I/ Giới thiệu chung

1. Mục tiêu nghiên cứu

- Đặt vấn đề: Trong bối cảnh kinh doanh hiện đại, các công ty cung cấp dịch vụ có mô hình đăng ký (subscription model) ngày càng chú trọng đến việc tăng tỷ lệ khách hàng đăng ký sử dụng dịch vụ dài hạn. Tuy nhiên, nhiều yếu tố ảnh hưởng đến quyết định đăng ký của khách hàng, chẳng hạn như nhân khẩu học (tuổi tác, giới tính), tần suất mua sắm, loại sản phẩm và dịch vụ sử dụng, phương thức thanh toán,...
- Mục tiêu báo cáo: Báo cáo này sẽ phân tích các nhân tố ảnh hưởng đến tình trạng đăng ký (Subscription Status) của khách hàng dựa trên dữ liệu khách hàng thu thập được. Mục tiêu cụ thể là xác định các **yếu tố có tác động mạnh** mẽ nhất đến khả năng khách hàng sẽ có trạng thái đăng ký, từ đó giúp công ty xây dựng chiến lược tiếp thị và chăm sóc khách hàng hiệu quả hơn.
- Ý nghĩa thực tiễn: Kết quả phân tích sẽ cung cấp cái nhìn sâu sắc về hành vi của khách hàng, giúp công ty tối ưu hóa các chiến dịch thu hút khách hàng đăng ký dịch vụ, cải thiện trải nghiệm người dùng và xây dựng chương trình khách hàng trung thành.

2. Phạm vi báo cáo

- Nguồn dữ liệu:
- Phạm vi phân tích: Chỉ tập trung vào việc xác định và phân tích các yếu tố tác động trực tiếp đến Subscription Status mà không đi sâu vào các yếu tố gián tiếp hoặc ngoài dữ liệu hiện có.
- Hạn chế của báo cáo: Do giới hạn về dữ liệu, báo cáo chỉ có thể phân tích dựa trên các thông tin hiện có mà không bao gồm các yếu tố khác như văn hóa hay tâm lý khách hàng.

II/ Mô tả dữ liệu nghiên cứu

- Quy mô dữ liệu: Dữ liệu gồm khoảng 3900 bản ghi khách hàng, với mỗi bản ghi chứa các thông tin chi tiết về khách hàng như ID, tuổi, giới tính, các yếu tố hành vi mua hàng và trạng thái đăng ký.
- Khái niệm Subscription Status: Subscription Status là trạng thái đăng ký dịch vụ của khách hàng, thường được phân loại thành hai trạng thái "Yes" (đã đăng ký) và "No" (chưa đăng ký). Subscription Status là chỉ báo quan trọng để đánh giá mức độ trung thành và tiềm năng giá trị dài hạn của khách hàng đối với doanh nghiệp.

- Các yếu tố có thể ảnh hưởng đến Subscription Status được phân tích trong báo cáo:
    + Age, Gender, Item Purchased, Category, Purchase Amount, Location, Size, Color, Season, Review Rating, Shipping Type, Discount Applied, Promo Code Used, Previous Purchases, Payment Method, Frequency of Purchases.

III/ Phương pháp phân tích

1. Khám phá và Thống kê Mô tả (Exploratory Data Analysis - EDA)

- Mục tiêu: Khám phá các đặc điểm cơ bản của dữ liệu và các yếu tố tiềm năng ảnh hưởng đến Subscription Status trước khi tiến hành phân tích sâu hơn.
- Phương pháp:
    + Thống kê mô tả là tính toán các số liệu cơ bản như trung bình, độ lệch chuẩn, phân phối của các biến nhằm có cái nhìn tổng quan về dữ liệu.
    + Đưa ra các nhận định sơ bộ về mối quan hệ giữa các yếu tố với Subscription Status dựa trên quan sát trực quan từ biểu đồ và thống kê mô tả.

2. Phân tích Tương Quan (Correlation Analysis)

- Mục tiêu: Xác định mối quan hệ giữa các biến độc lập (Age, Gender, Purchase Amount, Discount Applied, Frequency of Purchases, v.v.) với Subscription Status và đánh giá mức độ ảnh hưởng của từng yếu tố
- Phương pháp:
    + Ma trận tương quan (Correlation Matrix): Tính toán hệ số tương quan giữa các biến số liên tục, sử dụng heatmap để hiển thị mức độ tương quan và phát hiện những yếu tố có khả năng ảnh hưởng cao nhất.
    + Phân tích tương quan phân loại (Chi-square Test): Đối với các biến phân loại (như `Gender`, `Category`), sử dụng kiểm định chi-square để kiểm tra sự liên quan giữa các biến và Subscription Status.

