**TRƯỜNG ĐẠI HỌC SƯ PHẠM HÀ NỘI**

**Khoa Công Nghệ Thông Tin**

**🙞🕮🙜**



**BÁO CÁO  
THỰC HÀNH DỰ ÁN**

**CHUYÊN ĐỀ TỐT NGHIỆP DỰ ÁN KHOA HỌC** **DỮ LIỆU**

***Đề tài: Phân tích dự đoán sự rời bỏ của khách hàng***

|  |  |  |
| --- | --- | --- |
| **Giảng viên** | **:** | Phạm Thọ Hoàn |
| **Nhóm sinh viên** | **:** | Nguyễn Đức Mậu - 715105146 |
|  |  |  |
|  |  |  |
|  |  |  |

*Hà Nội, 2024*

MỤC LỤC

[Chương 1. Giới thiệu và chuẩn bị dữ liệu 2](#_Toc184928359)

[**1.1 Mục tiêu của dự án** 2](#_Toc184928360)

[**1.2 Phạm vi và giới hạn của dự án** 2](#_Toc184928361)

[***1.2.1. Phạm vi phân tích dữ liệu*** 2](#_Toc184928362)

[***1.2.2*** ***Giới hạn và giả định của dự án*** 2](#_Toc184928363)

[**1.3** **Tổng quan về dữ liệu** 3](#_Toc184928364)

[**1.4. Tiền xử lí dữ liệu** 4](#_Toc184928365)

[Chương 2. Khám phá dữ liệu và lựa chọn mô hình 9](#_Toc184928366)

[**2.1. Trực quan hóa dữ liệu** 9](#_Toc184928367)

[***2.1.1. Báo cáo khám phá dữ liệu từ trực quan hóa*** 9](#_Toc184928368)

[***2.1.2.*** ***Kết quả khám phá từ việc trực quan hóa*** 10](#_Toc184928369)

[**2.2. Lựa chọn mô hình** 14](#_Toc184928370)

[***2.2.1. Trình bày mô hình và thuật toán được chọn để phân tích dữ liệu*** 14](#_Toc184928371)

[***2.2.2. Lý do chọn mô hình*** 14](#_Toc184928372)

[***2.2.3 Quá trình chia dữ liệu thành tập huấn luyện, tập kiểm tra*** 15](#_Toc184928373)

[Chương 3. Huấn luyện, tối ưu hóa và đánh giá mô hình 16](#_Toc184928374)

[**3.1 Mô hình sử dụng: XGBoost Classifier** 16](#_Toc184928375)

[***3.1.1 Huấn luyện mô hình*** 16](#_Toc184928376)

[***3.1.2*** ***Đánh giá hiệu suất mô hình*** 17](#_Toc184928377)

# Chương 1. Giới thiệu và chuẩn bị dữ liệu

## **1.1 Mục tiêu của dự án**

Mục tiêu chính của dự án này là phân tích và dự đoán khả năng rời bỏ dịch vụ (churn) của khách hàng tại công ty viễn thông. Bằng cách hiểu rõ hơn về hành vi và đặc điểm của khách hàng có xu hướng rời bỏ, công ty có thể áp dụng các chiến lược giữ chân khách hàng hiệu quả hơn, giảm thiểu tổn thất doanh thu và chi phí liên quan đến việc giữ chân khách hàng.

Vấn đề rời bỏ dịch vụ là một trong những thách thức lớn trong ngành viễn thông, nơi khách hàng thường có xu hướng thay đổi nhà cung cấp dịch vụ khi không hài lòng. Khi khách hàng rời bỏ, công ty không chỉ mất đi doanh thu từ các hợp đồng dịch vụ mà còn phải đối mặt với chi phí tiếp thị để thu hút khách hàng mới, thường cao hơn rất nhiều so với chi phí giữ chân khách hàng hiện tại.

Dự án tập trung vào việc xây dựng một mô hình dự đoán churn dựa trên dữ liệu khách hàng, từ đó đưa ra cảnh báo sớm cho đội ngũ dịch vụ khách hàng và phòng tiếp thị để có biện pháp can thiệp phù hợp.

## **1.2 Phạm vi và giới hạn của dự án**

### ***1.2.1. Phạm vi phân tích dữ liệu***

Dự án tập trung phân tích dữ liệu về hành vi và các yếu tố liên quan đến dịch vụ của khách hàng trong ngành viễn thông, bao gồm các đặc điểm như:

Thời gian sử dụng dịch vụ, số phút gọi trong ngày, buổi tối và ban đêm.

Dịch vụ đăng kí (e.g, gói cước quốc tế, hộp thư thoại).

Mức độ tương tác với dịch vụ khách hàng và số cuộc gọi dịch vụ.

Phân tích sẽ được thực hiện trên tập dữ liệu của khách hàng nhằm xác định các yếu tố ảnh hưởng đến khả năng rời bỏ. Các yếu tố này sẽ hỗ trợ xây dựng mô hình dự đoán churn để đánh giá khả năng rời bỏ dịch vụ của khách hàng.

