

* Các thuộc tính có mối quan hệ với tỉ lệ churn bao gồm:
  + contacts\_ct\_12\_mon
  + months\_inactive\_12\_mon
  + total\_trans\_ct.
  + total\_ct\_chng\_q4\_q1.
  + total\_revolving\_bal.
* Thuộc tính credit\_limt có mối tương quan tuyến tính mạnh với avg\_open\_to\_buy.
* Hầu hết khách hàng của ngân hàng là những người có thu nhập thấp, và tệp khách hàng thu nhập thấp cũng có số lượng rời bỏ nhiều hơn

### **3.3.7 Ưu điểm và nhược điểm của dataset**

**3.3.7.1 Ưu điểm**

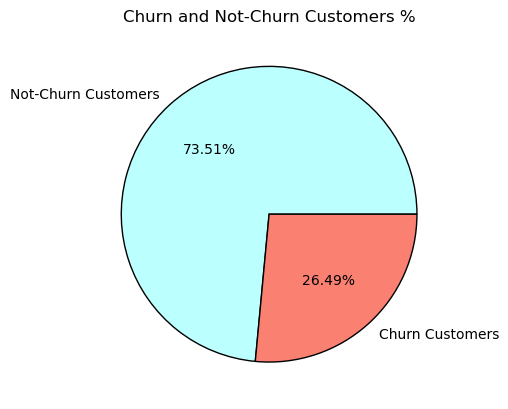
* *Bộ dữ liệu với tỉ lệ sai lệch thấp*: Bộ dữ liệu bao gồm 10127 hàng và 23 cột, với số lượng giá trị không quá ít tuy nhiên bộ dữ liệu lại có thế mạnh với tỉ lệ valid là 100%, không có giá trị null hay giá trị mismatch.
* *Tỉ lệ hầu như đồng đều ở các thuộc tính phân loại*: Tỉ lệ giữa các giá trị của các thuộc tính phân loại (category column) tương đối đều, không có sự chênh lệch quá lớn (trừ một số thuộc tính như loại thẻ (Card\_Category)).
* *Đa dạng các thuộc tính dùng để phân tích*: Với 23 cột thuộc tính giúp cho người phân tích có thể phân tích kỹ hơn ảnh hưởng của các yếu tố đối với tỉ lệ churn từ đó có thể đưa ra được các dự đoán chính xác hơn.

**3.3.7.2 Nhược điểm**

* *Sự không chênh lệch về tỉ lệ churn*: Tỉ lệ giữa khác hàng ở lại và khách hàng từ bỏ tương đôi chênh lệch (83.93% và 16.05%). Điều này có thể ảnh hưởng đến sự tương quan với các thuộc tính phân loại khi vẽ biểu đồ tương quan giữa nhưng thuộc tính phân loại đó với thuộc tính xác định churn
* *Giá trị thuộc tính không rõ ràng*: Vẫn tồn tại những giá trị Unknown trong các thuộc tính như Education\_Level, Marital\_Status, Income\_Category.

