

# PHÂN TÍCH VÀ DỰ ĐOÁN KHẢ NĂNG RỜI BỎ CỦA KHÁCH HÀNG TRONG THƯƠNG MẠI ĐIỆN TỬ

IE313.Q11 - Phân tích và trực quan dữ liệu  
GVHD: ThS. Phạm Nguyễn Phúc Toàn

# Nhóm 3



1

Phạm Huỳnh Tấn Khang

22520624

2

Huỳnh Ngọc Trang

22521510

3

Nguyễn Huỳnh Xuân Nghi

23521004

4

Nguyễn Thị Ngọc Phước

23521235

• • •

# Nội dung

- 01 **Tổng quan đề tài**
- 02 **Công trình liên quan**
- 03 **Phương pháp nghiên cứu**
- 04 **Kết quả & Bàn luận**
- 05 **Kết luận & Giải pháp**
- 06 **Hướng phát triển**

# 1. Tổng quan đề tài



## Bối cảnh

- **Cạnh tranh gay gắt** giữa các doanh nghiệp trong ngành thương mại điện tử đòi hỏi chiến lược giữ chân khách hàng hiệu quả.
- **Chi phí thu hút khách hàng mới cao hơn** đáng kể so với chi phí giữ chân khách hàng hiện có.



## Vấn đề giải quyết

**Phát hiện sớm** khả năng rời bỏ của khách hàng là một bài toán có ý nghĩa thực tiễn cao, giúp doanh nghiệp chủ động xây dựng các chiến lược can thiệp.

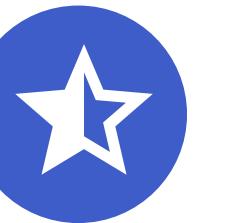
## Tác động khách hàng rời bỏ



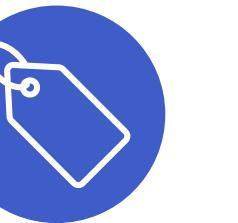
Giảm doanh thu



Trải nghiệm  
người dùng



Chất lượng  
dịch vụ



Khuyến mãi &  
nhu cầu hóa



# 1. Tổng quan đề tài



## Mục tiêu

01

### Phân tích và xác định các yếu tố ảnh hưởng

Thông qua các phương pháp thống kê và trực quan hóa dữ liệu

02

### Xây dựng và so sánh các mô hình học máy

Để dự đoán khả năng rời bỏ của khách hàng, từ đó lựa chọn mô hình có hiệu suất tối ưu nhất

03

### Đề xuất các giải pháp kinh doanh

Để xây dựng chiến lược giữ chân khách hàng một cách chủ động và hiệu quả



## Phạm vi và công cụ nghiên cứu

### Phạm vi

Nghiên cứu tập trung vào việc áp dụng các kỹ thuật khoa học dữ liệu trên bộ dữ liệu khách hàng thương mại điện tử.

### Công cụ và kỹ thuật



Python



PowerBI



Flourish



React & TypeScript

## 2. Công trình liên quan

Bảng I: Tóm tắt các công trình liên quan

Năm	Công trình	Tác giả	Phương pháp
2024	Dự đoán churn trong viễn thông [3]	S. K. Wagh <i>et al.</i>	Logistic Regression, RF, ML tổng hợp
2022	Churn trong TMĐT [8]	S. Baghla	Feature Engineering + ML
2016	XGBoost [6]	T. Chen, C. Guestrin	Tree Boosting tối ưu
2002	Xử lý mất cân bằng dữ liệu [7]	N. V. Chawla <i>et al.</i>	SMOTE
2001	Random Forest [4]	L. Breiman	Random Forest (ensemble cây quyết định)
2001	Gradient Boosting Machine [5]	J. H. Friedman	Gradient Boosting

## 2. Công trình liên quan

### Sự dịch chuyển mô hình (Model Evolution)

- **Cũ:** Các mô hình thống kê (**Logistic Regression**)  
=> Hạn chế với dữ liệu phi tuyến.
- **Mới:** Ensemble Learning (**Random Forest**, **XGBoost**)
- **Lý do:** Tối ưu hiệu suất, giảm Overfitting.

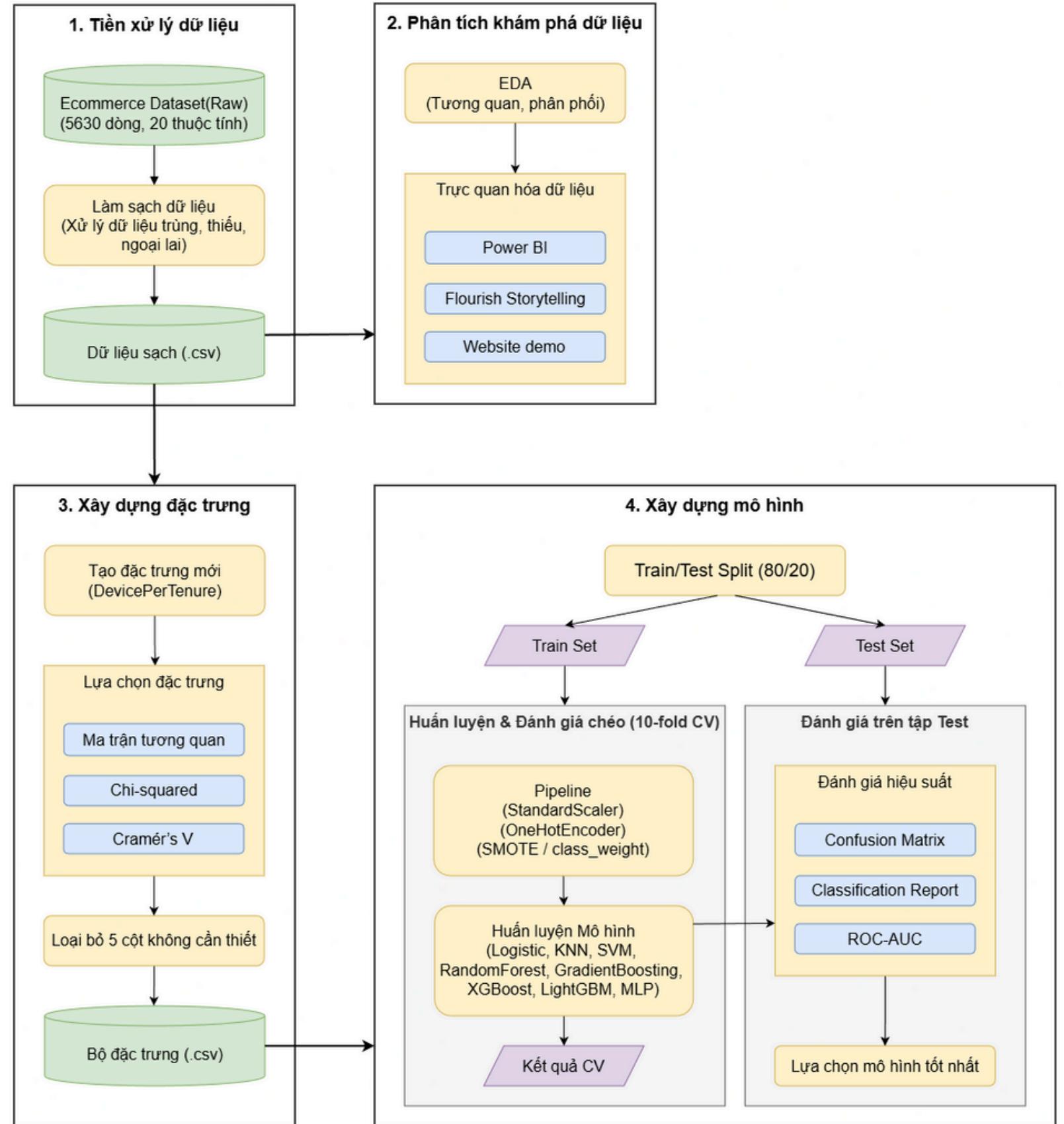
### Xử lý dữ liệu (Data Strategy)

- **Thách thức:** Mất cân bằng lớp (Imbalanced Class).
- **Giải pháp:** Kỹ thuật sinh mẫu **SMOTE & Class Weight**

### Điểm mới (Novelty)

- Kết hợp **Feature Engineering**.
- Tích hợp **Dashboard** tương tác & **Website Demo**.

# 3. Phương pháp nghiên cứu



*Sơ đồ qui trình toàn hệ thống*

# Dữ liệu

- **Tên bộ dữ liệu:** Ecommerce Customer Churn Analysis and Prediction
- **Nguồn:** Được thu thập từ Kaggle [9].
- **Quy mô:**
  - 5630 khách hàng (bản ghi).
  - 20 thuộc tính mô tả (hành vi mua sắm, tương tác,...).
- **Biến mục tiêu:** Churn
  - **Ý nghĩa:** Dự đoán khách hàng có rời bỏ hay không.
    - Giá trị 1: Khách hàng rời bỏ.
    - Giá trị 0: Khách hàng ở lại.

# Dữ liệu

STT	Biến	Mô tả	Giá trị	Kiểu dữ liệu
1	CustomerID	Mã định danh duy nhất của khách hàng		Định tính
2	<b>Churn*</b>	Cờ đánh dấu khách hàng rời bỏ	1: Rời bỏ, 0: Ở lại	Định tính
3	Tenure	Thời gian gắn bó của khách hàng với tổ chức (tháng)	[0, 61]	Định lượng
4	PreferredLoginDevice	Thiết bị đăng nhập ưa thích của khách hàng	{Computer, Mobile Phone}	Định tính
5	CityTier	Cấp độ thành phố của khách hàng (cấp 1, 2, 3)	{1, 2, 3}	Định tính
6	WarehouseToHome	Khoảng cách từ kho hàng đến nhà của khách hàng (km)	[5, 127]	Định lượng
7	PreferredPaymentMode	Phương thức thanh toán ưa thích của khách hàng	{Debit Card, Credit Card, UPI, ...}	Định tính
8	Gender	Giới tính của khách hàng	{Male, Female}	Định tính
9	HourSpendOnApp	Số giờ khách hàng sử dụng ứng dụng di động/website	{0, 1, 2, 3, 4, 5}	Định lượng
10	NumberOfDeviceRegistered	Tổng số thiết bị đã được đăng ký bởi khách hàng	{1, 2, 3, 4, 5, 6}	Định lượng

\*Biến mục tiêu

# Dữ liệu

STT	Biến	Mô tả	Giá trị	Kiểu dữ liệu
11	PreferredOrderCat	Ngành hàng ưa thích của khách hàng trong tháng cuối	{Mobile Phone, Fashion, Grocery, ...}	Định tính
12	SatisfactionScore	Điểm hài lòng của khách hàng về dịch vụ (1-5)	{1, 2, 3, 4, 5}	Định tính
13	MaritalStatus	Tình trạng hôn nhân của khách hàng	{Single, Married, Divorced}	Định tính
14	NumberOfAddress	Tổng số địa chỉ mà khách hàng đã thêm	[1, 22]	Định lượng
15	Complain	Khách hàng có khiếu nại trong tháng cuối	(1: Có, 0: Không)	Định tính
16	OrderAmountHikeFromlastYear	Tỷ lệ % gia tăng giá trị đơn hàng so với năm trước	[11, 26]	Định lượng
17	CouponUsed	Tổng số coupon đã sử dụng trong tháng cuối	[0, 16]	Định lượng
18	OrderCount	Tổng số đơn hàng đã đặt trong tháng cuối	[1, 16]	Định lượng
19	DaySinceLastOrder	Số ngày kể từ lần đặt hàng cuối cùng	[0, 46]	Định lượng
20	CashbackAmount	Lượng cashback trung bình trong tháng cuối	[0, 324.99]	Định lượng

# Dữ liệu

- **Phân bố dữ liệu ban đầu cho thấy sự mất cân bằng rõ rệt:**
  - **Ở lại (Churn=0):** 4682 khách hàng, chiếm 83.2%.
  - **Rời bỏ (Churn=1):** 948 khách hàng, chỉ chiếm 16.8%.
- **Định hướng giải quyết:** Mô hình xây dựng được áp dụng các kỹ thuật đặc biệt để xử lý mất cân bằng, tránh dự đoán thiên lệch về nhóm khách hàng ở lại.

# Tiền xử lý dữ liệu

**Mục tiêu:** Đảm bảo dữ liệu nhất quán và loại bỏ các thông tin không cần thiết.

- **Loại bỏ thuộc tính định danh**

- Cột CustomerID được loại bỏ vì không mang giá trị dự đoán

- **Chuẩn hóa dữ liệu không nhất quán**

- **PreferredLoginDevice** và **PreferredOrderCat**: Các giá trị như “Phone”, “Mobile” và “Mobile Phone” đều chỉ cùng một loại thiết bị/ngành hàng. Chúng được chuẩn hóa và gộp chung thành giá trị “Mobile Phone”.
- **PreferredPaymentMode**: Các cách viết tắt như “CC” và “COD” được ánh xạ về dạng đầy đủ là “Credit Card” và “Cash on Delivery” để đảm bảo tính đồng nhất.

# Tiền xử lý dữ liệu

Áp dụng các phương pháp điền dữ liệu phù hợp với đặc điểm của từng thuộc tính để **xử lý giá trị thiếu:**

- **Tenure:** xử lý dựa trên mức độ hoạt động của khách hàng. Nếu khách hàng chưa có hoạt động (OrderCount, HourSpendOnApp bằng 0 hoặc thiếu), Tenure được điền bằng 0. Nếu khách hàng đã có hoạt động, Tenure được điền bằng 1, phản ánh đây là khách hàng mới nhưng đã bắt đầu tương tác.
- **WarehouseToHome, HourSpendOnApp, và OrderAmountHikeFromlastYear:** Do phân phối của các biến này tương đối tập trung và không quá lệch, các giá trị thiếu được điền bằng giá trị trung vị (median).

# Tiền xử lý dữ liệu

- **CouponUsed** và **OrderCount**: Hai biến này có mối quan hệ logic chặt chẽ ( $\text{CouponUsed} \leq \text{OrderCount}$ ). Nếu cả hai giá trị đều thiếu, chúng được điền bằng 0. Nếu chỉ một trong hai bị thiếu, giá trị của nó sẽ được điền dựa trên giá trị của cột còn lại và phân phối chênh lệch giữa chúng.
- **DaySinceLastOrder**: Dựa trên phân tích từ biến OrderCount (cho thấy tất cả khách hàng đều có đơn hàng trong tháng cuối), các giá trị của DaySinceLastOrder không hợp lệ ( $\geq 30$ ) được xác định và chuyển thành giá trị thiếu, sau đó được điền bằng median của cột.

# Tiền xử lý dữ liệu

## Xử lý giá trị ngoại lai (Outliers)

- **Phương pháp:** Sử dụng IQR (Interquartile Range) để xác định ngưỡng.
- **Kỹ thuật:** Áp dụng Capping/Winsorizing (thay thế giá trị ngoại lai bằng giá trị ngưỡng).
  - Xử lý ở cột **NumberOfAddress**.

# Phân tích khám phá dữ liệu

## Mục tiêu EDA:

- Hiểu phân bố và đặc điểm các biến trong dữ liệu.
- Khám phá hành vi và đặc trưng khách hàng.
- Xác định các yếu tố liên quan đến hành vi rời bỏ (Churn).

## Phương pháp:

- Histogram, KDE (Kernel Density Estimation).
- Bar chart.
- Boxplot theo Churn.

# Phân tích khám phá dữ liệu

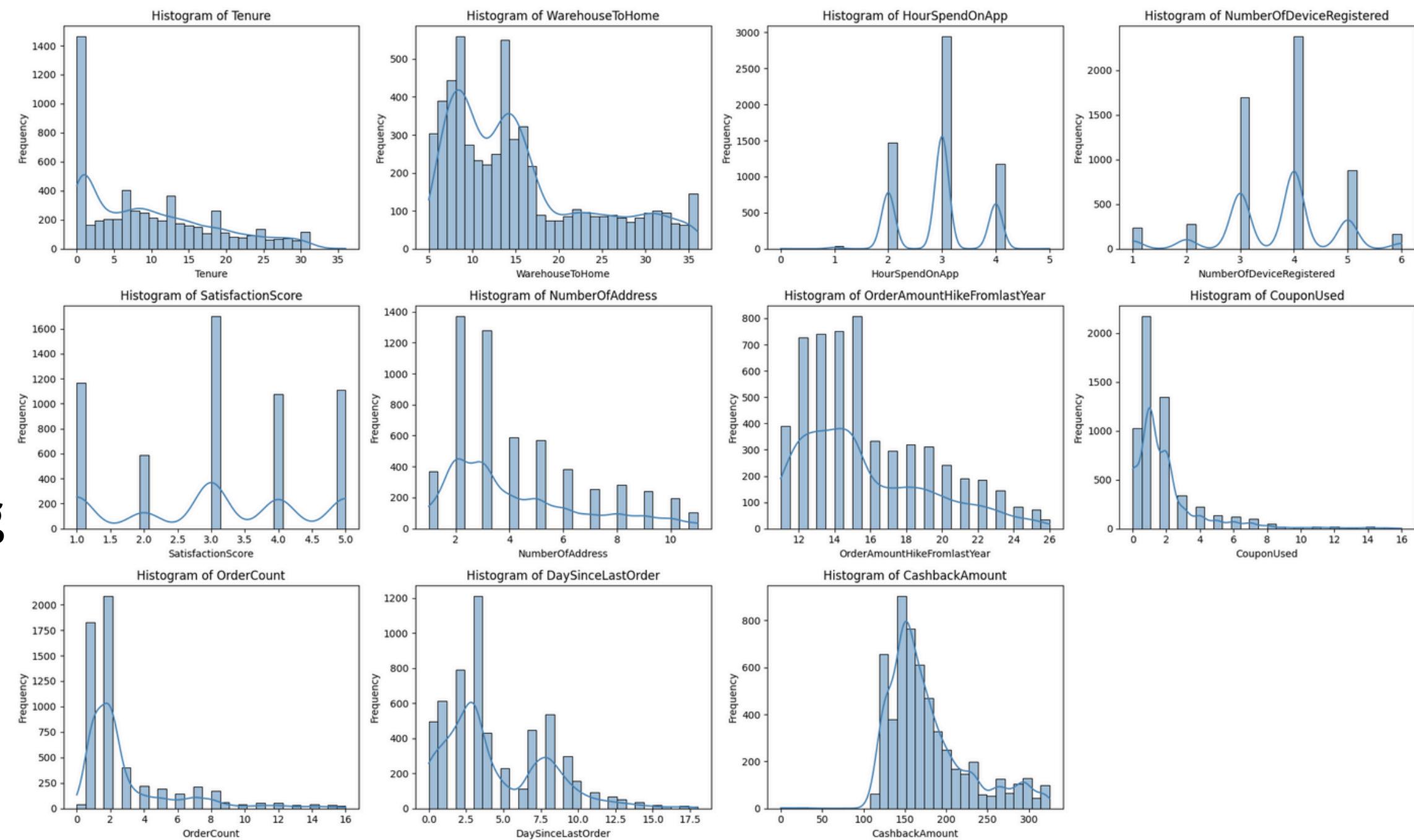
## Phân bố các biến định lượng:

Phần lớn biến không phân phối chuẩn.

Chủ yếu lệch phải.

## Đặc trưng dữ liệu hành vi:

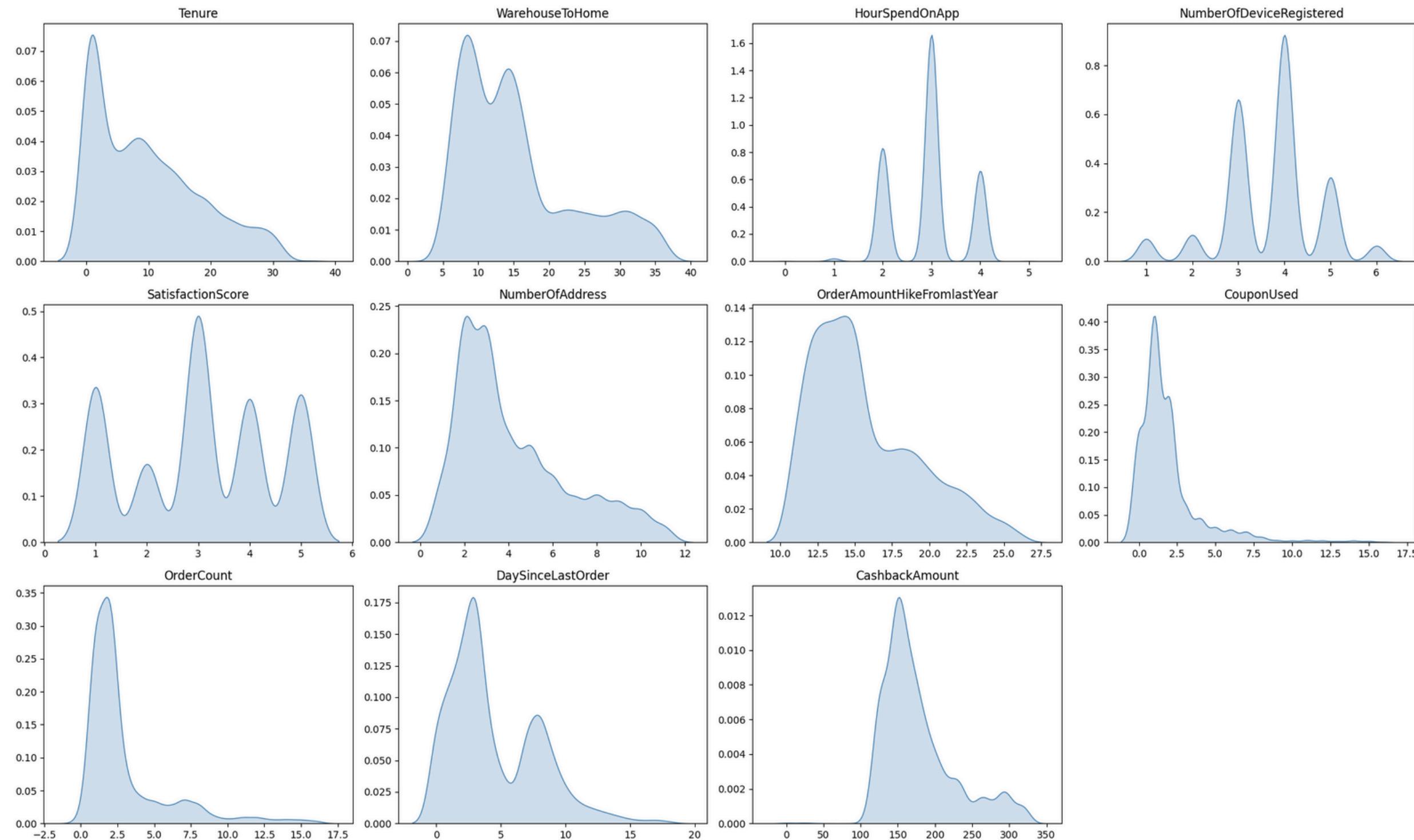
- Ít khách hàng giá trị rất cao
- Đa số tập trung ở mức thấp – trung bình
- Xuất hiện outlier



# Phân tích khám phá dữ liệu

## Phân bố mật độ các biến:

- Dữ liệu phi tuyến.
- Nhiều đỉnh cho nên còn tồn tại nhiều nhóm hành vi.
- Phù hợp cho các mô hình phi tuyến.



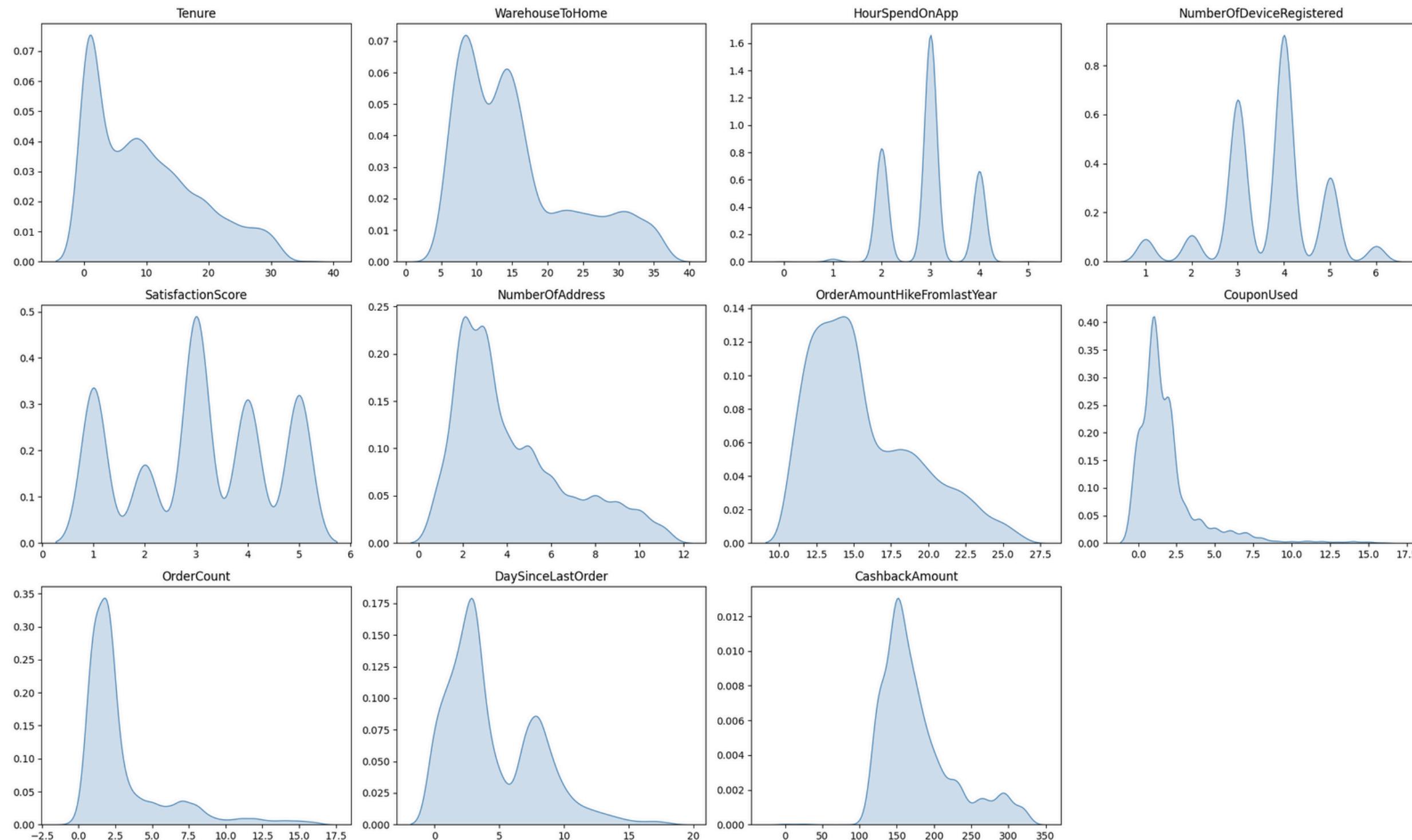
# Phân tích khám phá dữ liệu

## Phân bố mật độ các biến:

### Tenure

- Tập trung ở giá trị thấp.
- Ít khách hàng gắn bó lâu dài.