* + 1. ***Giới hạn và giả định của dự án***

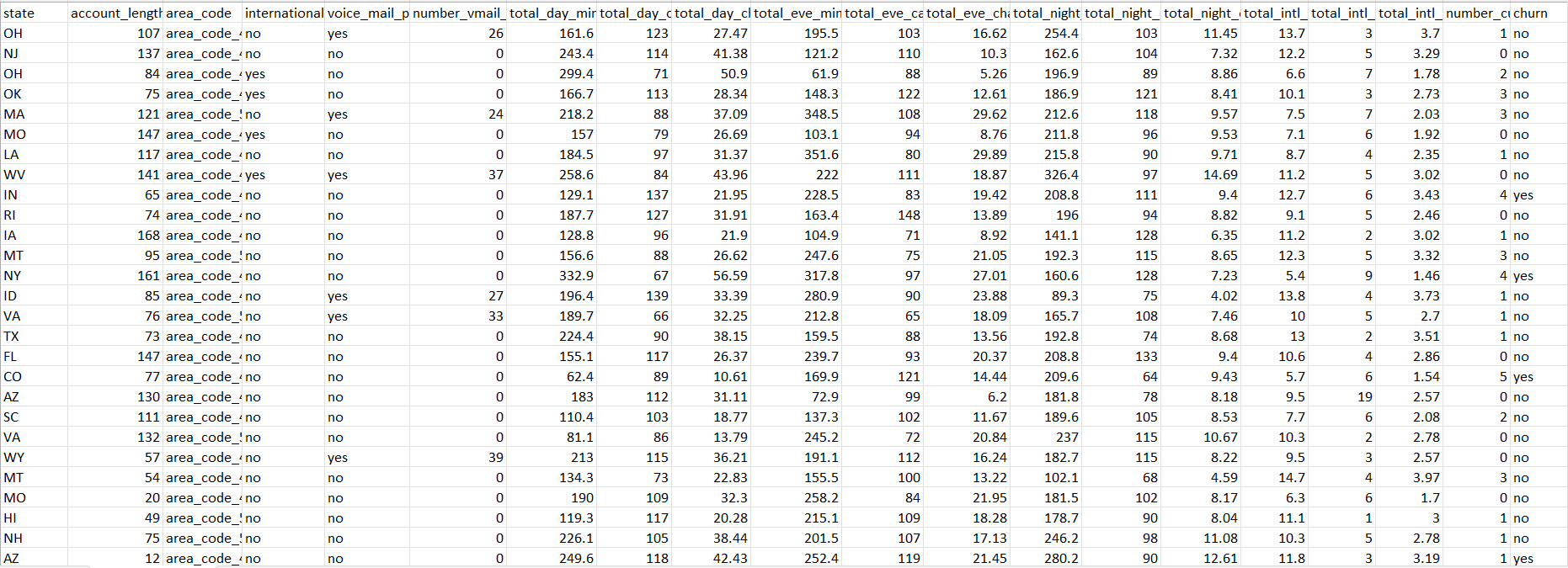
Dữ liệu hiện tại:Dự án giới hạn trong phạm vi dữ liệu hiện có, với các biến được cung cấp ban đầu. Các biến không có sẵn hoặc các yếu tố có khả năng tác động khác sẽ không được đưa vào phân tích.

Độ chính xác của dữ liệu:Giả định rằng dữ liệu được thu thập là đầy đủ và chính xác. Nếu có giá trị thiếu hoặc không nhất quán thì cần phải được xử lí để không ảnh hưởng đến kết quả của mô hình.

Sự ổn định của dữ liệu: Dự án giả định rằng các yếu tố quan trọng trong tập dữ liệu sẽ không thay đổi đột ngột hoặc có những biến động lớn trong thời gian phân tích.

## **Tổng quan về dữ liệu**

Nguồn dữ liệu: Kaggle

****

**state:** Trạng thái (tiểu bang) nơi khách hàng cư trú, thường là mã viết tắt của bang, ví dụ “CA” cho California.

**account\_length:** Thời gian (số ngày hoặc số tháng) khách hàng đã có tài khoản tại công ty.

**area\_code:** Mã vùng của khách hàng, cho biết vùng địa lí và khách hàng đang ở.

**international\_plan:** Kế hoạch cuộc gọi quốc tế, chỉ ra liệu khách hàng có đăng ký dịch vụ gọi quốc tế hay không (Yes hoặc No)

**voice\_mail\_plan:** Kế hoạch hộp thư thoại, cho biết khách hàng có dịch vụ hộp thư thoại hay không (Yes hoặc No)

**number\_vmail\_messages:** Số lượng tin nhắn hộp thư thoại mà khách hàng có trong tài khoản của mình.

**total\_day\_minutes:** Tổng số phút gọi trong ngày của khách hàng trong tháng, thường là 7 giờ sáng đến 7 giờ tối.

**total\_day\_calls:** Tổng số cuộc gọi trong ngày của khách hàng trong tháng.

**total\_day\_charge:** Tổng chi phí cuộc gọi trong ngày của khách hàng trong tháng, tính toán dựa trên thời gian gọi và cước phí.

**total\_eve\_minutes:** Tổng số phút gọi vào buổi tối của khách hàng trong tháng, thường là từ 7 giờ tối đến 11 giờ đêm.

**total\_eve\_calls:** Tổng số cuộc gọi vào buổi tối của khách hàng trong tháng.

**total\_eve\_charge:** Tổng chi phí cuộc gọi vào buổi tối của khách hàng trong tháng.

**total\_night\_minutes:** Tổng số phút gọi vào ban đêm của khách hàng trong tháng, thường là từ 11 giờ đêm đến 7 giờ sáng.