## **3.3 DATASET *“Telecom Company Churn Rate, Call Center Data”***

### **3.3.1 Phân tích thuộc tính “churn”**

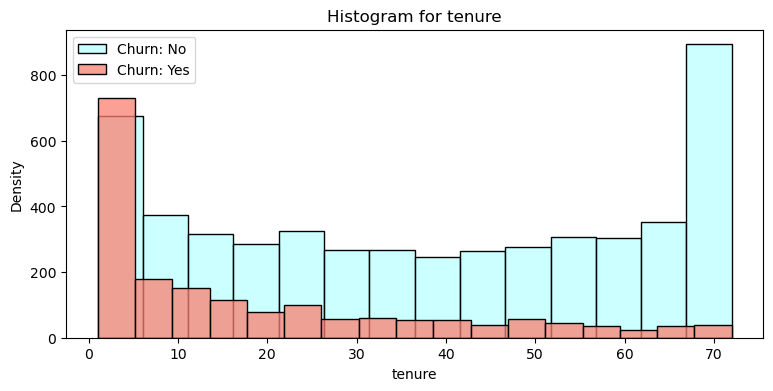


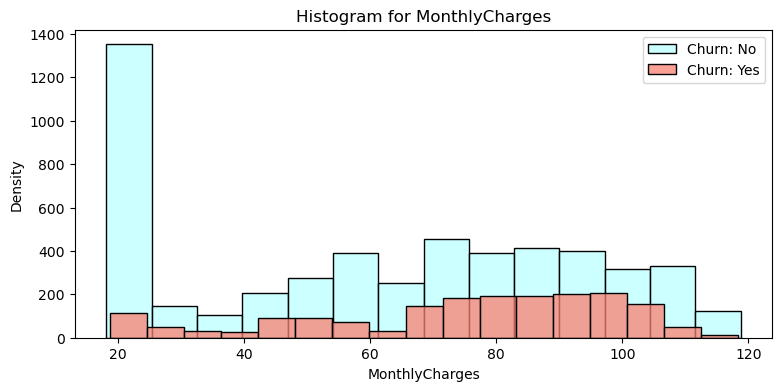
Tỷ lệ khách hàng rời bỏ dịch vụ viễn thông đạt tới 26.49%, trong khi phần lớn tỷ lệ khách hàng vẫn tiếp tục sử dụng dịch vụ vẫn chiếm phần lớn(73.51%)

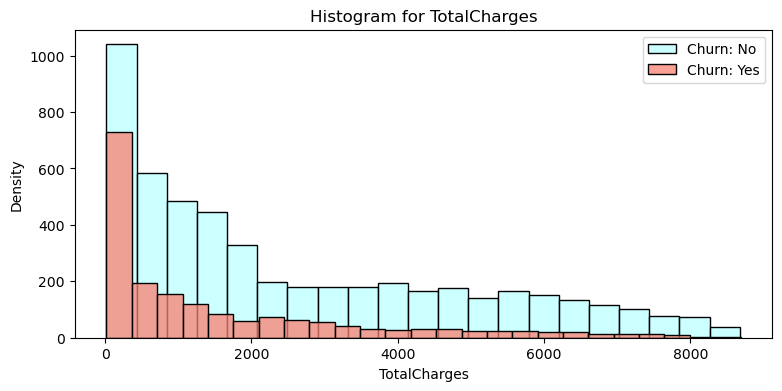
Tỉ lệ churn và không churn lệch với tỷ lệ gần 1:4 cho thấy biểu đồ thiên vị về phía khách hàng không rời bỏ dịch vụ.

### **3.3.2 Phân tích yếu tố ảnh hưởng “churn”**

Tensure, Monthly Charges, Total Changes



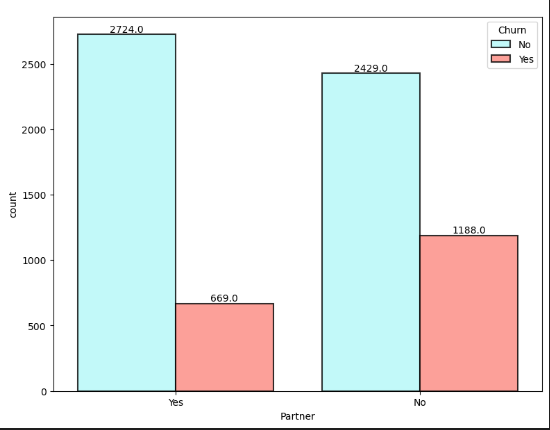
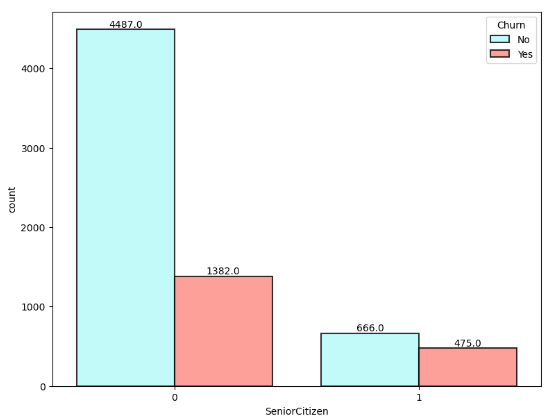
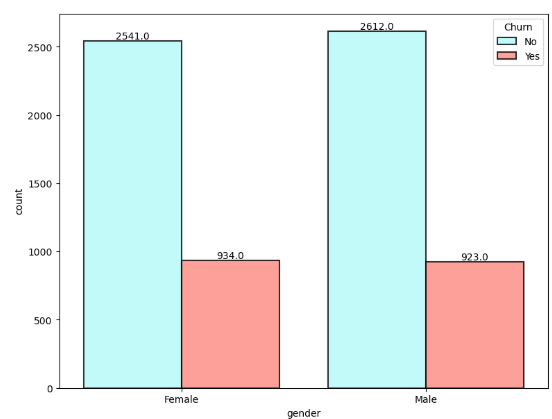