→ Có liên quan đến khả năng Churn.



### WarehouseToHome

- Phân bố rộng, đuôi dài.
- Trải nghiệm giao hàng không đồng đều.

# Phân tích khám phá dữ liệu

**Phân bố mật độ các biến:**

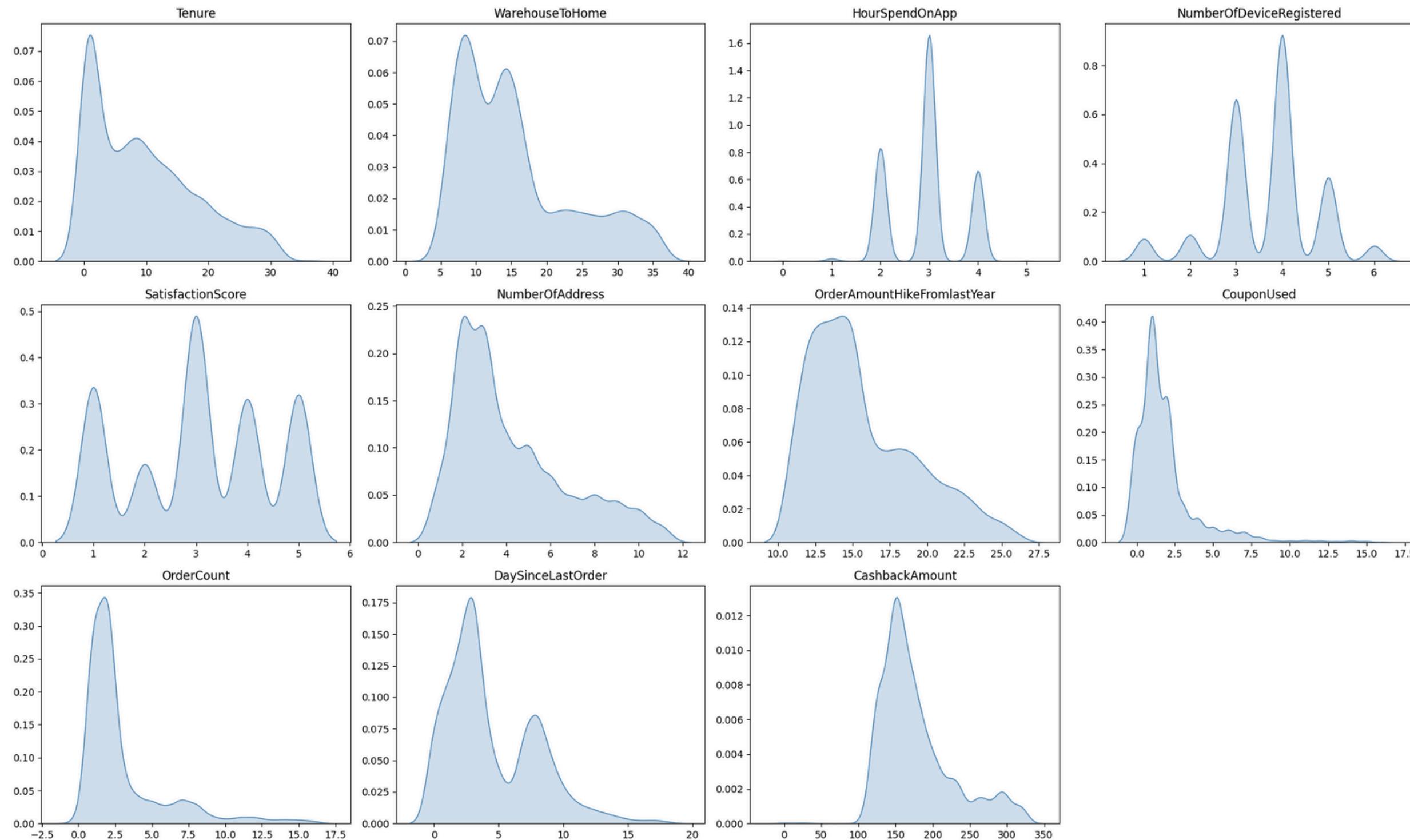
**HourSpendOnApp &**

**NumberOfDeviceRegistered**

- Phân bố rời rạc.
- Thể hiện nhiều nhóm hành vi.

**SatisfactionScore**

- Phân bố theo mức rời rạc.
- Phản ánh sự khác biệt rõ ràng về trải nghiệm.



# Phân tích khám phá dữ liệu

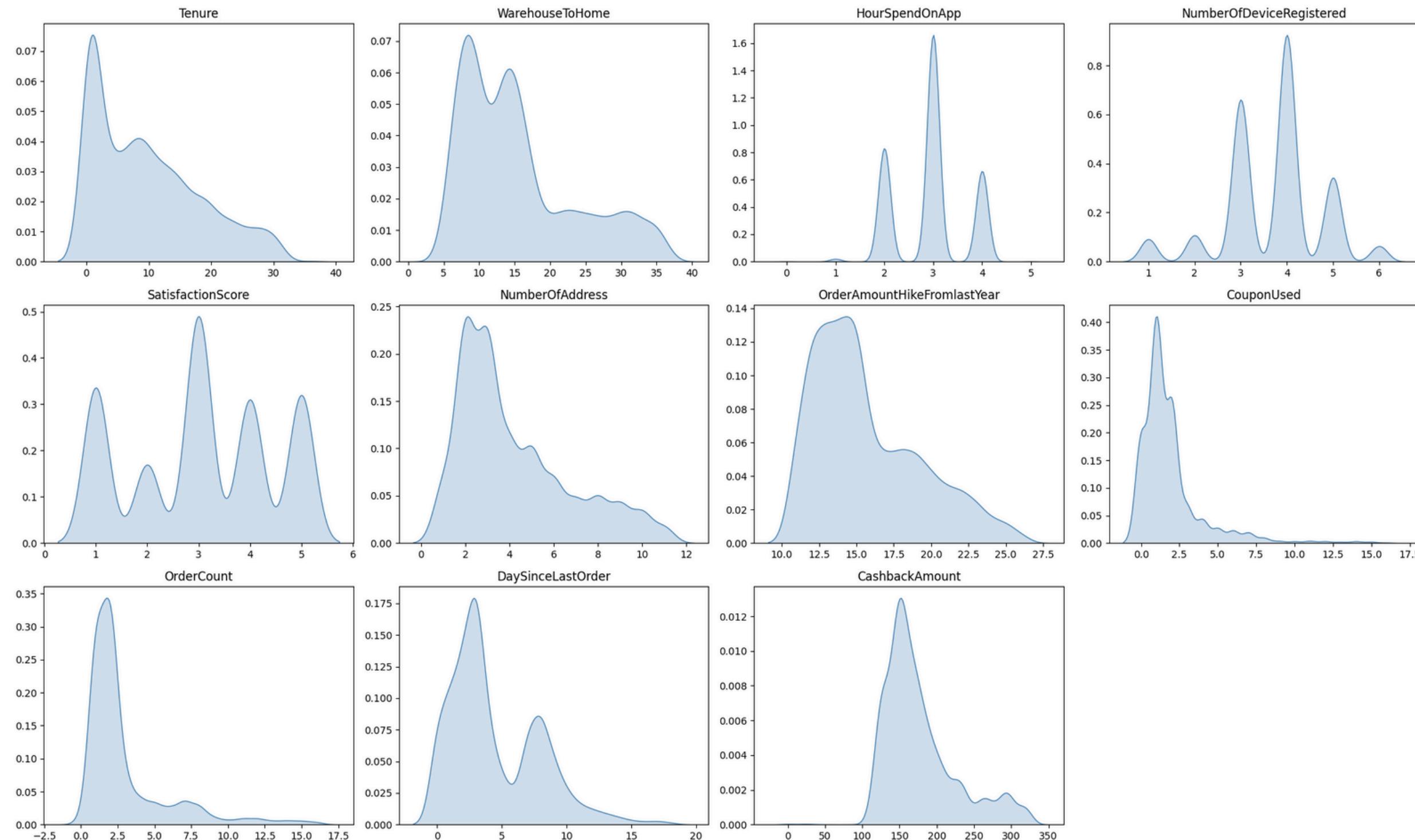
## Phân bố mật độ các biến:

**OrderCount, CouponUsed,  
DaySinceLastOrder**

- Lệch phải mạnh.
- Phần lớn khách hàng tương tác thấp.

## CashbackAmount

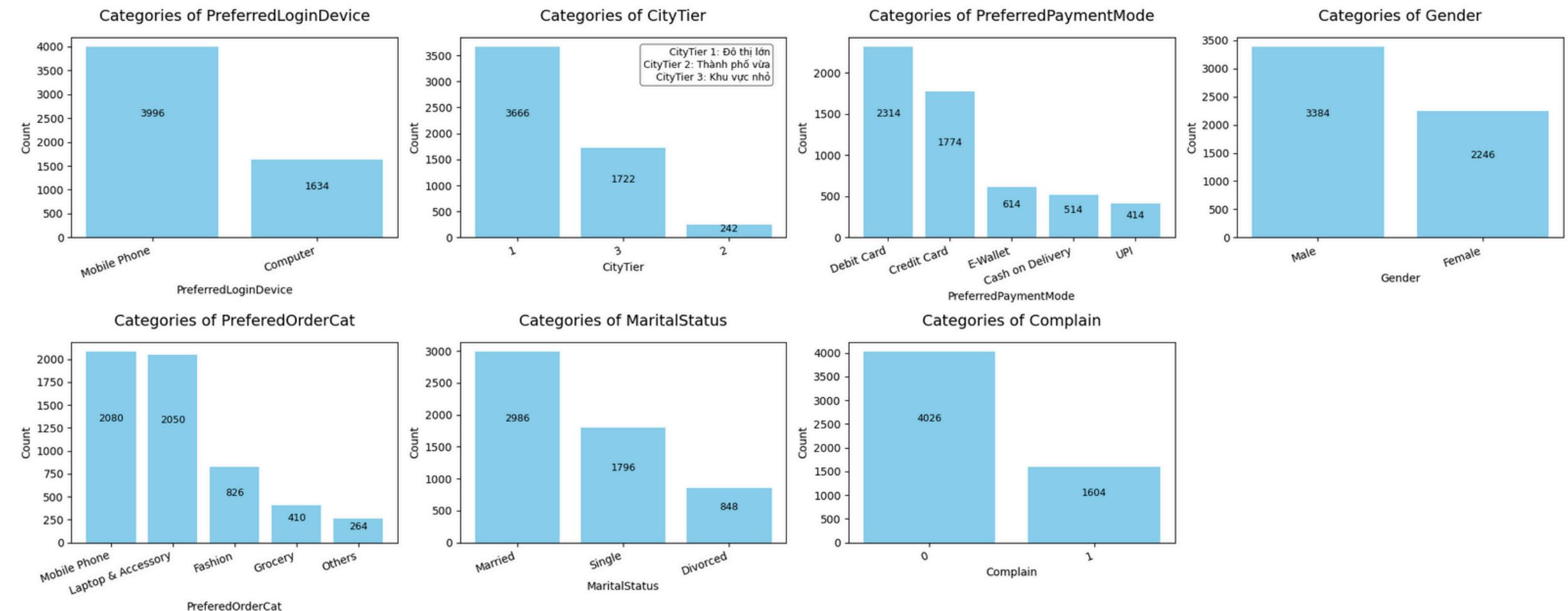
- Ổn định hơn nhưng vẫn có đuôi dài.
- Mức ưu đãi không đồng đều.



# Phân tích khám phá dữ liệu

## Phân tích biến định tính:

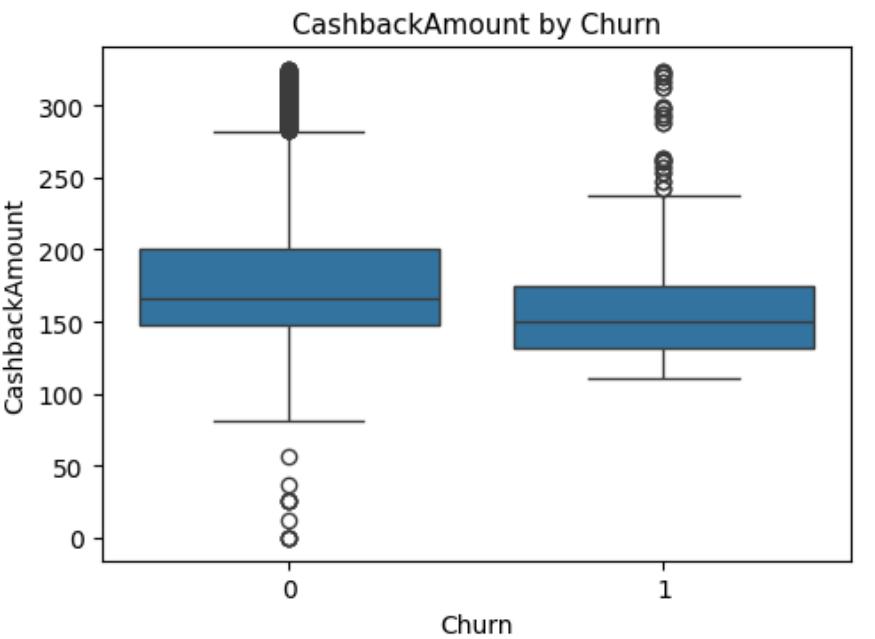
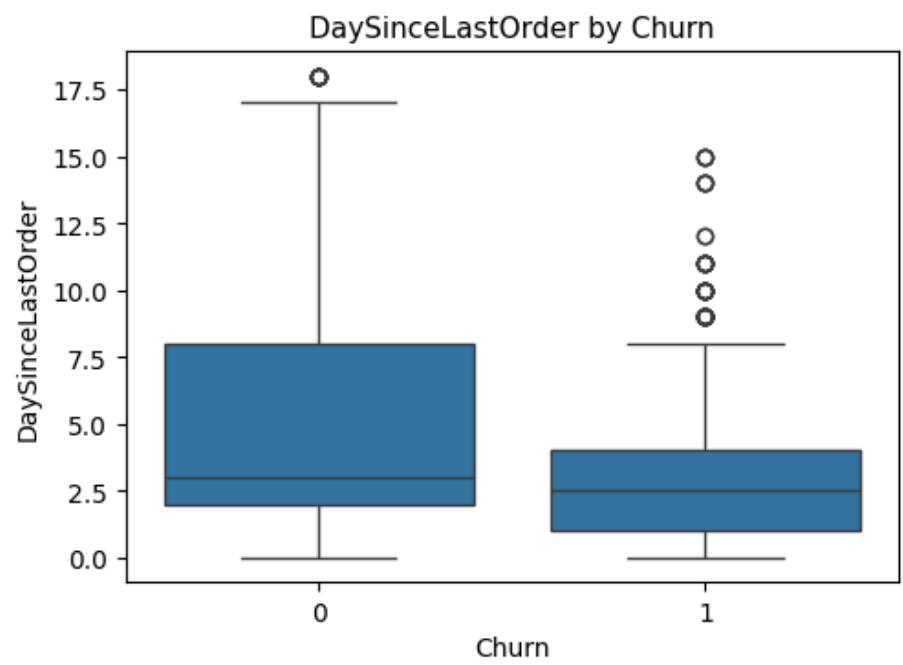
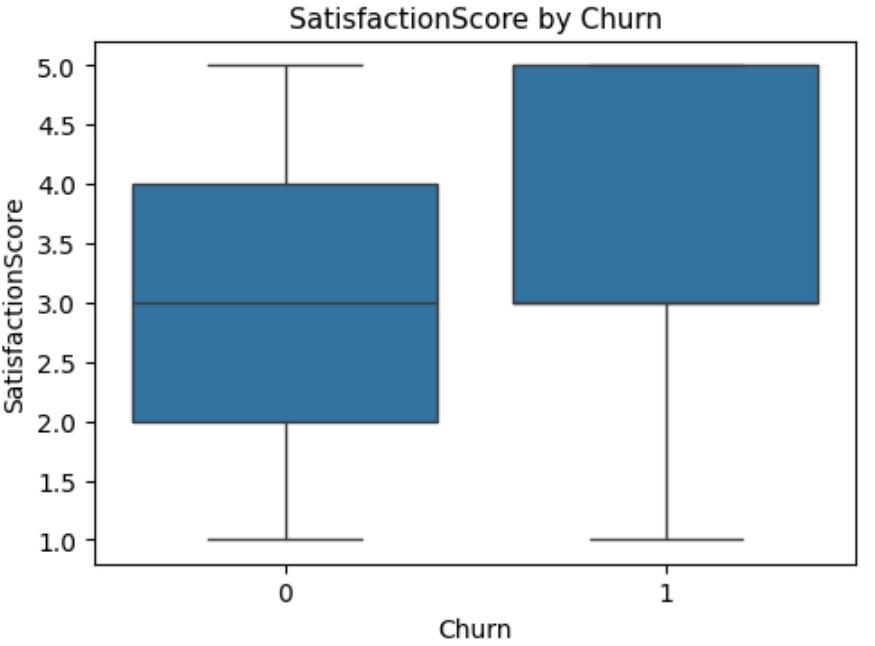
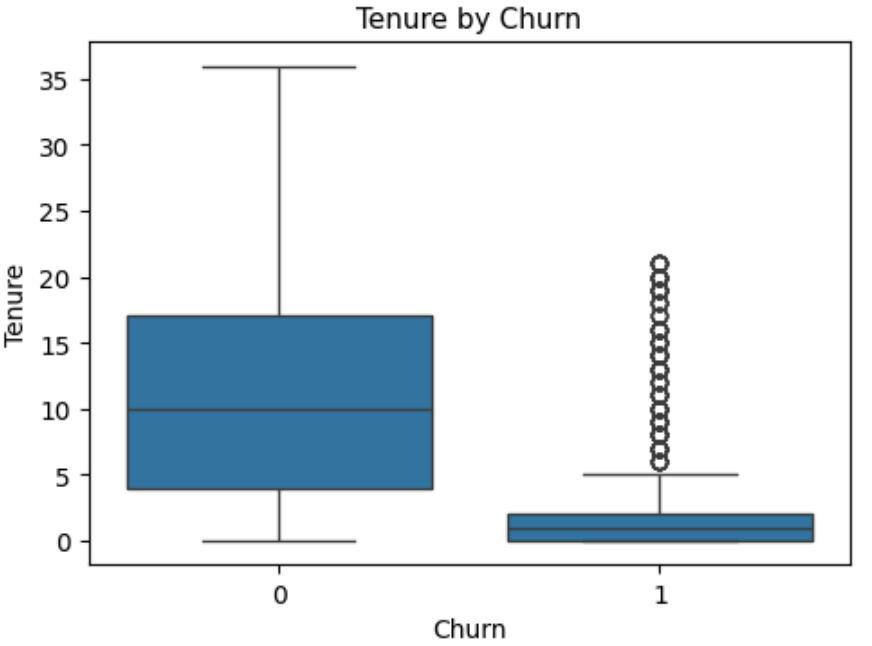
- PreferredLoginDevice: Mobile Phone chiếm đa số.
- CityTier: Khách hàng tập trung ở vài nhóm thành phố.
- PreferredPaymentMode: Debit & Credit Card phổ biến.
- PreferredOrderCat: Mobile Phone, Laptop & Accessory chiếm ưu thế.



# Phân tích khám phá dữ liệu

## Mối quan hệ với Churn:

- Tenure: Non-Churn cao hơn rõ rệt.
- SatisfactionScore: Non-Churn hài lòng hơn.
- DaySinceLastOrder: Churn cao hơn.
- CashbackAmount: Non-Churn nhận ưu đãi nhiều hơn.



# Phân tích khám phá dữ liệu

## Tổng hợp insight

**Churn chịu ảnh hưởng chính bởi:**

- Tenure.
- SatisfactionScore.
- DaySinceLastOrder.
- CashbackAmount.

Dữ liệu có tính phi tuyến rõ rệt, đặc biệt ở các biến: *Tenure*, *DaySinceLastOrder*, *OrderCount*, *CouponUsed*, *CashbackAmount*, *HourSpendOnApp*, *SatisfactionScore*.

Các biến này có phân bố lệch, nhiều outlier và mối quan hệ với Churn không tuân theo xu hướng tuyến tính, Churn chủ yếu liên quan đến hành vi sử dụng và trải nghiệm khách hàng, phù hợp với các mô hình học máy phi tuyến.

# Trực quan hóa dữ liệu

## Các công cụ được sử dụng:

- **Power BI:** Xây dựng Dashboard tương tác để phân tích đa chiều.
- **Flourish Studio:** "Kể chuyện" dữ liệu (Data Storytelling) một cách sinh động.
- **Website Demo (React):** Tạo ra một sản phẩm demo linh hoạt, cho phép người dùng cuối tự khám phá dữ liệu.

# Trực quan hóa dữ liệu

Power BI

## Tổng quan về Dashboard:

- Một hệ thống báo cáo toàn diện, được thiết kế để dẫn dắt người xem đi từ tổng quan đến chi tiết.
- Bao gồm 5 trang báo cáo chuyên biệt, cung cấp một cái nhìn đa chiều về bài toán khách hàng rời bỏ
  - Trang 1 - Tổng quan
  - Trang 2 - Phân tích hành vi khách hàng
  - Trang 3 - Phân tích mua sắm & giá trị
  - Trang 4 - Phân tích trải nghiệm khách hàng
  - Trang 5 - So sánh hiệu suất mô hình

# Trực quan hóa dữ liệu

Trang 1

Tổng khách hàng

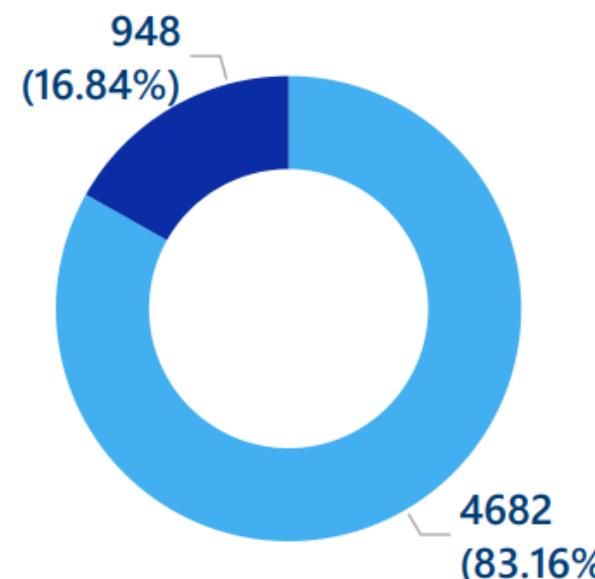
5630

Tỉ lệ Churn

16.8%

Phân bố khách hàng theo Churn

Churn ● 0 ● 1



Tổng khách hàng rời bỏ

948

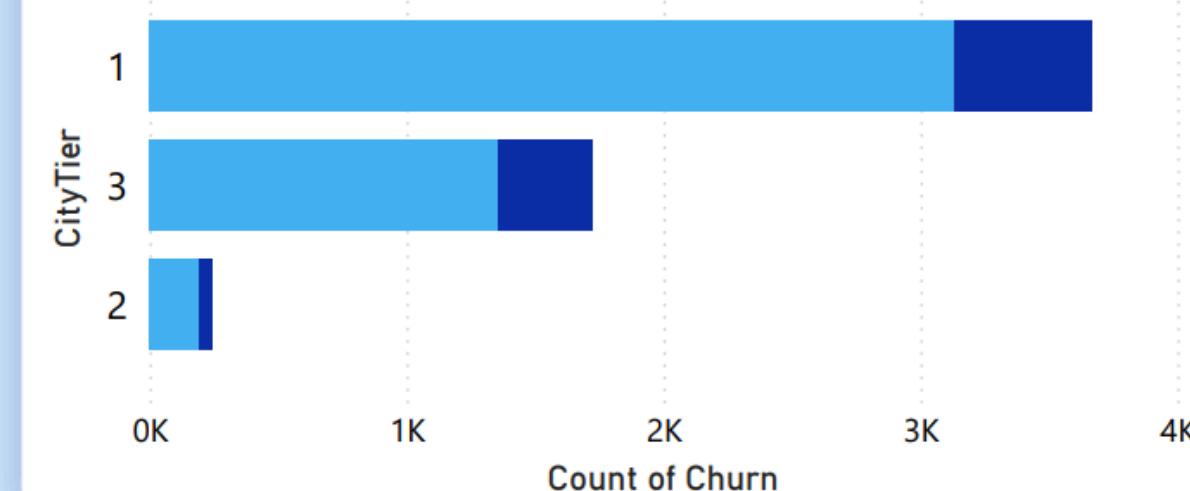
Trung bình thời gian gắn bó

(tháng)