**total\_night\_calls:** Tổng số cuộc gọi vào ban đêm của khách hàng trong tháng.

**total\_night\_charge:** Tổng chi phí cuộc gọi vào ban đêm của khách hàng trong tháng.

**total\_intl\_minutes:** Tổng số phút gọi quốc tế của khách hàng trong tháng.

**total\_intl\_calls:** Tổng số cuộc gọi quốc tế của khách hàng trong tháng.

**total\_intl\_charge:** Tổng chi phí cuộc gọi quốc tế của khách hàng trong tháng.

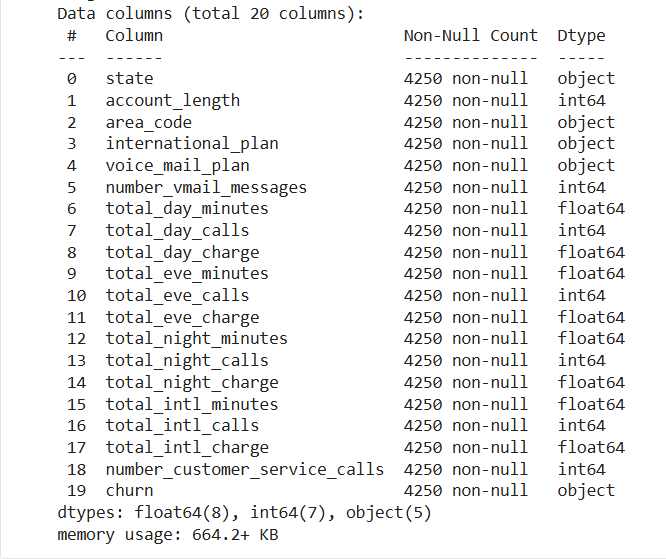
**number\_customer\_service\_calls:** Số lần khách hàng gọi đến dịch vụ khách hàng trong tháng, có thể là một yếu tố đo lường mức độ hài lòng của khách hàng.

**churn:** Biến mục tiêu cho biết khách hàng có rời bỏ (churn) công ty hay không(no: không rời bỏ, yes: rời bỏ)

## **1.4. Tiền xử lí dữ liệu**

**Số bản ghi là 4250**

**Xem thông tin các trường dữ liệu của dataset:**



Cần mã hóa các trường dữ liệu object như ‘state’, ‘area\_code’, ‘international\_plan’, ‘voice\_mail\_plan’, ‘churn’ để huấn luyện mô hình.

Do cột “state” có nhiều giá trị nên sẽ sử dụng HashingEncoder để mã hóa.

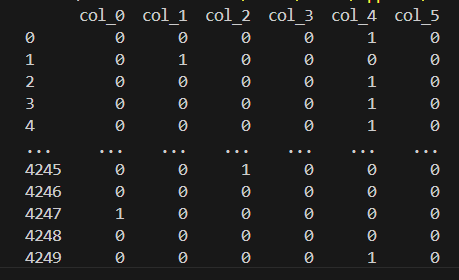
Đầu tiên khởi tạo HashingEncoder cho cột ‘state’:

**he = ce.HashingEncoder(cols='state')**

Sau đó thực hiện mã hóa lên dữ liệu trên

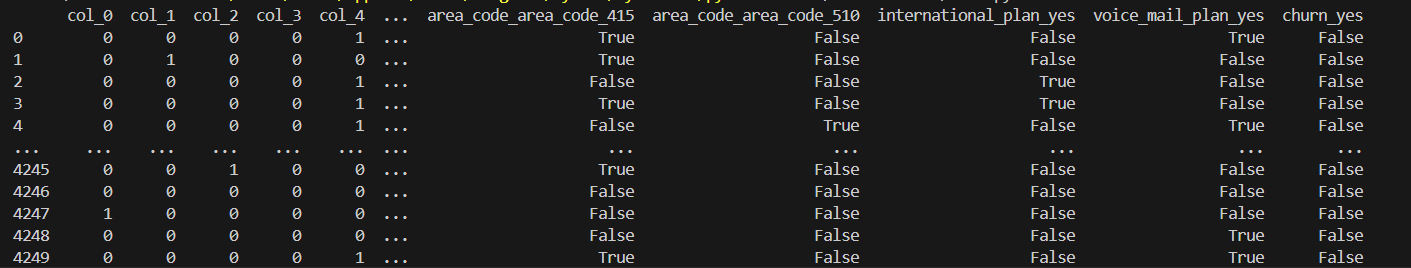
**dataset\_hash = he.fit\_transform(dataset)**

fit\_transform() áp dụng HashingEncoder() lên dataset để chuyển đổi cột 'state' thành các cột số mới, chứa dữ liệu mã hóa dạng nhị phân hoặc giá trị số khác, và trả về dataset\_hash, phiên bản đã mã hóa của dataset. Sau khi đã mã hóa ta được:

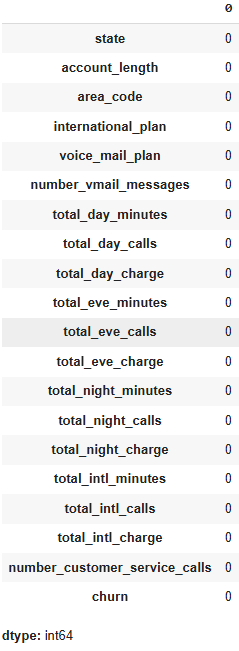
****

Với những trường dữ liệu object còn lại sử dụng OneHotEncoder của pandas, chuyển thành các cột nhị phân.