* Khách hàng có Phí hàng tháng cao hơn có nhiều khả năng rời bỏ hơn;
* Rất nhiều khách hàng đã rời bỏ chỉ sau một tháng;
* Hầu hết các hợp đồng hàng tháng kéo dài từ 1-2 tháng, trong khi hợp đồng 2 năm có xu hướng kéo dài khoảng hơn 70 tháng.
* Điều này cho thấy những khách hàng ký hợp đồng dài hạn sẽ trung thành hơn với công ty và có xu hướng gắn bó với công ty lâu hơn.
* Đây cũng là những gì chúng ta đã thấy trong biểu đồ trước về mối tương quan với tỷ lệ churn.

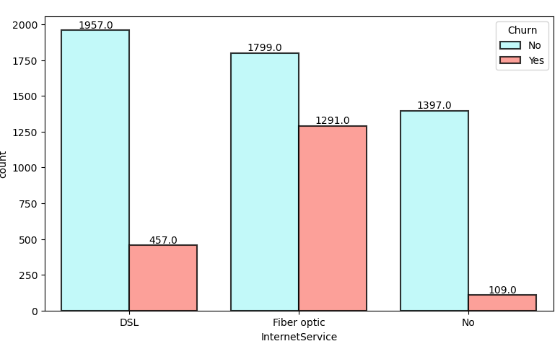
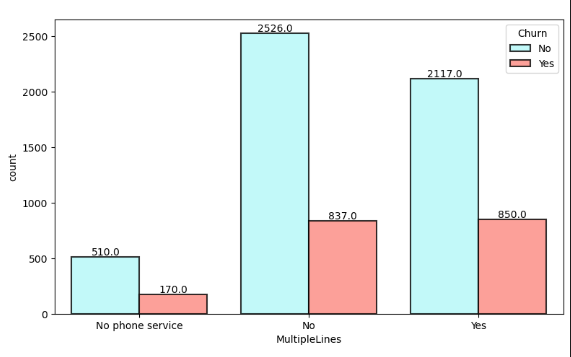
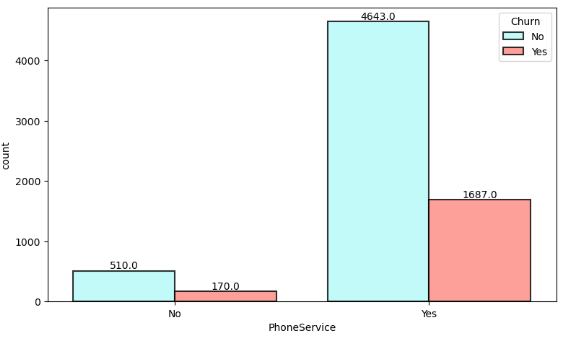
Mặc dù tổng chi phí hàng tháng cao, nhưng nếu dịch vụ hoặc sản phẩm cung cấp giá trị lớn và độc đáo cho khách hàng, họ có thể cảm thấy hài lòng và tin tưởng hơn. Trong trường hợp này, dù tổng phí hàng tháng nhưng tỉ lệ churn có thể thấp vì khách hàng nhận thấy giá trị lớn từ dịch vụ đó.

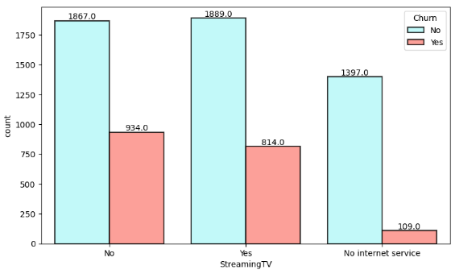
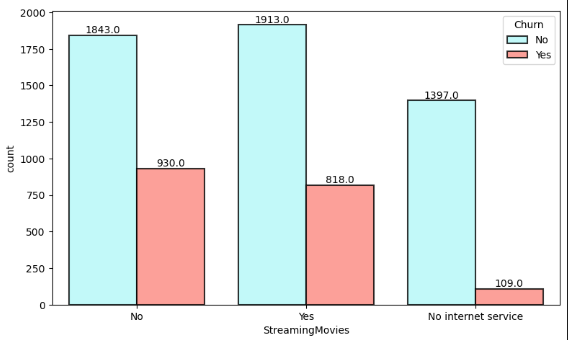
Gender, SeniorCitizen, Partner, Dependents