3.18

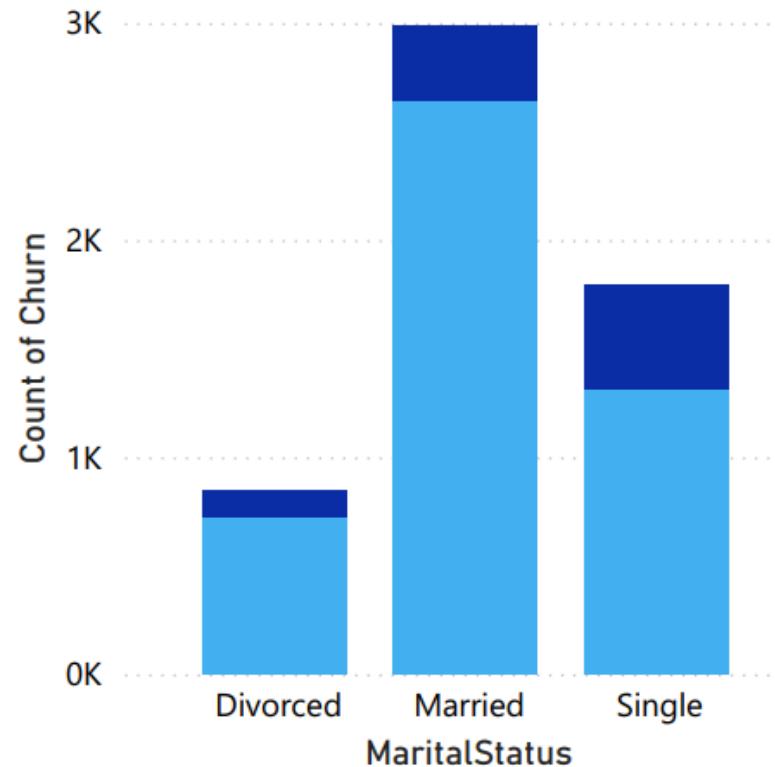
Phân bố Churn theo cấp độ thành phố

Churn ● 0 ● 1

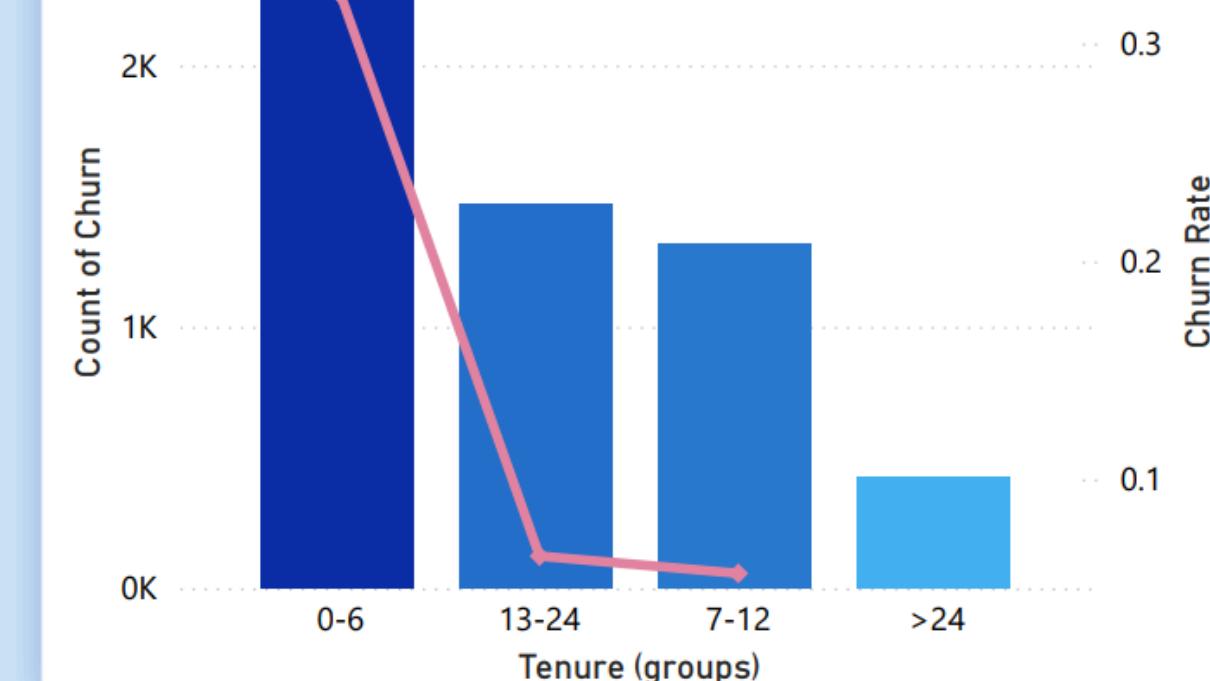


Phân bố Churn theo hôn nhân

3K



Phân bố Churn theo thời gian gắn bó



# Trực quan hóa dữ liệu

Trang 1

## Nhận xét:

### Các chỉ số KPI cốt lõi:

- Tỷ lệ rời bỏ (Churn Rate): 16.8% (948 trên 5630 khách hàng).
- Thời gian gắn bó trung bình: Chỉ 3.18 tháng, cho thấy tính cấp thiết của vấn đề.

### Những phát hiện chính:

- Báo động ở khách hàng mới: Tỷ lệ rời bỏ cao đột biến (trên 30%) ở nhóm có thời gian gắn bó từ 0-6 tháng. Đây là giai đoạn "vàng" nhưng cũng thách thức nhất để giữ chân.

### Phân khúc rủi ro:

- Thành phố cấp 1: Thị trường lớn nhất, nhưng cũng có số lượng khách hàng rời bỏ tuyệt đối cao nhất.
- Tình trạng hôn nhân: Nhóm "Độc thân" có tỷ lệ rời bỏ cao nhất.

# Trực quan hóa dữ liệu

Trang 2



# Trực quan hóa dữ liệu

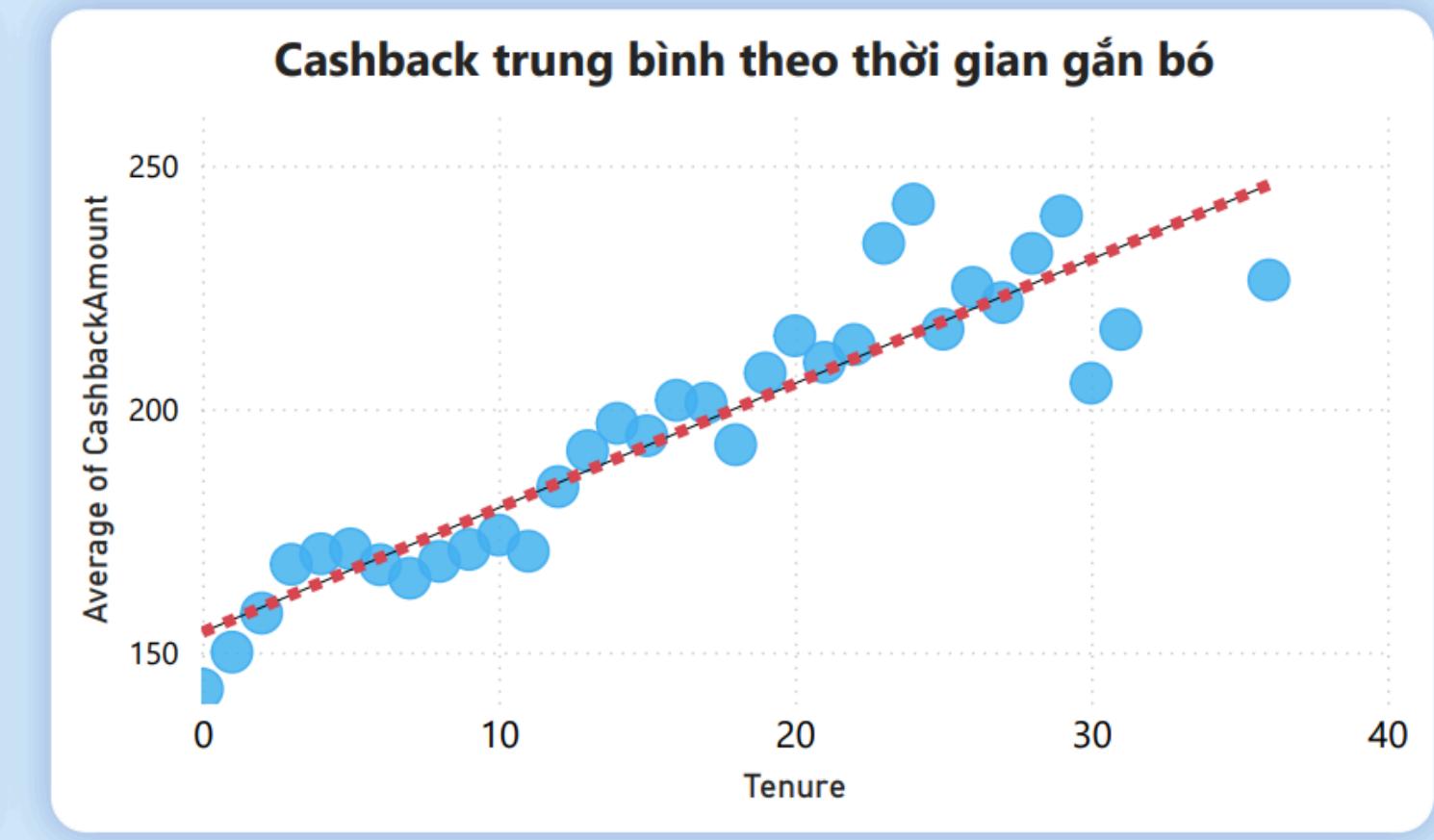
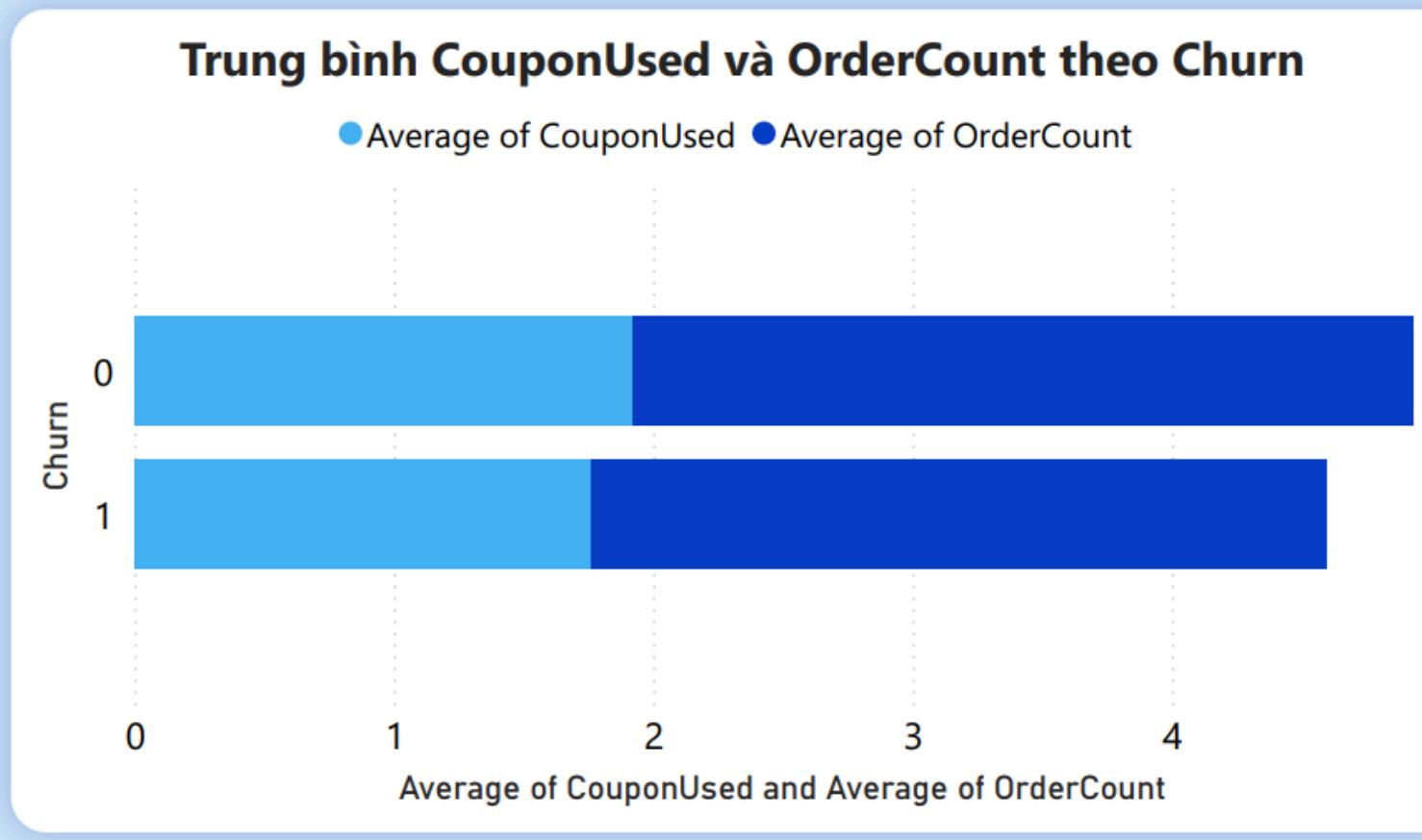
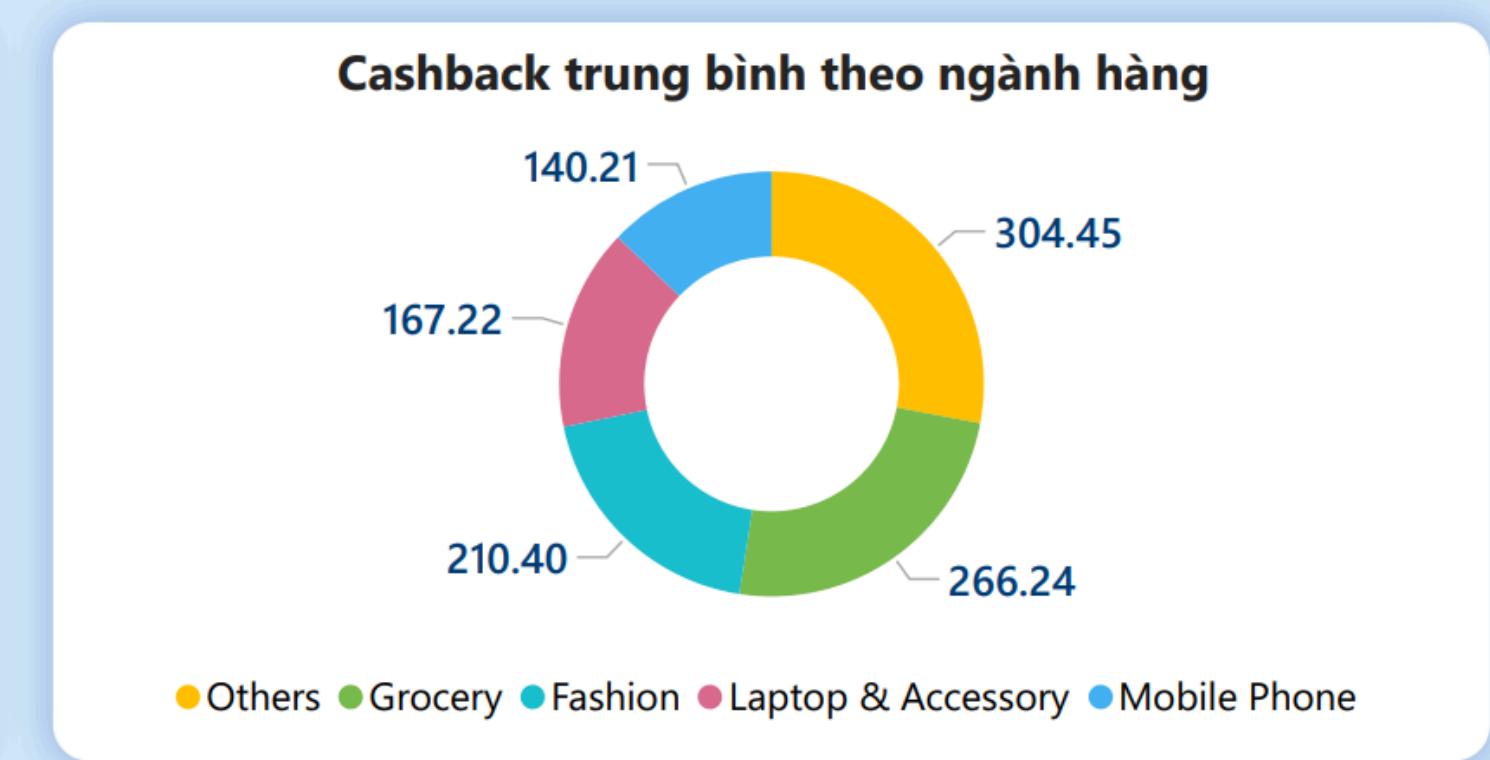
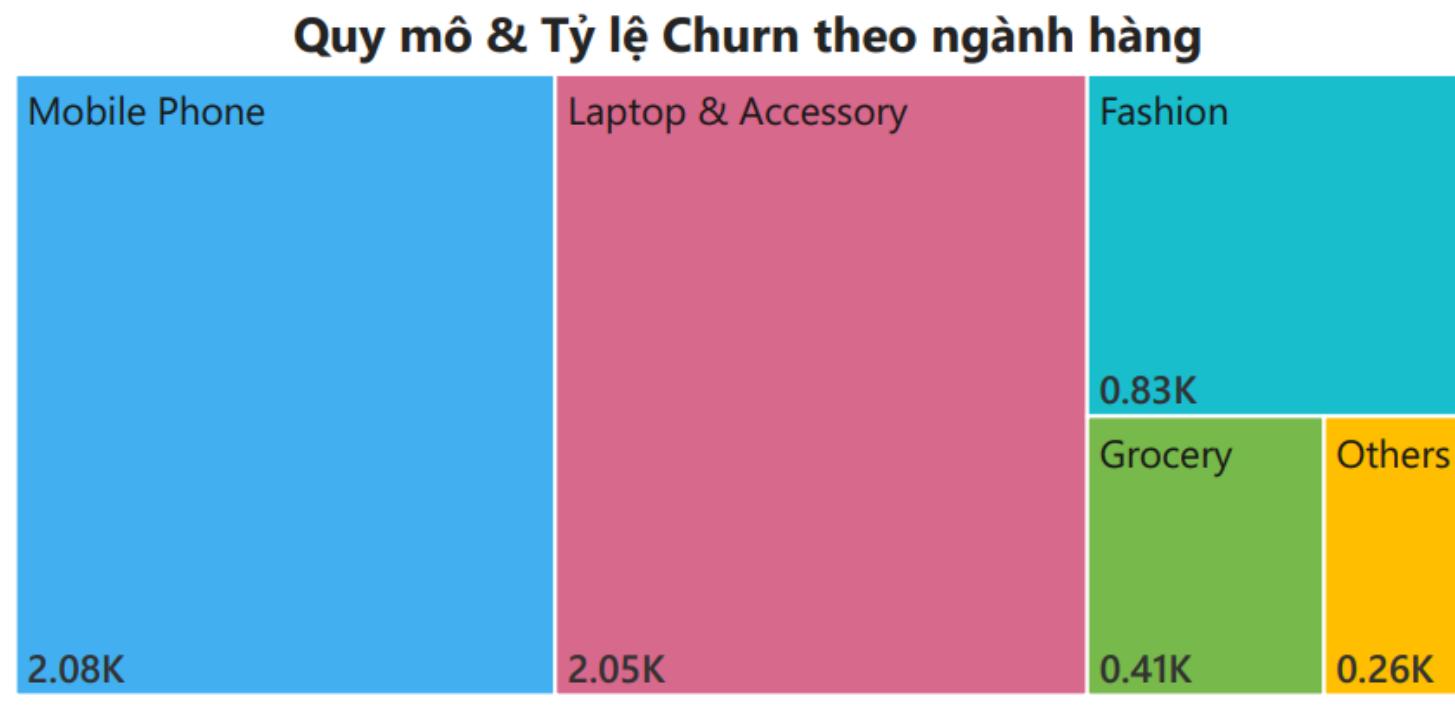
Trang 2

## Nhận xét:

- **Mức độ gắn kết cao:** Phần lớn là người dùng tích cực, dành 3 giờ trên app và đăng ký 3-4 thiết bị.
- **Thanh toán:** Thẻ tín dụng (Credit Card) và Ghi nợ (Debit Card) là phương thức chủ đạo.
- **Sử dụng đa mục đích:** Đỉnh điểm khách hàng có 3 địa chỉ, gợi ý họ dùng dịch vụ cho cả nhà riêng, công ty, hoặc gửi quà.

# Trực quan hóa dữ liệu

Trang 3



# Trực quan hóa dữ liệu

Trang 3

## Nhận xét:

- Hoạt động tháng cuối (**số đơn hàng, coupon**) của nhóm rời bỏ và nhóm ở lại *gần như không khác biệt* => Quyết định rời bỏ đến từ trải nghiệm lâu dài, không phải sự sụt giảm hoạt động tức thời.
- **Chính sách khách hàng trung thành hiệu quả:** Thời gian gắn bó càng dài, lượng cashback nhận được càng cao.
- **Phân hóa Cashback:** Ngành hàng "Others" và "Grocery" có cashback cao nhất, trong khi "Mobile Phone" thấp nhất.

# Trực quan hóa dữ liệu

Trang 4

Điểm hài lòng trung bình

3.07

Mức tăng giá trung bình (%)

15.67

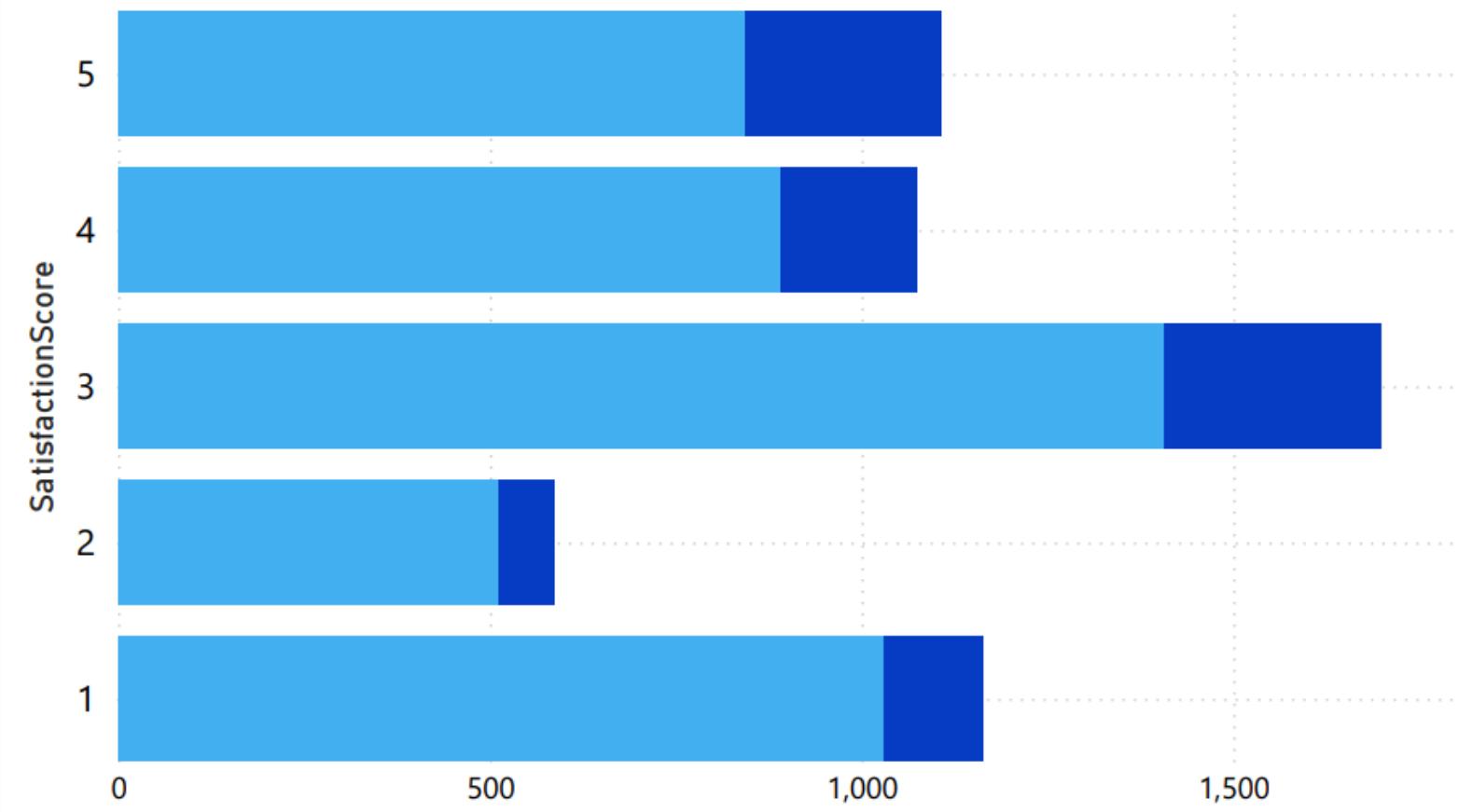
Phân bố khiếu nại khách hàng

Complain ● 0 ● 1



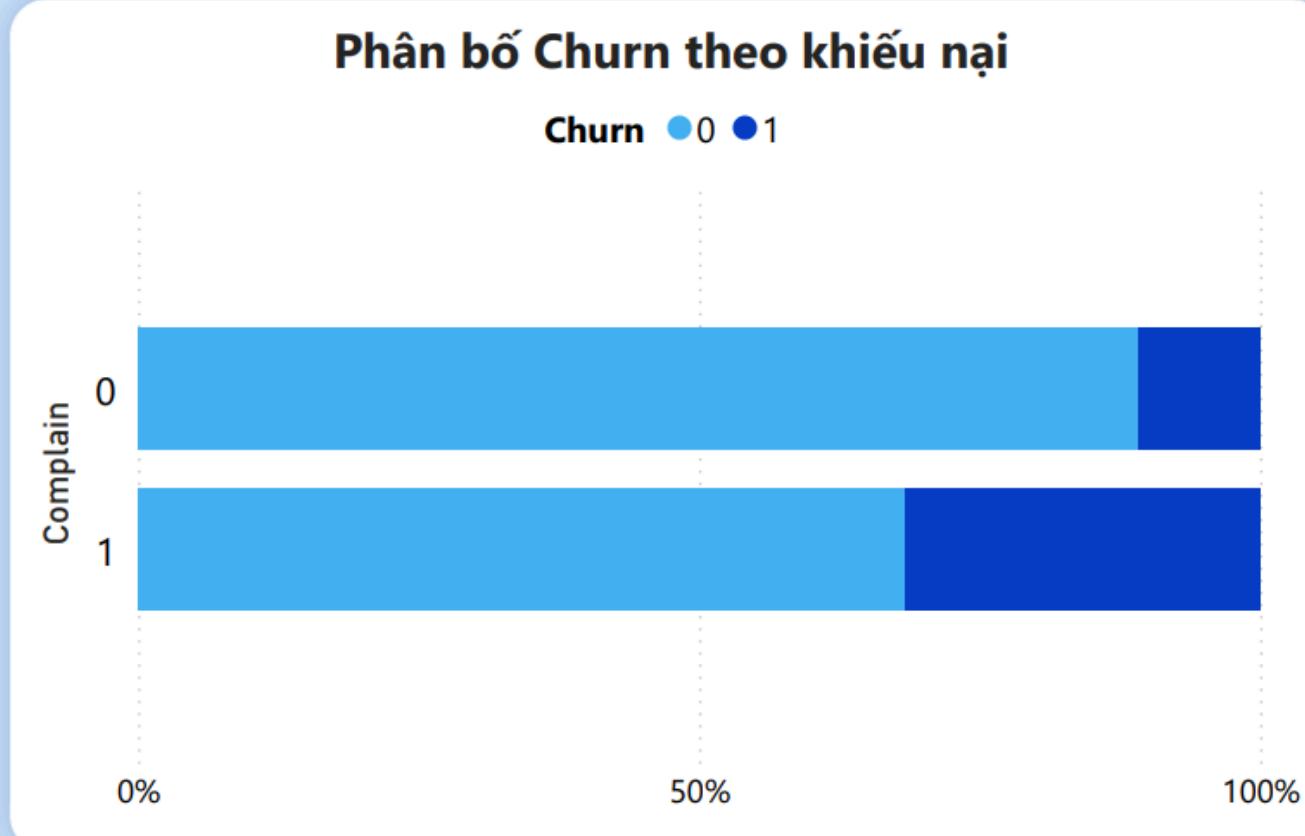
Phân bố Churn theo độ hài lòng

Churn ● 0 ● 1



Phân bố Churn theo khiếu nại

Churn ● 0 ● 1



# Trực quan hóa dữ liệu

Trang 4

## Nhận xét:

- **Điểm hài lòng (Satisfaction Score):**

- Mỗi quan hệ ngược chiều rõ ràng: Hài lòng càng thấp, tỷ lệ rời bỏ càng cao.
- Điểm hài lòng trung bình chỉ 3.07/5, cho thấy vấn đề tiềm ẩn về chất lượng dịch vụ.