**dataset\_hash\_dummy = pd.get\_dummies(dataset\_hash, drop\_first=True)**

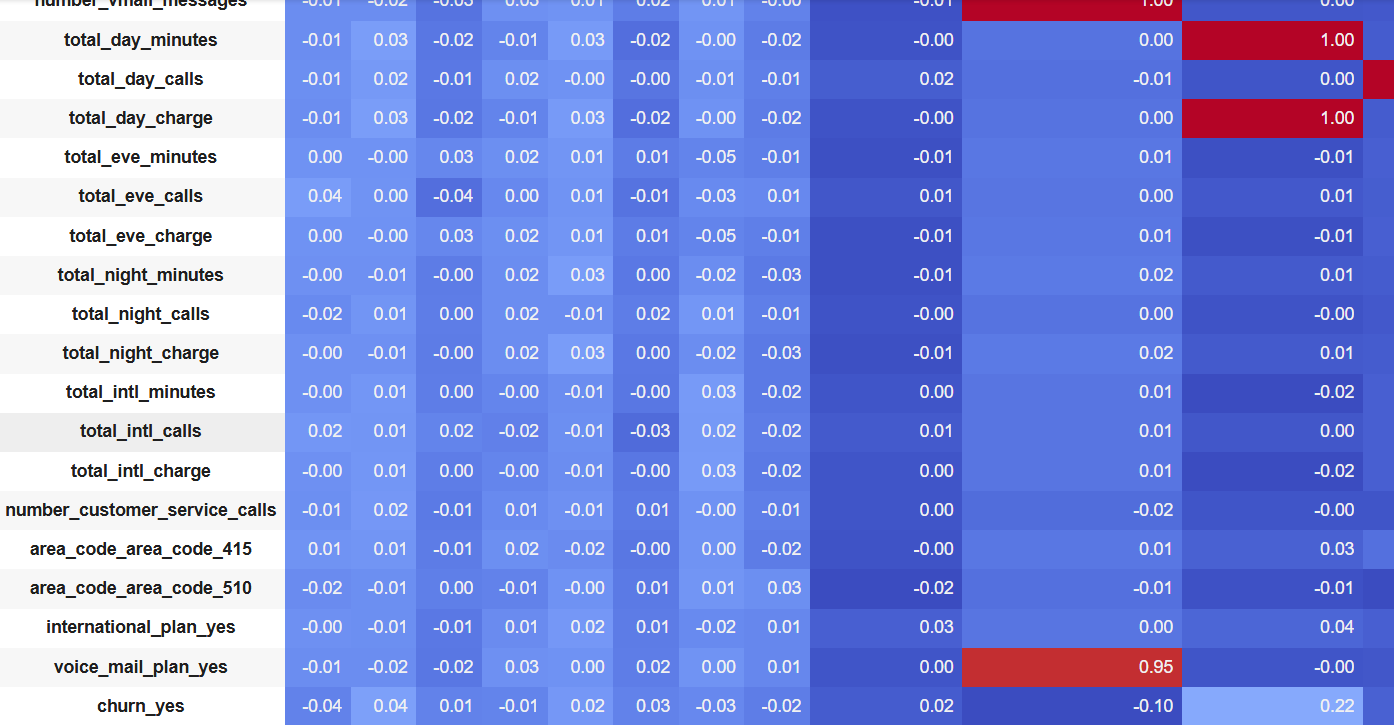


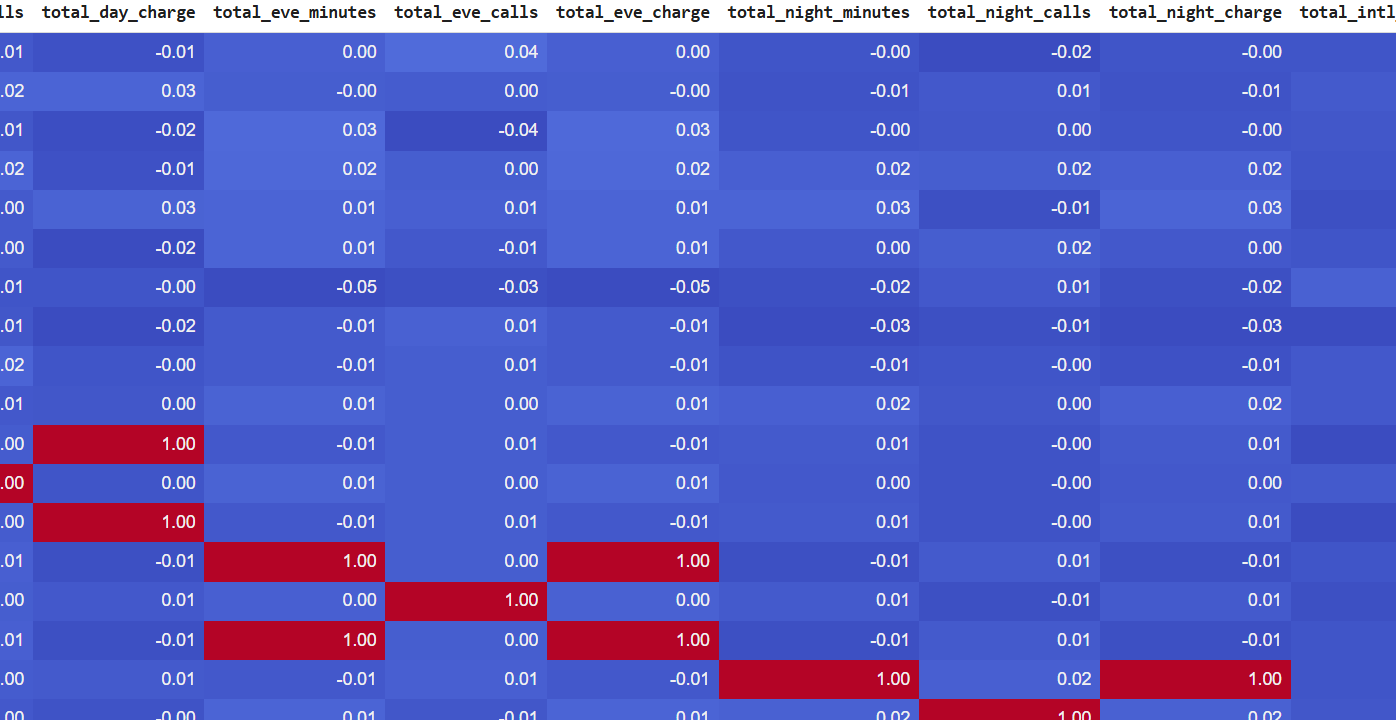
**Kiểm tra dữ liệu null:**

****

**Sử dụng correlation để tính toán độ tương quan giữa các cột trong dataset**

Nếu như có hai cột bất kì có độ tương quan cao (> 0.9 hoặc < -0.9), có thể xem xét loại bớt để tránh đa cộng tuyến, giúp cho mô hình đạt được kết quả tốt.





Sau khi chạy correlation nhận thấy có những trường dữ liệu có mối tương quan cao như:  
 **voice\_mail\_plan\_yes** và **number\_vmail\_messages**: 0.95

**total\_day\_charge** và **total\_day\_minutes**: 1.00

**total\_night\_minutes** và **total\_night\_charge**: 1.00

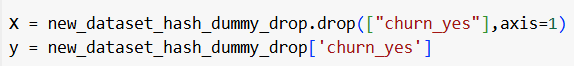
**total\_eve\_minutes** và **total\_eve\_charge**: 1.00

**total\_intl\_minutes** và **total\_intl\_charge**: 1.00

Drop những cột voice\_mail\_plan\_yes, total\_day\_charge, total\_eve\_charge, total\_night\_charge, total\_intl\_charge.

**new\_dataset\_hash\_dummy\_drop=dataset\_hash\_dummy.drop(columns=["voice\_mail\_plan\_yes","total\_day\_charge","total\_eve\_charge","total\_night\_charge","total\_intl\_charge"]).**

**Tách dữ liệu đầu vào**



Tách dữ liệu thành biến đầu vào (features) và biến mục tiêu (target).

Biến X lấy toàn bộ dữ liệu trừ cột churn\_yes để sử dụng làm biến đầu vào.

Biến y lấy riêng cột churn\_yes để làm biến mục tiêu.

Tách riêng features và target giúp dễ dàng áp dụng các thuật toán học máy.

Biến X được chuẩn hóa hoặc xử lý thêm, trong khi biến y giữ nguyên.

# Chương 2. Khám phá dữ liệu và lựa chọn mô hình

## **2.1. Trực quan hóa dữ liệu**

### ***2.1.1. Báo cáo khám phá dữ liệu từ trực quan hóa***

2.1.1.1. Mục tiêu phân tích