* Sự phụ thuộc giữa giới tính và khả năng rời bỏ là rất nhỏ.
* Đối với những thuộc tính SeniorCitizen, Partner, Dependents , do số lượng khách hàng churn vẫn nhỏ số khách hàng không no\_chunrn nên các thuộc tính trên không ảnh hưởng đến nguyên nhân gây ra churn

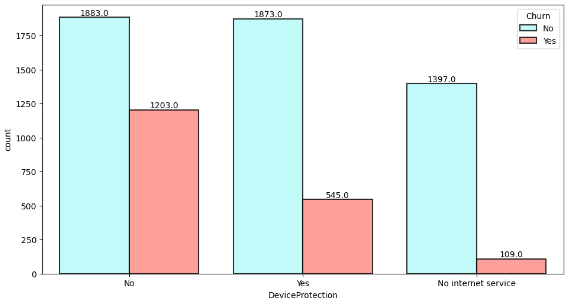
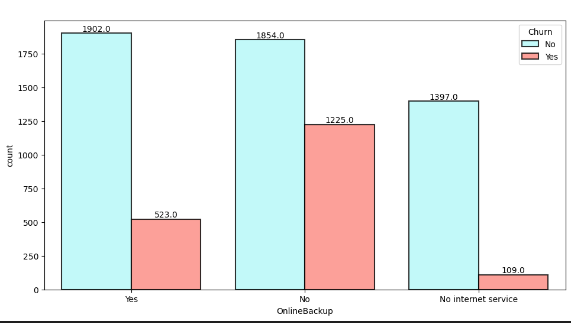
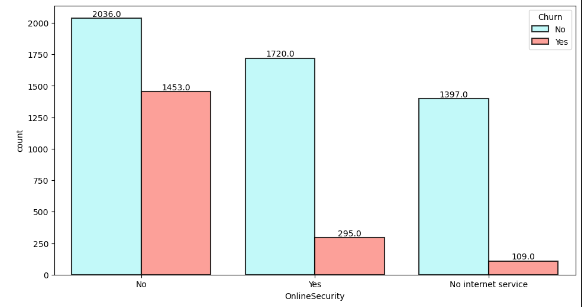
PhoneService, MultipleLines, InternetService, StreamingTV, StreamingMovies

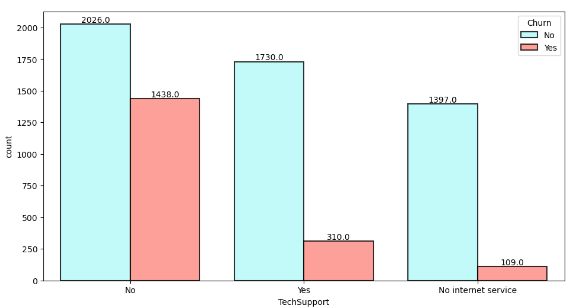




* Một số lượng lớn khách hàng sử dụng Fiber để cung cấp Internet đã từ bỏ dịch vụ. Ngược lại, những khách hàng đang sử dụng DSL không rời bỏ dịch vụ nhiều.
* Khách hàng không có internet có tỷ lệ rời bỏ rất thấp.
* Tổng quan đối với khách hàng có hoặc không đăng ký PhoneService, MultipleLines, InternetService, StreamingTV, StreamingMovies , số lượng khách hàng churn vẫn nhỏ số khách hàng không no\_chunrn nên các thuộc tính trên không ảnh hưởng đến nguyên nhân gây ra khách hàng rời bỏ dịch vụ.

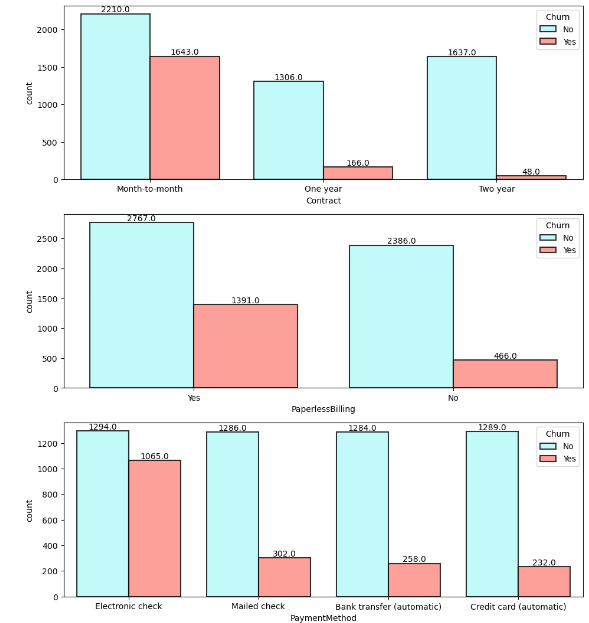
OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport





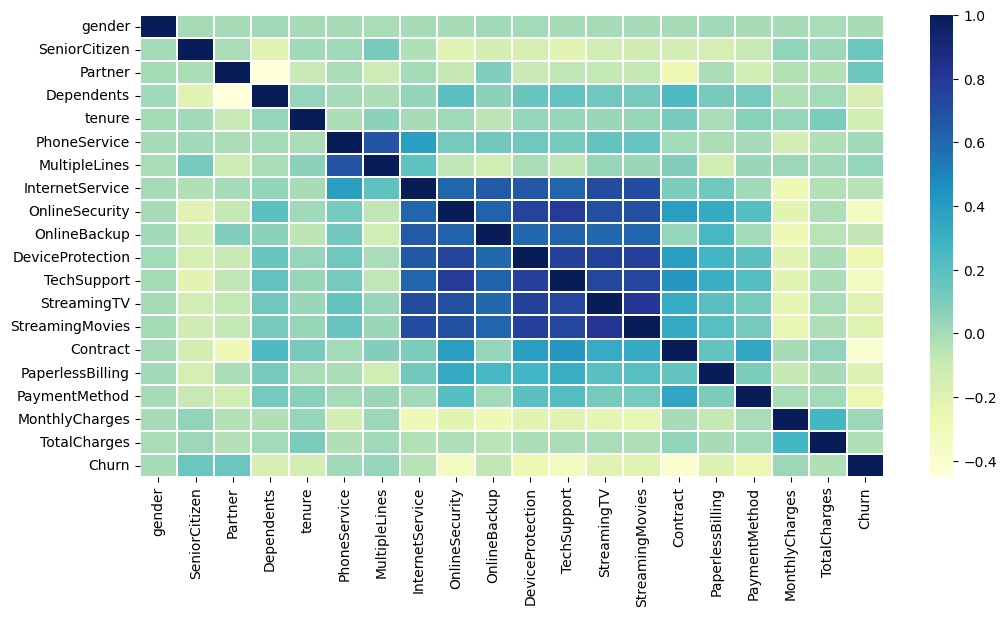
* Những khách hàng không đăng ký dịch vụ hỗ trợ có nhiều khả năng rời bỏ hơn.
* Khách hàng không có đăng ký OnlineSecurity và TechSupport có số lượng lớn nên tuy số người no\_churn vẫn lớn hơn số người churn nhìn vào biểu đồ ta có thể thấy điều này có thể ảnh hưởng đến tỉ lệ churn.
* Tổng quan đối với khách hàng có hoặc không đăng ký OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport , số lượng khách hàng churn vẫn nhỏ số khách hàng không no\_chunrn nên các thuộc tính trên không ảnh hưởng đến nguyên nhân gây ra khách hàng rời bỏ dịch vụ

Contract, PaperlessBilling, PaymentMethod



* Hợp đồng ngắn hạn có tỷ lệ rời bỏ cao hơn nhiều.
* Những khách hàng sử dụng phương thức thanh toán không cần giấy tờ có nhiều khả năng rời bỏ hơn.
* Khách hàng sử dụng séc điện tử làm phương thức thanh toán có nhiều khả năng rời bỏ hơn.

### **3.3.3 Bản đồ nhiệt cho sự tương quan giữa các thuộc tính và mối tương quan giữa tỉ lệ rời bỏ và mỗi giá trị trong các cột**