- **Lịch sử khiếu nại (Complain):**

- Đây là "tín hiệu cảnh báo đỏ" mạnh mẽ nhất. Tỷ lệ rời bỏ ở nhóm có khiếu nại cao hơn đáng kể so với nhóm không khiếu nại.

# Trực quan hóa dữ liệu

Trang 5

Recommend Model

XGBoost

Accuracy

0.99

F1 (C1)

0.96

Recall (C1)

0.95

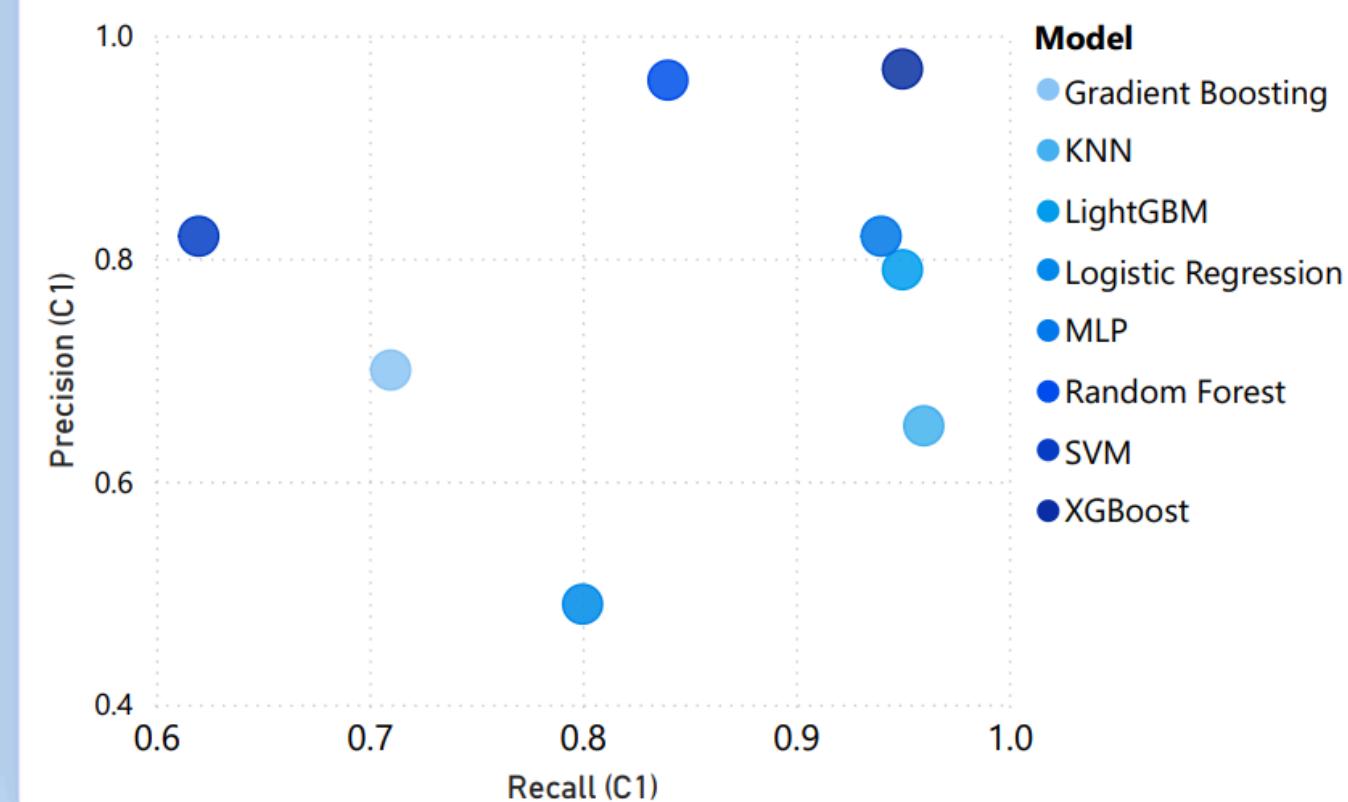
AUC

0.99528

Hiệu suất của 8 model

Model	Accuracy	ROC-AUC	Macro F1	Precision (C1)	Recall (C1)
Logistic Regression	0.83	0.88571	0.75	0.49	0.80
Gradient Boosting	0.90	0.91902	0.82	0.70	0.71
KNN	0.91	0.97367	0.86	0.65	0.96
SVM	0.91	0.92792	0.83	0.82	0.62
LightGBM	0.95	0.98507	0.91	0.79	0.95
MLP	0.95	0.98513	0.92	0.82	0.94
Random Forest	0.97	0.99549	0.94	0.96	0.84
XGBoost	0.99	0.99528	0.98	0.97	0.95

Phân tích đánh đổi Precision-Recall



XGBoost là mô hình được đề xuất với hiệu suất vượt trội và cân bằng nhất giữa việc phát hiện (Recall) và độ tin cậy (Precision).

# Trực quan hóa dữ liệu

## KỂ CHUYỆN VỚI FLOURISH

Câu chuyện được xây dựng bao gồm 3 phần chính:

- **Lượng hóa tác động:** Dùng Parliament Chart để cụ thể hóa con số 16.8% Churn thành 948 khách hàng thực tế, nhấn mạnh quy mô vấn đề.
- **So sánh sự khác biệt:** Dùng 100% Stacked Bar Chart để làm nổi bật sự chênh lệch "khổng lồ" về tỷ lệ rời bỏ giữa nhóm có và không có khiếu nại.
- **Truy vết "dòng chảy" rời bỏ:** Dùng Sankey Diagram để trực quan hóa "lộ trình rời bỏ" phổ biến nhất của khách hàng (ví dụ: từ ngành hàng Fashion -> dùng thiết bị Mobile -> Churned).

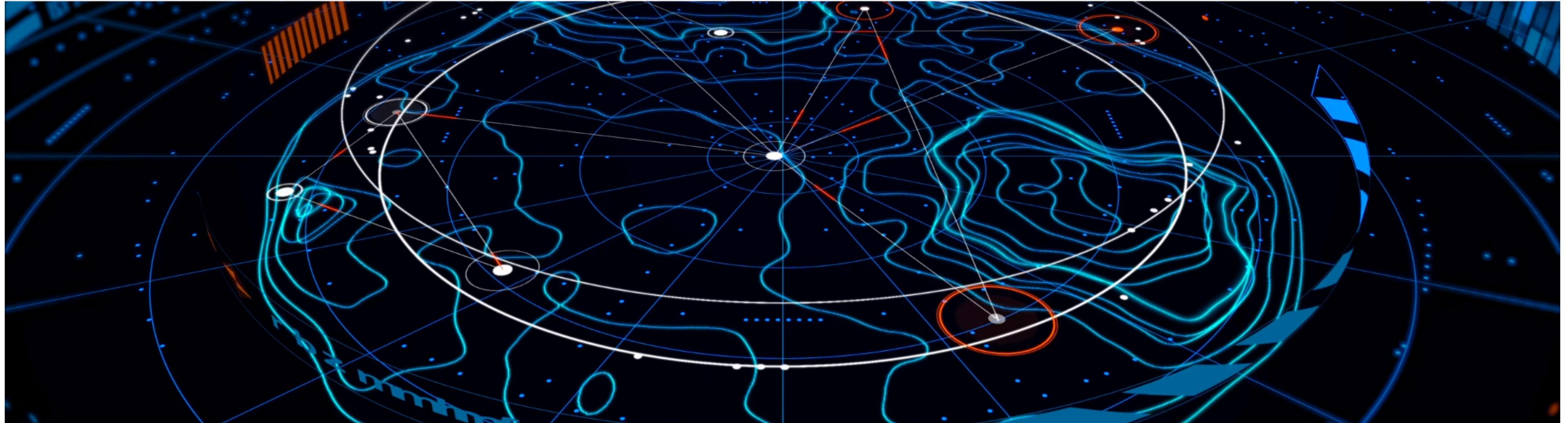
# Trực quan hóa dữ liệu

Slide 1



1 of 5

Mỗi khách hàng rời bỏ là một câu chuyện không được kể và một cơ hội tăng trưởng đã mất. Trong phân tích này, chúng ta sẽ không chỉ nhìn vào con số, mà sẽ đi sâu vào việc "bóc tách" dữ liệu để tìm hiểu chân dung, hành vi và những "điểm nóng" đã dẫn đến quyết định rời bỏ của họ.



Giải mã "Sự rời bỏ thầm lặng" của khách hàng

- Lượng hóa tác động của Churn
- Xác định các yếu tố ảnh hưởng mạnh mẽ nhất
- Đề xuất các hành động dựa trên dữ liệu để cải thiện tỷ lệ giữ chân

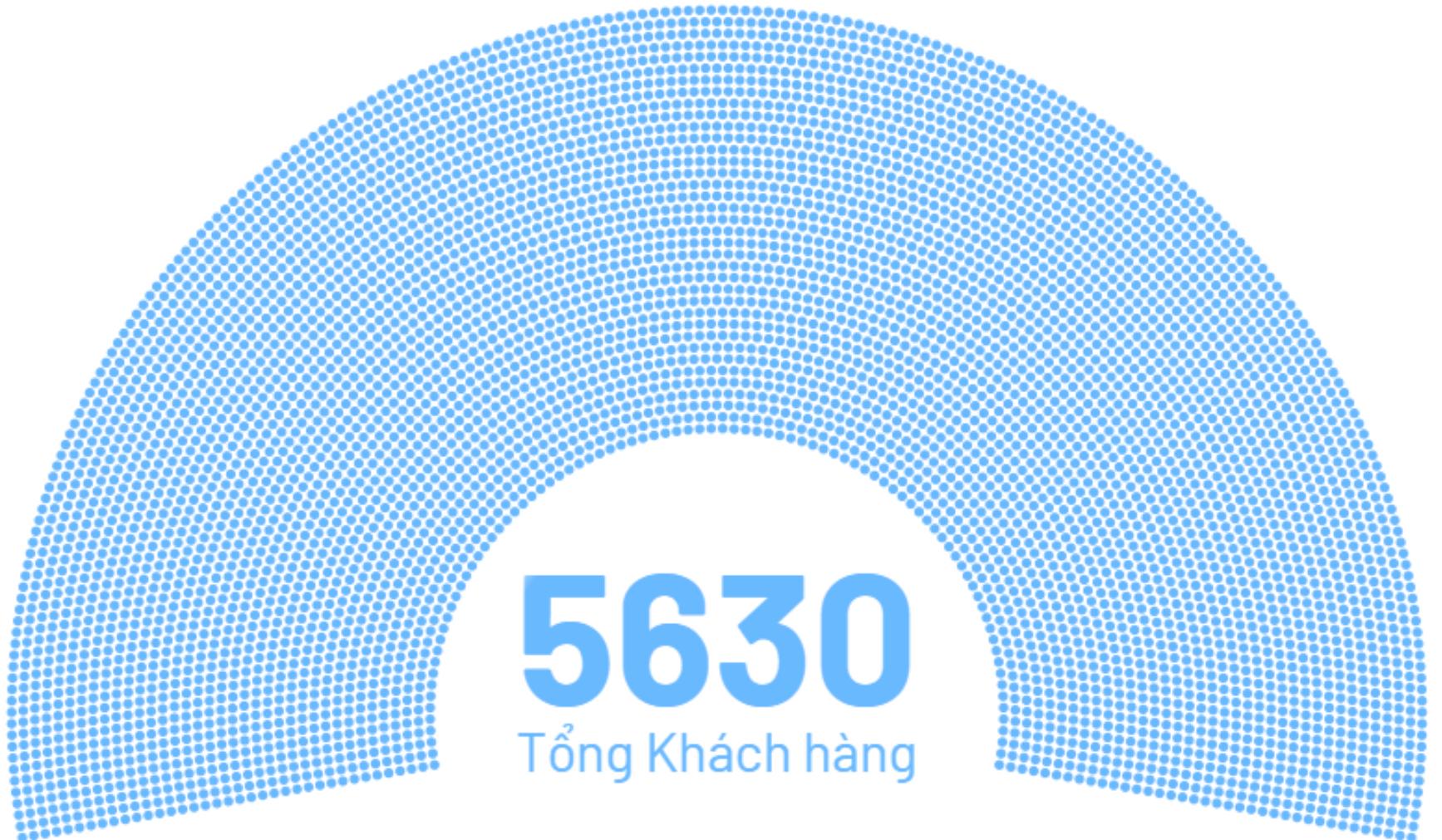
# Trực quan hóa dữ liệu

Slide 2

## Mức độ khách hàng rời bỏ trên thực tế

So sánh trực quan giữa quy mô khách hàng ban đầu và trạng thái giữ chân thực tế

Tổng quy mô    Hiện trạng



### Trạng thái

- Khách hàng ở lại (Retained)
- Khách hàng rời bỏ (Churned)
- Tổng Khách hàng

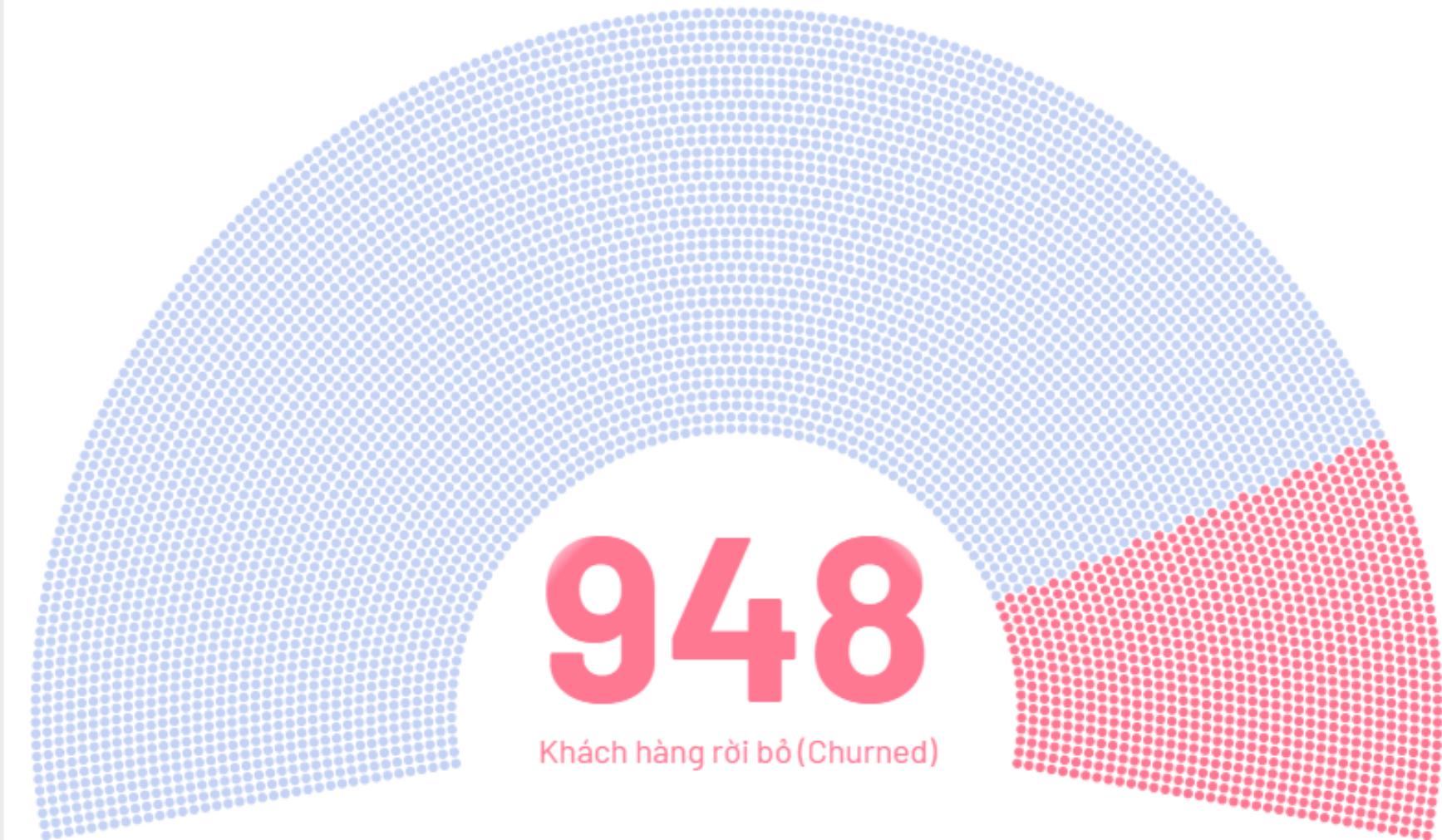
Source: Data

	Tổng quy mô	Hiện trạng	Seat change
Khách hàng ở lại (Retained)	0	4682	↑ 4682
Khách hàng rời bỏ (Churned)	0	948	↑ 948
Tổng Khách hàng	5630		↓ 5630

## Mức độ khách hàng rời bỏ trên thực tế

So sánh trực quan giữa quy mô khách hàng ban đầu và trạng thái giữ chân thực tế

Tổng quy mô    Hiện trạng



### Trạng thái

- Khách hàng ở lại (Retained)
- Khách hàng rời bỏ (Churned)
- Tổng Khách hàng

Source: Data

	Tổng quy mô	Hiện trạng	Seat change
Khách hàng ở lại (Retained)	0	4682	↑ 4682
Khách hàng rời bỏ (Churned)	0	948	↑ 948
Tổng Khách hàng	5630		↓ 5630

# Trực quan hóa dữ liệu

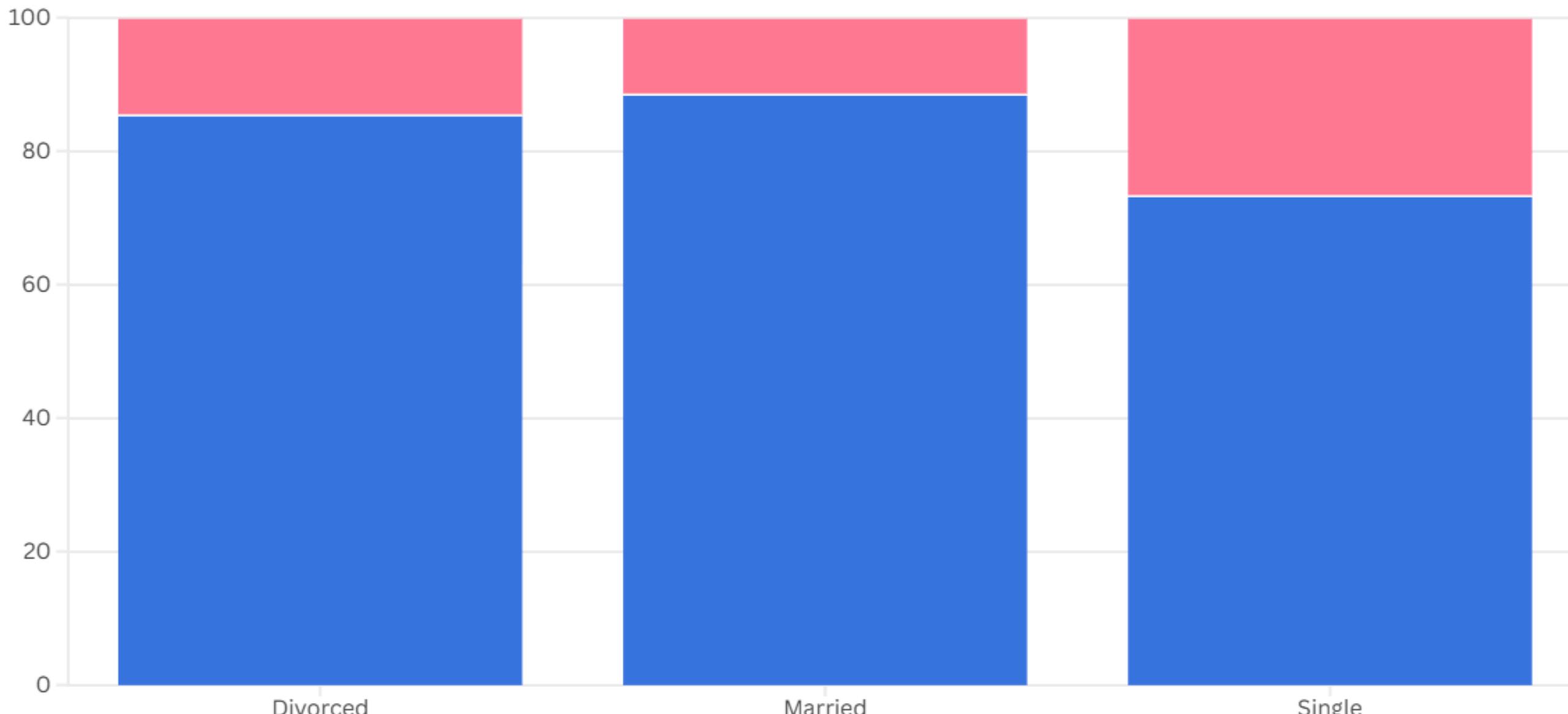
Slide 3

## So sánh trạng thái ở lại và rời bỏ theo nhóm đặc trưng khách hàng

Phân tích phân bố churn theo từng nhóm như Giới tính, Tình trạng hôn nhân, Cấp độ thành phố và Hành vi tiêu dùng.

