|  |
| --- |
| **TRƯỜNG ĐẠI HỌC SƯ PHẠM HÀ NỘI**  **Khoa Công Nghệ Thông Tin**  **-----  -----**    **BÁO CÁO: THỰC HÀNH DỰ ÁN VÀ CHUYÊN ĐỀ TỐT NGHIỆP DỰ ÁN**    **Đề tài:**  ***Phân tích dự đoán sự rời bỏ của khách hàng***  Giảng viên: Phạm Thọ Hoàn  Sinh viên: Nguyễn Đức Mậu - 715105146    *Hà Nội, 2024* |

**PHÂN TÍCH DỰ ĐOÁN SỰ RỜI BỎ CỦA KHÁCH HÀNG**

**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 cần giải quyết:**

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**

**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.

**Giới hạn và giả định của dự án**

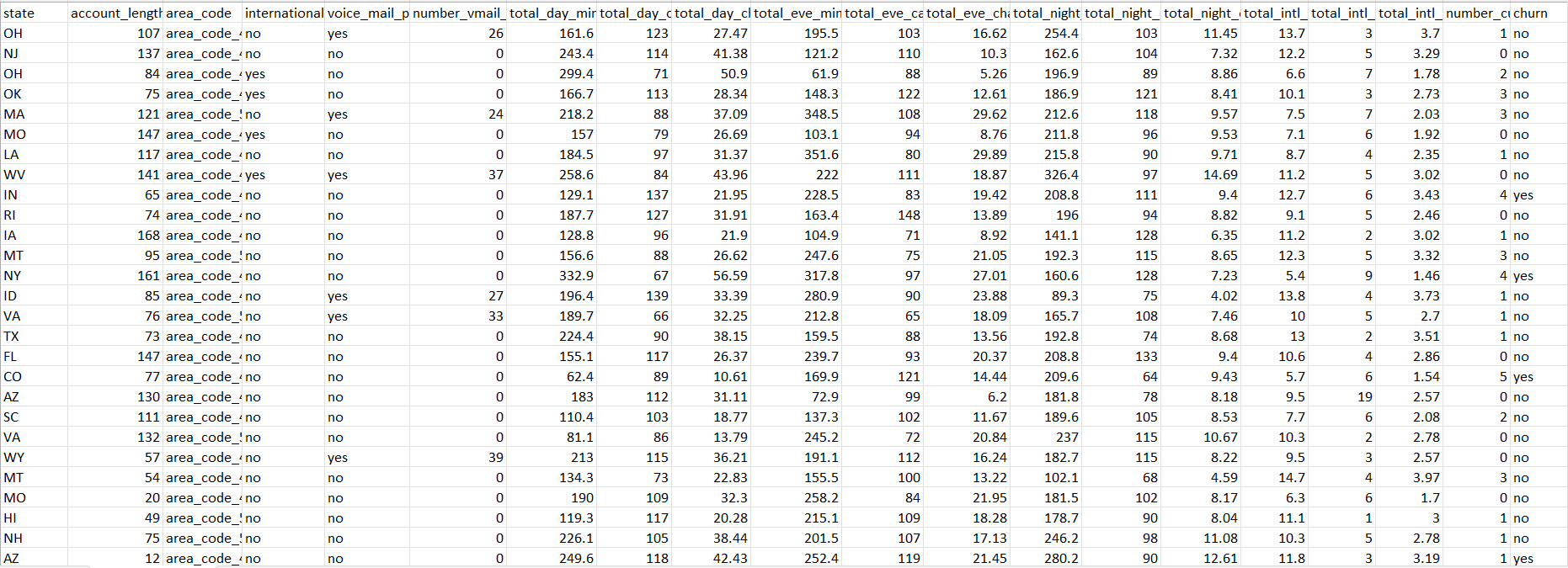
**Dữ liệu hiện tại:** Dự án giới hạn trong phạm vi dữ liệu hiện có, với các biến được cung cấp ban đầu. Các biến không có sẵn hoặc các yếu tố có khả năng tác động khác sẽ không được đưa vào phân tích.