Trình bày kết quả khám phá từ dữ liệu trực quan hóa về hành vi rời bỏ của khách hàng.

Xác định các yếu tố quan trọng tác động đến quyết định rời bỏ dịch vụ.

Nhận diện nhóm khách hàng có nguy cơ cao để đưa ra các chiến lược giữ chân hiệu quả.

2.1.1.2. Dữ liệu và công cụ sử dụng

**Nguồn dữ liệu:** Kaggle

**Các biến số chính:**

**state:** Trạng thái (tiểu bang) nơi khách hàng cư trú, thường là mã viết tắt của bang, ví dụ “CA” cho California.

**account\_length:** Thời gian (số ngày hoặc số tháng) khách hàng đã có tài khoản tại công ty.

**area\_code:** Mã vùng của khách hàng, cho biết vùng địa lí và khách hàng đang ở.

**international\_plan:** Kế hoạch cuộc gọi quốc tế, chỉ ra liệu khách hàng có đăng ký dịch vụ gọi quốc tế hay không (Yes hoặc No).

**voice\_mail\_plan:** Kế hoạch hộp thư thoại, cho biết khách hàng có dịch vụ hộp thư thoại hay không (Yes hoặc No).

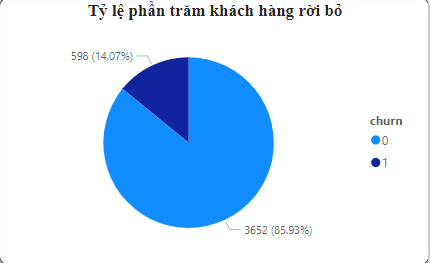
**number\_customer\_service\_calls:** Số lần khách hàng gọi đến dịch vụ khách hàng trong tháng, có thể là một yếu tố đo lường mức độ hài lòng của khách hàng.

**churn:** Biến mục tiêu cho biết khách hàng có rời bỏ (churn) công ty hay không(no: không rời bỏ, yes: rời bỏ) .

**Công cụ sử dụng**: Power BI.

* + 1. ***Kết quả khám phá từ việc trực quan hóa***

2.1.2.1 Phân tích về khách hàng

****

Số lượng khách hàng thống kê được là 4250 người. Trong đó có 598 người rời bỏ, chiếm 14%.

Số lượng khách hàng rời bỏ chỉ chiếm tỉ lệ khá nhỏ so với tổng số khách hàng, cho thấy mức độ hài lòng của phần lớn khách hàng hoặc sự thành công trong việc giữ chân khách hàng, Tuy nhiên, tỉ lệ 14% vẫn có thể là đáng kể, cần tìm hiểu sâu hơn về lí do khiến cho 14% khách hàng rời bỏ để có thể cải thiện các chiến lược giữ chân khách hàng.