Tình trạng hôn nhân

Giữ chân Rời bỏ

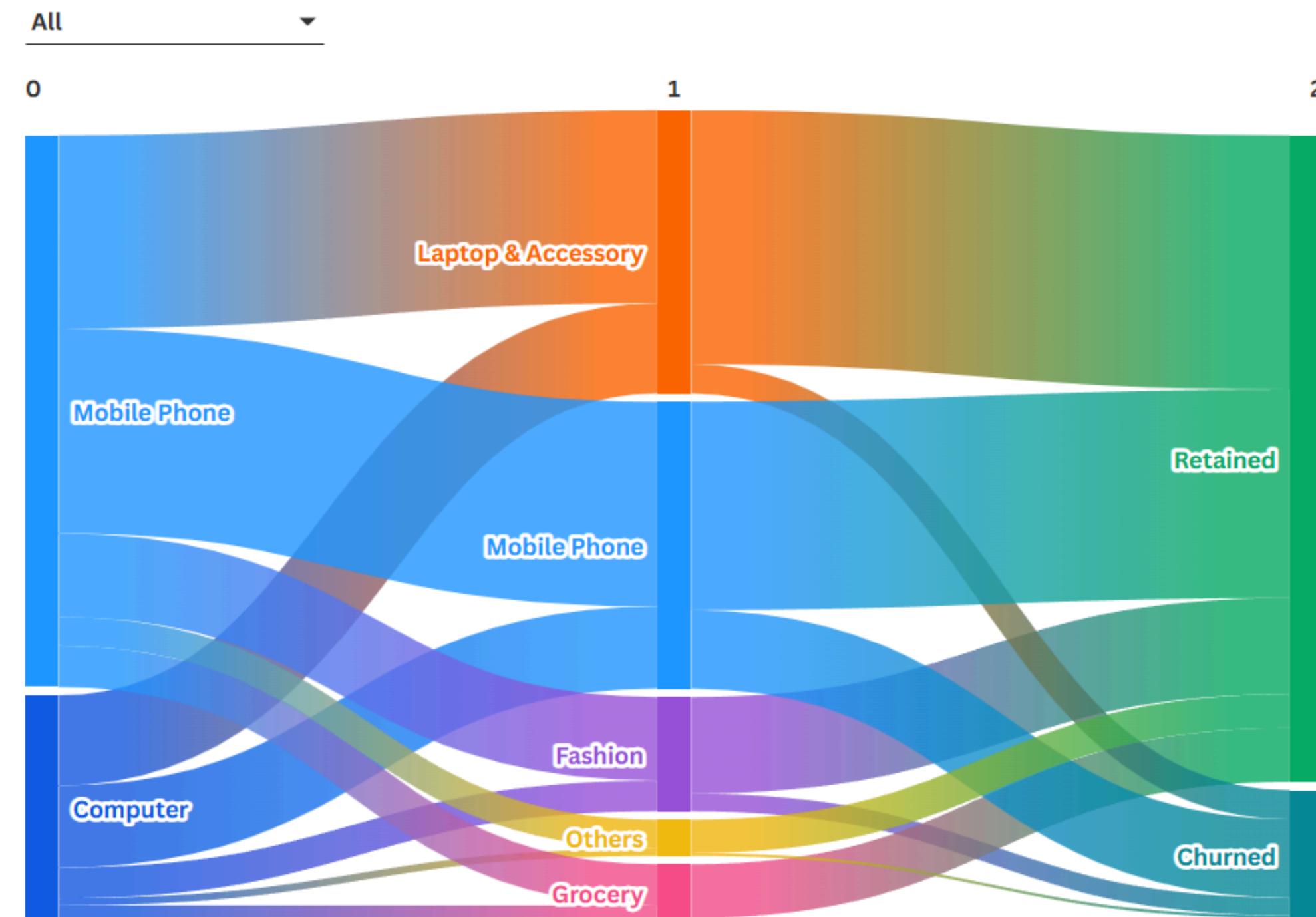


Source: Data

# Trực quan hóa dữ liệu

## Truy vết dòng chảy rời bỏ của khách hàng

Từ Thiết bị đăng nhập → Ngành hàng tiêu dùng → Trạng thái Churn



Source: Data

# Trực quan hóa dữ liệu



Từ Insight đến hành động: Xây dựng chiến lược giữ chân thông minh

Tóm tắt các Phát hiện Chính:

5 of 5

- **Quy mô:** Chúng ta đang mất đi một lượng khách hàng đáng kể (948 người).
- **Tín hiệu mạnh nhất:** Khiếu nại là một yếu tố dự báo Churn cực kỳ mạnh.
- **Phân khúc ủi ro cao:** Khách hàng mua sắm "Thời trang" qua "Máy tính" là nhóm có nguy cơ rời bỏ cao nhất.



## Đề xuất hành động

1. **Ưu tiên xử lý khiếu nại:** Thiết lập quy trình phản hồi nhanh chóng và ưu tiên cho tất cả các khách hàng có khiếu nại. Biến mỗi lời phàn nàn thành một cơ hội để củng cố lòng trung thành.
2. **Triển khai chiến dịch mục tiêu:** Xây dựng các chiến dịch marketing/chăm sóc khách hàng cá nhân hóa nhắm vào phân khúc "Thời trang", đặc biệt là những người dùng website, với các ưu đãi độc quyền hoặc khảo sát trải nghiệm.
3. **Tận dụng mô hình dự đoán:** Sử dụng mô hình Machine Learning đã xây dựng để xác định sớm các cá nhân có nguy cơ rời bỏ cao và chủ động can thiệp trước khi quá muộn.

# Trực quan hóa dữ liệu

[Website demo](#)

## Các tính năng nổi bật:

- **Phân tích linh hoạt:** Cho phép người dùng tải lên tệp dữ liệu CSV của riêng mình và tự khám phá.
- **Tùy biến biểu đồ:** Người dùng có thể tự lựa chọn loại biểu đồ, trục X, Y để tạo ra các phân tích theo ý muốn.
- **Tích hợp hầu như tất cả trong một:** Kết hợp biểu đồ phân tích và Flourish trong cùng một nền tảng.

# Trực quan hóa dữ liệu

Trang thông tin

X Thông tin đồ án

 **PHÂN TÍCH & DỰ ĐOÁN TỈ LỆ RỜI BỎ CỦA KHÁCH HÀNG**  
Demo trực quan hóa dữ liệu

**Trường:** Đại học Công nghệ Thông Tin - ĐHQG TPHCM  
**Khoa:** Khoa học và Kỹ Thuật Thông tin  
**Môn học:** Phân tích và Trực quan hóa dữ liệu (IE313.Q11)

**Giảng viên hướng dẫn**  
ThS. Phạm Nguyễn Phúc Toàn  
toanpnp@uit.edu.vn

**Tổng quan đồ án**  
Đồ án tập trung vào việc **phân tích và trực quan hóa hành vi rời bỏ của khách hàng** trong thương mại điện tử.  
Hệ thống hỗ trợ khám phá sự khác biệt giữa khách hàng rời bỏ và trung thành.

**Thành viên nhóm**  
Phạm Huỳnh Tấn Khang  
22520624@gm.uit.edu.vn

**Huỳnh Ngọc Trang**  
22521510@gm.uit.edu.vn

**Nguyễn Huỳnh Xuân Nghi**  
23521004@gm.uit.edu.vn

**Nguyễn Thị Ngọc Phước**  
23521235@gm.uit.edu.vn

**Nguồn dữ liệu**

Ecommerce Customer Churn Analysis and Prediction (Kaggle)  
Dataset đã xử lý (Google Drive)

Story  
Biểu đồ 1  
Biểu đồ 2  
Biểu đồ 3

Dashboard phân tích

TRỰC QUAN (FLOURISH)

# Trực quan hóa dữ liệu

Trang phân tích

PHÂN TÍCH & DỰ ĐOÁN  
TỈ LỆ RỜI BỎ  
CỦA KHÁCH HÀNG

DANH MỤC HỆ THỐNG

Thông tin đồ án

Dashboard phân tích

TRỰC QUAN (FLOURISH)

Story

Biểu đồ 1

Biểu đồ 2

Biểu đồ 3

X Dashboard phân tích

↑ Tải lên dữ liệu

1. Tải dataset (CSV)

Choose File

ECommerce\_Dataset\_cleaned\_visualization.csv

Đã tải: ECommerce\_Dataset\_cleaned\_visualization.csv

2. CẤU HÌNH BIỂU ĐỒ

Vertical Bar (Comparison)

Horizontal Bar (Long Labels)

Doughnut (Proportions)

Scatter Plot (Correlation)

Stacked Bar (Churn Analysis)

Radar Chart (Profile)

NHÓM THEO (CATEGORY)

PreferredLoginDevice

GIÁ TRỊ PHÂN TÍCH (TÙY CHỌN)

(Mặc định: Đếm số lượng khách hàng)

Biểu đồ sẽ phân tách dữ liệu thành 2 phần: **Rời bỏ** và **Ở lại**.

VẼ BIỂU ĐỒ →



Sẵn sàng trực quan hóa

Tải dữ liệu và cấu hình biểu đồ để xem kết quả phân tích

# Trực quan hóa dữ liệu

Trang phân tích

## PHÂN TÍCH & DỰ ĐOÁN TỈ LỆ RỜI BỎ CỦA KHÁCH HÀNG

### DANH MỤC HỆ THỐNG

Thông tin đồ án

Dashboard phân tích

### TRỰC QUAN (FLOURISH)

Story

Biểu đồ 1

Biểu đồ 2

Biểu đồ 3

X Dashboard phân tích

#### 2. CẤU HÌNH BIỂU ĐỒ

- Vertical Bar (Comparison)
- Horizontal Bar (Long Labels)
- Doughnut (Proportions)
- Scatter Plot (Correlation)
- Stacked Bar (Churn Analysis)**
- Radar Chart (Profile)

#### NHÓM THEO (CATEGORY)

PreferredPaymentMode

#### GIÁ TRỊ PHÂN TÍCH (TÙY CHỌN)

(Mặc định: Đếm số lượng khách hàng)

Biểu đồ sẽ phân tách dữ liệu thành 2 phần: Rời bỏ và Ở lại.

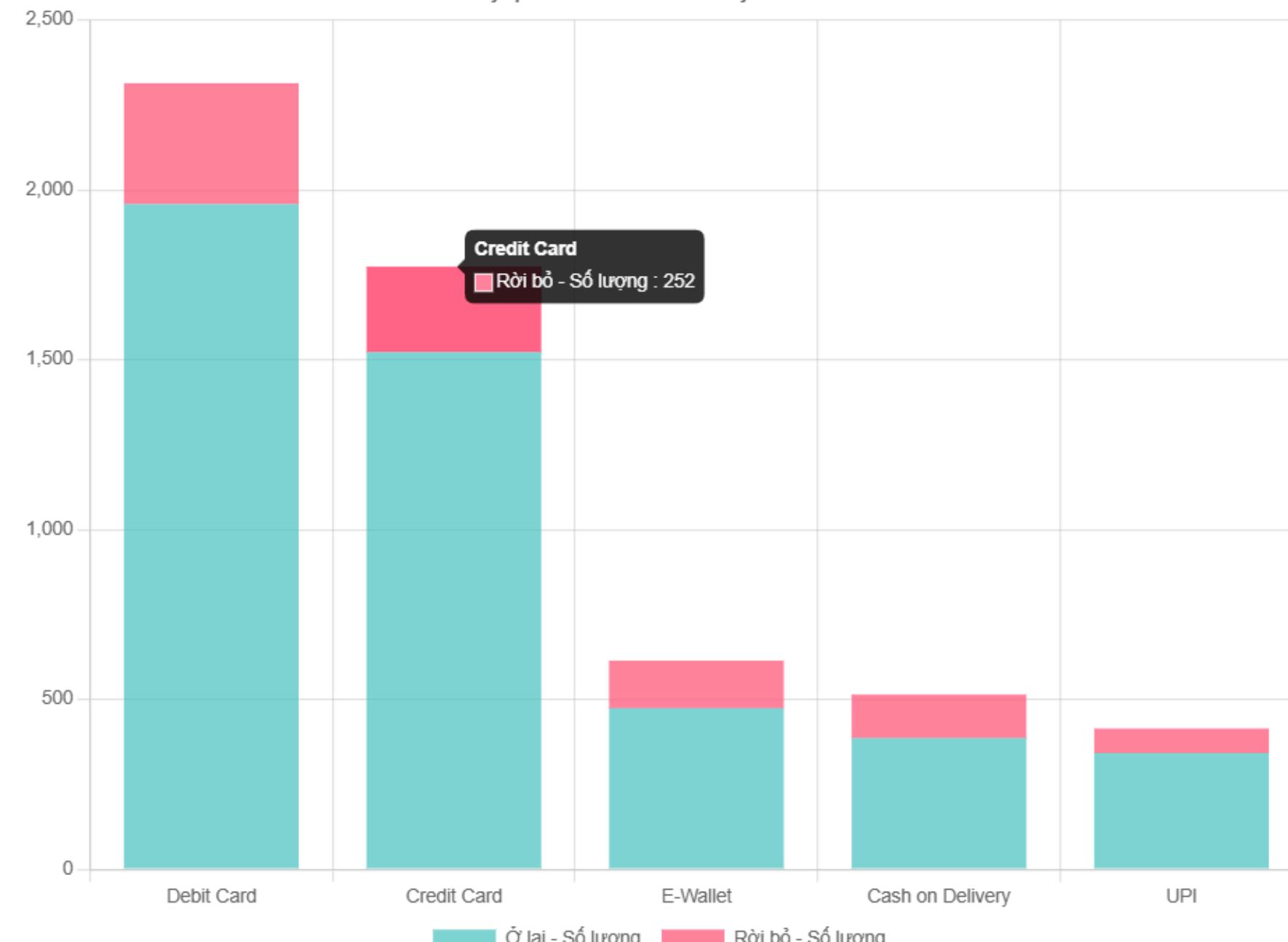
VẼ BIỂU ĐỒ →

#### Kết quả phân tích

Tổng số dòng dữ liệu: 5630

PreferredPaymentMode vs Tân suất

Tỷ lệ Churn theo PreferredPaymentMode



# Xây dựng đặc trưng

## Ý tưởng

- Cần một chỉ số đo lường mức độ gắn kết của khách hàng với hệ sinh thái công nghệ của công ty ngay từ giai đoạn đầu.

## Tạo đặc trưng mới

- $\text{DevicePerTenure} = \text{Số thiết bị đã đăng ký} / \text{Thời gian gắn bó (tháng)}$

## Ý nghĩa

- Đo lường tốc độ đăng ký thiết bị trung bình mỗi tháng.
- *Giá trị cao:* Khách hàng nhanh chóng tích hợp dịch vụ trên nhiều thiết bị, dấu hiệu của sự cam kết cao.
- *Giá trị thấp:* Khách hàng ít tương tác, tiềm ẩn nguy cơ rời bỏ.

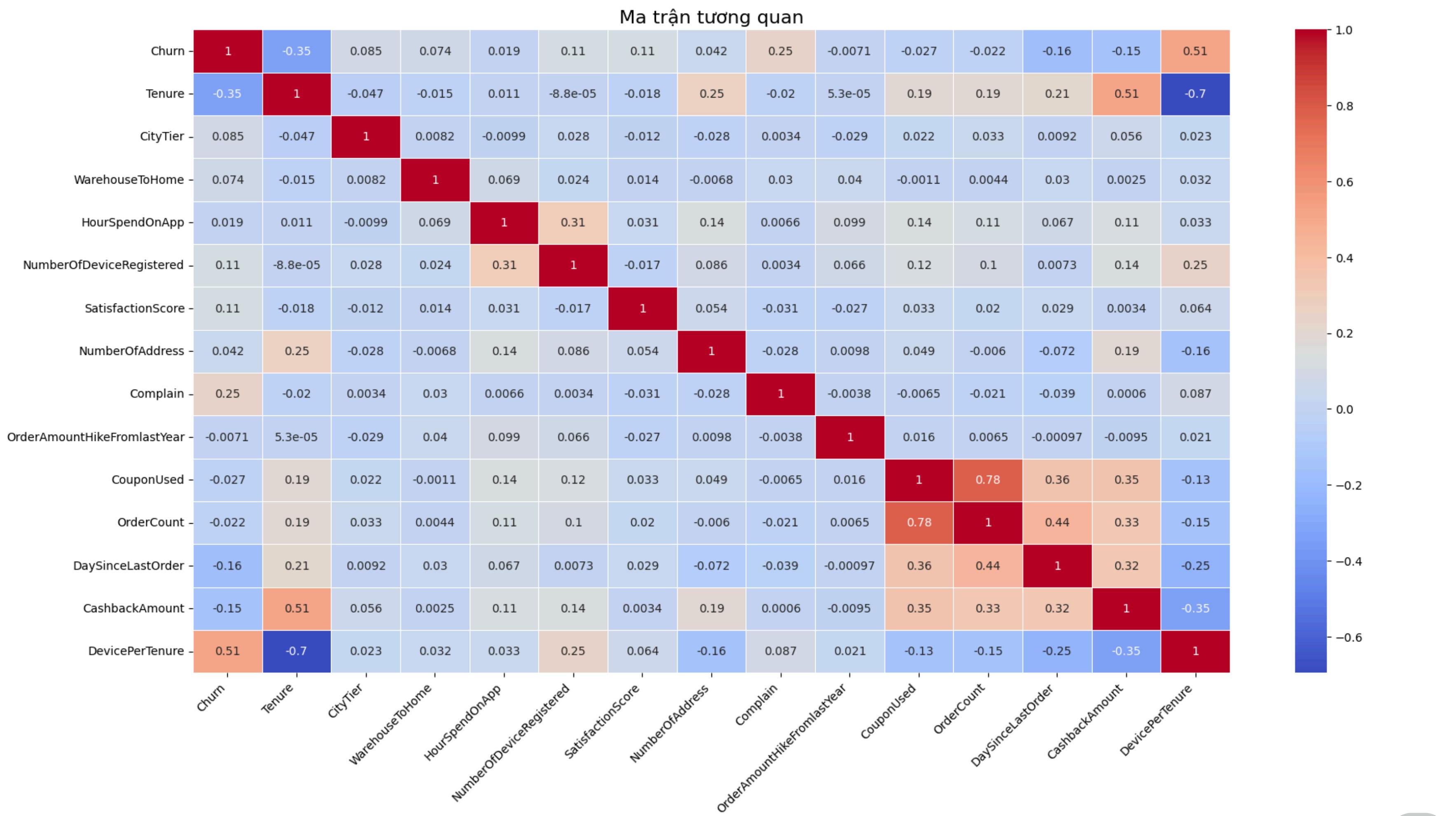
# Xây dựng đặc trưng

## Phân tích biến định lượng

- **Phương pháp:** Sử dụng Ma trận tương quan (Correlation Matrix)
- **Phát hiện chính:**
  - **OrderCount** và **CouponUsed** có tương quan dương mạnh ( $r = 0.78$ ), cung cấp thông tin trùng lặp.
  - **Tenure** và **DevicePerTenure** có tương quan âm mạnh ( $r = -0.70$ ), phân tích sâu hơn cho thấy DevicePerTenure có sức mạnh dự đoán Churn tốt hơn.

Ngưỡng phân loại mức độ tương quan tham khảo [10]:

- 0.00–0.19: rất yếu
- 0.20–0.39: yếu
- 0.40–0.59: trung bình
- 0.60–0.79: mạnh
- 0.80–1.00: rất mạnh



Lưu ý rằng các biến định tính có thứ bậc như CityTier và SatisfactionScore cũng được đưa vào ma trận này để có một cái nhìn tổng quan ban đầu về xu hướng của chúng

# Xây dựng đặc trưng

**Phân tích biến định tính:** Kết hợp 2 kiểm định (Chi-squared & Hệ số Cramér's V) [11], [12]

- **Kiểm định Chi-squared ( $\chi^2$ )** được sử dụng để xác định mối liên quan có ý nghĩa thống kê giữa biến định tính và biến mục tiêu Churn thông qua p-value (< 0.05), nhưng không phản ánh độ mạnh của mối quan hệ.
- **Hệ số Cramér's V** được dùng bổ sung để đo lường cường độ liên hệ giữa hai biến định tính, với giá trị trong khoảng [0,1]. Giá trị càng lớn thể hiện mối liên hệ càng mạnh giữa các biến.

Việc kết hợp Chi-squared và Cramér's V cho phép vừa đánh giá ý nghĩa thống kê, vừa xác định mức độ ảnh hưởng thực tế của các biến định tính đến Churn.

## Ngưỡng Cramér's V

- 0.00 – 0.10: Mối liên hệ rất yếu / không đáng kể
- 0.10 – 0.30: Mối liên hệ yếu
- 0.30 – 0.50: Mối liên hệ trung bình
- > 0.50: Mối liên hệ mạnh

# Xây dựng đặc trưng

**Phân tích biến định tính:** Kết hợp 2 kiểm định

- **Chi-squared:** Xác nhận tất cả các biến đều có liên quan ( $p < 0.05$ ).
- **Hệ số Cramér's V:** Đo lường sức mạnh của mối liên quan đó

```
PreferredLoginDevice : chi2 = 14.4 ; p-value = 0.00015
PreferredPaymentMode : chi2 = 51.83 ; p-value = 0.0
Gender : chi2 = 4.66 ; p-value = 0.03082
PreferredOrderCat : chi2 = 288.6 ; p-value = 0.0
MaritalStatus : chi2 = 188.67 ; p-value = 0.0
```

*Kết quả kiểm định Chi-squared*

# Xây dựng đặc trưng

**Phân tích biến định tính:** Kết hợp 2 kiểm định

- **Chi-squared:** Xác nhận tất cả các biến đều có liên quan ( $p < 0.05$ ).
- **Hệ số Cramér's V:** Đo lường sức mạnh của mối liên quan đó.
  - Liên quan yếu: **PreferredOrderCat** (0.226), **MaritalStatus** (0.183)
  - Mối liên hệ rất yếu / không đáng kể: **PreferredLoginDevice** (0.051), **Gender** (0.029), **PreferredPaymentMode** (0.096) → Ít giá trị dự đoán

PreferredLoginDevice 0.051  
PreferredPaymentMode 0.096  
Gender 0.029  
PreferredOrderCat 0.226  
MaritalStatus 0.183

Kết quả chỉ số Cramer's V

# Xây dựng đặc trưng

## Loại bỏ 5 đặc trưng

- Lý do đa cộng tuyến & có biến thay thế tốt hơn:
  - OrderCount (thay bằng các biến hành vi khác)
  - Tenure (thay bằng DevicePerTenure mạnh hơn)
- Lý do sức mạnh dự đoán rất yếu:
  - PreferredLoginDevice
  - PreferredPaymentMode
  - Gender

# Xây dựng mô hình

- Để đảm bảo tính nhất quán và khả năng tái lặp của kết quả thực nghiệm, toàn bộ quy trình tiền xử lý dữ liệu và huấn luyện mô hình được chuẩn hóa thông qua việc sử dụng **Pipeline**.
- **Pipeline** cho phép kết hợp các bước xử lý như chuẩn hóa dữ liệu, mã hóa đặc trưng và huấn luyện mô hình vào một quy trình thống nhất.
- Giúp tự động hóa quá trình huấn luyện, đảm bảo các bước tiền xử lý được áp dụng đồng nhất trên cả tập huấn luyện và tập kiểm tra, đồng thời hạn chế hiện tượng rò rỉ dữ liệu (data leakage), từ đó nâng cao độ tin cậy của kết quả đánh giá mô hình.

# Xây dựng mô hình

- Tập **train-test** được chia theo tỉ lệ **8:2**
- Sử dụng phương pháp *train\_test\_split* kết hợp phân tầng (stratify = y).
- Việc phân tầng giúp đảm bảo tỷ lệ Churn và Non-Churn trong hai tập dữ liệu tương đồng với phân phối ban đầu.
- **Tiền xử lý dữ liệu trong Pipeline:**
  - Một *ColumnTransformer* được đưa vào Pipeline để xử lý các loại biến khác nhau.
  - Các bước tiền xử lý thực hiện đồng thời và nhất quán trong quá trình huấn luyện.

# Xây dựng mô hình

## Xử lý biến đầu vào

- Biến định lượng:
  - Chuẩn hóa bằng StandardScaler
  - Đưa dữ liệu về cùng thang đo với trung bình 0 và phương sai 1
- Biến định tính:
  - Mã hóa bằng OneHotEncoder
  - Chuyển đổi dữ liệu dạng danh mục sang dạng số

# Xây dựng mô hình

## Xử lý mất cân bằng dữ liệu (class\_weight )

- Áp dụng kỹ thuật class\_weight đối với các mô hình hỗ trợ trực tiếp.
- Trọng số lớp được điều chỉnh trong hàm mất mát, giúp:
  - Tăng mức phạt cho các lỗi dự đoán sai trên lớp thiểu số (Churn)
  - Cải thiện khả năng nhận diện khách hàng rời bỏ

## Xử lý mất cân bằng dữ liệu (SMOTE)

- Đối với các mô hình không hỗ trợ class\_weight trực tiếp (KNN, Gradient Boosting, MLP), kỹ thuật SMOTE được tích hợp vào Pipeline.
- SMOTE tạo ra các mẫu dữ liệu tổng hợp mới cho lớp thiểu số.
- Giúp cân bằng phân phối lớp trước khi đưa dữ liệu vào huấn luyện mô hình.

# Xây dựng mô hình

- **Mô hình Logistic Regression:** Là một mô hình tuyến tính cơ sở, được sử dụng để thiết lập một ngưỡng hiệu suất ban đầu. Mô hình này ước tính xác suất xảy ra của một sự kiện (chẳng hạn như Churn) dựa trên một tổ hợp tuyến tính của các biến đầu vào, sau đó ánh xạ kết quả này thông qua hàm sigmoid để đưa ra xác suất thuộc về lớp mục tiêu. [13]
- **Cấu hình:**
  - class\_weight = 'balanced'
  - max\_iter = 1000

# Xây dựng mô hình

- **KNN (K-Nearest Neighbors):** Là một thuật toán học phi tham số, phân loại một điểm dữ liệu mới dựa trên “phiếu bầu” của K điểm dữ liệu gần nhất trong không gian đặc trưng. Kết quả dự đoán được xác định bởi lớp chiếm đa số trong số K láng giềng, do đó KNN phản ánh trực tiếp mức độ tương đồng giữa các mẫu dữ liệu. Thuật toán này không yêu cầu quá trình huấn luyện rõ ràng, tuy nhiên hiệu suất phụ thuộc mạnh vào việc lựa chọn giá trị K và cách chuẩn hóa dữ liệu đầu vào. [14]
- **Cấu hình:** Các siêu tham số được tìm kiếm tự động thông qua GridSearch. Dữ liệu đầu vào được cân bằng với SMOTE.

```
Tham số tốt nhất: {'classifier_metric': 'euclidean', 'classifier_n_neighbors': 7, 'classifier_weights': 'distance'}
Best Cross-Validation Score: 0.9564
```

# Xây dựng mô hình

- **SVM (Support Vector Machine):** Là một mô hình mạnh mẽ tìm kiếm một siêu phẳng trong không gian nhiều chiều để phân tách tốt nhất các lớp dữ liệu. Mô hình đạt được điều này bằng cách tối đa hóa khoảng cách giữa ranh giới phân loại và các điểm dữ liệu gần nhất của mỗi lớp. [15]
- **Cấu hình:**
  - kernel='rbf'
  - probability=True
  - class\_weight='balanced'

# Xây dựng mô hình

- **Random Forest:** Là một thuật toán học tổ hợp theo phương pháp bagging. Kết quả dự đoán của mô hình được xác định thông qua cơ chế bỏ phiếu của các cây thành phần, qua đó làm giảm sự phụ thuộc vào một mô hình đơn lẻ. Cách tiếp cận này giúp Random Forest giảm phương sai, tăng tính ổn định và hạn chế hiện tượng quá khớp khi làm việc với dữ liệu phức tạp. [16]
- **Cấu hình:**
  - n\_estimators=300
  - class\_weight='balanced'

# Xây dựng mô hình

- **Gradient Boosting:** Là một phương pháp học máy tổ hợp dựa trên kỹ thuật boosting, trong đó các cây quyết định được huấn luyện theo trình tự thay vì độc lập. Mỗi mô hình mới được xây dựng nhằm tập trung vào các sai số còn tồn tại của mô hình trước, từ đó dần dần tối ưu hàm mất mát tổng thể. Cơ chế học nối tiếp này cho phép Gradient Boosting cải thiện hiệu suất dự đoán và nắm bắt tốt hơn các mẫu dữ liệu phức tạp. [17]
- **Cấu hình:**
  - n\_estimators=100
  - learning\_rate=0.1
  - max\_depth=3
- Dữ liệu đầu vào được cân bằng với SMOTE

# Xây dựng mô hình

- **XGBoost (Extreme Gradient Boosting):** Là một thuật toán học máy tổ hợp được phát triển từ Gradient Boosting với nhiều cải tiến về hiệu suất và khả năng mở rộng. Mô hình tích hợp các kỹ thuật tối ưu như huấn luyện song song, kiểm soát độ phức tạp thông qua regularization và xử lý dữ liệu hiệu quả, nhờ đó rút ngắn thời gian huấn luyện đồng thời nâng cao độ chính xác dự đoán. [6]
- **Cấu hình:**
  - n\_estimators=300
  - subsample=0.8
- Trọng số *scale\_pos\_weight* được tính toán tự động để xử lý mất cân bằng.