**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.

**1.3 Tổng quan về dữ liệu**

Dữ liệu được sử dụng trong dự án này là tập dữ liệu lịch sử của khách hàng từ một công ty viễn thông, ghi nhận thông tin liên quan đến các đặc điểm sử dụng dịch vụ, mức độ tương tác và thông tin tài khoản của khách hàng. Dữ liệu này đại diện cho nhiều loại khách hàng khác nhau, từ đó phân tích các yếu tố dẫn đến việc rời bỏ dịch vụ của khách hàng.

****

**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.

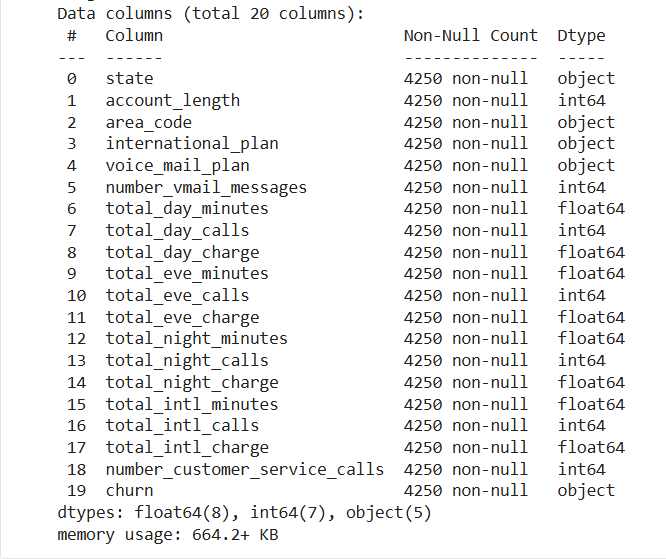
**number\_customer\_service\_calls:** Số lần khách hàng gọi đến dịch vụ khách hàng trong tháng, có thể là một yếu tố đo lường mức độ hài lòng của khách hàng.

**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 Thu thập và tiền xử lí dữ liệu**

**Nguồn dữ liệu: Kaggle**

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



Có những trường object như “state”,“area\_code”,“international\_plan”, “voice\_mail\_plan” và “churn ” cần được mã hóa để có thể huấn luyện mô hình

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

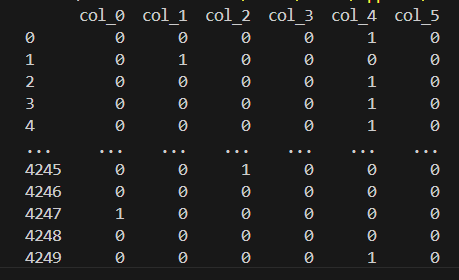
Đầu tiên khởi tạo HashingEncoder cho cột ‘state’:

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

Sau đó thực hiện mã hóa lên dữ liệu trên

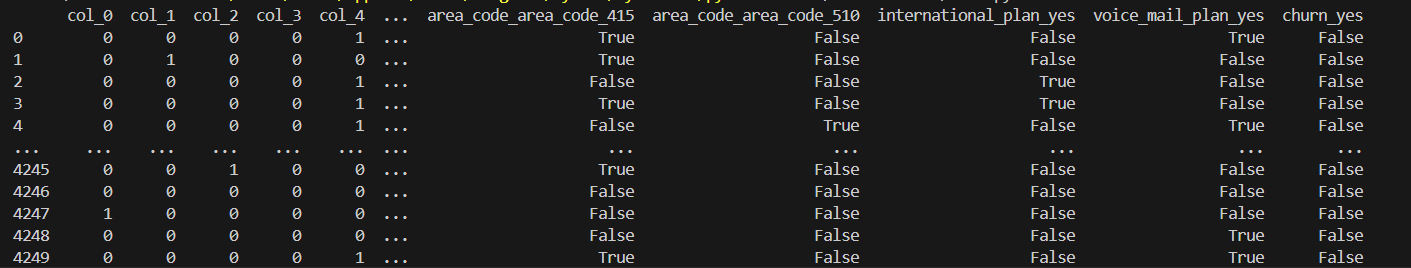
**dataset\_hash = he.fit\_transform(dataset)**