3. Phân tích Nhân Tố Ảnh Hưởng Đến Subscription Status

- Mục tiêu: Đánh giá chi tiết từng nhân tố có thể ảnh hưởng đến Subscription Status, đặc biệt là các yếu tố nhân khẩu học, hành vi mua sắm và các yếu tố kinh tế.
- Phương pháp:
    + Phân tích nhân khẩu học: Khám phá mối liên hệ giữa các yếu tố nhân khẩu học (`Age`, `Gender`) với Subscription Status để xác định những nhóm khách hàng có khả năng đăng ký cao hơn.

+ Phân tích hành vi mua sắm: Phân tích các biến liên quan đến hành vi khách hàng (`Frequency of Purchases`, `Item Purchased`, `Category`) để xem yếu tố nào ảnh hưởng đến việc đăng ký.
+ Phân tích các yếu tố kinh tế: Đánh giá tác động của `Purchase Amount`, `Discount Applied`, `Promo Code Used` và `Payment Method` đến Subscription Status nhằm xem xét yếu tố tài chính nào thúc đẩy việc đăng ký.

4. Xây dựng Mô hình Dự đoán (Predictive Modeling)

- Mục tiêu: Xây dựng mô hình dự đoán Subscription Status dựa trên các yếu tố có liên quan, từ đó giúp xác định yếu tố quan trọng và cải thiện độ chính xác của dự báo.

- Phương pháp:

+ Hồi quy logistic (Logistic Regression): Sử dụng hồi quy logistic để kiểm tra mức độ ảnh hưởng của từng biến đến Subscription Status. Đây là phương pháp lý tưởng cho dữ liệu phân loại với biến phụ thuộc dạng nhị phân (Yes/No).

Quy trình:

Xác định biến đầu vào: Lựa chọn các biến quan trọng nhất từ kết quả phân tích tương quan và phân tích các yếu tố ảnh hưởng.

Kiểm định ý nghĩa thống kê: Kiểm tra các hệ số hồi quy để xác định biến nào có tác động lớn nhất và ý nghĩa thống kê.

Đánh giá mô hình: Sử dụng các thước đo như accuracy, precision, recall, và F1 score để kiểm tra hiệu suất mô hình và đảm bảo tính chính xác của dự đoán.

+ Decision Tree và Random Forest: Sử dụng Decision Tree hoặc Random Forest để hiểu rõ hơn về các yếu tố quan trọng nhất ảnh hưởng đến Subscription Status và đánh giá độ quan trọng của từng biến. Phương pháp này hỗ trợ xác định các yếu tố chính yếu một cách trực quan và dễ hiểu.

5. Kiểm định Giả thuyết (Hypothesis Testing)

- Mục tiêu: Kiểm định các giả thuyết ban đầu về mối quan hệ giữa các biến độc lập và Subscription Status.
- Phương pháp:
  + T-test: Sử dụng kiểm định t-test cho các biến liên tục như `Purchase Amount` hoặc `Age` để so sánh trung bình giữa hai nhóm khách hàng có và không có Subscription Status.

+ Chi-square test: Đối với các biến phân loại như `Gender`, `Category`, sử dụng kiểm định chi-square để xác nhận mối quan hệ với Subscription Status có ý nghĩa thống kê.

6. Tổng hợp và Đưa ra Kết quả

- Mục tiêu: Tóm tắt các kết quả và rút ra những yếu tố có ảnh hưởng nhất đến Subscription Status, từ đó đề xuất các chiến lược kinh doanh

- Phương pháp:

+ Diễn giải ý nghĩa thực tiễn: Giải thích cách các yếu tố quan trọng ảnh hưởng đến Subscription Status có thể được tận dụng để tăng tỷ lệ đăng ký, đồng thời tối ưu hóa chiến lược marketing và chăm sóc khách hàng.
+ Đề xuất hành động: Từ kết quả phân tích, đưa ra các đề xuất cụ thể nhằm giúp doanh nghiệp tối ưu tỷ lệ đăng ký Subscription và tạo lợi thế cạnh tranh.