# Xây dựng mô hình

- **LightGBM (Light Gradient Boosting Machine):** Là một thuật toán boosting được phát triển nhằm tối ưu tốc độ huấn luyện và hiệu quả sử dụng bộ nhớ. Mô hình áp dụng các kỹ thuật như gradient-based one-side sampling (GOSS) để giảm số lượng mẫu cần xử lý và exclusive feature bundling (EFB) để gom các đặc trưng ít khi xuất hiện đồng thời. Nhờ đó, LightGBM đạt hiệu suất cao và đặc biệt phù hợp với các tập dữ liệu có kích thước lớn. [18]
- **Cấu hình:**
  - n\_estimators=300
  - learning\_rate=0.05
  - num\_leaves=20
  - max\_depth=5
  - class\_weight='balanced'.

# Xây dựng mô hình

- **MLP (Multi-layer Perceptron):** Là mô hình mạng nơ-ron nhân tạo truyền thẳng, gồm nhiều lớp nơ-ron được sắp xếp theo cấu trúc nhiều tầng. Thông qua các hàm kích hoạt phi tuyến, mô hình học và biểu diễn các mối quan hệ phi tuyến phức tạp trong dữ liệu. Nhờ tính linh hoạt trong cấu trúc, MLP thường được sử dụng cho các bài toán phân loại và hồi quy khi dữ liệu không tuân theo quan hệ tuyến tính. [19]
- **Cấu hình:**
  - Mạng nơ-ron có 2 lớp ẩn với số lượng nơ-ron lần lượt là 32 và 16, sử dụng hàm kích hoạt ReLU, thuật toán tối ưu Adam và hệ số điều chỉnh  $\alpha=0.01$ .
  - Dữ liệu đầu vào được cân bằng bằng phương pháp SMOTE.

# Xây dựng mô hình

## Các độ đo đánh giá [20]

**Ma trận nhầm lẫn (Confusion Matrix):** Là một bảng tóm tắt hiệu suất của một mô hình phân loại. Đối với bài toán phân loại nhị phân, ma trận này có dạng 2x2, gồm bốn giá trị:

- **True Positives (TP):** Số trường hợp được dự đoán là Churn và thực tế Churn.
- **True Negatives (TN):** Số trường hợp được dự đoán là Non-Churn và thực tế Non-Churn.
- **False Positives (FP):** Lỗi loại I - Số trường hợp được dự đoán là Churn nhưng thực tế Non-Churn.
- **False Negatives (FN):** Lỗi loại II - Số trường hợp được dự đoán là Non-Churn nhưng thực tế Churn.

*Trong bài toán dự đoán Churn, việc giảm thiểu False Negatives (FN), tức lỗi loại II, thường được ưu tiên, vì bỏ sót một khách hàng sắp rời bỏ gây tổn thất kinh doanh lớn hơn so với việc tiếp cận nhầm một khách hàng trung thành.*

# Xây dựng mô hình

## Các độ đo đánh giá [20]

Classification Report:

- **Accuracy (Độ chính xác):** Tỷ lệ tổng số dự đoán đúng trên tổng số mẫu
- **Precision (Độ chuẩn xác):** Mức độ tin cậy của các dự đoán dương
- **Recall (Độ phủ / Độ nhạy):** Khả năng của mô hình phát hiện tất cả các trường hợp dương
- **F1-Score:** Trung bình điều hòa của Precision và Recall, cung cấp chỉ số cân bằng

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

# Xây dựng mô hình

## Các độ đo đánh giá [20]

### Đường cong ROC và chỉ số AUC:

- **ROC (Receiver Operating Characteristic):** Biểu đồ thể hiện khả năng phân loại của mô hình nhị phân tại các ngưỡng khác nhau. Trục tung là tỷ lệ True Positive (Recall), trục hoành là tỷ lệ False Positive. Mô hình tốt có đường cong tiến gần góc trên trái của biểu đồ.
- **AUC (Area Under the Curve):** Diện tích dưới đường cong ROC, nằm trong khoảng 0.5 (phân loại ngẫu nhiên) đến 1.0 (phân loại hoàn hảo). AUC là thước đo tổng thể, không phụ thuộc ngưỡng, thể hiện năng lực phân biệt giữa hai lớp Churn và Non-Churn. Đây là chỉ số quan trọng để so sánh hiệu suất tổng thể của các mô hình, đặc biệt trên dữ liệu mất cân bằng.

# 4. Kết quả & bàn luận

Bảng II: Bảng so sánh hiệu suất của 8 mô hình

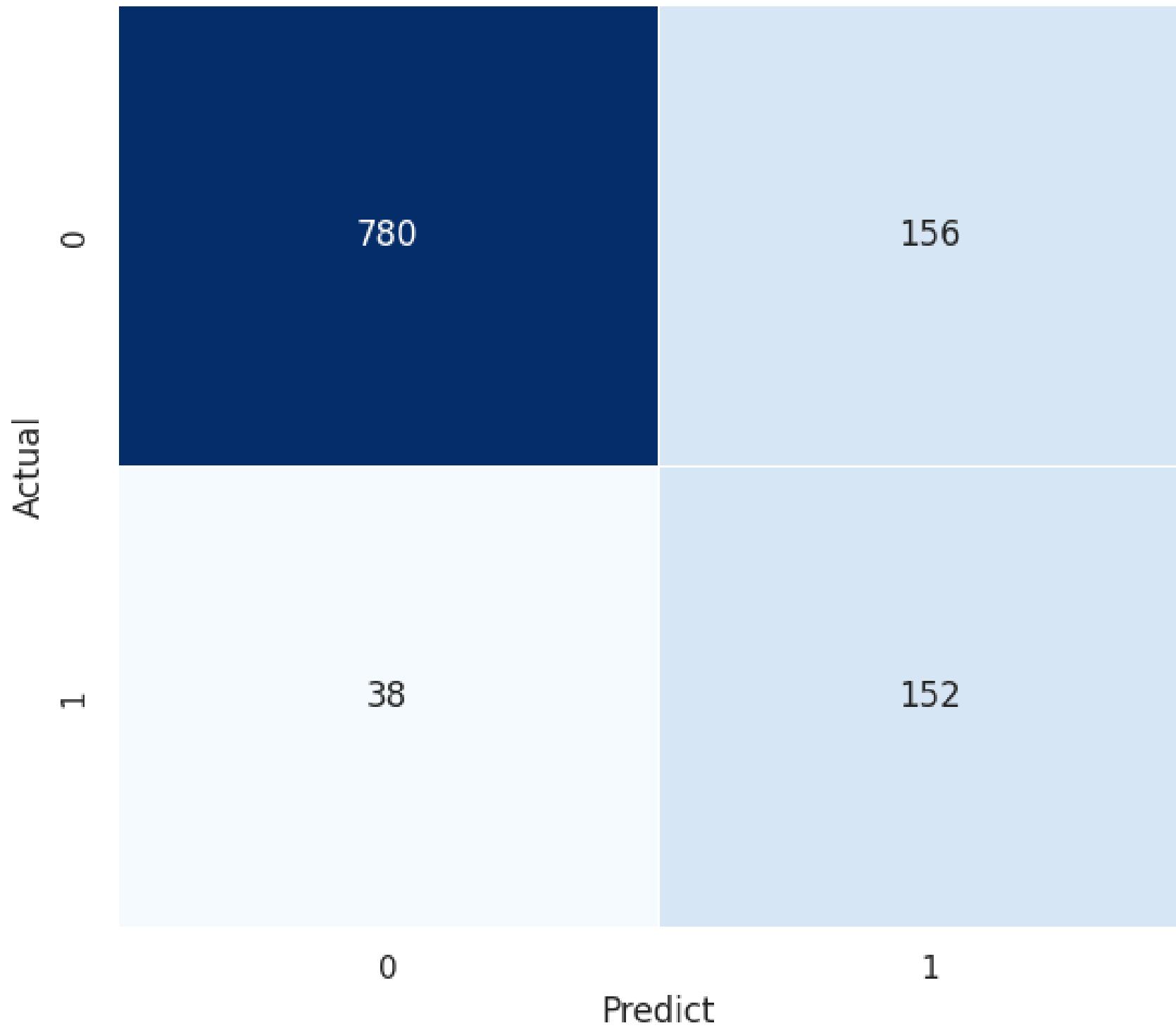
Model	Accuracy	ROC-AUC	F1-Score (C1)	Precision (C1)	Recall (C1)
XGBoost	<b>0.99</b>	0.99528	<b>0.96</b>	<b>0.97</b>	0.95
Random Forest	0.97	<b>0.99549</b>	0.89	0.96	0.84
MLP	0.95	0.98513	0.87	0.82	0.94
LightGBM	0.95	0.98507	0.86	0.79	0.95
KNN	0.91	0.97367	0.78	0.65	<b>0.96</b>
SVM	0.91	0.92792	0.71	0.82	0.62
Gradient Boosting	0.90	0.91902	0.70	0.70	0.71
Logistic Regression	0.83	0.88571	0.61	0.49	0.80

Giá trị tốt nhất được in đậm.

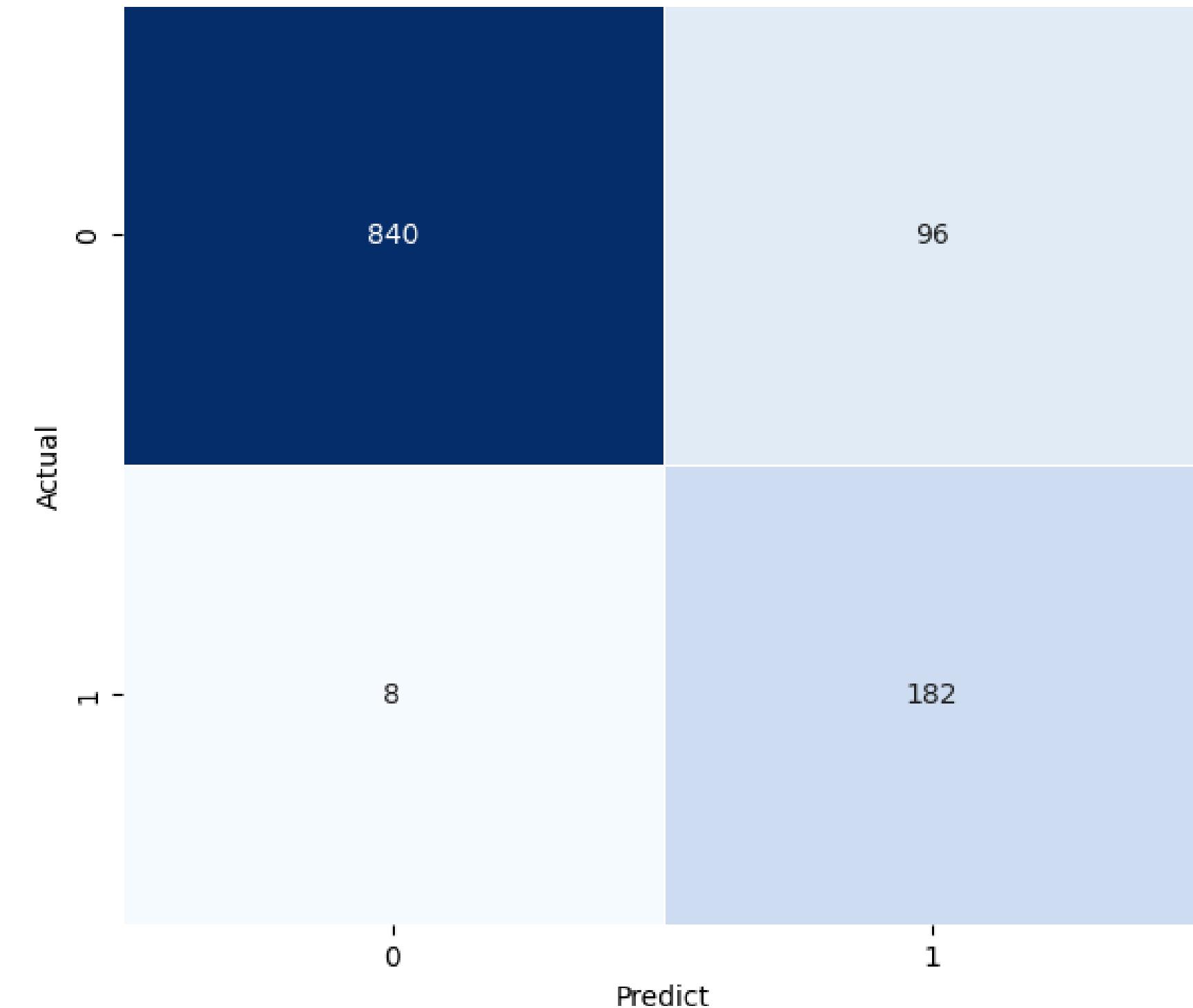
Chú thích: Các chỉ số F1-Score, Precision và Recall được báo cáo cho lớp Churn (C1) do đây là lớp thiểu số và có ý nghĩa kinh doanh quan trọng, trong đó khả năng phát hiện chính xác khách hàng rời bỏ là ưu tiên hàng đầu nhằm hỗ trợ chiến lược giữ chân hiệu quả.

# Kết quả

Confusion Matrix - Logistic Regression

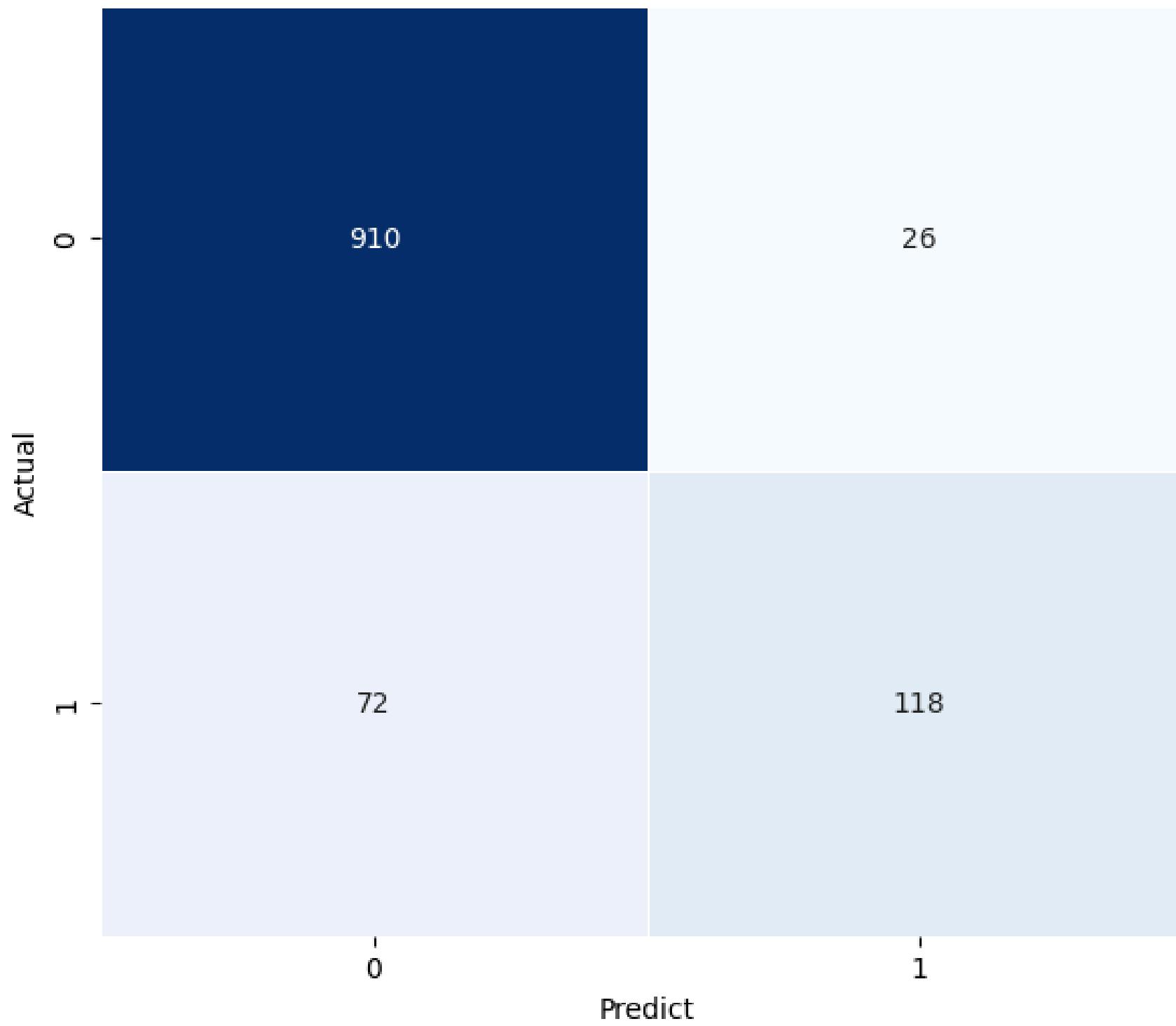


Confusion Matrix - KNN (Optimized)