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:

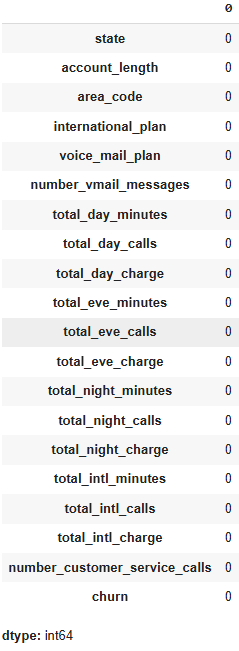
****

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)**

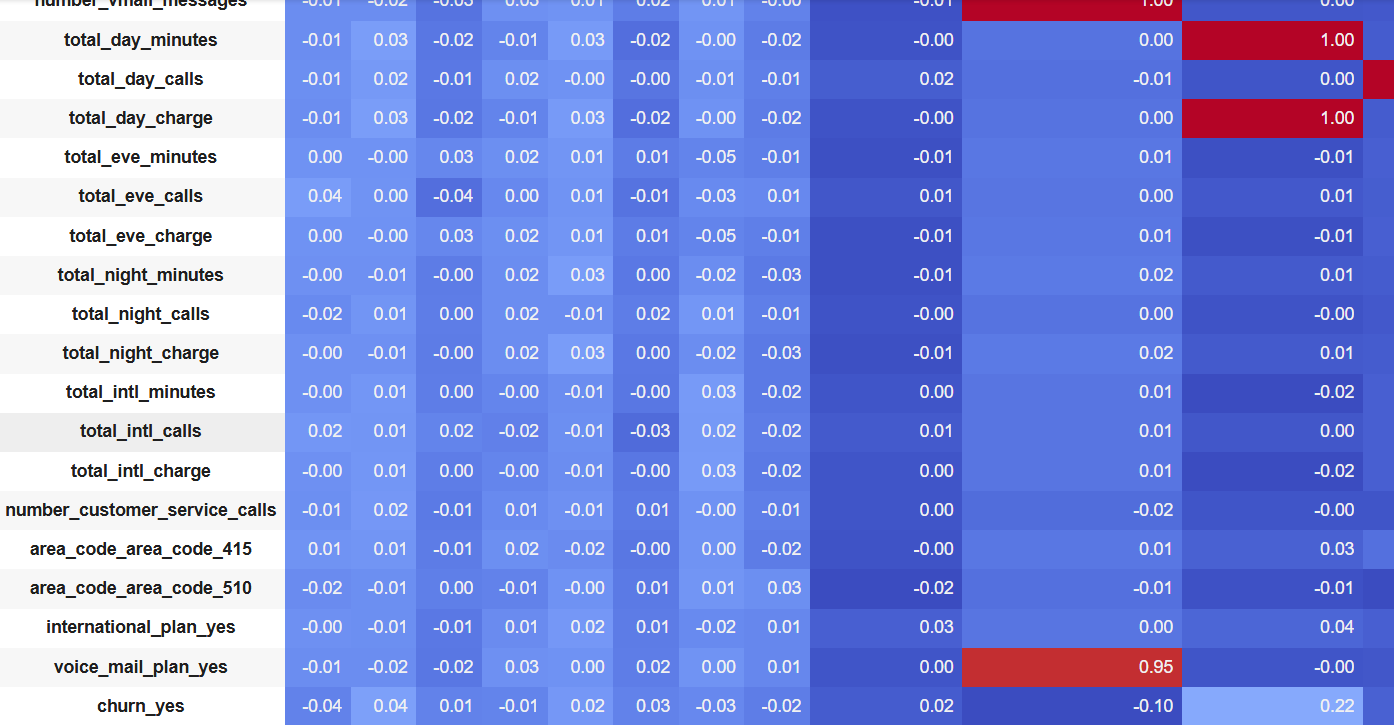


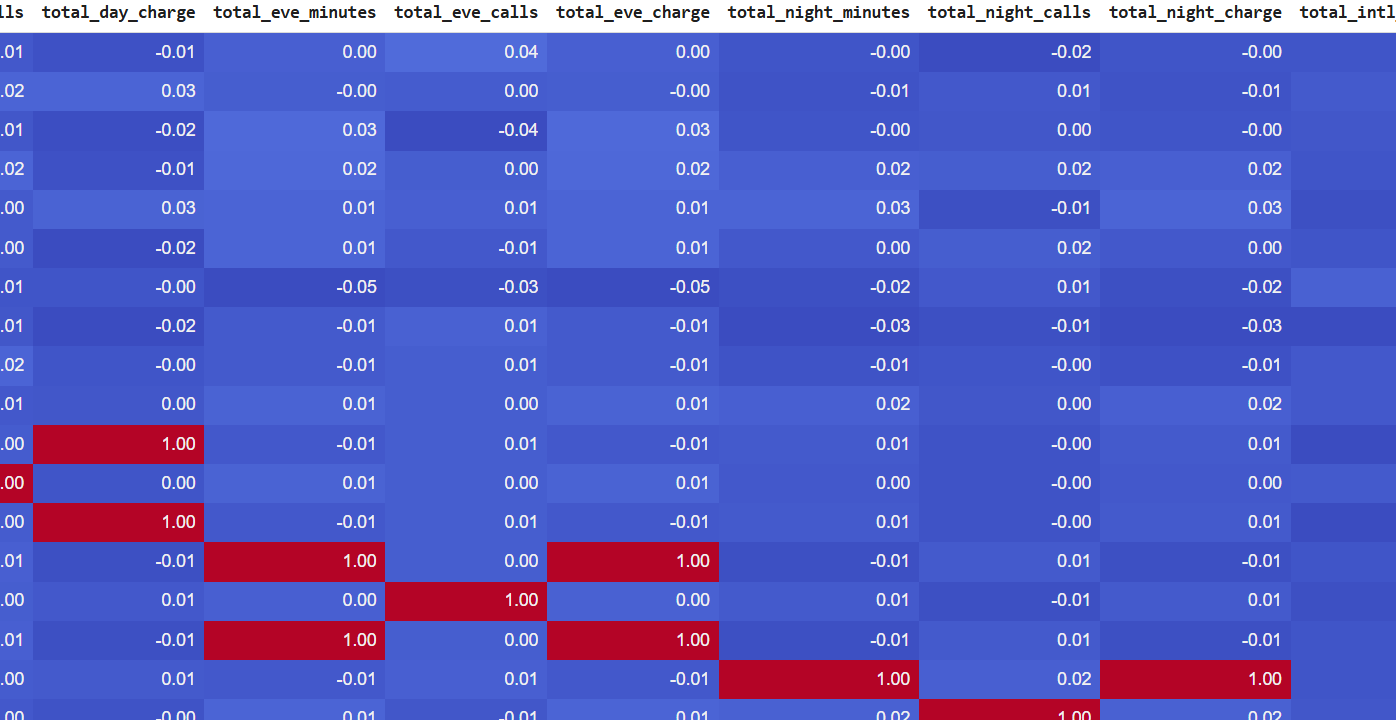
**Kiểm tra dữ liệu null:**

****

**\* 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"])**

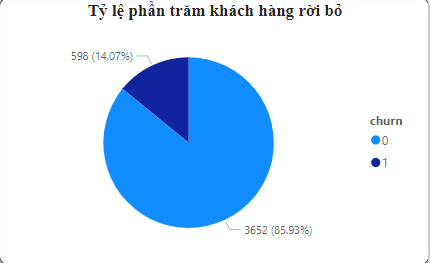
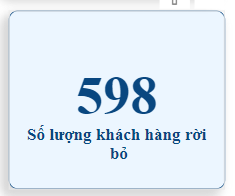
**Chương 2: Khám phá dữ liệu và lựa chọn mô hình**