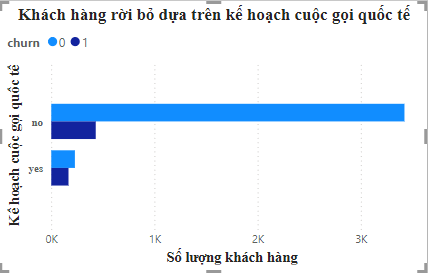
2.1.2.2. Phân tích thời gian sử dụng trung bình



Thời gian trung bình (mean của account\_length) khách hàng sử dụng dịch vụ của công ty là 100, cho thấy công ty có một lượng khách hàng duy trì dịch vụ trong thời gian tương đối ổn định.

Đối với những khách hàng có account\_length cao hơn giá trị trung bình này, họ có thể là những người gắn bó lâu dài và có xu hướng ít rời bỏ. Ngược lại, những khách hàng có thời gian sử dụng ít hơn có thể là nhóm dễ rời bỏ và cần được chú ý thêm.

2.1.2.3. Phân tích lượng khách hàng rời bỏ dựa trên dịch vụ quốc tế



Biểu đồ trên thể hiện mối quan hệ giữa kế hoạch cuộc gọi quốc tế (international\_plan) và khách hàng rời bỏ dịch vụ (churn). Trong đó:

**Trục Y (Kế hoạch cuộc gọi quốc tế)**: Hiển thị hai nhóm khách hàng:

"Yes" (có đăng ký kế hoạch cuộc gọi quốc tế).

"No" (không đăng ký kế hoạch cuộc gọi quốc tế).

**Trục X (Số lượng khách hàng)**: Thể hiện số lượng khách hàng trong từng nhóm.

**Màu sắc (churn)**:

Màu xanh nhạt (churn = 0): Khách hàng không rời bỏ.

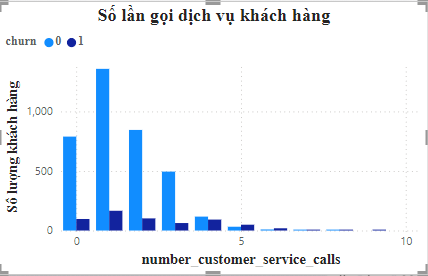
Màu xanh đậm (churn = 1): Khách hàng rời bỏ.

Biểu đồ trên cho thấy:

Khách hàng không đăng kí kế hoạch gọi quốc tế chiếm đa số nhưng lượng khách hàng rời bỏ thấp, cho thấy tỉ lệ liên quan đến rời bỏ dịch vụ là không cao.

Khách hàng đăng kí kế hoạch gọi quốc tế ít hơn nhưng tỉ lệ rời bỏ cao hơn rõ rệt. Cần tập trung vào nhóm khách hàng trên để tìm hiểu thêm lí do rời bỏ và đưa ra chiến lược cải thiện trải nghiệm nhằm giảm tỉ lệ rời bỏ ở nhóm này.

2.1.2.4 Phân tích số lần gọi dịch vụ khách hàng



Biểu đồ trên là **biểu đồ cột** thể hiện mối quan hệ giữa **số lần gọi dịch vụ khách hàng (number\_customer\_service\_calls)** và tỷ lệ **khách hàng rời bỏ dịch vụ (churn)**. Các yếu tố chính của biểu đồ bao gồm:

**Trục Y (Số lượng khách hàng)**: Hiển thị số lượng khách hàng tương ứng với từng số lần gọi dịch vụ khách hàng.

**Trục X (number\_customer\_service\_calls)**: Thể hiện số lần gọi đến dịch vụ khách hàng, từ 0 đến 10 lần.

**Màu sắc (churn)**:

Màu xanh nhạt (churn = 0): Khách hàng không rời bỏ.

Màu xanh đậm (churn = 1): Khách hàng rời bỏ.

**Nhóm khách hàng gọi từ 0 đến 1 lần**:

Chiếm số lượng lớn nhất và đa phần thuộc nhóm không rời bỏ (màu xanh nhạt).

Điều này cho thấy khách hàng ít gọi đến dịch vụ khách hàng có xu hướng không rời bỏ dịch vụ.

**Nhóm khách hàng gọi từ 2 đến 4 lần**:

Số lượng khách hàng có xu hướng giảm dần khi số lần gọi dịch vụ khách hàng tăng lên.

Trong nhóm này, tỷ lệ rời bỏ bắt đầu tăng dần (màu xanh đậm tăng lên), cho thấy có mối quan hệ giữa việc gọi nhiều lần và khả năng rời bỏ.

**Nhóm khách hàng gọi từ 5 lần trở lên**:

Số lượng khách hàng trong nhóm này rất ít, nhưng tỷ lệ rời bỏ lại rất cao (màu xanh đậm chiếm ưu thế).

Điều này cho thấy rằng khi khách hàng gọi dịch vụ khách hàng nhiều lần, khả năng họ rời bỏ công ty là cao hơn.

Biểu đồ cho thấy xu hướng: càng gọi nhiều lần đến dịch vụ khách hàng, tỷ lệ rời bỏ càng cao.

Điều này có thể phản ánh rằng những khách hàng gọi nhiều lần có thể đang gặp vấn đề với dịch vụ và không hài lòng với trải nghiệm của mình.