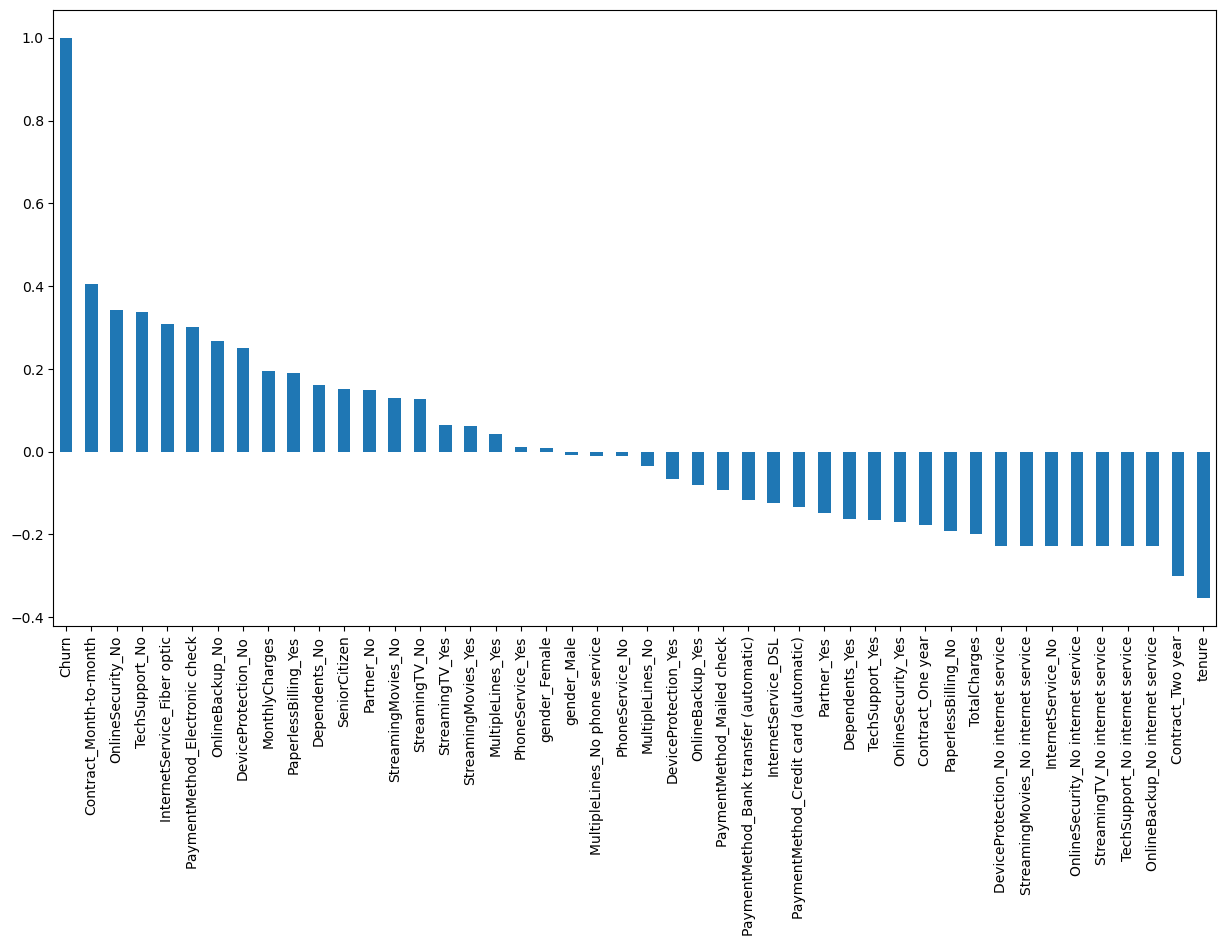
Từ biểu đồ nhiệt trên ta có thể thấy được đặc điểm sau:

Đa phần các thuộc tính có mối tương quan thấp.

Các vấn đề về giới tính, tuổi tác,... dường như không có ảnh hưởng đến quyết định rời bỏ dịch vụ

Các thuộc tính InternetService, StreamingTV, StreamingMovies, , InternetService, StreamingTV, StreamingMovies có mối tương quan cao, điều đó cho thấy ở về dịch vụ về Internet có vấn đề khiến cho tỷ lệ rời bỏ dịch vụ khách hàng tăng.

Để có thể rõ hơn về mối tương quan trên ta sẽ tìm hiểu sâu hơn về mối tương quan giữa tỉ lệ rời bỏ và mỗi giá trị trong các cột ở biểu đồ dưới



Biểu đồ cột thể hiện sự quan trọng của các thuộc tính

Qua biểu đồ trên ta thấy rõ:

Contract\_Month-to-month, OnlineSecurity\_No,TechSupport\_No, InternetService\_ Fiber Obtic, … dường như có mối tương quan tích cực với khả năng rời bỏ.