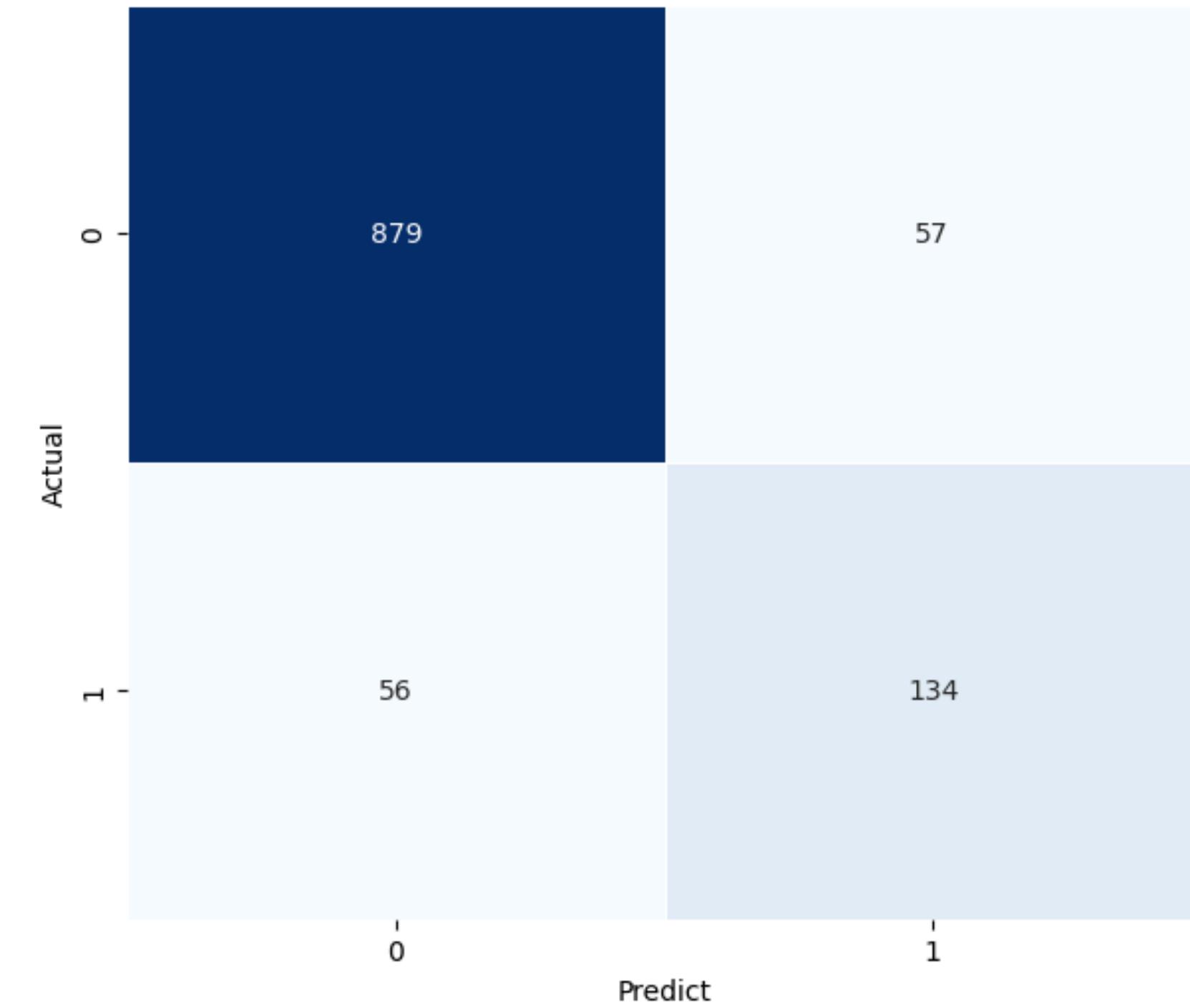


# Kết quả

Confusion Matrix - SVM

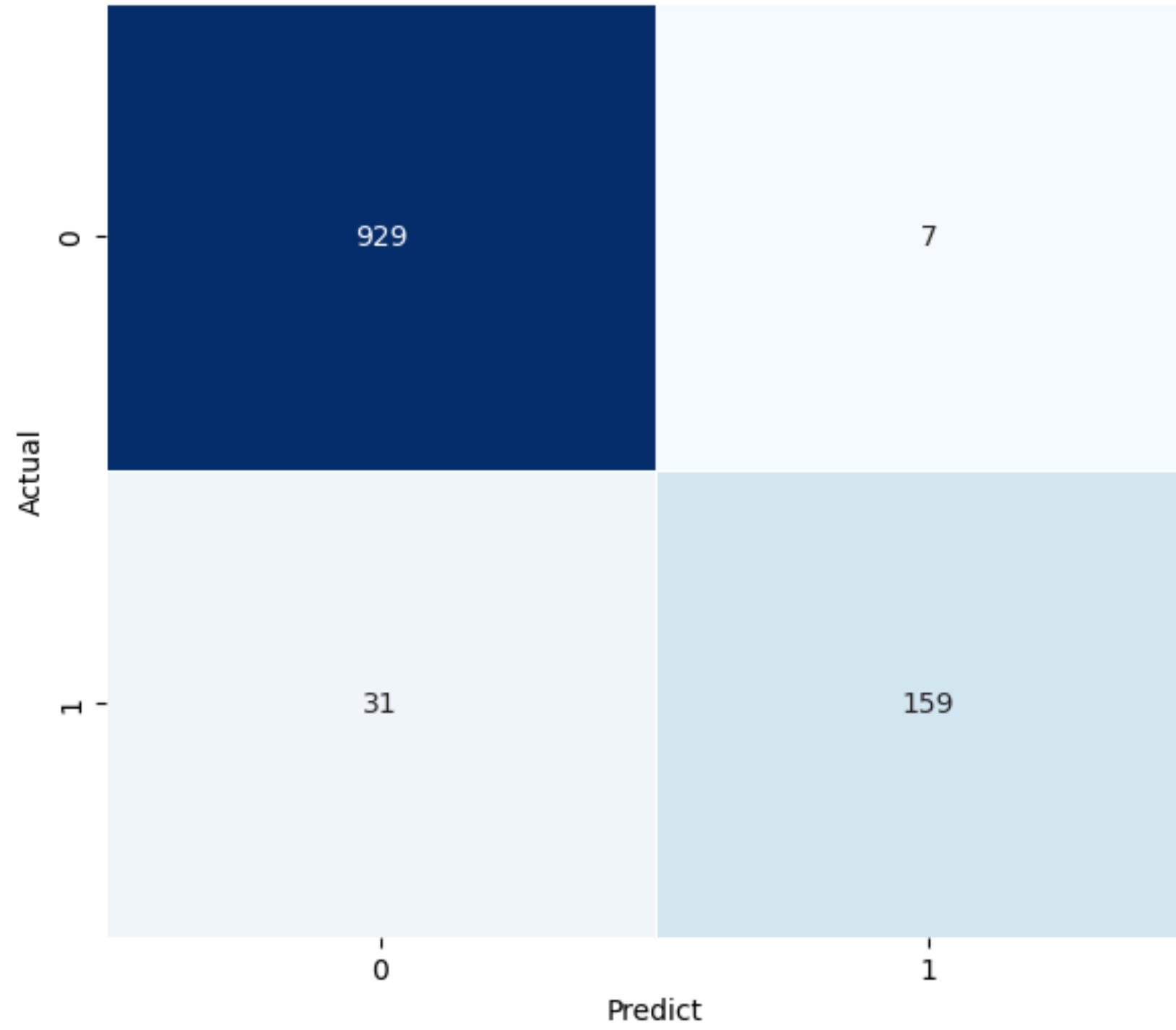


Confusion Matrix - Gradient Boosting

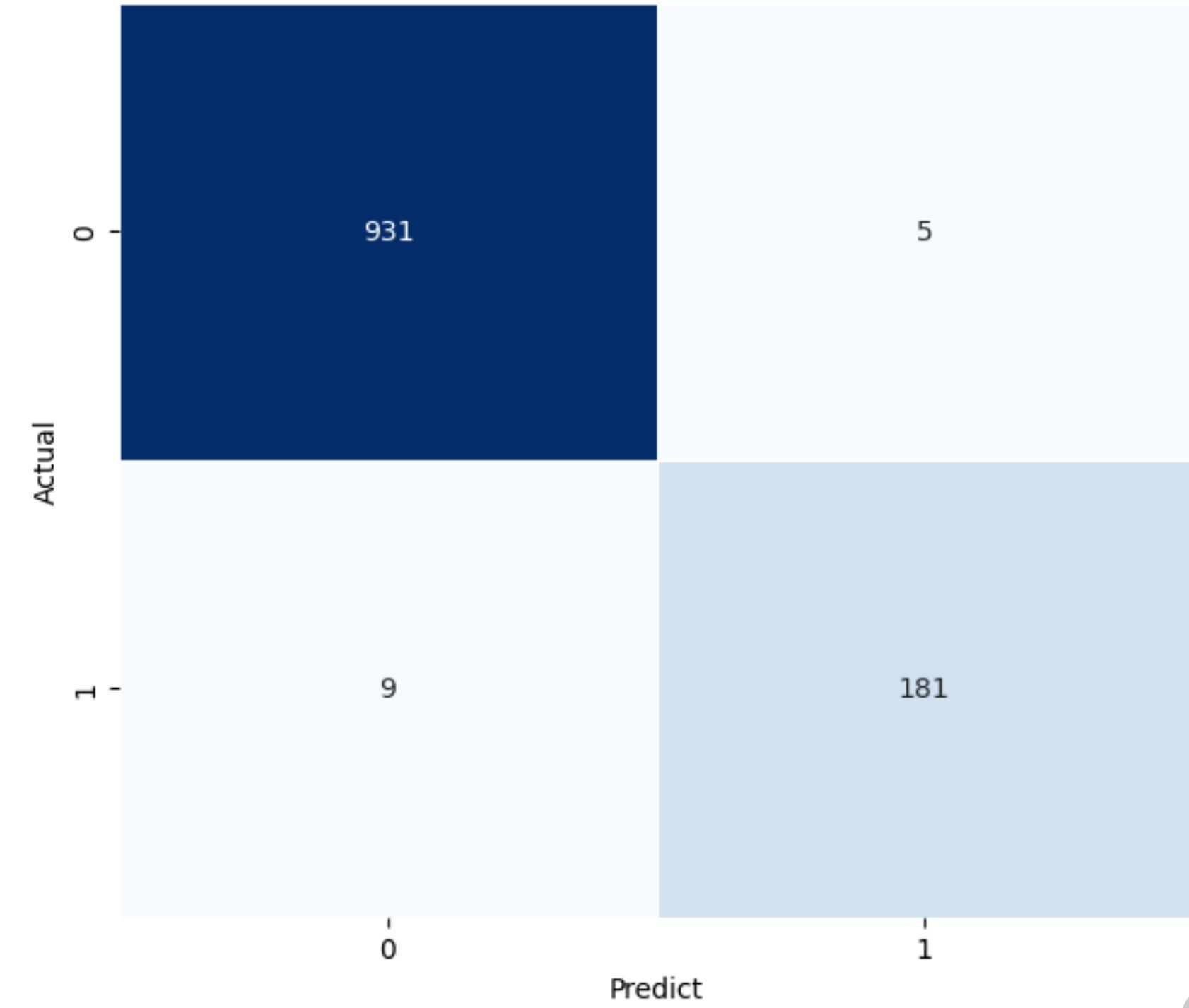


# Kết quả

Confusion Matrix - Random Forest

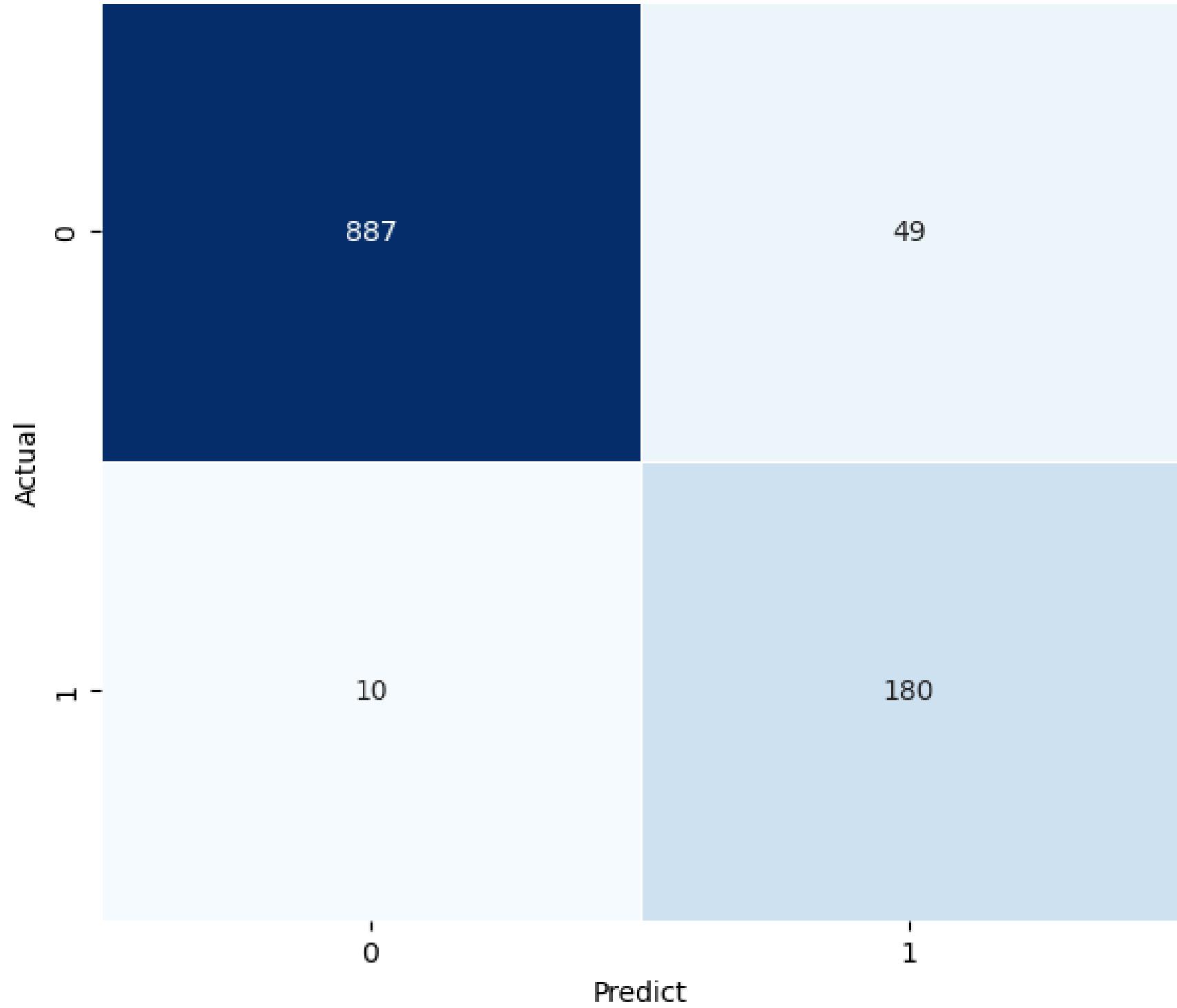


Confusion Matrix - XGBoost

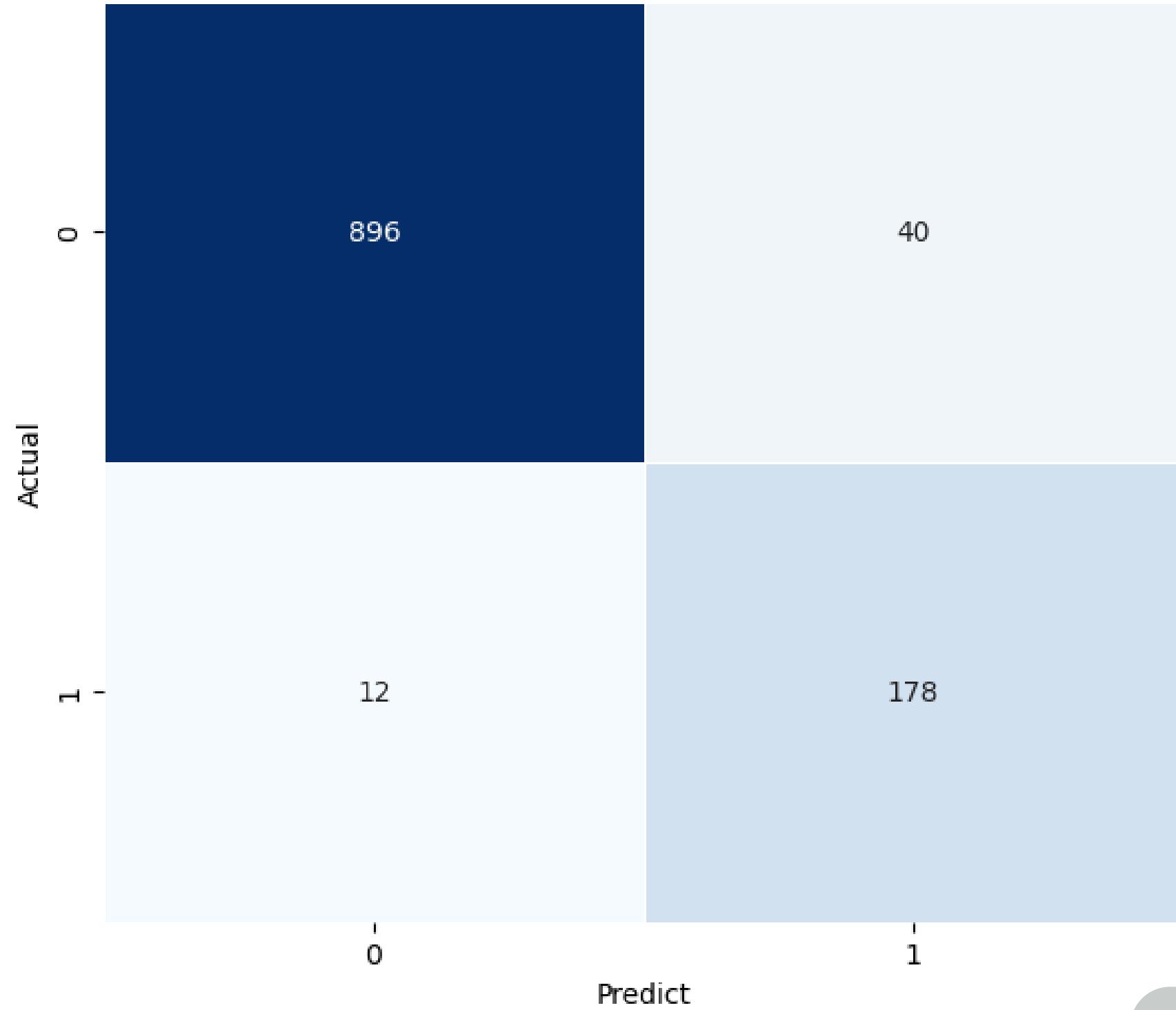


# Kết quả

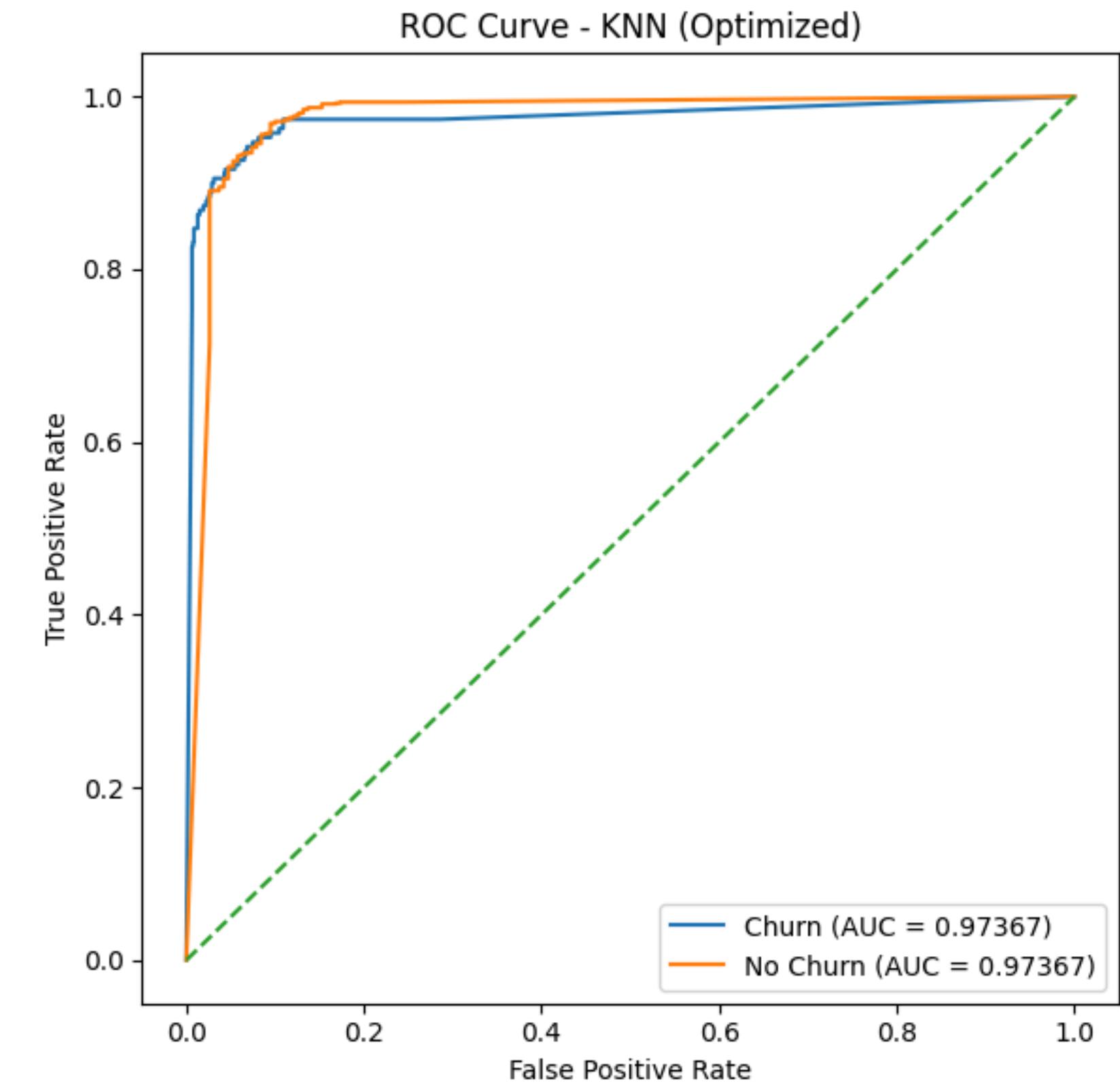
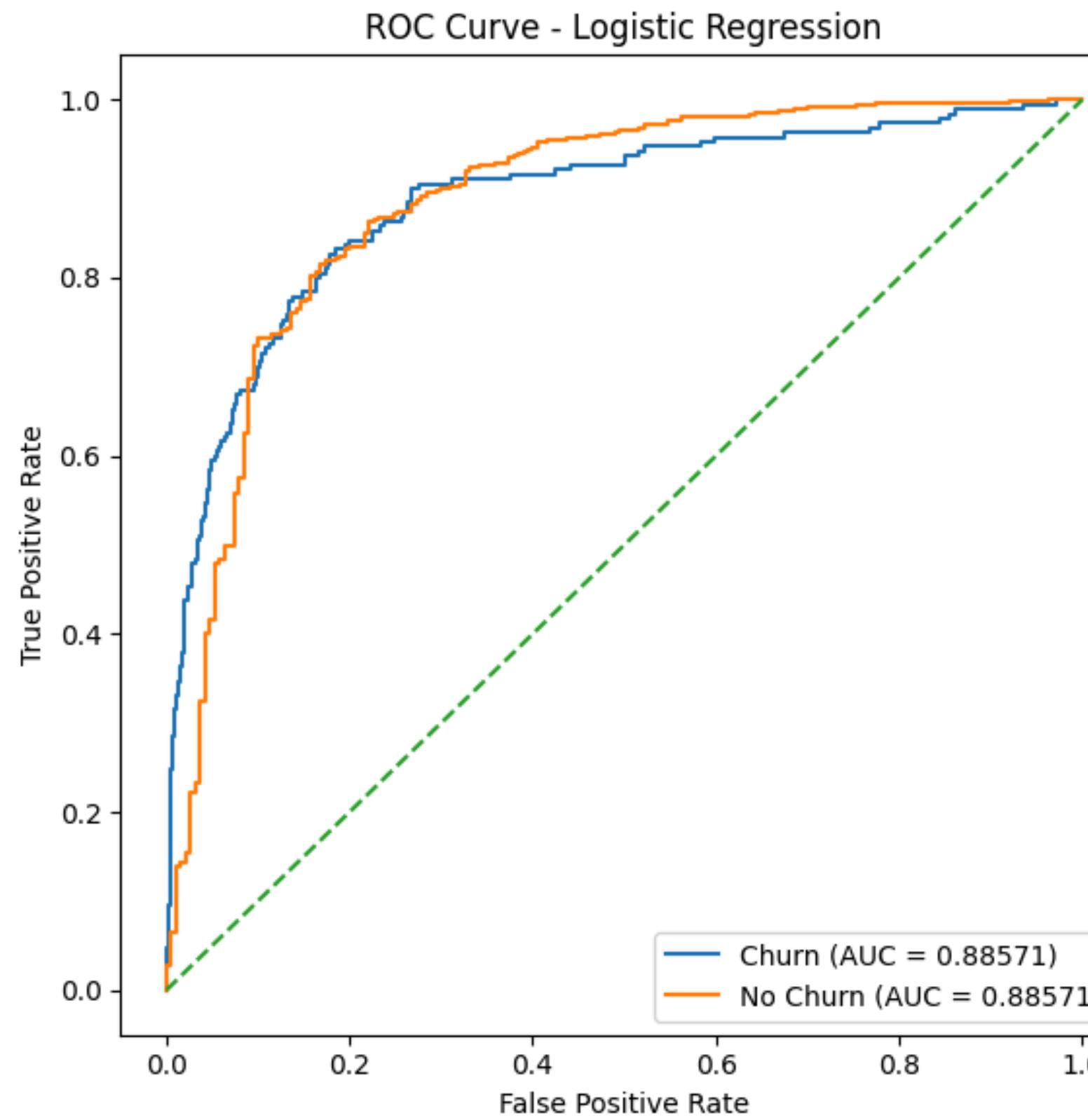
Confusion Matrix - LightGBM



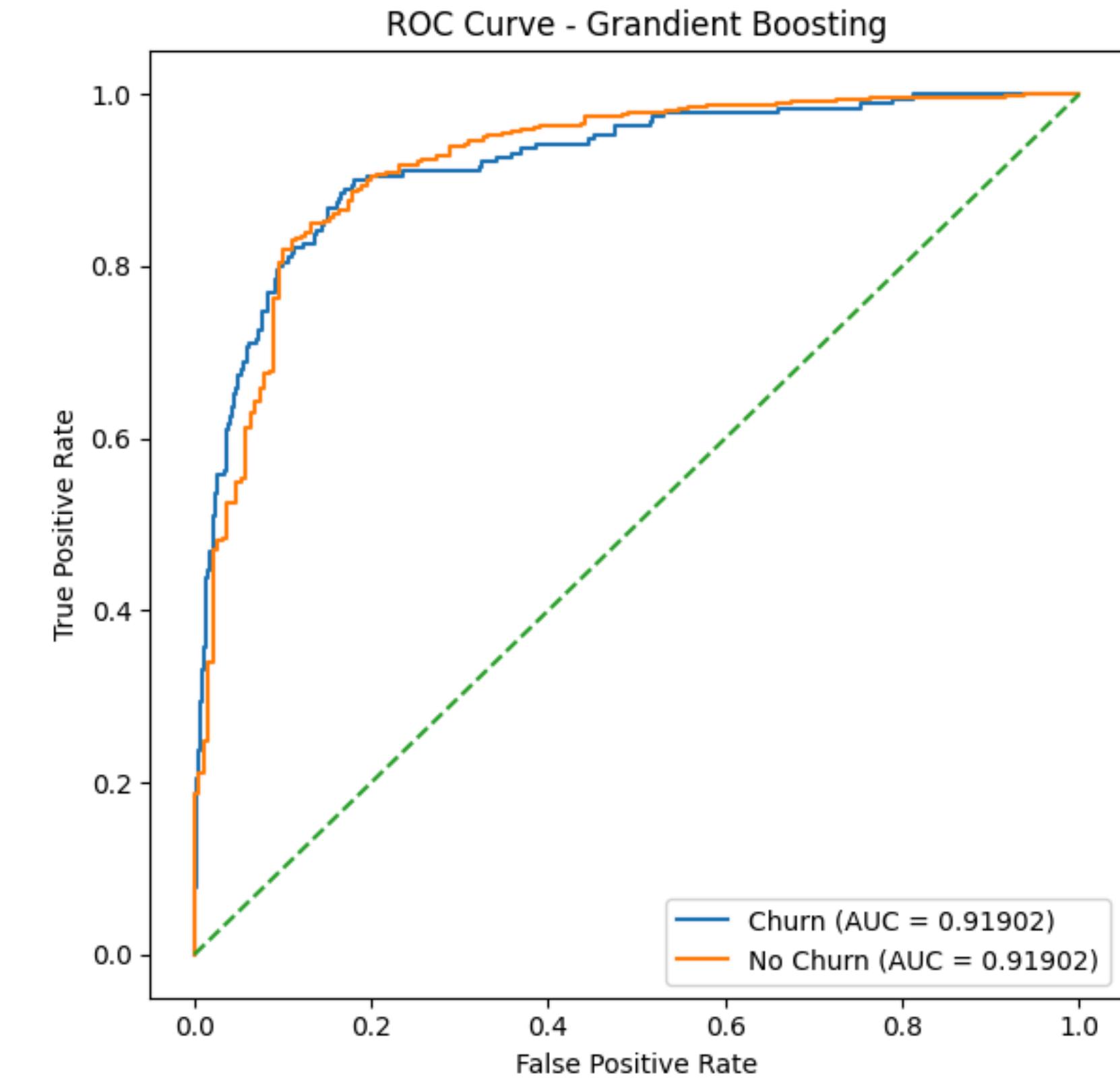
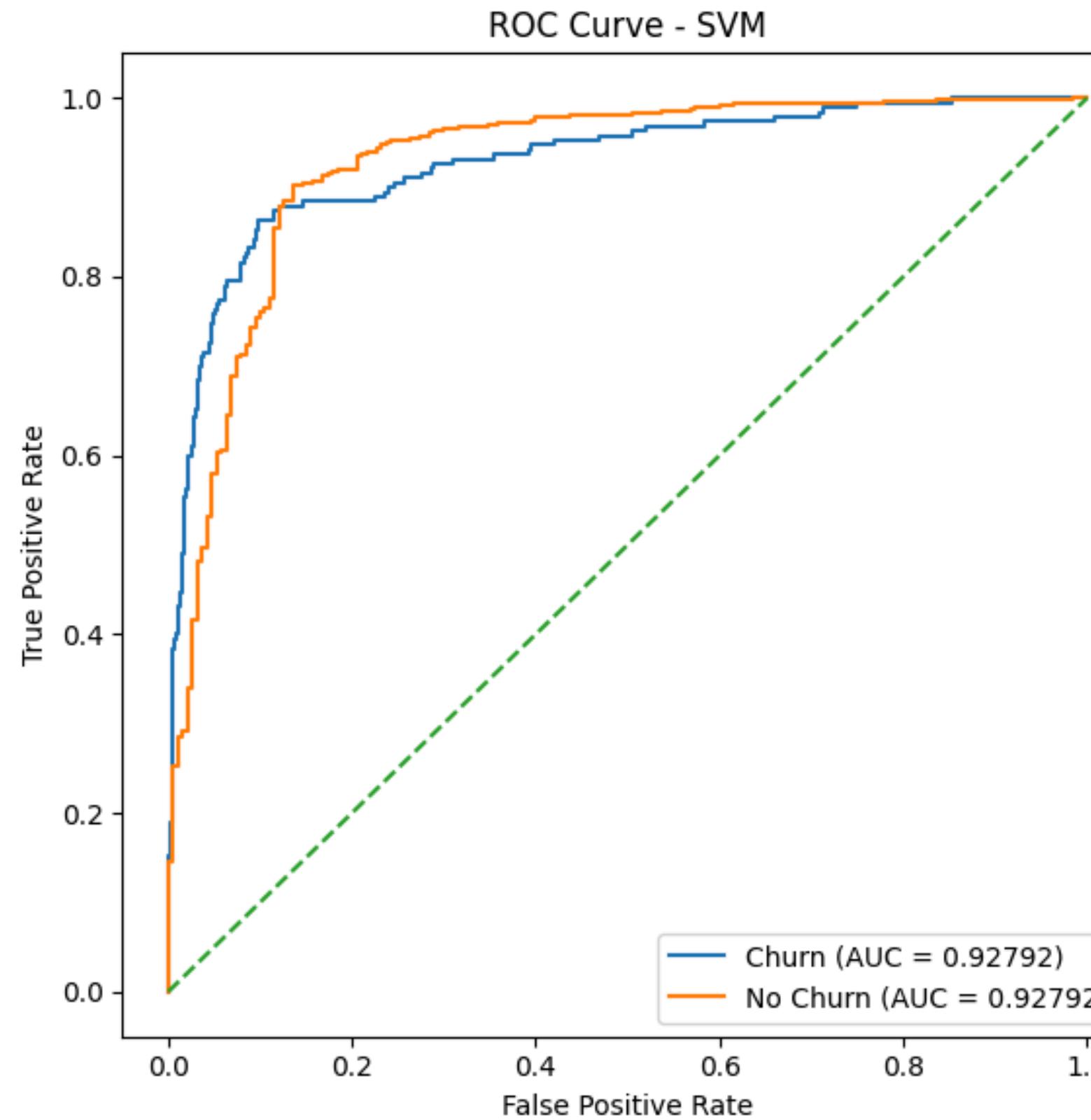
Confusion Matrix - MLP