Để giảm tỷ lệ rời bỏ, công ty có thể tập trung cải thiện chất lượng dịch vụ khách hàng và giảm số lần khách hàng cần phải gọi hỗ trợ.

## **2.2. Lựa chọn mô hình**

### ***2.2.1. Trình bày mô hình và thuật toán được chọn để phân tích dữ liệu***

Mô hình sử dụng: XGBoost, một thuật toán học máy mạnh mẽ với hiệu suất cao cho các bài toán phân loại và hồi quy.

### ***2.2.2. Lý do chọn mô hình***

Hiệu quả cao trong xử lý dữ liệu phi tuyến tính:

Các yếu tố ảnh hưởng đến sự rời bỏ thường có quan hệ phi tuyến, như tương tác giữa chi phí, loại hợp đồng và thời gian gắn bó.

XGBoost giúp khai thác tối ưu mối quan hệ này.

Độ chính xác cao:

XGBoost sử dụng kỹ thuật gradient boosting, cho phép cải thiện độ chính xác của dự đoán qua nhiều vòng lặp (iterations).

Xử lý tốt dữ liệu không đồng nhất:

XGBoost hiệu quả với các tập dữ liệu có đặc điểm khác biệt, như dữ liệu định lượng (chi phí, thời gian) và dữ liệu danh mục (loại hợp đồng, phương thức thanh toán).

Khả năng chống overfitting:

Sử dụng các tham số kiểm soát như learning\_rate, max\_depth để tránh việc mô hình học quá chi tiết vào dữ liệu huấn luyện.

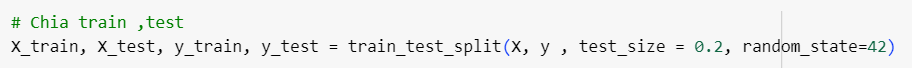
Tối ưu hóa thời gian xử lý:

XGBoost được tối ưu hóa cho hiệu suất cao, đặc biệt với các tập dữ liệu lớn.

### ***2.2.3 Quá trình chia dữ liệu thành tập huấn luyện, tập kiểm tra***

**Chia dữ liệu thành tập huấn luyện (training sets) và tập kiểm tra (datasets)**

Sử dụng hàm train\_test\_split từ thư viện scikit-learn để chia dữ liệu đầu vào thành hai tập riêng biệt:



Tập huấn luyện (X\_train, y\_train): dữ liệu được dùng để huấn luyện mô hình.

Tập kiểm tra (X\_test, y\_test): Dữ liệu được dữ lại để đánh giá hiệu suất của mô hình trên dữ liệu chưa từng thấy trước đó.

Các tham số đầu vào:

**X:** Tập dữ liệu đặc trưng (features). Kích thước ban đầu: (4250, 22).

**y:** Biến mục tiêu (cột churn\_yes), kích thước (4250,).

**test\_size=0.2:** 20% dữ liệu được đưa vào tập X\_test và y\_test. Phần còn lại (80%) đưa vào X\_train và y\_train.

**random\_state=42:** giá trị này giúp chia dữ liệu cùng một cách mỗi lần chạy, đảm bảo kết quả ổn đỉnh khi huấn luyện mô hình. Nếu không đặt, dữ liệu sẽ được ngẫu nhiên khác nhau.

# Chương 3. Huấn luyện, tối ưu hóa và đánh giá mô hình

## **3.1 Mô hình sử dụng: XGBoost Classifier**

### ***3.1.1 Huấn luyện mô hình***

Quá trình huấn luyện mô hình

Khởi tạo mô hình XGBoost

****

Đây là lớp phân loại (classifier) của XGBoost, được thiết kế để giải quyết các bài toán phân loại

Tham số random\_state=42: Đảm bảo kết quả có thể tái lập được bằng cách cố định trạng thái khởi tạo ngẫu nhiên.

n\_estimator=200: số lượng cây quyết định trong mô hình XGBoost

****

X\_train: là tập dữ liệu đầu vào (đặc trưng) dùng để huấn luyện mô hình.

y\_train: là tập nhãn mục tiêu tương ứng với từng mẫu trong X\_train.

fit: phương thức dùng để huấn luyện mô hình. Mô hình sẽ học cách phân loại dựa trên các đặc trưng đầu vào X\_train và nhãn mục tiêu y\_train.

****

**predict()**

Đây là phương thức dùng để dự đoán nhãn (đầu ra) của dữ liệu mới.

Mỗi mẫu trong X\_test được đưa qua các cây quyết định của mô hình.

Kết quả của từng cây sẽ được tổng hợp để tính toán xác suất của mỗi nhãn.

Nhãn có xác suất cao nhất sẽ được gán làm kết quả dự đoán cho mẫu đó.

Kết quả sẽ là:

0: Nếu dự đoán mẫu không thuộc lớp khách hàng rời bỏ.

1: Nếu dự đoán mẫu thuộc lớp khách hàng rời bỏ.

Đầu vào:

X\_test:

Dữ liệu kiểm tra bao gồm các đặc trưng đầu vào.

Mỗi hàng đại diện cho một mẫu, mỗi cột là một đặc trưng.

Đầu ra:

y\_pred:

Mảng chứa các nhãn dự đoán cho tập kiểm tra (X\_test).