Tensure, Contract\_Two-year, …dường như có mối tương quan nghịch với khả năng rời bỏ.

Từ đó, ta có thể giả định rằng những người đăng ký dịch vụ Internet từ Fiber Obtic cùng với các dịch vụ đi kèm có khả năng rời bỏ cao hơn.

Điều này khớp với những biểu đã phân tích trên.

### **3.3.4 Ưu điểm và nhược điểm của dataset**

**3.3.4.1 Ưu điểm**

- Dữ liệu đa dạng: Dataset có hơn 10 thuộc tính, đa dạng về thông tin khách hàng của nhà cung cấp dịch vụ cũng như có nhiều dữ liệu hơn cho việc phân tích và dự đoán tỷ lệ churn của khách hàng.

- Không có dữ liệu trống:Dataset không có thuộc tính nào có dữ liệu trống giúp cho việc phân tích và dự đoán tỷ lệ churn của khách hàng gần đúng hơn với thực tế.

- Kết hợp nhiều loại dịch vụ, nhóm khách hàng: Việc tích hợp nhiều loại dịch vụ, nhóm khách hàng giúp cho dữ liệu càng chi tiết hơn.

**3.3.4.2 Nhược điểm**

**-** Không có thông tin về địa lý: Không có thông tin về địa lý của khách hàng làm giảm đi khả năng phân tích tác động của vị trí địa lý lên tỷ lệ churn.

- Dữ liệu phân bố không đều: Tỷ lệ khách hàng không churn quá cao khiến cho việc dự đoán và phân tích tỷ lệ churn của khách hàng bị ảnh hưởng không ít.

## **3.4 NHẬN XÉT CHUNG**

Dựa trên mục tiêu là phân tích khả năng khách hàng có thể từ bỏ dịch vụ, ta có thể đánh giá ba bộ dữ liệu như sau:

|  |  |  |
| --- | --- | --- |
| **Dataset** | **Ưu điểm** | **Nhược điểm** |
| Internet Service Provider Customer Churn | Bộ dữ liệu này cung cấp thông tin đa dạng về các khách hàng Internet, bao gồm thông tin hợp đồng, dịch vụ đăng ký và hóa đơn trung bình. Điều này làm cho nó trở thành một cơ sở dữ liệu phong phú cho phân tích và dự đoán sự churn của khách hàng Internet. | Mặc dù dữ liệu đa dạng, nhưng thiếu thông tin quan trọng như đánh giá từ khách hàng và nguyên nhân gây ra sự churn. Điều này có thể làm hạn chế khả năng phân tích và dự đoán. |
| Credit Card Churn Prediction | Bộ dữ liệu này có tỉ lệ sai lệch thấp và đa dạng các thuộc tính, giúp cho việc phân tích chi tiết hơn về ảnh hưởng của các yếu tố đến tỷ lệ churn. | Tỷ lệ giữa khách hàng ở lại và khách hàng từ bỏ chênh lệch lớn, có thể ảnh hưởng đến dự đoán và phân tích. |
| Telecom Company Churn Rate, Call Center Data | Bộ dữ liệu này cung cấp thông tin đa dạng và không có dữ liệu trống, giúp cho việc phân tích và dự đoán tỷ lệ churn gần đúng hơn với thực tế. | Dữ liệu phân bố không đều, điều này có thể làm giảm độ chính xác của dự đoán. |

Bộ dữ liệu Internet Service Provider Customer Churn có vẻ phù hợp hơn vì đa dạng và chi tiết của thông tin khách hàng Internet, mặc dù cần lưu ý về các thiếu sót như thiếu thông tin về nguyên nhân gây ra sự churn. Tuy nhiên, việc kết hợp các bộ dữ liệu có thể cung cấp góc nhìn toàn diện hơn về vấn đề này.

# **TÀI LIỆU THAM KHẢO**

1. EDWIN GEEVARUGHESE, *“Internet Service churn Analysis”*, link: <https://www.kaggle.com/code/edwingeevarughese/internet-service-churn-analysis>
2. JOYCLYN UJUNWA OGBONNA, *“Credit Card Churn Prediction”*, link: <https://www.kaggle.com/code/ujunwajoyclyn/credit-card-churn>
3. MARCIN RUTECKI, “Telco Churn - EDA & Model (Voting + boosting)”, link: <https://www.kaggle.com/code/marcinrutecki/telco-churn-eda-model-voting-boosting>