**2.1 Trực quan hóa dữ liệu**

****

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

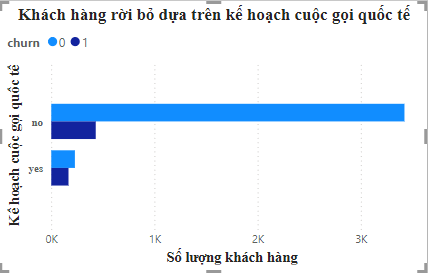
Số lượng khách hàng rời bỏ chỉ chiếm tỉ lệ khá nhỏ so với tổng số khách hàng, cho thấy mức độ hài lòng của phần lớn khách hàng hoặc sự thành công trong việc giữ chân khách hàng, Tuy nhiên, tỉ lệ 14% vẫn có thể là đáng kể, cần tìm hiểu sâu hơn về lí do khiến cho 14% khách hàng rời bỏ để có thể cải thiện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

****



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.



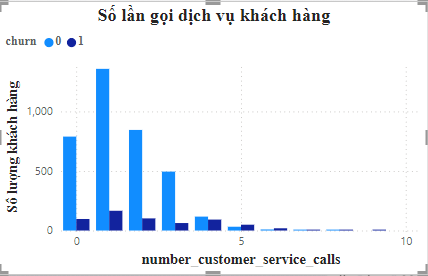
Biểu đồ trên thể hiện mối quan hệ giữa kế hoạch cuộc gọi quốc tế (international\_plan) và khách hàng rời bỏ dịch vụ (churn).Trong đó:

* **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.



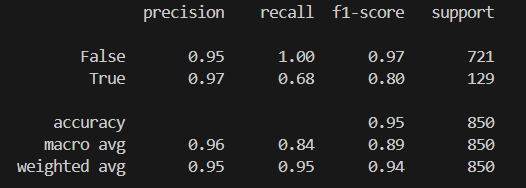
Biểu đồ trên là **biểu đồ cột** thể hiện mối quan hệ giữa **số lần gọi dịch vụ khách hàng (number\_customer\_service\_calls)** và tỷ lệ **khách hàng rời bỏ dịch vụ (churn)**. Các yếu tố chính của biểu đồ bao gồm:

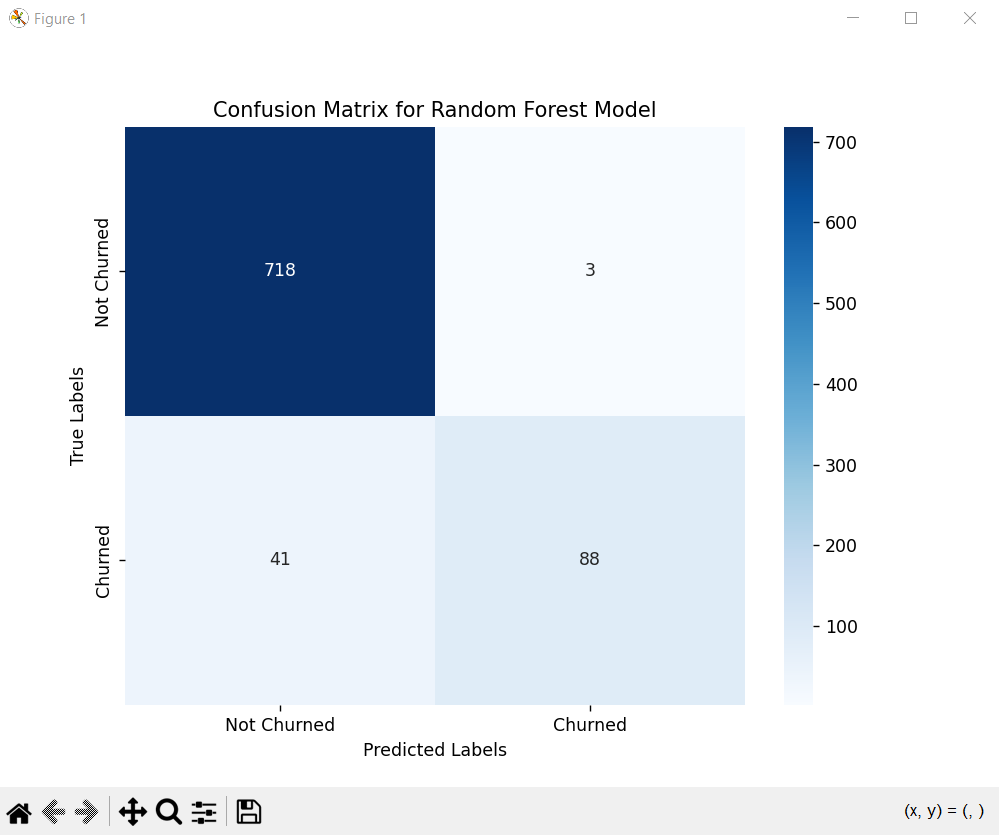
* **Trục Y (Số lượng khách hàng)**: Hiển thị số lượng khách hàng tương ứng với từng số lần gọi dịch vụ khách hàng.
* **Trục X (number\_customer\_service\_calls)**: Thể hiện số lần gọi đến dịch vụ khách hàng, từ 0 đến 10 lần.
* **Màu sắc (churn)**:
  + Màu xanh nhạt (churn = 0): Khách hàng không rời bỏ.
  + 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.

**Chương 3: Huấn luyện, tối ưu hóa và đánh giá mô hình**

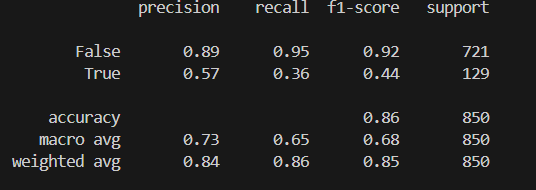
**Lựa chọn mô hình:** RandomForestClassifier, Hồi quy Logistic và XGBoost