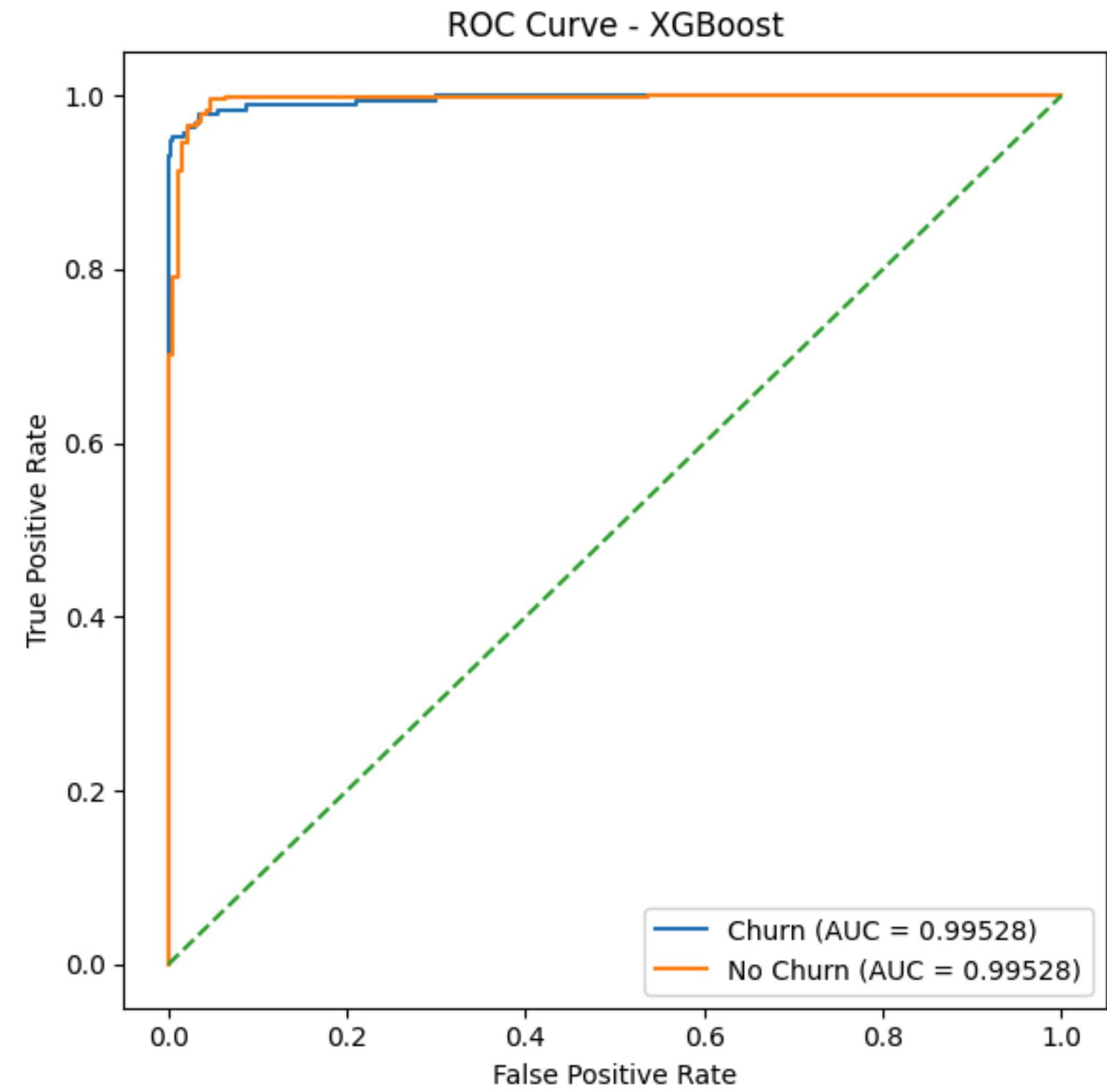
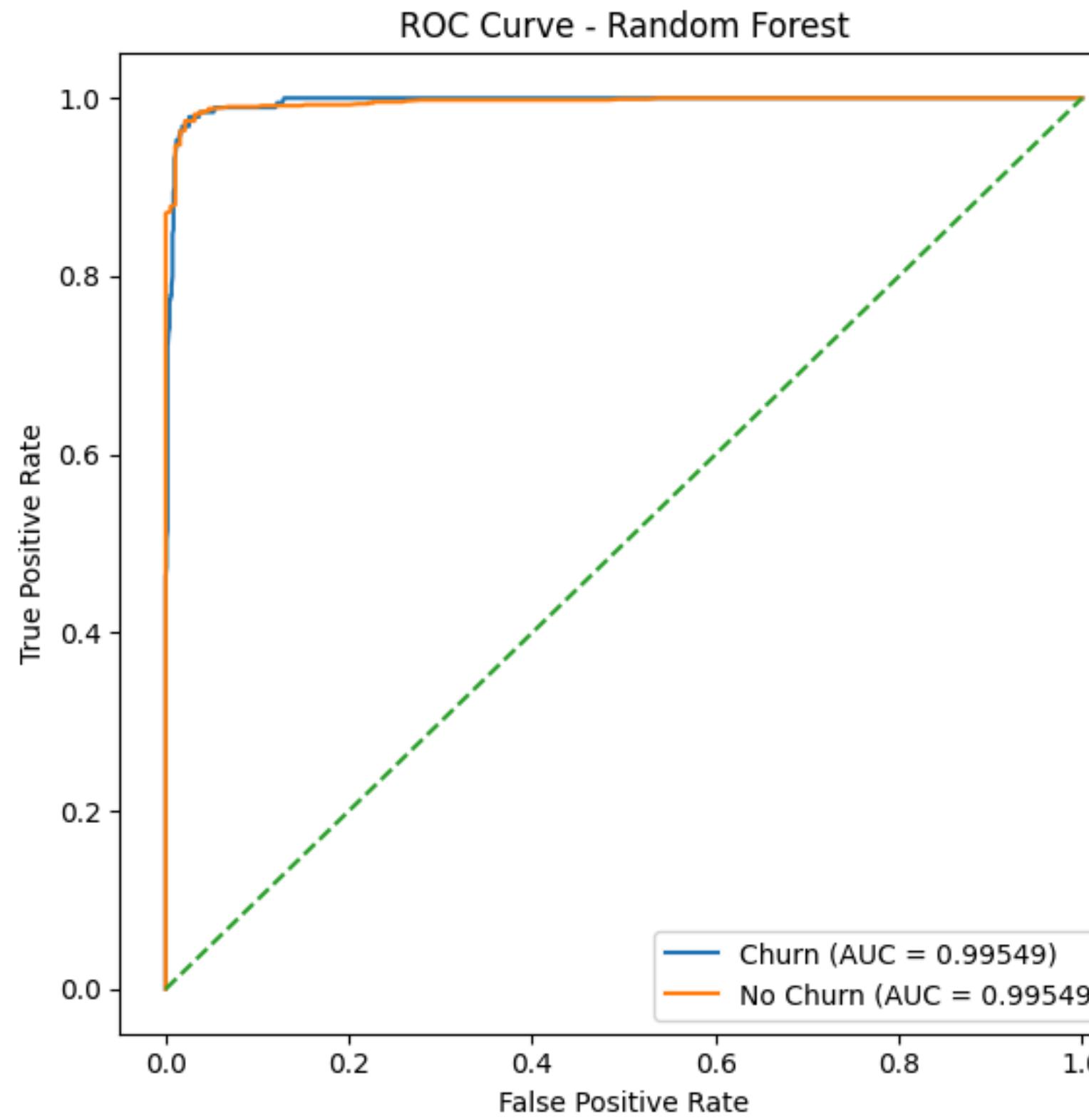
# Kết quả



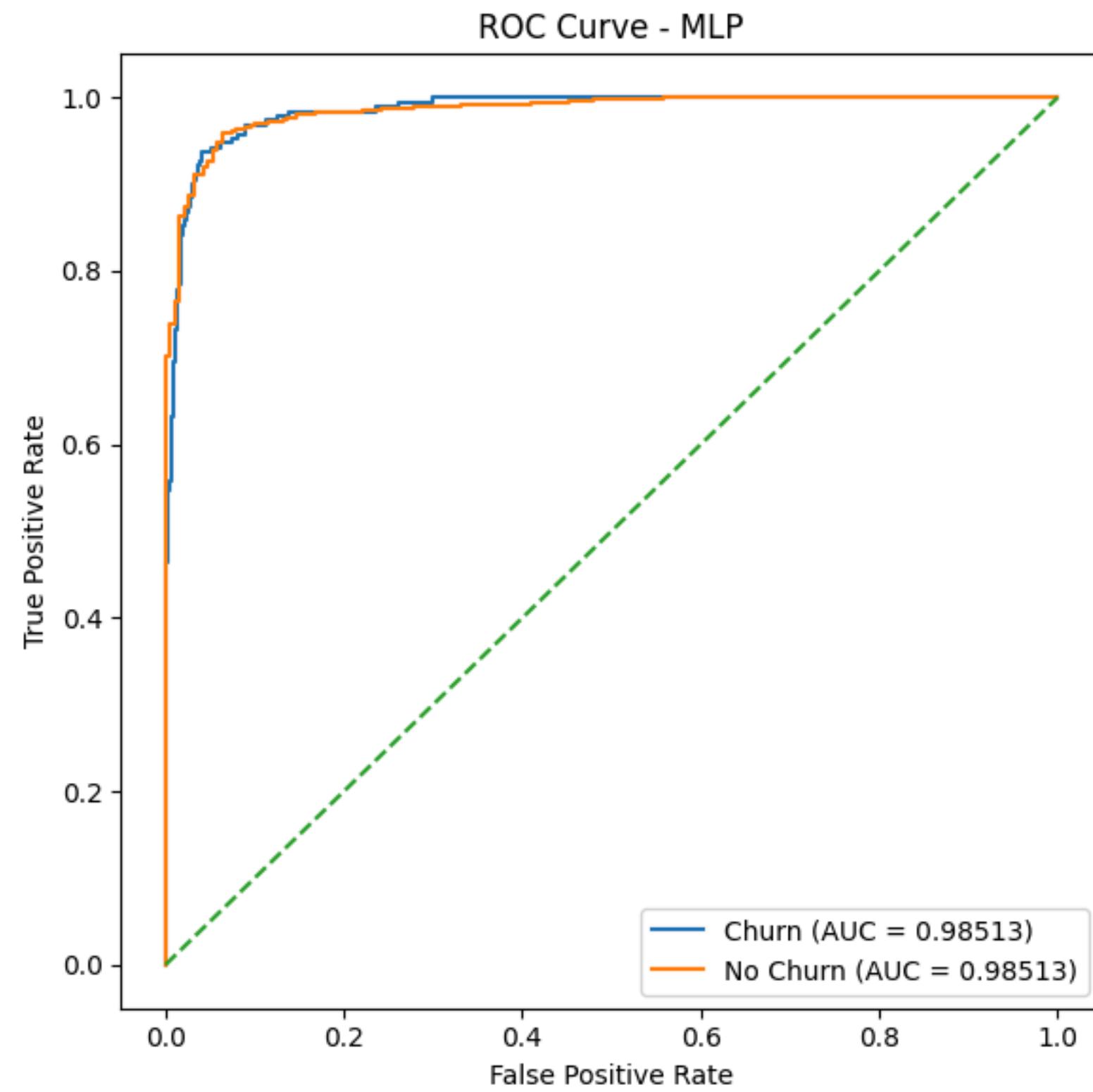
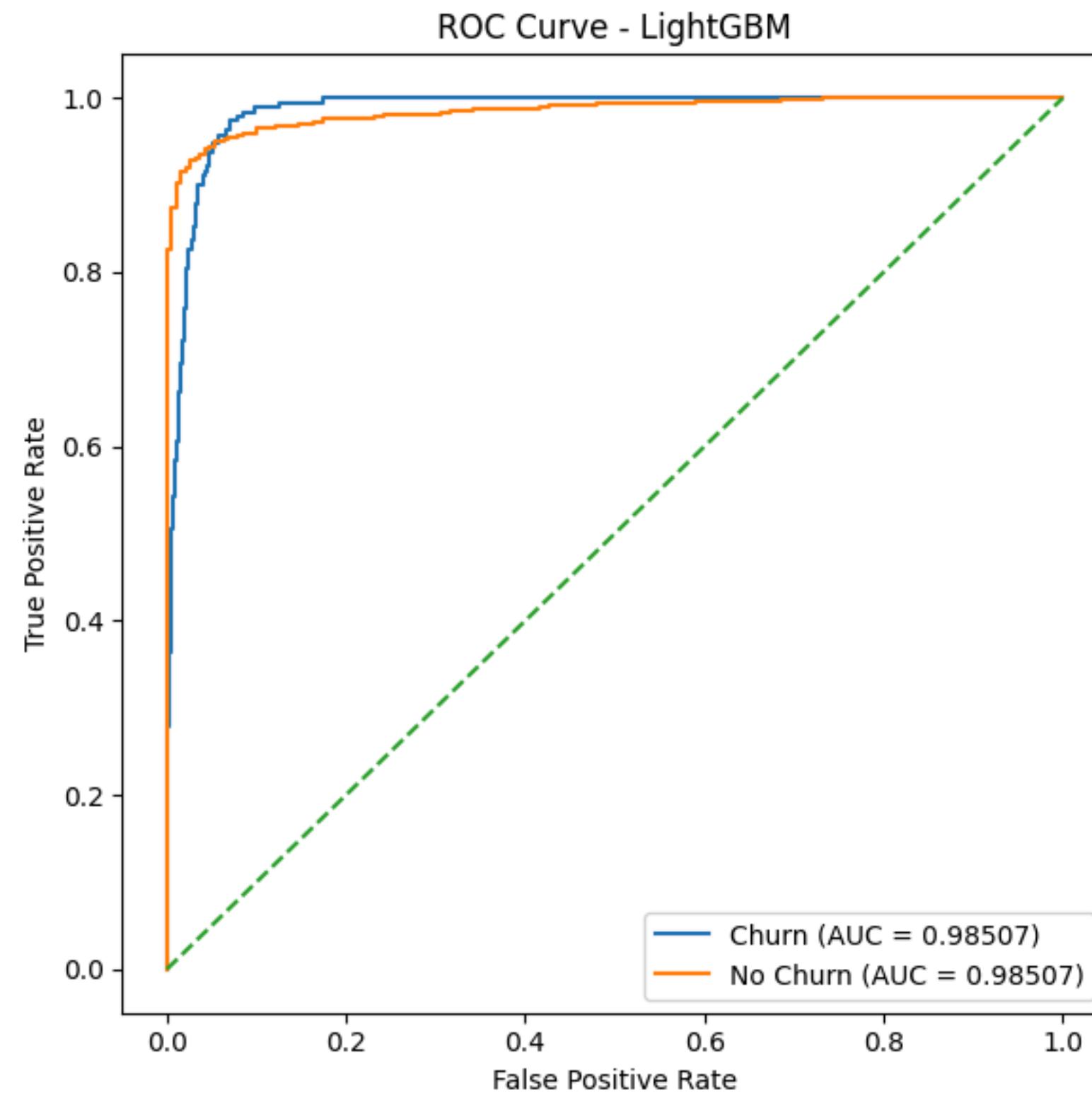
# Kết quả



# Kết quả

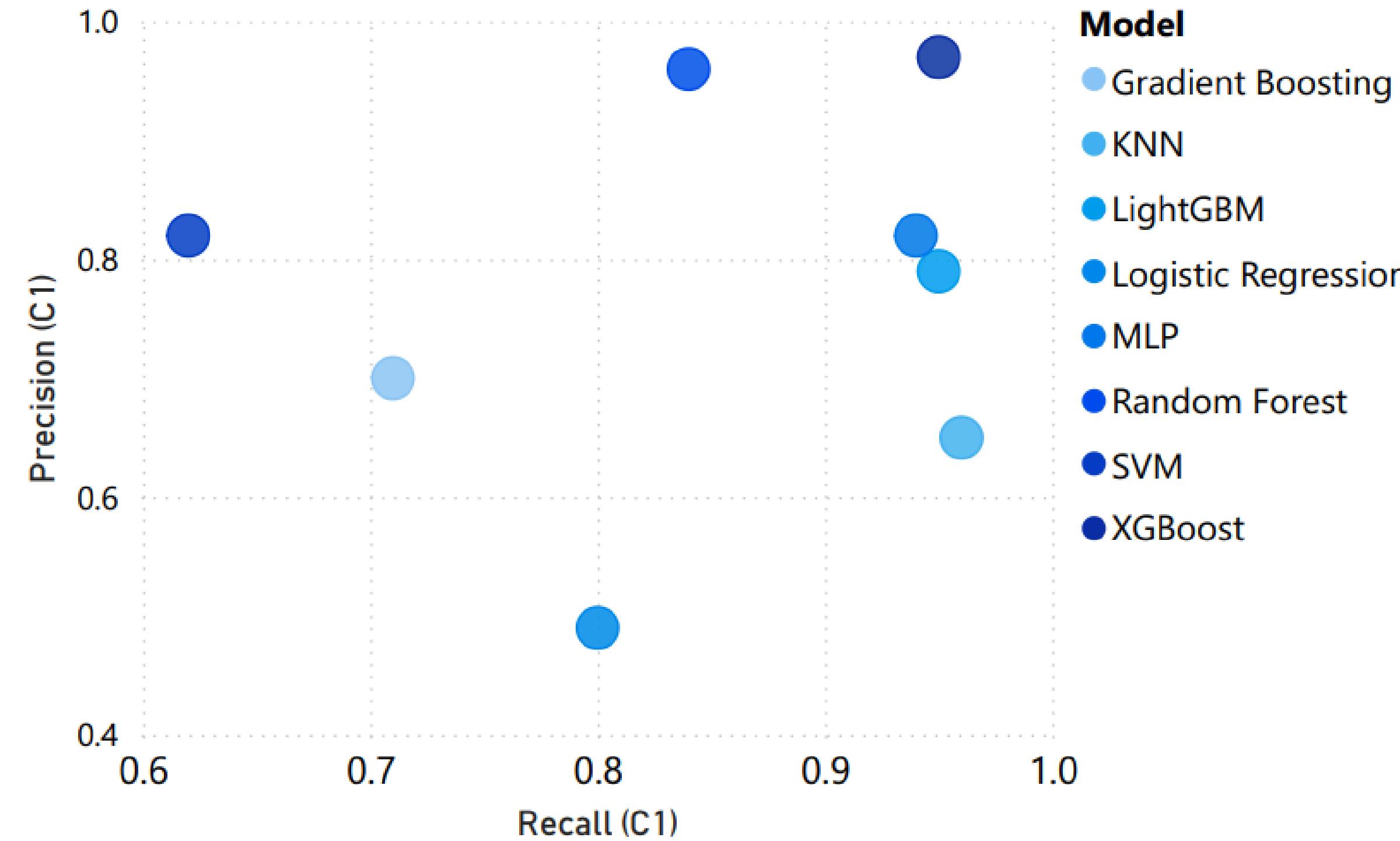


# Kết quả



# Kết quả

## Phân tích đánh đổi Precision-Recall



# Bàn luận

- Mục tiêu là cân bằng Recall (phát hiện khách rời bỏ) và Precision (độ tin cậy cảnh báo); mô hình lý tưởng nằm ở **góc trên bên phải của biểu đồ Precision–Recall**.
- KNN và LightGBM đạt Recall cao ( $\approx 0.96$ ) nhưng Precision thấp (0.65–0.79), tiềm ẩn rủi ro gán nhầm nhãn Churn và lãng phí nguồn lực.
- **Random Forest** và **XGBoost** duy trì Precision cao ( $> 0.95$ ) ngay cả khi dữ liệu mất cân bằng.
- **XGBoost** được lựa chọn là mô hình tối ưu nhất (Precision = 0.97, Recall = 0.95, F1-score = 0.96), vừa phát hiện hiệu quả khách hàng có nguy cơ rời bỏ vừa đảm bảo độ chính xác cao, hỗ trợ tối ưu hóa chiến dịch giữ chân khách hàng.

# Bàn luận

- Nhóm dẫn đầu: XGBoost, Random Forest, LightGBM, MLP
  - Hiệu suất vượt trội thể hiện qua các chỉ số ROC-AUC và F-Score.
  - Kết quả ấn tượng: Random Forest và XGBoost đạt đường cong ROC tiệm cận hoàn hảo với AUC xấp xỉ 0.995.
  - Lý do: Khả năng học các mối quan hệ phi tuyến phức tạp và tương tác giữa các đặc trưng tốt hơn.
- Hạn chế của các mô hình đơn giản (Logistic Regression, KNN)
  - Chỉ thiết lập được ngưỡng hiệu suất cơ sở.
  - Vấn đề "Học vẹt" (Overfitting lớp thiểu số): Mặc dù đạt Recall cao, nhưng Precision rất thấp (0.49 - 0.65).

# 5. Kết luận & Giải pháp

## Kết quả đạt được

- Xây dựng thành công quy trình phân tích và dự đoán Churn toàn diện, đạt được mục tiêu đề ra
- Hệ thống sản phẩm đa dạng: Dashboard Power BI, Data Storytelling (Flourish) và Website Demo tương tác

**XGBoost** được xác định là giải pháp **tối ưu nhất**, với hiệu suất dự đoán xuất sắc, đạt **Accuracy** là **0.99**, **ROC-AUC** là **0.99528** và chỉ số cân bằng **F1-Score** (cho lớp Churn) là **0.96**.

# 5. Kết luận & Giải pháp

## Đề xuất giải pháp cho doanh nghiệp

Xác định nhóm rủi ro cao: Khách hàng mới (0-6 tháng), Nhóm khách hàng độc thân, Thành phố Cấp 1, Khách hàng có trải nghiệm tiêu cực.

### Đề xuất giải pháp chiến lược:

- **Chăm sóc khách hàng mới (0-6 tháng)**
  - Hướng dẫn sử dụng & ưu đãi sớm
  - Thu thập phản hồi định kỳ
  - Ứng dụng mô hình dự đoán để can thiệp cá nhân hóa
- **Phản ứng nhanh với trải nghiệm tiêu cực**
  - Xử lý khiếu nại trong 24h
  - Chủ động chăm sóc khách hàng có mức hài lòng thấp

# 5. Kết luận & Giải pháp

## Đề xuất giải pháp cho doanh nghiệp

Xác định nhóm rủi ro cao: Khách hàng mới (0-6 tháng), Nhóm khách hàng độc thân, Thành phố Cấp 1, Khách hàng có trải nghiệm tiêu cực.

### Đề xuất giải pháp chiến lược:

- **Tối ưu chiến lược tại thành phố cấp 1**
  - Phân tích cạnh tranh, cải thiện vận hành
  - Triển khai chiến dịch marketing địa phương hóa
- **Cá nhân hóa cho nhóm khách hàng độc thân**
  - Thông điệp và ưu đãi theo lối sống
  - Gợi ý sản phẩm & chương trình trung thành theo tần suất

# 6. Hướng phát triển



## Làm giàu dữ liệu

- Tích hợp lịch sử giao dịch chi tiết từng sản phẩm.
- Thu thập dữ liệu tương tác thời gian thực (Real-time) trên Web/App



## Kỹ thuật nâng cao

- **Survival Analysis:** Chuyển từ dự đoán "Có/Không" sang dự đoán "Khi nào" khách hàng rời bỏ.
- **SHAP:** Áp dụng kỹ thuật diễn giải mô hình để hiểu rõ lý do ("Why") cho từng dự đoán.



## Triển khai thực tế

- Tích hợp vào hệ thống **CRM** để tự động hóa cảnh báo rủi ro.
- Thực hiện **A/B Testing** để đo lường hiệu quả thực tế của các chiến dịch giữ chân.

# Tài liệu

- [1] F. F. Reichheld and W. E. Sasser, "Zero defections: Quality comes to services," Harvard Business Review, vol. 68, no. 5, pp. 105–111, 1990.
- [2] P. Kotler and K. L. Keller, Marketing Management, 15th ed. Pearson Education, 2016.
- [3] S. K. Wagh et al., "Customer churn prediction in telecom sector using machine learning techniques," Results in Control and Optimization, 2024.
- [4] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
- [5] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
- [6] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD), 2016, pp. 785–794.
- [7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
- [8] S. Baghla, "Prediction of customer churning in e-commerce applications using machine learning," Procedia Computer Science, vol. 198, pp. 531–538, 2022.
- [9] A. Verma, "E-Commerce Customer Churn Analysis and Prediction," Kaggle.
- [10] J. D. Evans, Straightforward Statistics for the Behavioral Sciences. Pacific Grove, CA, USA: Thomson Brooks/Cole Publishing Co., 1996.
- [11] A. Agresti, An Introduction to Categorical Data Analysis, 2nd ed. Hoboken, NJ, USA: Wiley, 2007, ch. 3.
- [12] H. Cramér, Mathematical Methods of Statistics. Princeton, NJ, USA: Princeton University Press, 1946, pp. 212–214.
- [13] D. W. Hosmer Jr., S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.
- [14] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," The American Statistician, vol. 46, no. 3, pp. 175–185, 1992.
- [15] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
- [16] T. K. Ho, "Random decision forests," in Proc. 3rd Int. Conf. on Document Analysis and Recognition, 1995, pp. 278–282.
- [17] J. H. Friedman, "Stochastic gradient boosting," Computational Statistics & Data Analysis, vol. 38, no. 4, pp. 367–378, 2002.
- [18] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Proc. 31st Int. Conf. on Neural Information Processing Systems (NIPS), 2017, pp. 3146–3154.
- [19] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.
- [20] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Morgan Kaufmann, 2011, ch. 8.
- [21] Microsoft, "Power BI documentation," Microsoft Learn.
- [22] Flourish, "Flourish Studio Homepage."

# THANK YOU