* + 1. ***Đánh giá hiệu suất mô hình***

****

**accuracy\_score():**

Tính độ chính xác tổng quát (Accuracy) của mô hình

Kết quả: Accuracy: 0.9611764705882353

Mô hình dự đoán đúng ~96.11% số mẫu trong tập kiểm tra

****

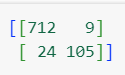
**Confusion\_matrix():**

Tính ma trận nhầm lẫn (Confusion Matrix) để đánh giá kết quả phân loại

Ma trận này cho biết số lượng dự đoán đúng và sai của mô hình theo từng lớp

Đầu vào là y\_test: nhãn thực tế từ tập kiểm tra, y\_pred: Nhãn dự đoán từ mô hình

Kết quả:

****

**True Positive (TP) = 712:**

Mô hình dự đoán chính xác rủi ro tín dụng cho 712 khách hàng. Đây là kết quả mong muốn, giúp cho công ty tối ưu hóa nguồn lực để tập trung vào nhóm khách hàng có tỉ lệ rời bỏ .

**False Negative (FN) = 9:**

9 khách hàng không rời bỏ đã bị mô hình dự đoán nhầm lẫn là rời bỏ, khiển công ty có thể triển khai các chương trình giữ chân khách hàng không cần thiết, gây ra lãng phí tài nguyên hoặc chi phí không cần thiết.

Tuy nhiên, việc nhầm lẫn có mức độ ảnh hưởng nhỏ so với dự đoán đúng khách hàng không rời bỏ.

**False Positive (FP) = 24:**

24 khách hàng rời bỏ đã bị mô hình dự đoán sai là không rời bỏ.

Đây là một dự đoán sai nghiêm trọng.

Ngân hàng không nhận diện được những khách hàng có nguy cơ rời bỏ, dẫn đến:

Mất doanh thu từ nhóm khách hàng này.

Tăng chi phí để tìm kiếm khách hàng mới thay thế.

**True Negative (TN) = 105:**

105 khách hàng rời bỏ đã được mô hình dự đoán chính xác là rời bỏ.

Đây là kết quả mong muốn.

Ngân hàng có thể chủ động triển khai các biện pháp giữ chân trước khi khách hàng này thực sự rời bỏ, ví dụ:

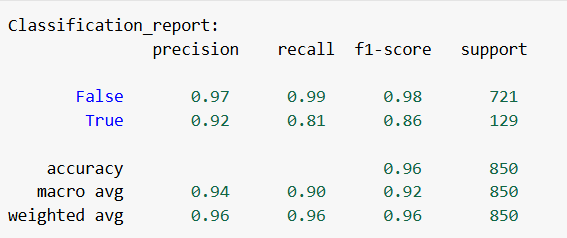
Giảm lãi suất, tăng hạn mức vay, cải thiện chất lượng dịch vụ.

Phân tích nguyên nhân rời bỏ để đưa ra các cải tiến phù hợp.

****

**classification\_report():** Tạo một bảng báo cáo tóm tắt các chỉ số đánh giá hiệu năng phân loại của mô hình.

Kết quả:



**Lớp False (Không rời bỏ khách hàng)**

***Precision = 0.97:***

Trong số các dự đoán là "Không rời bỏ", 97% là chính xác.

Chỉ có 3% bị nhầm lẫn thành "Rời bỏ".

Điều này cho thấy mô hình rất tốt trong việc không "gắn nhãn sai" nhóm không rời bỏ.

***Recall = 0.99:***

Trong số các khách hàng thực sự "Không rời bỏ", mô hình nhận diện đúng 99%.

Chỉ 1% bị bỏ sót và nhầm lẫn thành "Rời bỏ".

***F1-Score = 0.98:***

Mô hình cân bằng tốt giữa Precision và Recall cho nhóm này.

Hiệu suất rất cao, phản ánh mô hình có độ chính xác ổn định.

***Support = 721:***

Số lượng khách hàng thực tế thuộc lớp "Không rời bỏ" là 721, chiếm phần lớn trong tập kiểm tra.

**Lớp True (Rời bỏ khách hàng)**

***Precision = 0.92:***

Trong số các dự đoán là "Rời bỏ", 92% là chính xác.

8% còn lại bị nhầm lẫn thành "Không rời bỏ".

Precision ở mức tốt, nhưng có thể cải thiện thêm.

***Recall = 0.81:***

Trong số các khách hàng thực sự "Rời bỏ", mô hình nhận diện đúng 81%.

19% bị bỏ sót và nhầm lẫn thành "Không rời bỏ".

Độ nhạy của lớp này thấp hơn so với lớp "Không rời bỏ", có thể gây rủi ro nếu ưu tiên nhận diện khách hàng rời bỏ.

***F1-Score = 0.86:***

Hiệu suất cân bằng giữa Precision và Recall là khá tốt, nhưng thấp hơn so với lớp "Không rời bỏ".

***Support = 129:***

Số lượng khách hàng thực tế thuộc lớp "Rời bỏ" là 129, nhỏ hơn nhiều so với lớp "Không rời bỏ".

***Accuracy = 0.96:***

Tỷ lệ dự đoán đúng trên toàn bộ tập dữ liệu kiểm tra là 96%. Đây là một kết quả rất cao.

***Macro Average:***

Trung bình cộng của Precision, Recall, và F1-Score trên các lớp.

***Precision = 0.94, Recall = 0.90, F1-Score = 0.92.***

Thể hiện hiệu suất tổng thể của mô hình, không cân nhắc độ lệch về số lượng giữa các lớp.

***Weighted Average:***

Trung bình có trọng số của Precision, Recall, và F1-Score, cân nhắc số lượng mẫu của từng lớp.

***Precision = 0.96, Recall = 0.96, F1-Score = 0.96.***

Phù hợp để đánh giá trong các tập dữ liệu mất cân bằng (như ở đây, lớp "Không rời bỏ" chiếm ưu thế).