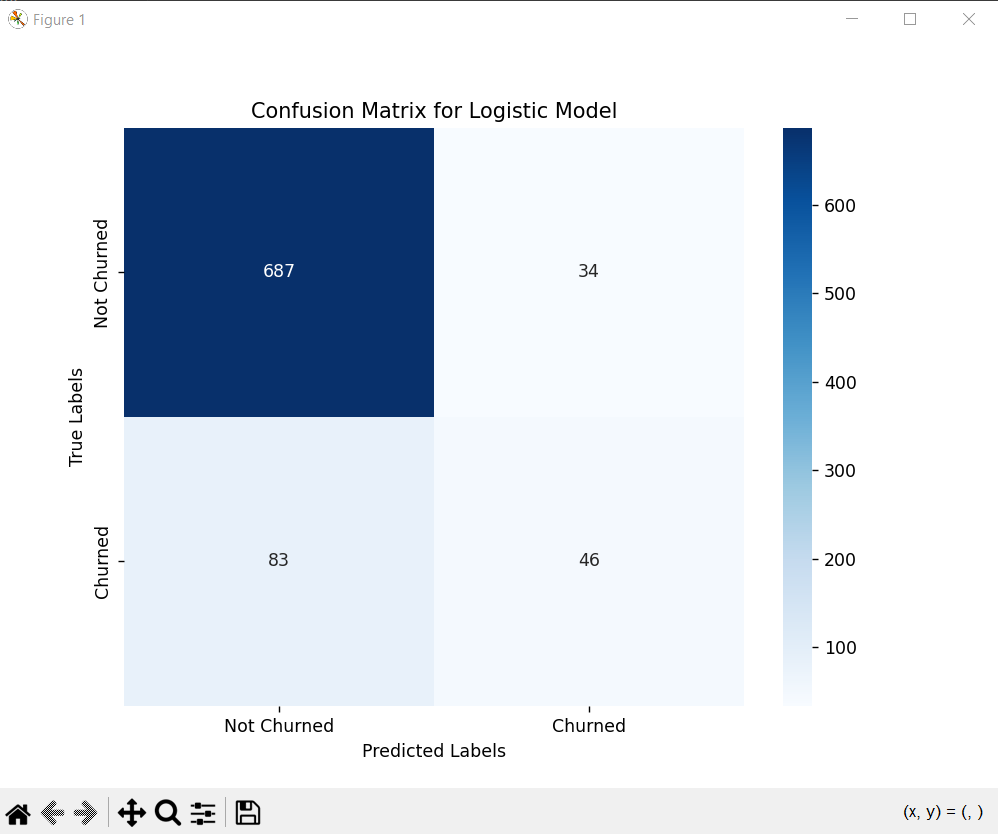
Kết quả chạy RandomForestClassifier



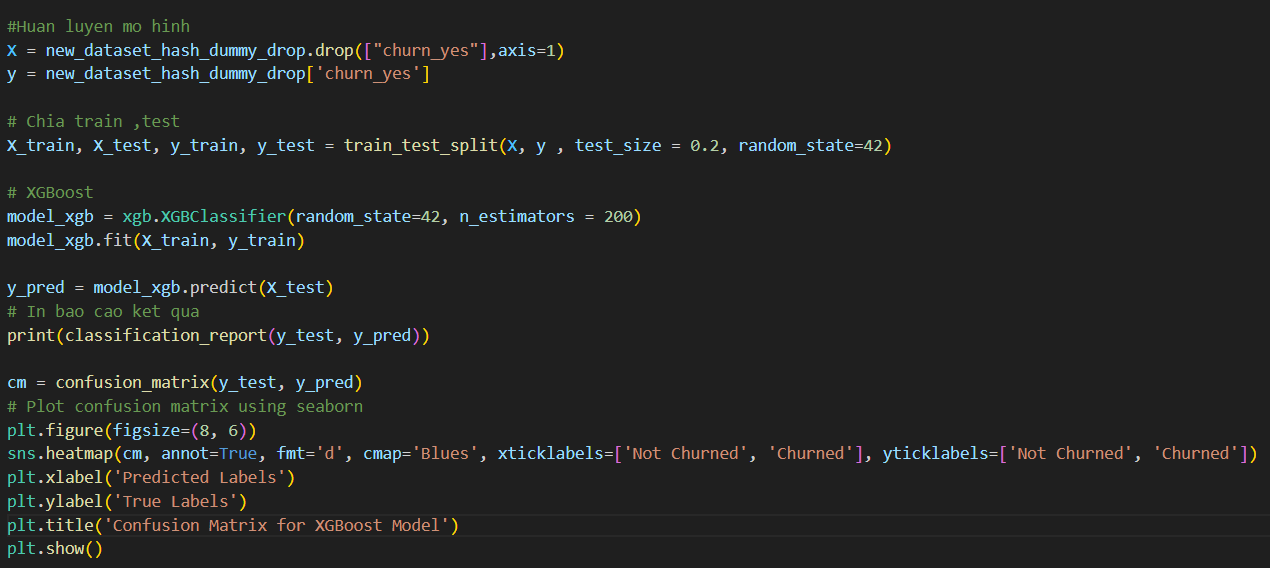
****

Kết quả chạy Hồi quy Logistic





**Mô hình XGBoost**

****

Kết quả chạy XGBoost:

