

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

Phạm Huỳnh Tấn Khang

Trường Đại học Công nghệ Thông tin, ĐHQG-HCM  
22520624@gm.uit.edu.vn

Nguyễn Huỳnh Xuân Nghi

Trường Đại học Công nghệ Thông tin, ĐHQG-HCM  
23521004@gm.uit.edu.vn

Huỳnh Ngọc Trang

Trường Đại học Công nghệ Thông tin, ĐHQG-HCM  
22521510@gm.uit.edu.vn

Nguyễn Thị Ngọc Phước

Trường Đại học Công nghệ Thông tin, ĐHQG-HCM  
23521235@gm.uit.edu.vn

**Tóm tắt nội dung**— Báo cáo trình bày nghiên cứu phân tích và dự đoán hành vi rời bỏ khách hàng trong thương mại điện tử dựa trên bộ dữ liệu *Ecommerce Customer Churn Analysis and Prediction*. Dữ liệu được tiền xử lý và xây dựng thêm đặc trưng để phản ánh sâu hơn hành vi của khách hàng. Thông qua phân tích khám phá dữ liệu và trực quan hóa bằng Power BI và Flourish, các yếu tố quan trọng ảnh hưởng đến quyết định rời bỏ được xác định, trong đó nổi bật là thời gian gắn bó, mức độ hài lòng và lịch sử khiếu nại. Tám mô hình phân loại được huấn luyện và đánh giá, bao gồm Logistic Regression, KNN, SVM, Random Forest, Gradient Boosting, XGBoost, LightGBM và MLP, kết hợp các chiến lược xử lý mất cân bằng lớp như class-weight và SMOTE. Kết quả thực nghiệm chỉ ra hiệu suất của các mô hình có sự khác biệt đáng kể, với F1-Score cho lớp rời bỏ dao động từ 0.61 đến 0.96. Trong đó, mô hình XGBoost đạt hiệu quả cao nhất, với Accuracy 0.99, ROC (AUC) 0.99528 và F1-Score cho lớp rời bỏ là 0.96, thể hiện khả năng dự đoán vượt trội và cân bằng. Nghiên cứu cung cấp một công cụ dự báo mạnh mẽ và các insight sâu sắc, hỗ trợ doanh nghiệp xây dựng các chiến lược giữ chân khách hàng chủ động và hiệu quả.

**Từ khóa**—Dự đoán khách hàng rời bỏ, Thương mại điện tử, Học máy, Phân tích dữ liệu, Trực quan hóa dữ liệu.

## I. GIỚI THIỆU

Trong thương mại điện tử, việc duy trì khách hàng hiện tại có vai trò đặc biệt quan trọng do chi phí thu hút khách hàng mới thường cao hơn đáng kể so với chi phí giữ chân [1], [2]. Hiện tượng khách hàng rời bỏ dịch vụ không chỉ làm giảm doanh thu mà còn phản ánh những vấn đề tiềm ẩn liên quan đến trải nghiệm người dùng, chất lượng dịch vụ, chính sách khuyến mãi và khả năng đáp ứng nhu cầu cá nhân hóa của doanh nghiệp. Thách thức này không chỉ tồn tại trong thương mại điện tử mà còn phổ biến ở nhiều lĩnh vực khác, đặc biệt là viễn thông, và đã được nghiên cứu rộng rãi với nhiều kỹ thuật học máy khác nhau [3]. Do đó, dự đoán sớm khả năng rời bỏ của khách hàng là một bài toán có ý nghĩa thực tiễn cao, hỗ trợ doanh nghiệp chủ động xây dựng các chiến lược can thiệp và giữ chân hiệu quả.

Với sự phát triển của các nền tảng thương mại điện tử, lượng dữ liệu hành vi người dùng ngày càng phong phú và đa dạng. Tuy nhiên, dữ liệu thực tế thường tồn tại nhiều thách thức như định dạng không đồng nhất, thiếu hoặc nhiễu, đặc biệt là hiện tượng mất cân bằng lớp khi nhóm khách hàng không rời bỏ chiếm ưu thế so với nhóm rời bỏ. Bên cạnh đó, mối quan hệ giữa các đặc trưng hành vi và khả năng rời bỏ thường mang tính phi tuyến, khiến các phương pháp phân tích truyền thống khó đạt được hiệu quả dự đoán cao.

Đề tài này sử dụng bộ dữ liệu “Ecommerce Customer Churn Analysis and Prediction” để xây dựng một quy trình phân tích và dự đoán churn toàn diện. Quy trình bao gồm các bước tiền xử lý dữ liệu nhằm đảm bảo tính nhất quán về kiểu dữ liệu và định dạng, trích xuất và xây dựng thêm đặc trưng, cũng như phân tích khám phá dữ liệu (EDA) để nhận diện các xu hướng và mối quan hệ quan trọng. Trên cơ sở đó, tiến hành xây dựng và so sánh hiệu suất của tám mô hình học máy khác nhau, gồm: Logistic Regression, KNN, SVM, Random Forest, Gradient Boosting, XGBoost, LightGBM và MLP. Việc huấn luyện và đánh giá các mô hình được thực hiện trong một pipeline thống nhất nhằm đảm bảo tính nhất quán và so sánh công bằng giữa các mô hình.

Mục tiêu của đề tài này bao gồm:

- Xác định các yếu tố chính ảnh hưởng đến hành vi rời bỏ của khách hàng thông qua phân tích thống kê và trực quan hóa dữ liệu.
- Xây dựng và so sánh các mô hình học máy để dự đoán khả năng rời bỏ và lựa chọn mô hình tối ưu.
- Đề xuất giải pháp kinh doanh hỗ trợ chiến lược giữ chân khách hàng cho doanh nghiệp.

Phạm vi đề tài tập trung vào việc áp dụng các kỹ thuật khoa học dữ liệu trên bộ dữ liệu khách hàng thương mại điện tử. Ngôn ngữ Python được sử dụng trong xử lý dữ liệu và xây dựng mô hình. Bên cạnh đó, các công cụ Power BI, Flourish và một website demo được triển khai nhằm hỗ trợ phân tích và trình bày kết quả một cách trực quan.

Báo cáo được tổ chức thành các chương như sau: Chương II trình bày các công trình liên quan trong lĩnh vực dự đoán churn. Chương III mô tả chi tiết phương pháp nghiên cứu. Chương IV trình bày kết quả và bàn luận chi tiết. Cuối cùng, Chương V đưa ra kết luận và đề xuất giải pháp kinh doanh, trong khi Chương VI trình bày các hướng phát triển trong tương lai.

## II. CÁC 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

Bài toán dự đoán khách hàng rời bỏ đã được nghiên cứu rộng rãi trong nhiều lĩnh vực như viễn thông, tài chính và thương mại điện tử. Các nghiên cứu ban đầu thường tiếp cận bài toán này bằng cách phương pháp thống kê và mô hình tuyến tính, trong đó Logistic Regression là lựa chọn phổ biến nhờ tính đơn giản, khả năng diễn giải cao và dễ triển khai trong thực tế [3]. Tuy nhiên, các mô hình này gặp hạn chế khi dữ liệu có quan hệ phi tuyến và tương tác phức tạp giữa các đặc trưng. Do đó, nhiều nghiên cứu đã chuyển sang các mô hình học máy dựa trên cây quyết định và phương pháp tổ hợp. Random Forest được chứng minh là có khả năng cải thiện hiệu năng dự đoán nhờ cơ chế kết hợp nhiều cây quyết định độc lập, giúp giảm hiện tượng overfitting và học tốt hơn các mối quan hệ phi tuyến trong dữ liệu [4]. Bên cạnh đó, các thuật toán boosting như Gradient Boosting, XGBoost và LightGBM cũng được áp dụng rộng rãi do khả năng học tuần tự, tập trung cải thiện các trường hợp dự đoán sai và tối ưu hiệu năng trên dữ liệu thực tế. Các nghiên cứu cho thấy các mô hình boosting thường đạt kết quả vượt trội so với các mô hình đơn lẻ khi đánh giá bằng các chỉ số như F1-score và ROC-AUC [5], [6].

Một thách thức lớn trong dự đoán churn là hiện tượng mất cân bằng lớp, khi nhóm khách hàng không rời bỏ chiếm ưu thế, khiến mô hình dễ thiên lệch và bỏ sót các trường hợp churn quan trọng nếu không được xử lý phù hợp. Để giải quyết vấn đề này, các kỹ thuật tái cân bằng dữ liệu như Synthetic Minority Over-sampling Technique (SMOTE) đã được đề xuất và áp dụng nhằm tăng số lượng mẫu của lớp thiểu số [7]. Ngoài ra, các chiến lược học nhạy chi phí như sử dụng class-weight trong quá trình huấn luyện cũng được sử dụng để tăng mức phạt đối với các lỗi dự đoán sai ở lớp churn.

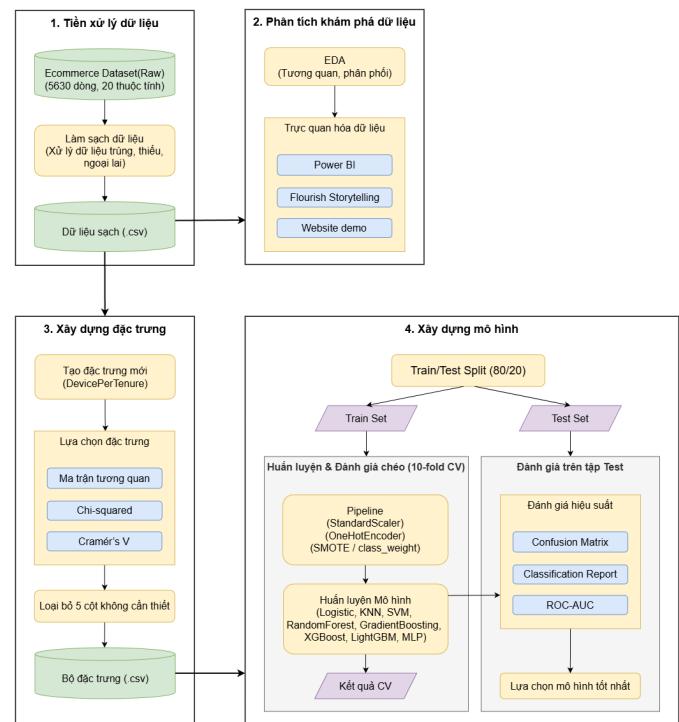
Gần đây, các nghiên cứu cũng nhấn mạnh vai trò của feature engineering trong việc nâng cao hiệu quả dự đoán churn [8].

Việc kết hợp feature engineering với các mô hình học máy hiện đại và chiến lược xử lý mất cân bằng lớp đã trở thành một hướng tiếp cận phổ biến và hiệu quả.

Kết hợp các phương pháp trên, đề tài này tiến hành so sánh một cách có hệ thống các mô hình học máy khác nhau, kết hợp các chiến lược xử lý dữ liệu mất cân bằng và kỹ thuật đặc trưng để tối ưu hóa hiệu suất dự đoán. Điểm mới của đề tài là việc tích hợp kết quả phân tích và dự đoán vào một hệ thống dashboard tương tác, nhằm chuyển hóa các insight phức tạp thành những thông tin trực quan, dễ hiểu và có tính ứng dụng cao cho doanh nghiệp.

## III. PHƯƠNG PHÁP NGHIÊN CỨU

Để minh họa một cách chi tiết và có hệ thống toàn bộ quy trình kỹ thuật đã được thực hiện, Hình 1 dưới đây phác họa luồng xử lý dữ liệu và xây dựng mô hình, từ dữ liệu thô ban đầu cho đến khi lựa chọn được mô hình dự đoán cuối cùng.



Hình 1: Sơ đồ quy trình nghiên cứu tổng thể

### A. Dữ liệu

Đề tài này sử dụng bộ dữ liệu công khai có tên “Ecommerce Customer Churn Analysis and Prediction”, được thu thập từ nền tảng Kaggle [9]. Bộ dữ liệu bao gồm 5630 bản ghi, tương ứng với 5630 khách hàng, và 20 thuộc tính mô tả thông tin nhân khẩu học, hành vi mua sắm, và mức độ tương tác của họ với dịch vụ thương mại điện tử. Mô tả chi tiết được trình bày ở Hình 2.

Biến mục tiêu của bài toán là Churn, một biến nhị phân nhận giá trị 1 nếu khách hàng rời bỏ và 0 nếu ngược lại. Phân tích ban đầu cho thấy bộ dữ liệu có tính chất mất cân bằng rõ rệt, với 948 khách hàng (chiếm 16.8%) được gán nhãn rời

bỏ và 4682 khách hàng (chiếm 83.2%) ở lại. Đặc điểm này là một thách thức quan trọng cần được giải quyết trong quá trình xây dựng mô hình để tránh hiện tượng mô hình bị thiên lệch về lớp đa số.

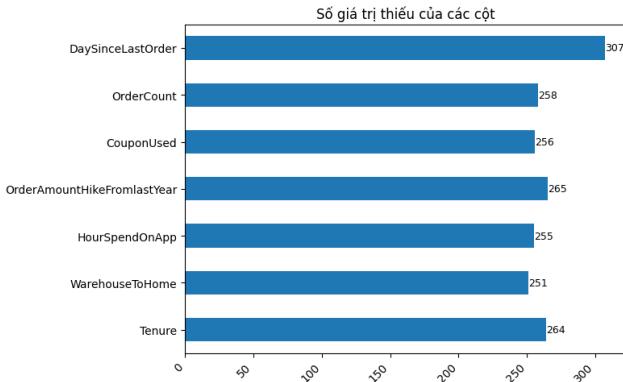
STT	Biến	Mô tả	Giá trị	Kiểu dữ liệu
1	CustomerID	Mã định danh duy nhất của khách hàng		Định tính
2	Churn*	Cờ đánh dấu khách hàng rời bỏ	1: Rời bỏ, 0: Ở lại	Định tính
3	Tenure	Thời gian gắn bó của khách hàng với tổ chức (tháng)	[0, 61]	Định lượng
4	PreferredLoginDevice	Thiết bị đăng nhậpưa thích của khách 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
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 % giá 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

Hình 2: Mô tả Dataset Ecommerce Customer Churn Analysis and Prediction

### B. Tiền xử lý dữ liệu

1) Xử lý dữ liệu không nhất quán: Tiền hành kiểm tra tính nhất quán của các biến hạng mục. Cột CustomerID được loại bỏ do là một biến định danh, không mang lại giá trị dự đoán. Phân tích các giá trị duy nhất cho thấy một số cột tồn tại sự không đồng nhất trong cách nhập liệu:

- 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.

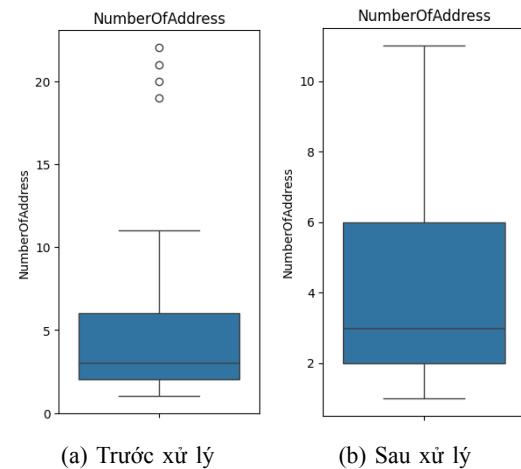


Hình 3: Số giá trị thiểu của các cột

2) Xử lý dữ liệu thiếu và ngoại lai: Phân tích ban đầu ở Hình 3 cho thấy nhiều cột định lượng chứa các giá trị thiểu.

Tiền hành áp dụng các chiến lược điền dữ liệu khác nhau tùy thuộc vào đặc điểm và ý nghĩa của từng biến, đồng thời xử lý các giá trị ngoại lai được phát hiện bằng phương pháp IQR (Interquartile Range) thông qua kỹ thuật capping/winsorizing (thay thế giá trị ngoại lai bằng giá trị ngưỡng trên/dưới).

- Cột Tenure: Dữ liệu thiếu ở cột này được 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.
- Cột WarehouseToHome và HourSpendOnApp: Do phân phối của hai biến này tương đối tập trung và không quá lệch, các giá trị thiếu được điền bằng giá trị trung vị (median).
- Cột CouponUsed và OrderCount: Hai biến này có mối quan hệ logic chặt chẽ ( $\text{CouponUsed} \leq \text{OrderCount}$ ). Do đó, việc điền dữ liệu thiếu được thực hiện một cách có điều kiện để bảo toàn mối quan hệ này. 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.
- Cột 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 giá trị trung vị của cột.
- Các cột khác: Các giá trị thiếu còn lại như ở OrderAmountHikeFromlastYear cũng được điền bằng trung vị. Các ngoại lai ở cột NumberOfAddress được xử lý bằng kỹ thuật capping, minh họa ở Hình 4.



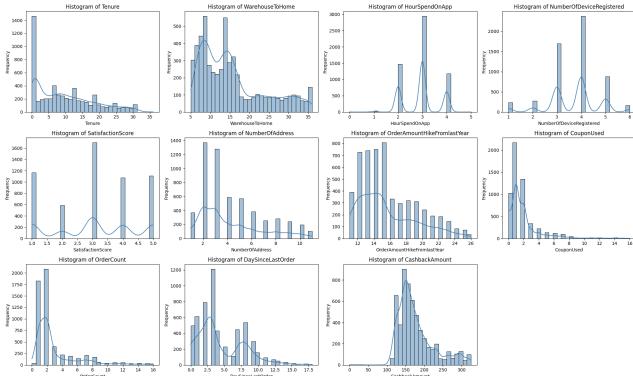
Hình 4: Biến NumberofAddress trước và sau xử lý ngoại lai

Sau khi hoàn tất các bước trên, tập dữ liệu đã được làm sạch hoàn toàn, không còn giá trị thiếu và các giá trị không nhất quán, sẵn sàng cho giai đoạn phân tích và xây dựng mô hình tiếp theo.

### C. Phân tích khám phá dữ liệu (EDA)

Giai đoạn EDA được thực hiện nhằm khám phá phân bố dữ liệu, hành vi khách hàng và xác định các yếu tố liên quan đến Churn.

#### 1) Phân bố các biến định lượng: Hình 5



Hình 5: Biểu đồ histogram và KDE

Phần lớn các biến định lượng không tuân theo phân phối chuẩn, mà chủ yếu có dạng lệch phải. Điều này phản ánh đặc trưng phổ biến của dữ liệu hành vi khách hàng, trong đó chỉ một nhóm nhỏ khách hàng có giá trị rất cao, trong khi đa số tập trung ở mức thấp đến trung bình.

Cụ thể, biến Tenure có mật độ cao ở các giá trị thấp và giảm dần khi thời gian gắn bó tăng, cho thấy phần lớn khách hàng có thời gian sử dụng dịch vụ chưa dài, trong khi chỉ một số ít khách hàng gắn bó lâu dài. Điều này gợi ý Tenure có thể là một yếu tố quan trọng liên quan đến khả năng rời bỏ.

Biến WarehouseToHome có phân bố khá rộng và xuất hiện đuôi dài, phản ánh sự khác biệt đáng kể về khoảng cách giao hàng giữa các nhóm khách hàng. Khoảng cách này có thể ảnh hưởng gián tiếp đến trải nghiệm mua sắm của khách hàng.

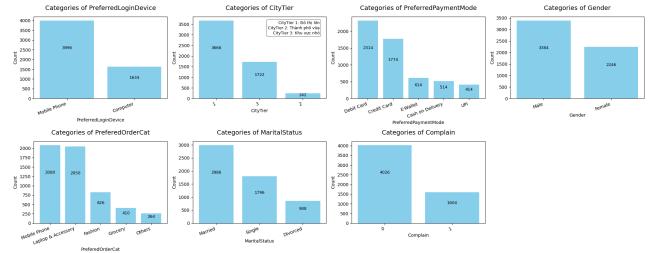
Đối với HourSpendOnApp và NumberofDeviceRegistered, các biểu đồ cho thấy phân bố rời rạc và xuất hiện nhiều đỉnh, phản ánh sự phân hóa rõ rệt trong hành vi sử dụng ứng dụng và thiết bị. Một số khách hàng dành nhiều thời gian trên ứng dụng hoặc sử dụng nhiều thiết bị, trong khi phần lớn chỉ sử dụng ở mức cơ bản.

Biến SatisfactionScore có phân bố theo các mức rời rạc của thang điểm, cho thấy mức độ hài lòng của khách hàng được phân nhóm khá rõ ràng, thay vì phân bố liên tục. Điều này phản ánh sự khác biệt đáng kể trong trải nghiệm giữa các nhóm khách hàng.

Các biến OrderCount, CouponUsed và DaySinceLastOrder đều có phân bố lệch phải mạnh. Phần lớn khách hàng có số đơn hàng thấp, sử dụng ít coupon và có thời gian kể từ lần đặt hàng gần nhất ngắn, trong khi tồn tại một số ít khách hàng có giá trị rất cao. Riêng CashbackAmount có phân bố tương đối ổn định hơn so với các biến hành vi khác, tuy nhiên vẫn tồn tại đuôi phải dài, cho thấy mức độ ưu đãi giữa các khách hàng vẫn có sự chênh lệch.

#### 2) Phân tích các biến định tính: Hình 6

Biểu đồ phân bố cho thấy PreferredLoginDevice chủ yếu là Mobile Phone (3.996 khách hàng), phản ánh nền tảng di động



Hình 6: Biểu đồ cột thê hiện phân bố tần suất biến định tính

là kênh tương tác chính của khách hàng. Điều này cho thấy trải nghiệm trên thiết bị di động đóng vai trò quan trọng và cần được ưu tiên trong các chiến lược tối ưu hóa giao diện và dịch vụ.

Đối với biến CityTier, Khách hàng tập trung chủ yếu tại CityTier 1 (3.666 khách hàng), trong khi CityTier 2 (242 khách hàng) chiếm tỷ lệ rất thấp. Điều này cho thấy dịch vụ hiện đang tiếp cận tốt nhóm khách hàng ở các thành phố lớn, nhưng chưa khai thác hiệu quả nhóm khách hàng ở khu vực trung gian. Sự phân bố không đồng đều theo khu vực có thể ảnh hưởng đến hành vi sử dụng và khả năng rời bỏ của khách hàng.

Về PreferredPaymentMode, các phương thức thanh toán như Debit Card (2.314 khách hàng) và Credit Card (1.774 khách hàng) chiếm ưu thế rõ rệt so với các hình thức khác như E-Wallet (614 khách hàng), Cash on Delivery (514 khách hàng) hay UPI (414 khách hàng). Điều này cho thấy thói quen thanh toán của khách hàng có xu hướng tập trung vào một số phương thức phổ biến, trong khi các hình thức còn lại được sử dụng ở mức hạn chế.

Biến Gender phân bố giới tính tương đối cân bằng với 3.384 nam và 2.246 nữ, cho thấy giới tính không phải là yếu tố phân biệt rõ ràng trong tập dữ liệu và có thể không đóng vai trò quyết định trong hành vi khách hàng.

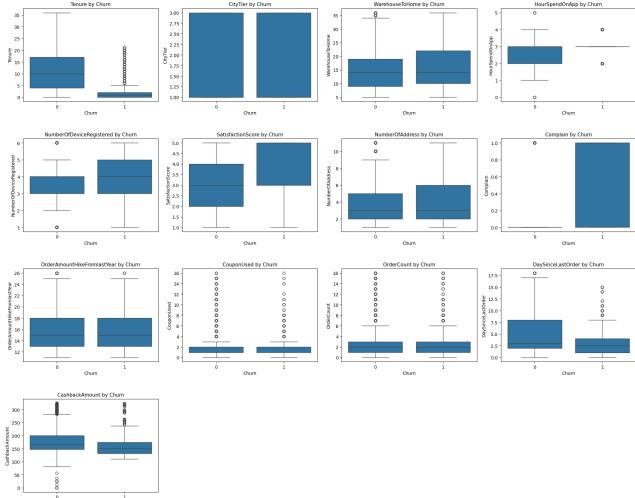
Cuối cùng, đối với PreferredOrderCat, các danh mục Mobile Phone và Laptop & Accessory chiếm tỷ trọng lớn nhất, phản ánh xu hướng tiêu dùng chủ đạo của khách hàng tập trung vào các sản phẩm công nghệ, trong khi các ngành hàng khác chiếm tỷ lệ thấp hơn.

Biểu đồ phân bố cho thấy tập dữ liệu có sự tập trung rõ rệt theo nền tảng, khu vực và Dữ liệu cho thấy sự tập trung rõ rệt theo thiết bị truy cập, khu vực địa lý, phương thức thanh toán và loại sản phẩm, trong khi các yếu tố nhân khẩu học như giới tính ít mang tính phân tách. Những đặc điểm này gợi ý rằng hành vi sử dụng và bối cảnh khách hàng nhiều khả năng ảnh hưởng mạnh mẽ đến churn so với các yếu tố cá nhân.

#### 3) Mối quan hệ giữa các biến và Churn: Hình 7

Phân tích các biểu đồ boxplot theo biến mục tiêu Churn cho thấy một số biến có sự khác biệt rõ ràng giữa hai nhóm khách hàng rời bỏ và không rời bỏ.

Cụ thể, Tenure của nhóm khách hàng không rời bỏ cao hơn đáng kể so với nhóm Churn, cho thấy khách hàng gắn bó lâu dài có xu hướng tiếp tục sử dụng dịch vụ. Tương tự, SatisfactionScore của nhóm Non-Churn cũng cao hơn, phản ánh vai trò quan trọng của mức độ hài lòng trong việc duy trì



Hình 7: Biểu đồ hộp

khách hàng. Biến DaySinceLastOrder có xu hướng cao hơn ở nhóm Churn, cho thấy khách hàng lâu không tương tác hoặc không phát sinh giao dịch dễ có khả năng rời bỏ hơn.

Ngoài ra, CashbackAmount cho thấy nhóm không rời bỏ thường nhận được mức hoàn tiền cao hơn, gợi ý rằng các chương trình ưu đãi và hoàn tiền có thể đóng vai trò tích cực trong việc giữ chân khách hàng. Ngược lại, các biến như CityTier, WarehouseToHome, OrderCount và CouponUsed không thể hiện sự phân tách rõ ràng giữa hai nhóm Churn và Non-Churn, cho thấy vai trò của các biến này trong việc phân biệt hành vi rời bỏ là hạn chế ở giai đoạn phân tích khám phá dữ liệu.

4) *Tổng hợp*: Kết quả EDA cho thấy hành vi sử dụng và trải nghiệm khách hàng là nhóm yếu tố có ảnh hưởng lớn nhất đến Churn như Tenure, SatisfactionScore, DaySinceLastOrder và CashbackAmount, vượt trội so với các yếu tố mang tính đặc điểm chung hay nhân khẩu học. Đồng thời, đặc điểm phân bố lệch và sự tồn tại của nhiều outlier cho thấy dữ liệu phù hợp hơn với các mô hình học máy phi tuyến, đặc biệt là các mô hình dựa trên cây quyết định và phương pháp tổ hợp. Những kết quả này cung cấp cơ sở quan trọng cho việc lựa chọn mô hình và chiến lược xử lý dữ liệu trong các bước phân tích và dự đoán tiếp theo.

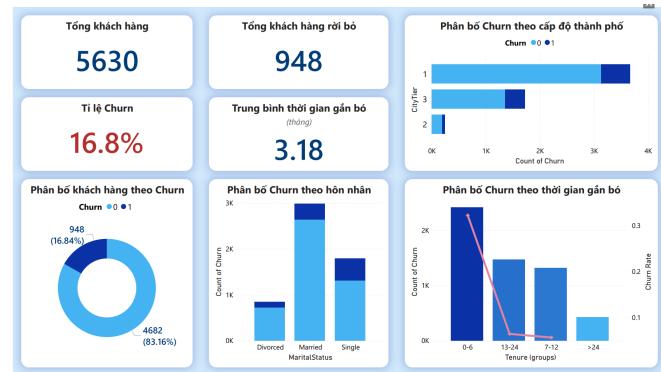
#### D. Trực quan hóa dữ liệu

1) *Dashboard phân tích tương tác với Power BI*: Power BI là nền tảng phân tích và trực quan hóa dữ liệu do Microsoft phát triển, cho phép người dùng kết nối với nhiều nguồn dữ liệu khác nhau, biến dữ liệu thô thành thông tin có thể đọc hiểu, tạo báo cáo và dashboard tương tác phục vụ business intelligence (BI) và ra quyết định kinh doanh [21].

Một bộ dashboard toàn diện đã được xây dựng bằng Power BI để cung cấp một cái nhìn đa chiều về bài toán khách hàng rời bỏ, được thiết kế thành 5 trang báo cáo chuyên biệt, dẫn dắt người dùng đi từ tổng quan đến chi tiết.

a) *Tổng quan (Hình 8)*: Cung cấp một cái nhìn toàn cảnh về tình hình kinh doanh với các chỉ số KPI cốt lõi như tỷ lệ Churn tổng thể và thời gian gắn bó trung bình.

- Biểu đồ “Phân bổ Churn theo thời gian gắn bó” cho thấy tỷ lệ rời bỏ là cao đột biến (trên 30%) ở nhóm khách hàng mới có Tenure từ 0-6 tháng. Sau giai đoạn này, tỷ lệ rời bỏ giảm mạnh. Điều này cho thấy việc giữ chân khách hàng trong giai đoạn 6 tháng đầu là thách thức lớn nhất.
- Cứ 6 khách hàng thì có 1 người rời bỏ. Với tỷ lệ Churn là 16.8%, tương đương 948 khách hàng, vấn đề rời bỏ đang gây ra một tổn thất đáng kể. Thời gian gắn bó trung bình của toàn bộ khách hàng cũng chỉ là 3.18 tháng, cung cấp thêm tính cấp thiết của vấn đề.
- Nhóm khách hàng độc thân có tỷ lệ rời bỏ cao nhất.
- Thành phố cấp 1 là thị trường lớn nhất và có số lượng khách hàng rời bỏ cao nhất về mặt tuyệt đối, cho thấy đây là phân khúc vừa quan trọng, vừa có rủi ro cao về mặt số lượng.

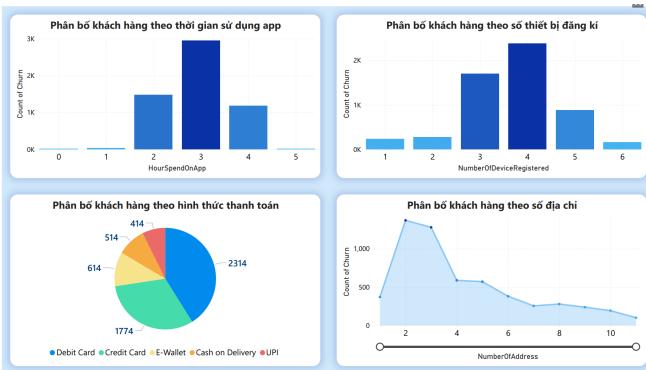


Hình 8: Trang 1 - Tổng quan

b) *Phân tích hành vi khách hàng (Hình 9)*: Tập trung vào việc giải mã các thói quen của khách hàng, trực quan hóa số giờ trên app, số thiết bị đăng ký và hình thức thanh toán, qua đó phác họa nên một bức tranh chung về cách khách hàng sử dụng dịch vụ.

- Phần lớn khách hàng là những người dùng tích cực, thường dành 3 giờ để sử dụng app/website và đã đăng ký 3-4 thiết bị khác nhau. Điều này cho thấy một mức độ gắn kết tương đối cao về mặt công nghệ trong cơ sở khách hàng.
- Thẻ ghi nợ (Debit Card) và thẻ tín dụng (Credit Card) chiếm thị phần áp đảo trong các phương thức thanh toán. Ngược lại, “Cash on Delivery” và “UPI” là hai hình thức ít phổ biến nhất, có thể chỉ ra các nhóm khách hàng đặc thù.
- Phân bố theo số địa chỉ cho thấy phần lớn khách hàng đã đăng ký từ 2 đến 4 địa chỉ, với đỉnh điểm là 3 địa chỉ. Điều này gợi ý rằng khách hàng có xu hướng sử dụng dịch vụ cho nhiều mục đích (ví dụ: nhà riêng, công ty, gửi quà).

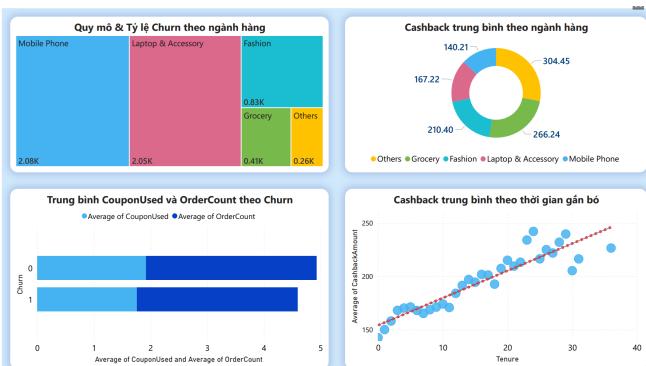
c) *Phân tích mua sắm & giá trị (Hình 10)*: Tập trung vào



Hình 9: Trang 2 - Phân tích hành vi khách hàng

các khía cạnh tài chính, phân tích quy mô và rủi ro của từng ngành hàng, cũng như giá trị cashback mà khách hàng mang lại.

- Một phát hiện đáng ngạc nhiên là số đơn hàng (OrderCount) và coupon sử dụng (CouponUsed) trung bình ở tháng cuối của nhóm rời bỏ và nhóm ở lại gần như không có sự khác biệt, cho thấy quyết định rời bỏ có thể đến từ các yếu tố trải nghiệm lâu dài hơn là sự sụt giảm hoạt động tức thời.
- Biểu đồ phân tán cho thấy một mối tương quan dương rất rõ ràng: thời gian gắn bó (Tenure) càng dài, lượng cashback trung bình khách hàng nhận được càng cao. Điều này khẳng định chính sách của công ty đang có xu hướng tương thường cho những khách hàng trung thành.
- Lượng cashback trung bình có sự chênh lệch lớn giữa các ngành hàng, với nhóm “Others” và “Grocery” nhận được giá trị cashback cao nhất, trong khi “Mobile Phone” có mức thấp nhất.



Hình 10: Trang 3 - Phân tích mua sắm & giá trị

- d) *Phân tích trải nghiệm khách hàng (Hình 11):* Khám phá các nguyên nhân gốc rễ dẫn đến Churn thông qua việc phân tích điểm hài lòng và lịch sử khiếu nại.

- Biểu đồ “Phân bố Churn theo khiếu nại” cho thấy tỷ lệ rời bỏ ở nhóm khách hàng có khiếu nại (Complain=1) cao hơn đáng kể so với nhóm không khiếu nại.

- Có một mối quan hệ ngược chiều rõ ràng: điểm hài lòng càng thấp, tỷ lệ rời bỏ càng cao. Nhóm khách hàng cho điểm 1, 2, và đặc biệt là 3 (nhóm lớn nhất) là những phân khúc có nguy cơ cao nhất.

- Với điểm hài lòng trung bình chỉ là 3.07 (trên thang 5) và có đến gần 30% khách hàng đã từng khiếu nại, điều này cho thấy có những vấn đề tiềm ẩn về chất lượng dịch vụ hoặc sản phẩm cần được xem xét và cải thiện.



Hình 11: Trang 4 - Phân tích trải nghiệm khách hàng

- e) *So sánh hiệu suất mô hình (Hình 12):* Dùng để tổng kết và trực quan hóa các kết quả đã được bàn luận chi tiết ở chương IV.



Hình 12: Trang 5 - So sánh hiệu suất mô hình

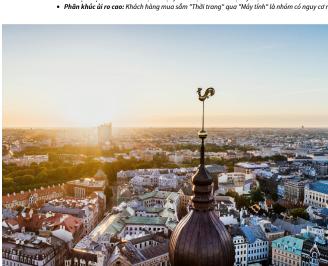
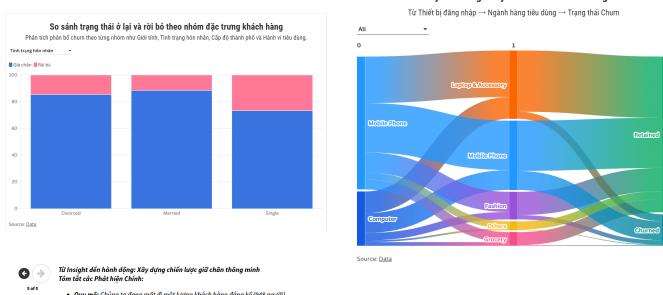
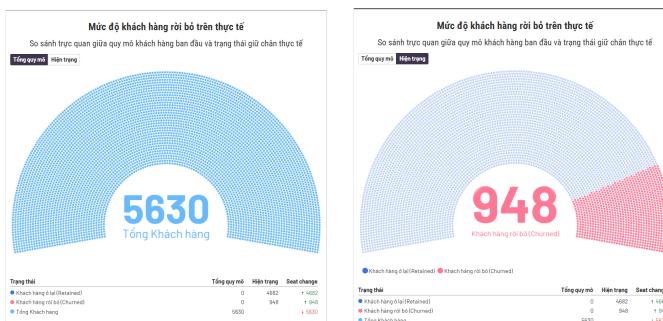
2) *Kể chuyện dữ liệu với Flourish:* Flourish Studio là một nền tảng trực quan hóa dữ liệu dựa trên nền tảng web, cho phép người dùng dễ dàng chuyển đổi dữ liệu thành các đồ thị, bản đồ, infographics và những câu chuyện dữ liệu (data storytelling) có tính tương tác cao mà không cần viết mã lập trình [22].

Để làm nổi bật các phát hiện quan trọng nhất một cách sinh động và hấp dẫn, đê tài đã sử dụng công cụ Flourish để xây dựng một câu chuyện dữ liệu như Hình 13. Câu chuyện này bao gồm các biểu đồ tương tác cao, tập trung vào việc truyền tải thông điệp một cách súc tích:

- Mức độ khách hàng rời bỏ trên thực tế:* Sử dụng biểu đồ Parliament Chart để cụ thể hóa con số 16.8% Churn

thành 948 khách hàng thực tế, nhán mạnh quy mô và tác động của vân đê.

- 2) *So sánh trạng thái theo đặc trưng:* Dùng biểu đồ 100% Stacked Bar Chart để làm nổi bật sự khác biệt về tỷ lệ Churn giữa các nhóm khách hàng, đặc biệt là nhóm có/không có khiếu nại.
- 3) *Truy vết dòng chảy rời bỏ:* Áp dụng biểu đồ Sankey Diagram để trực quan hóa “hành trình” của khách hàng từ các đặc điểm ban đầu (thiết bị, ngành hàng) đến trạng thái cuối cùng (Churned/Retained), giúp xác định các “lộ trình rời bỏ” phổ biến nhất.



### Đề xuất hành động

1. Tạo trang web và ứng dụng di động: Thiết lập ứng dụng phân tích hành động và ứng dụng di động để thu thập dữ liệu và phân tích hành động của khách hàng.
2. Tích hợp AI vào quy trình bán hàng: Sử dụng AI để phân tích hành vi mua sắm và cung cấp đề xuất sản phẩm.
3. Tạo ứng dụng quản lý khách hàng: Sử dụng ứng dụng quản lý khách hàng để quản lý thông tin khách hàng và cung cấp dịch vụ.

3) *Website demo phân tích dữ liệu:* Để nâng cao tính ứng dụng và khả năng tương tác, một sản phẩm demo dưới dạng ứng dụng web đã được phát triển bằng React và TypeScript.

Ứng dụng này phục vụ như một công cụ phân tích linh hoạt, cho phép người dùng cuối có thể tự mình khám phá dữ liệu.

Các chức năng chính của website bao gồm:

- *Tải và xử lý dữ liệu:* Người dùng có thể tải lên tệp dữ liệu CSV của riêng mình. Toàn bộ quá trình xử lý và tổng hợp dữ liệu được thực hiện trực tiếp trên trình duyệt (client-side), đảm bảo tính bảo mật và tốc độ.
- *Tùy biến biểu đồ:* Ứng dụng cung cấp nhiều loại biểu đồ tương tác (Bar, Scatter, Doughnut, Stacked Bar, Radar). Người dùng có thể dễ dàng lựa chọn loại biểu đồ, trục X, trục Y để tạo ra các phân tích theo ý muốn.
- *Tích hợp Data Storytelling:* Website tích hợp các câu chuyện dữ liệu đã được xây dựng từ Flourish, kết hợp sức mạnh của phân tích dashboard và kể chuyện dữ liệu trong cùng một nền tảng.

(a) Trang thông tin website

(b) Trang phân tích dữ liệu

(c) Trang trực quan từ Flourish

Hình 14: Giao diện chính của Website Demo

Sản phẩm demo này không chỉ là một công cụ trình diễn kết quả mà còn là một minh chứng cho khả năng triển khai

các giải pháp phân tích dữ liệu thành một ứng dụng thực tế, dễ tiếp cận và sử dụng.

### E. Xây dựng đặc trưng

Sau khi dữ liệu đã được tiền xử lý và khám phá, tiến hành các bước để xây dựng đặc trưng (Feature Engineering) nhằm tối ưu hóa bộ dữ liệu đầu vào cho các mô hình học máy. Mục tiêu của giai đoạn này là tạo ra biến mới có sức mạnh dự đoán cao hơn và loại bỏ các biến không cần thiết hoặc gây nhiễu.

1) *Tạo đặc trưng mới:* Để nắm bắt sâu hơn mức độ gắn kết của khách hàng với hệ sinh thái công nghệ của doanh nghiệp, một đặc trưng mới có tên DevicePerTenure đã được tạo ra. Đặc trưng này được tính bằng công thức:

$$\text{DevicePerTenure} = \frac{\text{NumberOfDeviceRegistered}}{\text{Tenure}} \quad (1)$$

Biến DevicePerTenure đo lường tốc độ đăng ký thiết bị trung bình mỗi tháng của khách hàng. Một giá trị cao cho thấy khách hàng nhanh chóng tiếp nhận và tích hợp dịch vụ trên nhiều thiết bị ngay từ khi bắt đầu, đây có thể là một chỉ báo sớm về mức độ cam kết và tiềm năng trở thành khách hàng trung thành. Ngược lại, một giá trị thấp có thể chỉ ra một người dùng ít tương tác hơn.

2) *Lựa chọn đặc trưng:* Để giảm thiểu hiện tượng đa cộng tuyến và loại bỏ các đặc trưng có ít giá trị dự đoán, nghiên cứu đã sử dụng kết hợp các phương pháp thống kê cho cả biến định lượng và định tính.

Một ma trận tương quan như Hình 15 đã được xây dựng để kiểm tra mối quan hệ tuyến tính giữa các biến định lượng. Nguồn 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.* Phân tích cho thấy một số cặp biến có mức độ tương quan cao, đáng chú ý là:

- OrderCount và CouponUsed ( $r = 0.78$ ): Tương quan dương mạnh, cho thấy hai biến này cung cấp thông tin tương tự nhau về hành vi mua sắm gần đây.
- Tenure và đặc trưng mới DevicePerTenure ( $r = -0.70$ ): Tương quan âm mạnh. Điều này là hợp lý về mặt logic, tuy nhiên, phân tích sâu hơn cho thấy DevicePerTenure có tương quan mạnh hơn với biến mục tiêu Churn.

Đối với các biến định tính, sử dụng hai phương pháp hỗ trợ cho nhau [11], [12]:

- *Kiểm định Chi-squared:* Kết quả cho thấy tất cả các biến định tính đều có mối liên hệ có ý nghĩa thống kê với biến Churn ( $p\text{-value} < 0.05$ ). Tuy nhiên, kiểm định này không đo lường được “sức mạnh” của mối liên hệ đó.
- *Hệ số Cramér's V:* Dùng để định lượng sức mạnh của mối liên hệ. Nguồn Cramér's V tham khảo:



Hình 15: Ma trận tương quan

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

Kết quả cho thấy:

- PreferredOrderCat (0.226) và MaritalStatus (0.183) có mức độ liên quan yếu.
- PreferredLoginDevice (0.051), Gender (0.029), và PreferredPaymentMode (0.096) có mức độ liên quan rất yếu, cho thấy chúng không có giá trị đáng kể trong việc dự đoán hành vi rời bỏ.

Qua các phân tích trên, nghiên cứu quyết định loại bỏ 5 đặc trưng sau đây khỏi bộ dữ liệu huấn luyện cuối cùng:

- OrderCount và Tenure: Bị loại bỏ để giải quyết vấn đề đa cộng tuyến. DevicePerTenure được giữ lại thay cho Tenure do có sức mạnh dự đoán cao hơn.
- PreferredLoginDevice, PreferredPaymentMode, Gender: Bị loại bỏ do có mức độ liên quan thống kê rất thấp với biến mục tiêu Churn, nhằm giảm nhiễu và độ phức tạp cho mô hình.

Quá trình này giúp tạo ra một bộ dữ liệu cuối cùng tinh gọn, tập trung vào các đặc trưng có giá trị dự đoán cao nhất, làm nền tảng vững chắc cho việc xây dựng các mô hình học máy hiệu quả.

### F. Xây dựng mô hình

Để đảm bảo tính nhất quán và khả năng tái lập, một quy trình chuẩn hóa đã được thiết lập bằng cách sử dụng Pipeline để tự động hóa các bước tiền xử lý và huấn luyện mô hình.

- *Phân chia dữ liệu:* tập huấn luyện (80%) và tập kiểm tra (20%) được chia bằng phương pháp train\_test\_split. Kỹ thuật phân chia có phân tầng (stratify=y) được áp dụng để đảm bảo tỷ lệ Churn và Non-Churn trong cả hai tập dữ liệu là tương đồng với tập dữ liệu gốc.
- *Tiền xử lý trong Pipeline:* Một ColumnTransformer được tích hợp vào pipeline để xử lý các loại biến khác nhau. Các biến định lượng được chuẩn hóa bằng StandardScaler để đưa về cùng một thang đo (phân phối chuẩn với trung bình 0 và phương sai 1). Các biến định tính được mã hóa bằng phương pháp OneHotEncoder.

- *Xử lý mất cân bằng:*

- Trọng số lớp (class\_weight): Kỹ thuật này tự động điều chỉnh trọng số của các lớp trong hàm mất mát, khiến mô hình phạt nặng hơn đối với các lỗi dự đoán sai trên lớp thiểu số (lớp Churn).
- SMOTE: Đối với các mô hình không hỗ trợ class\_weight một cách trực tiếp (như KNN, Gradient Boosting, MLP), kỹ thuật SMOTE được tích hợp vào pipeline. SMOTE hoạt động bằng cách tạo ra các mẫu dữ liệu tổng hợp mới cho lớp thiểu số, giúp cân bằng phân phối lớp trước khi đưa vào huấn luyện.

Mỗi mô hình phân loại đã được xây dựng đại diện cho một hướng tiếp cận khác nhau trong học máy. Các tham số chính của mỗi mô hình được lựa chọn dựa trên các thực nghiệm ban đầu và các khuyến nghị phổ biến trong tài liệu.

1) *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 của một sự kiện (Churn) xảy ra dựa trên một tổ hợp tuyến tính của các biến đầu vào [13]. Cấu hình: class\_weight='balanced', max\_iter=1000.

2) *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 [14]. Cấu hình: Các siêu tham số được tìm kiếm tự động thông qua GridSearch (kết quả cho thấy k tối ưu là 9). Dữ liệu đầu vào được cân bằng bằng SMOTE.

3) *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. Nghiên cứu sử dụng kernel Radial Basis Function (RBF) để xử lý các mối quan hệ phi tuyến [15]. Cấu hình: kernel='rbf', probability=True, class\_weight='balanced'.

4) *Random Forest:* Là một thuật toán học tổ hợp theo phương pháp bagging. Mô hình này xây dựng nhiều cây quyết định độc lập và đưa ra dự đoán dựa trên kết quả bỏ phiếu của đa số cây, giúp giảm phương sai và chống lại hiện tượng quá khớp (overfitting) [16]. Cấu hình: n\_estimators=300, class\_weight='balanced'.

5) *Gradient Boosting:* Là một thuật toán tổ hợp theo phương pháp boosting, xây dựng các cây quyết định một cách tuần tự. Mỗi cây mới được huấn luyện để sửa lỗi của cây trước đó, giúp mô hình dần dần cải thiện hiệu suất [17]. Cấu hình: n\_estimators=100, learning\_rate=0.1, max\_depth=3. Dữ liệu đầu vào được cân bằng bằng SMOTE.

6) *XGBoost (Extreme Gradient Boosting):* Là một phiên bản tối ưu hóa và hiệu quả cao của Gradient Boosting, được biết đến với tốc độ xử lý nhanh và hiệu suất vượt trội nhờ các kỹ thuật như song song hóa và điều chỉnh (regularization) [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.

7) *LightGBM (Light Gradient Boosting Machine):* Là một thuật toán boosting khác, sử dụng kỹ thuật gradient-based one-side sampling (GOSS) và exclusive feature bundling (EFB) để tăng tốc độ huấn luyện và giảm bộ nhớ sử dụng, đặc biệt hiệu quả trên các bộ dữ liệu lớn [18]. Cấu

hình: n\_estimators=300, learning\_rate=0.05, num\_leaves=20, max\_depth=5, class\_weight='balanced'.

8) *MLP (Multi-layer Perceptron):* Là một mô hình mạng nơ-ron nhân tạo truyền thống. Mô hình này bao gồm nhiều lớp nơ-ron được kết nối với nhau, có khả năng học các mối quan hệ phi tuyến phức tạp trong dữ liệu [19]. Cấu hình: mạng nơ-ron có 2 lớp ẩn với (32, 16) nơ-ron tương ứng, 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 SMOTE.

Hiệu suất của mỗi mô hình trên tập huấn luyện được đánh giá một cách khách quan bằng kỹ thuật kiểm định chéo phân tầng 10 lần (10-fold Stratified Cross-Validation). Kỹ thuật này chia tập huấn luyện thành 10 phần, lần lượt sử dụng 9 phần để huấn luyện và 1 phần để kiểm định, sau đó lấy kết quả trung bình. Điều này giúp đảm bảo rằng hiệu suất được đánh giá là ổn định và không phụ thuộc vào một lần chia dữ liệu ngẫu nhiên duy nhất.

#### G. Độ đo đánh giá

Việc lựa chọn các độ đo đánh giá phù hợp là yếu tố then chốt để so sánh và lựa chọn mô hình một cách khách quan, đặc biệt đối với bài toán phân loại có dữ liệu mất cân bằng. Nghiên cứu này sử dụng một bộ các độ đo toàn diện để đánh giá hiệu suất các mô hình từ nhiều khía cạnh khác nhau [20].

1) *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 (Churn và Non-Churn), ma trận này có dạng 2x2, bao gồm bốn giá trị cột lối:

- True Positives (TP): Số trường hợp khách hàng được dự đoán là Churn và thực tế họ Churn.
- True Negatives (TN): Số trường hợp khách hàng được dự đoán là Non-Churn và thực tế họ Non-Churn.
- False Positives (FP): Số trường hợp khách hàng được dự đoán là Churn nhưng thực tế họ Non-Churn.
- False Negatives (FN): Số trường hợp khách hàng được dự đoán là Non-Churn nhưng thực tế họ Churn.

Trong bài toán dự đoán Churn, việc giảm thiểu False Negatives (FN), 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.

2) *Classification Report:* Từ các giá trị trong ma trận nhầm lẫn, một loạt các chỉ số hiệu suất được tính toán và trình bày trong báo cáo phân loại:

- Accuracy (Độ chính xác): Tỷ lệ tổng số dự đoán đúng trên tổng số mẫu.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

Lưu ý: Trong dữ liệu mất cân bằng, Accuracy không phải thước đo đáng tin cậy, vì mô hình chỉ dự đoán lớp đa số cũng có thể đạt Accuracy cao.

- Precision (Độ chuẩn xác): Mức độ tin cậy của các dự đoán dương. Đối với lớp Churn:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

- Recall (Độ phủ / Độ nhạy): Khả năng của mô hình phát hiện tất cả các trường hợp dương. Đối với lớp Churn:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4)$$

Chỉ số này rất quan trọng vì Recall cao nghĩa là ít bỏ sót khách hàng có nguy cơ rời bỏ.

- F1-Score: Trung bình điều hòa của Precision và Recall, cung cấp chỉ số cân bằng:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

### 3) 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 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.

## IV. KẾT QUẢ & BÀN LUẬN

### A. Hiệu suất các mô hình

Tất cả mô hình đã được huấn luyện và đánh giá trên tập kiểm tra dựa trên các độ đo đã được định nghĩa. Bảng II tóm tắt kết quả chi tiết của từng mô hình.

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

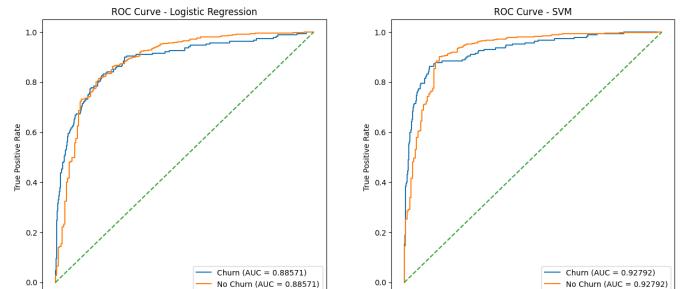
Giá trị tốt nhất được in đậm.

Để trực quan hóa khả năng phân loại tổng thể, Hình 16 so sánh đường cong ROC của bốn mô hình tiêu biểu: một mô hình cơ sở (Logistic Regression), một mô hình phi tuyến kinh điển (SVM), và hai mô hình ensemble hàng đầu (Random Forest và XGBoost).

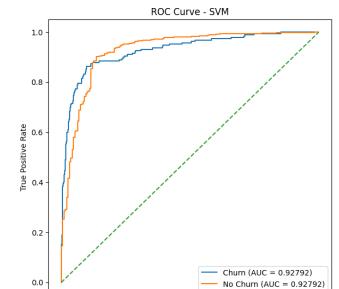
### B. Bàn luận

Từ kết quả trên, có thể rút ra một số nhận xét quan trọng:

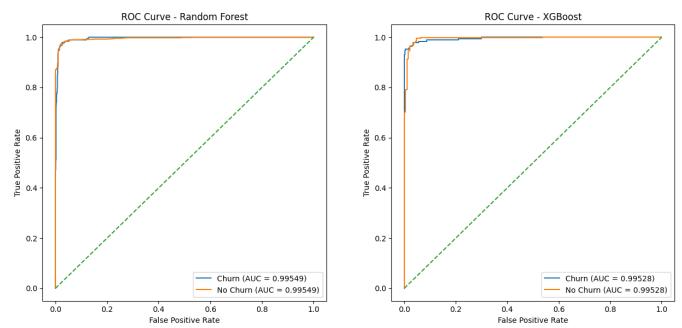
- Kết quả cho thấy một sự phân cấp rõ ràng về hiệu suất, trong đó các mô hình ensemble và boosting như XGBoost, Random Forest, LightGBM và MLP thể hiện



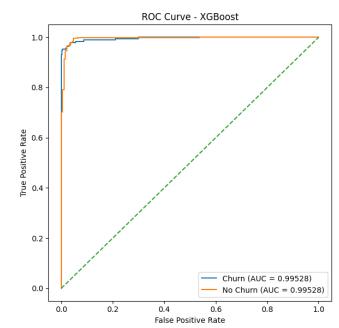
(a) Logistic Regression



(b) SVM



(c) Random Forest



(d) XGBoost

Hình 16: So sánh đường cong ROC của 4 mô hình tiêu biểu  
(Chú thích: Biểu đồ này so sánh khả năng phân loại tổng thể của bốn mô hình đại diện. Đường chéo châm châm thể hiện mô hình phân loại ngẫu nhiên (AUC=0.5))

sự thống trị hoàn toàn. Điều này được minh chứng qua cả bảng chỉ số, nơi chúng dẫn đầu về ROC-AUC và F-Score, lẫn qua phân tích trực quan. Như thể hiện trên đường cong ROC (Hình 16), trong khi Logistic Regression chỉ xác lập một ngưỡng hiệu suất cơ sở, thì Random Forest và XGBoost có các đường cong gần như tiệm cận điểm phân loại hoàn hảo (AUC ≈ 0.995). Sự vượt trội này khẳng định rằng khả năng học các mối quan hệ phi tuyến phức tạp và tương tác giữa các đặc trưng của các thuật toán ensemble là yếu tố then chốt để giải quyết hiệu quả bài toán này.

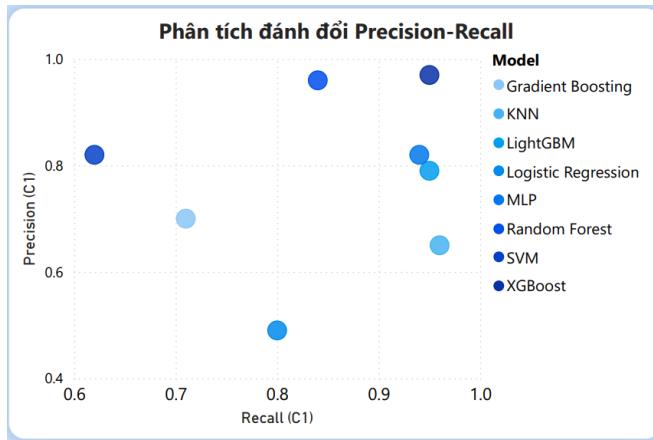
- Mặc dù tất cả các mô hình đều đã được áp dụng các kỹ thuật xử lý dữ liệu mất cân bằng, hiệu quả của chúng lại phụ thuộc nhiều vào bản chất của thuật toán.

- Các mô hình đơn giản hơn như Logistic Regression và KNN, dù đạt được Recall cao, nhưng phải đánh đổi bằng Precision rất thấp (0.49 và 0.65). Điều này cho thấy chúng có xu hướng “học vẹt” lớp thiểu số được tăng cường, dẫn đến các dự đoán kém tin cậy.

- Ngược lại, các mô hình ensemble và boosting lại tận dụng rất tốt lợi thế từ các kỹ thuật này. Chúng không chỉ cải thiện khả năng phát hiện lớp Churn (đạt Recall cao) mà còn giữ được Precision ở mức xuất sắc (trên 0.95 với XGBoost và Random Forest). Điều này chứng tỏ bản chất mạnh mẽ của chúng trong việc học các ranh giới quyết định phức tạp, giúp chúng không

bị “đánh lừa” bởi các mẫu tổng hợp hoặc việc điều chỉnh trọng số.

Để lựa chọn mô hình phù hợp nhất, việc đánh giá sự cân bằng giữa Precision và Recall cho lớp Churn là cực kỳ quan trọng. Biểu đồ phân tán ở Hình 17, được trích xuất từ dashboard phân tích hiệu suất mô hình (mục III-D), trực quan hóa sự đánh đổi này.



Hình 17: Biểu đồ đánh đổi Precision-Recall của 8 mô hình

Trong bài toán dự đoán Churn, một mô hình lý tưởng cần đạt cả hai yếu tố: phát hiện nhiều khách hàng sắp rời bỏ (Recall cao) và đảm bảo dự đoán đáng tin cậy (Precision cao). Vị trí lý tưởng trên biểu đồ là ở góc trên cùng bên phải.

Quan sát Hình 17, có thể thấy rõ các điểm đại diện cho XGBoost và Random Forest nằm ở vị trí vượt trội so với các mô hình khác. Trong khi đó, các mô hình như KNN và LightGBM, mặc dù có Recall rất cao (gần 0.96), nhưng Precision thấp hơn đáng kể (0.65 và 0.79), nghĩa là xác định nhiều khách hàng Churn nhưng có thể gán nhầm nhau “Churn” cho khách hàng trung thành, gây lãng phí nguồn lực.

Mô hình XGBoost thể hiện sự cân bằng tốt nhất với Precision = 0.97 và Recall = 0.95. Điều này mang lại F1-Score = 0.96, cao nhất trong số các mô hình được thử nghiệm. Do đó, nghiên cứu đề xuất XGBoost là mô hình tối ưu nhất, vì vừa phát hiện phần lớn khách hàng có nguy cơ rời bỏ, vừa đảm bảo tính chính xác của cảnh báo.

## V. KẾT LUẬN

### A. Kết luận

Đề tài này đã thực hiện thành công một quy trình phân tích và dự đoán toàn diện về vấn đề khách hàng rời bỏ trong lĩnh vực thương mại điện tử. Bằng cách áp dụng một loạt các kỹ thuật từ tiền xử lý dữ liệu, kỹ thuật đặc trưng, phân tích khám phá đến xây dựng và đánh giá mô hình, đề án đã đạt được các mục tiêu đề ra.

Với quá trình phân tích khám phá và trực quan hóa dữ liệu, đề tài này đã xác định được các yếu tố và phân khúc khách hàng có rủi ro cao. Các phát hiện chính bao gồm tỷ lệ rời bỏ đặc biệt cao ở nhóm khách hàng mới (0-6 tháng), các phân khúc nhân khẩu học như khách hàng độc thân và khách hàng

ở thành phố cấp 1, và đặc biệt là những người có trải nghiệm tiêu cực (diễn hài lòng thấp hoặc đã từng khiếu nại).

Nghiên cứu đã so sánh hiệu suất của tám mô hình học máy khác nhau. Kết quả cho thấy các mô hình ensemble và boosting thể hiện sự vượt trội rõ rệt. Trong đó, mô hình 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 0.995 và chỉ số cân bằng F1-Score (cho lớp Churn) là 0.96. Kết quả này khẳng định rằng việc kết hợp một thuật toán mạnh mẽ với quy trình xử lý dữ liệu bài bản có thể tạo ra một công cụ dự đoán churn với độ tin cậy rất cao.

Cuối cùng, các sản phẩm trực quan hóa trên Power BI, Flourish và Website Demo không chỉ trình bày kết quả một cách hiệu quả mà còn tạo ra các công cụ tương tác, giúp chuyên hóa dữ liệu phức tạp thành những insight dễ tiếp cận và có tính ứng dụng.

### B. Đề xuất giải pháp kinh doanh

Dựa trên các kết quả phân tích dữ liệu và hiệu suất của mô hình dự đoán, đề tài đề xuất một số giải pháp chiến lược mang tính mục tiêu nhằm giảm thiểu tỷ lệ khách hàng rời bỏ, tập trung trực tiếp vào các vấn đề đã được phát hiện.

- 1) Triển khai chương trình chăm sóc đặc biệt cho khách hàng mới, kéo dài 6 tháng, bao gồm:
  - Giai đoạn đầu (Tháng 1–2): Gửi email hướng dẫn sử dụng dịch vụ và cung cấp ưu đãi cho lần mua tiếp theo nhằm khuyến khích sự quay lại.
  - Giai đoạn giữa (Tháng 3–4): Thu thập phản hồi sớm thông qua các khảo sát ngắn để đánh giá mức độ hài lòng ban đầu.
  - Giai đoạn cuối (Tháng 5–6): Ứng dụng mô hình dự đoán để xác định những khách hàng mới vẫn còn nguy cơ rời bỏ cao và chủ động can thiệp bằng các ưu đãi được cá nhân hóa.
- 2) Thiết lập quy trình phản ứng nhanh với trải nghiệm tiêu cực:
  - Đổi với khách hàng khiếu nại: Thiết lập quy trình ưu tiên nhằm đảm bảo mọi khiếu nại được phản hồi trong vòng 24 giờ, đồng thời có bước theo dõi sau khi sự cố được giải quyết.
  - Đổi với khách hàng có điểm hài lòng thấp (1–2 điểm): Tự động gửi email hoặc tin nhắn hỏi thăm để tìm hiểu nguyên nhân, đề xuất hỗ trợ phù hợp, có thể kèm theo một voucher nhỏ như một hình thức xin lỗi thiện chí.
- 3) Tối ưu hóa chiến lược cho thị trường thành phố cấp 1, phân bổ thêm nguồn lực để phân tích chuyên sâu thị trường này, với các hướng triển khai cụ thể:
  - Phân tích cạnh tranh: Nghiên cứu các đối thủ chính trong khu vực nhằm xác định những chương trình và ưu đãi đang thu hút khách hàng.
  - Tối ưu vận hành: Đánh giá lại tốc độ giao hàng và chất lượng dịch vụ tại thành phố cấp 1, nơi khách hàng thường có kỳ vọng cao hơn.
  - Chiến dịch địa phương hóa: Triển khai các chiến dịch marketing hoặc khuyến mãi được thiết kế riêng cho

- khách hàng tại khu vực này nhằm gia tăng mức độ gắn kết và trung thành.
- 4) Triển khai các chiến dịch marketing và chương trình khách hàng thân thiết được thiết kế riêng cho lối sống và nhu cầu của nhóm khách hàng độc thân:
- Tạo các chiến dịch với thông điệp nhắm vào sở thích cá nhân, self-care, du lịch, hoặc nâng cấp không gian sống cá nhân thay vì các thông điệp hướng về gia đình.
  - Sử dụng thuật toán gợi ý các sản phẩm phù hợp với hành vi mua sắm của người độc thân. Xây dựng các combo ưu đãi hoặc gói sản phẩm dành cho một người.
  - Thay vì các chương trình tích điểm dựa trên chi tiêu lớn cho gia đình, có thể thiết kế các chương trình thường dựa trên tần suất mua hàng hoặc tương tác, với các phần thưởng là trải nghiệm hoặc các sản phẩm độc đáo, phù hợp với sở thích cá nhân.

## VI. HƯỚNG PHÁT TRIỂN

Hướng phát triển đầu tiên là làm giàu bộ dữ liệu bằng cách tích hợp các nguồn thông tin chi tiết hơn như lịch sử giao dịch từng sản phẩm hoặc dữ liệu tương tác thời gian thực trên website/app. Về mặt kỹ thuật, có thể khám phá các phương pháp mô hình hóa nâng cao hơn như phân tích sống còn (Survival Analysis) để dự đoán “khi nào” khách hàng sẽ rời bỏ thay vì chỉ “có hay không”, đồng thời kết hợp các kỹ thuật diễn giải mô hình như SHAP để hiểu rõ lý do đằng sau mỗi dự đoán riêng lẻ. Cuối cùng, hướng đi quan trọng nhất là triển khai mô hình vào hệ thống quản lý quan hệ khách hàng (CRM) thực tế, cho phép tự động hóa việc xác định khách hàng rủi ro và thực hiện các chiến dịch A/B testing nhằm đo lường hiệu quả kinh doanh thực tế của các chiến lược giữ chân dựa trên điểm số từ mô hình.

## LỜI CẢM ƠN

Nhóm xin trân trọng cảm ơn Khoa Khoa học và Kỹ thuật Thông tin, Trường Đại học Công nghệ Thông tin – ĐHQG TPHCM đã tạo điều kiện thuận lợi về môi trường học tập và các nguồn lực cần thiết trong môn học Phân tích và Trực quan Dữ liệu (IE313.Q11), giúp nhóm có cơ hội vận dụng kiến thức lý thuyết vào nghiên cứu thực tiễn. Đồng thời, nhóm xin chân thành cảm ơn giảng viên hướng dẫn – ThS. Phạm Nguyễn Phúc Toàn đã tận tình hướng dẫn, đóng góp những ý kiến chuyên môn quý báu và hỗ trợ nhóm trong suốt quá trình thực hiện đồ án.

## 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. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666720723001443>
- [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. 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. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1877050922003955>
- [9] A. Verma, “E-Commerce Customer Churn Analysis and Prediction,” Kaggle, [Online]. Available: <https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction>. Accessed: Dec. 23, 2025.
- [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: Wiley, 2007, ch. 3.
- [12] H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, 1946, pp. 212–214.
- [13] D. W. Hosmer Jr., S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. 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. 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. Neural Information Processing Systems (NIPS)*, 2017, pp. 3146–3154.
- [19] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: <http://www.deeplearningbook.org>
- [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*. [Online]. Available: <https://learn.microsoft.com/en-us/power-bi/>. [Accessed: Dec. 1, 2025].
- [22] Flourish, *Flourish Studio Homepage*, <https://flourish.studio/>. [Accessed: Dec. 1, 2025].