IV. Kết qủa phân tích

1. Khám phá và thống kê mô tả

	Customer ID	Age	Purchase Amount (USD)	Review Rating	Previous Purchases
count	3900.000000	3900.000000	3900.000000	3900.000000	3900.000000
mean	1950.500000	44.068462	59.764359	3.749949	25.351538
std	1125.977353	15.207589	23.685392	0.716223	14.447125
min	1.000000	18.000000	20.000000	2.500000	1.000000
25%	975.750000	31.000000	39.000000	3.100000	13.000000
50%	1950.500000	44.000000	60.000000	3.700000	25.000000
75%	2925.250000	57.000000	81.000000	4.400000	38.000000
max	3900.000000	70.000000	100.000000	5.000000	50.000000

-> Insight:

Age:

Phần lớn khách hàng nằm trong độ tuổi từ 31 đến 57, với độ tuổi trung bình là 44.

Có sự biến động đáng kể về độ tuổi của khách hàng, như được chỉ ra bởi độ lệch chuẩn cao là 15,21.

Trong số tất cả khách hàng, 25% có độ tuổi dưới 31, 50% có độ tuổi là 44, và 75% có độ tuổi dưới 57.

Purchase Amount:

Số tiền khách hàng chi tiêu dao động từ 20 USD đến 100 USD, với trung bình là 60 USD.

Trong số tất cả khách hàng, 25% chi tiêu dưới 39 USD, 50% chi tiêu 60 USD, và 75% chi tiêu dưới 81 USD.

Previous Purchase:

Khách hàng đã thực hiện từ 1 đến 50 giao dịch, với trung bình 25 giao dịch mỗi khách hàng.

25% khách hàng đã thực hiện 13 giao dịch hoặc ít hơn, 50% (trung vị) đã thực hiện 25 giao dịch, và 75% đã thực hiện 38 giao dịch hoặc ít hơn.

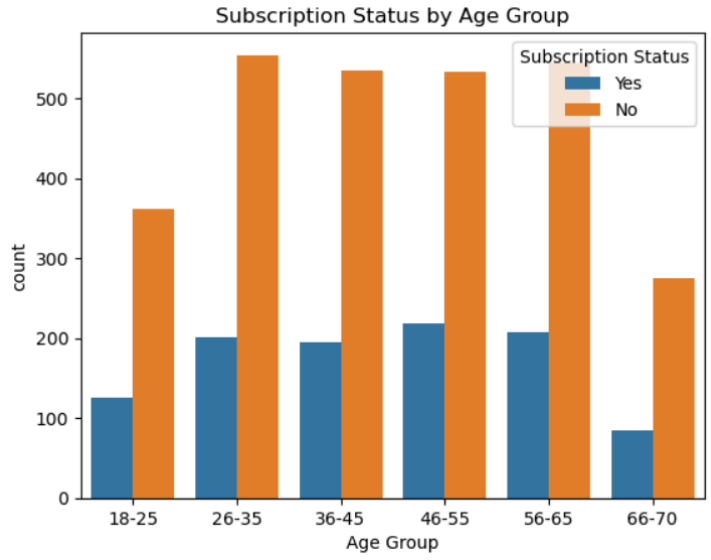
Tình trạng đăng ký

```
Subscription Status
No      2847
Yes     1053
Name: count, dtype: int64
```

-> **Dữ liệu**: Trong tổng số 3,900 khách hàng, có **2,847 khách hàng không đăng ký** (chiếm 73%) và **1,053 khách hàng đã đăng ký** (chiếm 27%).

Nhận định: Sự chênh lệch lớn giữa số lượng khách hàng không đăng ký và đã đăng ký cho thấy có một lượng lớn khách hàng tiềm năng mà công ty cần tập trung vào để cải thiện tỷ lệ đăng ký.

1.1. Tương quan giữa “Age” và “ Subscription Status”

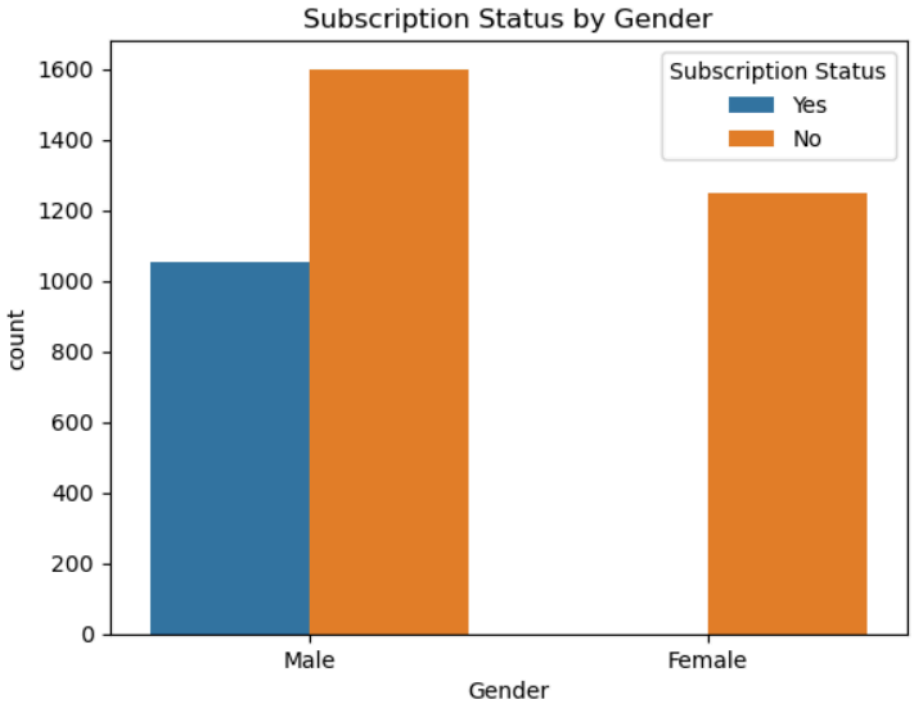


-> Insights:
Nhóm tuổi 18-25: Tỉ lệ khách hàng đăng ký có thể thấp hơn so với các nhóm tuổi lớn hơn, vì nhóm này có thể ít có khả năng tài chính hoặc ít quan tâm đến việc đăng ký dài hạn.

Nhóm tuổi 26-65: Đây có thể là nhóm có tỉ lệ đăng ký cao nhất, do nhóm này có thu nhập ổn định và có nhu cầu sử dụng nhiều dịch vụ.

Nhóm tuổi 66-70: Nhóm này có thể có tỉ lệ đăng ký thấp hơn vì họ ít có nhu cầu hoặc không quen với các dịch vụ trực tuyến.

1.2. Tương quan giữa "Gender" và " Subscription Status"



-> Insights:

Giới tính Nữ có vẻ như không có ai trong số họ đã đăng ký. Điều này có thể chỉ ra rằng sản phẩm hoặc dịch vụ đang phân tích không thu hút được phụ nữ hoặc có thể có vấn đề trong cách tiếp cận marketing đến họ.

Giới tính Nam có tỉ lệ đăng ký khá cao, với khoảng 39.71% trong số họ đã chọn đăng ký. Điều này cho thấy rằng có một số yếu tố có thể đang thúc đẩy nam giới tham gia vào dịch vụ hoặc sản phẩm này.

1.3. Tương quan giữa “Category” và “ Subscription Status”

Subscription Status	No	Yes
Category		
Accessories	73.064516	26.935484
Clothing	73.690271	26.309729
Footwear	71.452421	28.547579
Outerwear	71.913580	28.086420

-> Insights:

Trong tất cả các danh mục sản phẩm, tỷ lệ khách hàng không đăng ký ("No") luôn cao hơn tỷ lệ đăng ký ("Yes"). Footwear (giày dép) có tỷ lệ khách hàng đăng ký cao nhất (28.55%), trong khi Clothing (quần áo) có tỷ lệ đăng ký thấp nhất (26.31%) mặc dù có số lượng lớn nhất.

Điều này có thể cho thấy rằng các sản phẩm thuộc danh mục Footwear và Outerwear có khả năng thu hút người dùng đăng ký dịch vụ cao hơn so với Accessories và Clothing.

1.4. Tương quan giữa “Purchase Amount” và “ Subscription Status”

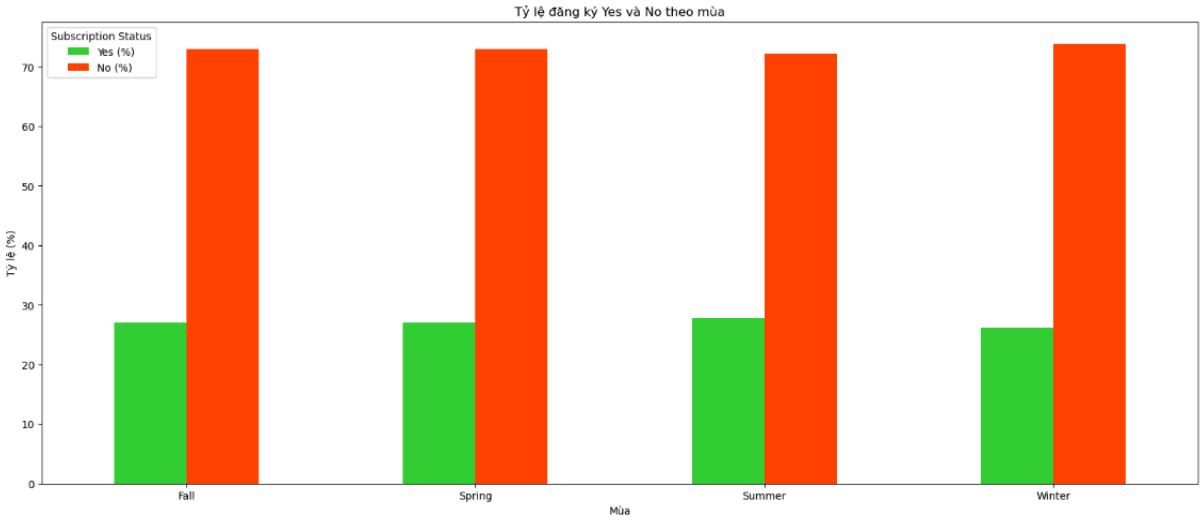
	mean	std
Subscription Status		
No	59.865121	23.775199
Yes	59.491928	23.449914

-> Insights:

Giá trị trung bình của Purchase Amount ở nhóm không đăng ký (No) là khoảng 59.87 USD, và ở nhóm đăng ký (Yes) là khoảng 59.49 USD. Sự chênh lệch này rất nhỏ (chỉ khoảng 0.37 USD), nên có thể nói rằng mức chi tiêu trung bình giữa hai nhóm khách hàng không có sự khác biệt đáng kể.

Độ lệch chuẩn (standard deviation) cho cả hai nhóm là tương tự nhau (khoảng 23.77 cho nhóm "No" và 23.45 cho nhóm "Yes"), cho thấy sự phân tán của dữ liệu là tương đương giữa hai nhóm.

1.5. Tương quan giữa “Season” và “Subscription Status”



Insights: Gần như không có sự khác biệt giữa các mùa.

1.6. Tương quan giữa “ Review Rating” và “ Subscription Status”

Subscription Status	No	Yes
Review Group		
Low (2 - 3)	72.963400	27.036600
Medium (3 - 4)	73.144654	26.855346
High (4 - 5)	72.863978	27.136022

-> Insights:

Tỷ lệ không đăng ký vẫn rất cao ở cả ba nhóm, dao động quanh mức 72-73%, cho thấy nhiều khách hàng vẫn quyết định không đăng ký dịch vụ, bất kể đánh giá của họ về sản phẩm.

Khách hàng đánh giá cao có xu hướng đăng ký nhiều hơn (tỷ lệ đăng ký cao nhất là 27.14% trong nhóm High), cho thấy rằng đánh giá tích cực có ảnh hưởng nhất định đến quyết định đăng ký, nhưng tác động này chưa đủ lớn để thay đổi rõ rệt tỷ lệ tổng thể.

1.7. Tương quan giữa “Payment Method/Prefered Payment Method” và “Subscription Status”

Subscription Status	No	Yes
Payment Match		
False	73.256168	26.743832
True	71.636953	28.363047

-> Insights:

Tỷ lệ khách hàng đăng ký (Subscription Status = 'Yes') cao hơn một chút khi Payment Method trùng với Preferred Payment Method (28.36% so với 26.74%). Tuy nhiên, sự khác biệt này không quá lớn, chỉ khoảng 1.6%.

Mặc dù có sự khác biệt nhỏ, việc khách hàng chọn cùng một phương thức thanh toán cho cả Payment Method và Preferred Payment Method chỉ ảnh hưởng nhẹ đến khả năng họ đăng ký. Điều này cho thấy rằng sự trùng khớp giữa hai phương thức này có thể là một yếu tố, nhưng không phải là yếu tố quyết định mạnh mẽ đến việc khách hàng có đăng ký hay không.

1.8. Tương quan giữa " Shipping Type" và "Subscription Status"

Subscription Status	No	Yes
Shipping Type		
2-Day Shipping	75.598086	24.401914
Express	70.588235	29.411765
Free Shipping	73.777778	26.222222
Next Day Air	74.074074	25.925926
Standard	73.241590	26.758410
Store Pickup	70.769231	29.230769

-> Insights:

Các phương thức vận chuyển như Express và Store Pickup có tỷ lệ khách hàng đăng ký cao hơn (gần 29%).

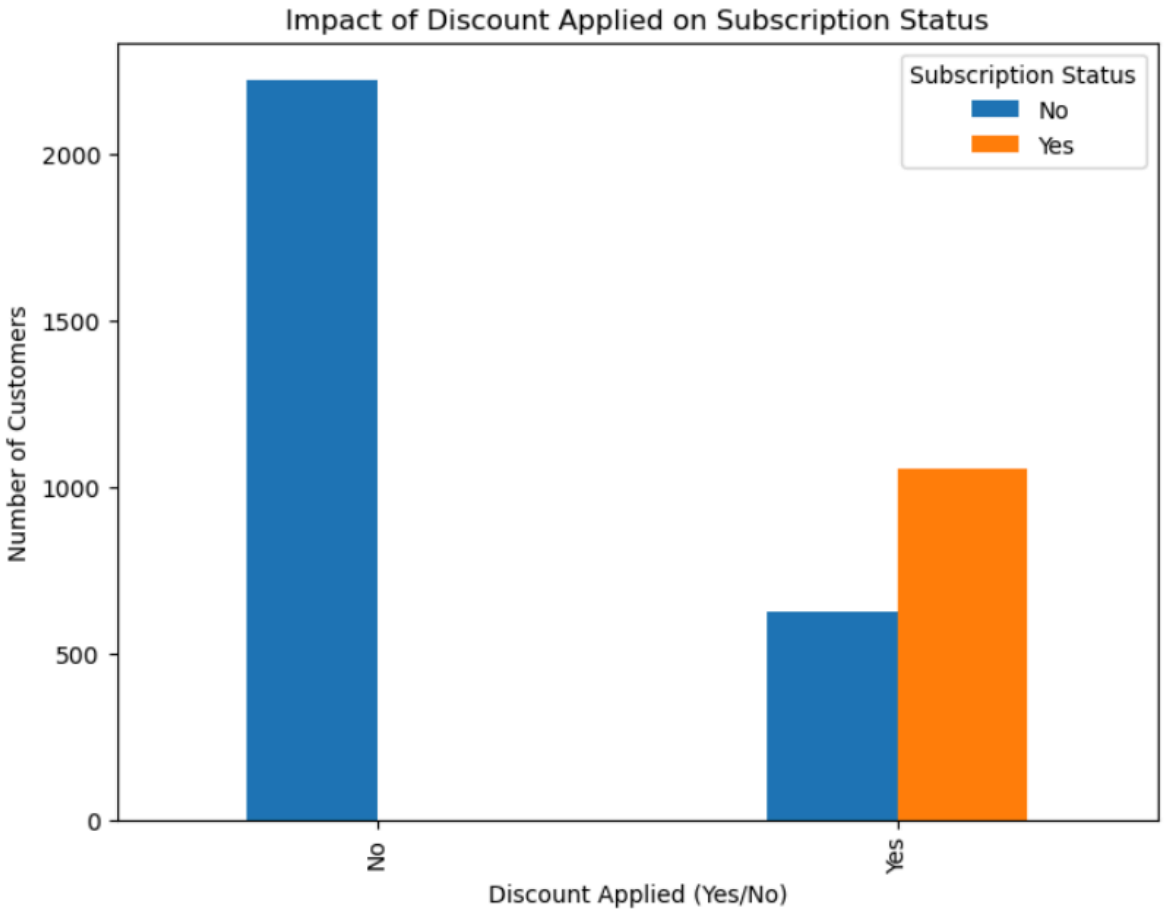
Ngược lại, 2-Day Shipping có tỷ lệ khách hàng đăng ký thấp nhất (24.40%), cho thấy phương thức này ít ảnh hưởng tích cực đến việc đăng ký.

Free Shipping, Standard, và Next Day Air có tỷ lệ khách hàng đăng ký ở mức trung bình, từ 25.93% đến 26.76%, không có sự khác biệt lớn.

Express Shipping và Store Pickup dường như có ảnh hưởng tích cực đến tỷ lệ đăng ký cao hơn so với các phương thức vận chuyển khác. Những khách hàng chọn hai phương thức này có khả năng đăng ký cao hơn khoảng 3-5% so với các phương thức khác. 2-Day Shipping có tỷ lệ đăng ký thấp hơn, cho thấy có thể phương thức này không được khách hàng ưa chuộng khi quyết định đăng ký.

Đề xuất tối ưu hóa các phương thức vận chuyển như Express hoặc Store Pickup để tăng tỷ lệ đăng ký

1.9. Tương quan giữa “Discount Applied/ Promo Cod Used” và “Subscription Status”



Subscription Status	No	Yes
Discount Applied		
No	100.000000	NaN
Yes	37.209302	62.790698

Subscription Status	No	Yes
Promo Code Used		
No	100.000000	NaN
Yes	37.209302	62.790698

-> Insights:

Tác động mạnh mẽ của giảm giá: Kết quả cho thấy rằng việc áp dụng giảm giá làm tăng khả năng đăng ký lên đến 62.79%, cho thấy khách hàng rất nhạy cảm với ưu đãi giảm giá.

Không có ưu đãi = Không có đăng ký: Tình trạng không có giảm giá dẫn đến tỷ lệ đăng ký bằng 0, điều này cho thấy rằng khách hàng có thể không thấy đủ động lực để đăng ký khi không có ưu đãi.

Kết luận: Giảm giá là một yếu tố quan trọng và hiệu quả trong việc thu hút khách hàng đăng ký. Chiến lược sử dụng ưu đãi giảm giá trong các chương trình khuyến mãi có thể tạo ra động lực lớn cho khách hàng. Đề xuất: Dựa trên phân tích trên, đây là một số đề xuất:

Tăng cường các chương trình giảm giá và đưa ra các chiến dịch marketing để thông báo cho khách hàng về các ưu đãi này.

Theo dõi phản hồi của khách hàng và điều chỉnh các ưu đãi để tối ưu hóa tỷ lệ đăng ký.

Kết hợp giảm giá với các yếu tố khác như dịch vụ khách hàng tốt hơn hoặc sản phẩm chất lượng cao để tăng cường sự hấp dẫn cho khách hàng.

Các chiến lược này có thể giúp nâng cao tỷ lệ chuyển đổi và tạo động lực cho khách hàng trong việc đăng ký dịch vụ.

Kết quả giống nhau 100% cho 2 yếu tố " Discount Applied" và "Promo Code Used" ảnh hưởng đến" Subscription Status"

Cả Discount Applied và Promo Code Used đều có tác động lớn đến khả năng đăng ký của khách hàng, điều này cho thấy rằng việc áp dụng các chiến lược ưu đãi sẽ là một yếu tố quan trọng trong việc tăng cường tỷ lệ chuyển đổi và tạo động lực cho khách hàng đăng ký.

1.10. Tương quan giữa “ Previous Purchases/Frequency of Purchase” và “Subscription Status”

Subscription Status		No	Yes
Previous Purchases Group	Frequency Group		
1-10 Purchases	Frequently	76.605505	23.394495
	Occasionally	75.776398	24.223602
	Rarely	77.459016	22.540984
11-20 Purchases	Frequently	71.144279	28.855721
	Occasionally	75.692308	24.307692
	Rarely	71.713147	28.286853
21-30 Purchases	Frequently	69.266055	30.733945
	Occasionally	73.255814	26.744186
	Rarely	71.929825	28.070175
31-40 Purchases	Frequently	68.240343	31.759657
	Occasionally	72.023810	27.976190
	Rarely	70.697674	29.302326
41+ Purchases	Frequently	71.563981	28.436019
	Occasionally	74.404762	25.595238
	Rarely	72.018349	27.981651

-> Insights:

-> Nhóm 1-10 Purchases:

Tần suất mua hàng: Frequently: 23.39% đăng ký Occasionally: 24.22% đăng ký Rarely: 22.54% đăng ký Nhận xét: Nhóm này có tỷ lệ đăng ký không cao, nhưng nhóm mua hàng Occasionally có tỷ lệ đăng ký cao nhất trong nhóm này.

-> Nhóm 11-20 Purchases:

Tần suất mua hàng: Frequently: 28.86% đăng ký Occasionally: 24.31% đăng ký Rarely: 28.29% đăng ký Nhận xét: Nhóm này có sự gia tăng đáng kể về tỷ lệ đăng ký ở cả hai nhóm Frequently và Rarely. Điều này cho thấy rằng những khách hàng đã có từ 11 đến 20 lần mua hàng có xu hướng đăng ký nhiều hơn.

-> Nhóm 21-30 Purchases:

Tần suất mua hàng: Frequently: 30.73% đăng ký Occasionally: 26.74% đăng ký Rarely: 28.07% đăng ký Nhận xét: Đây là nhóm có tỷ lệ đăng ký cao nhất cho nhóm Frequently so với các nhóm trước đó.

-> Nhóm 31-40 Purchases:

Tần suất mua hàng: Frequently: 31.76% đăng ký Occasionally: 27.98% đăng ký Rarely: 29.30% đăng ký Nhận xét: Tỷ lệ đăng ký tiếp tục tăng, đặc biệt ở nhóm Frequently.

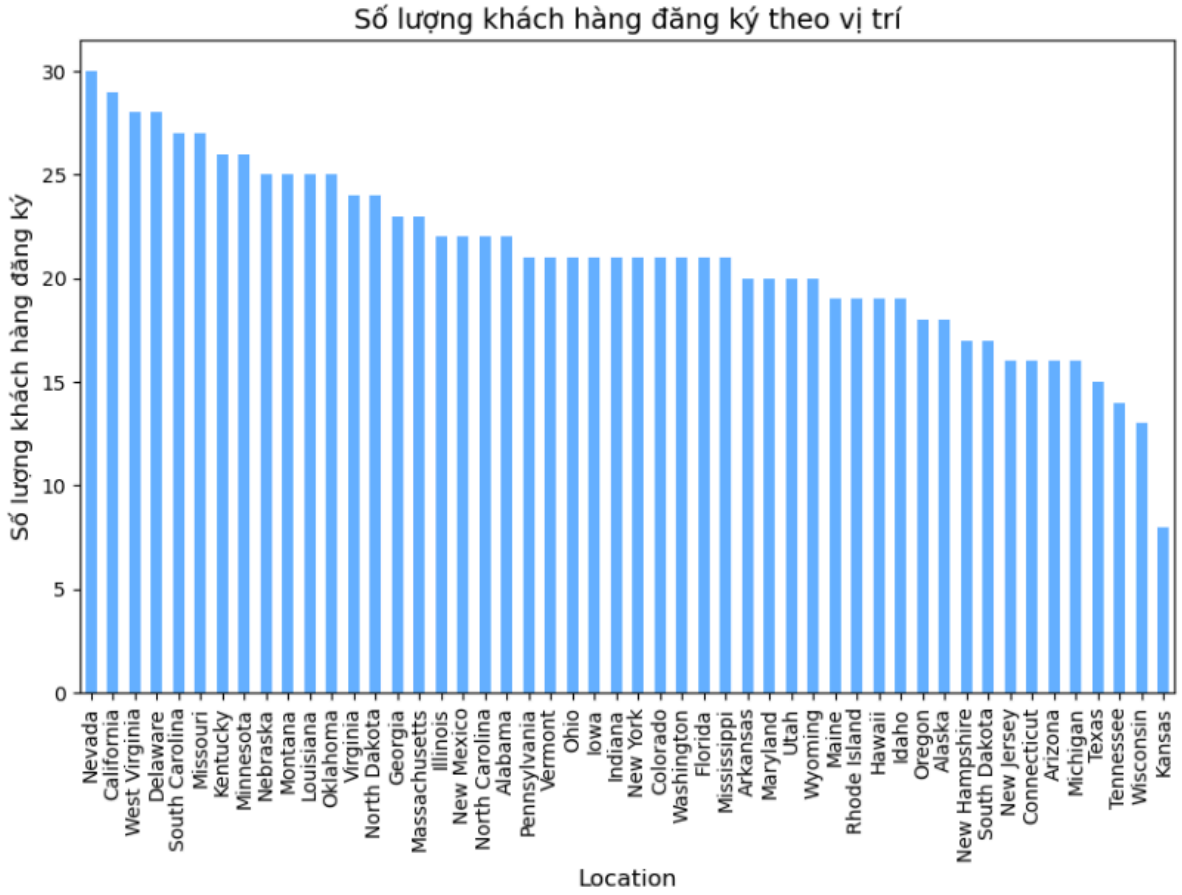
-> Nhóm 41+ Purchases:

Tần suất mua hàng: Frequently: 28.44% đăng ký Occasionally: 25.60% đăng ký Rarely: 27.98% đăng ký Nhận xét: Tỷ lệ đăng ký ở nhóm này thấp hơn so với nhóm trước, nhưng vẫn duy trì mức độ cao hơn 25%.

===> Tổng quát: Sự gia tăng trong số lần mua hàng trước đó dường như có mối liên hệ tích cực với tỷ lệ đăng ký. Các khách hàng đã thực hiện nhiều giao dịch hơn có xu hướng đăng ký cao hơn, đặc biệt là trong các nhóm tần suất mua hàng thường xuyên.

Tần suất Mua hàng: Khách hàng có tần suất mua hàng Frequently có tỷ lệ đăng ký cao hơn so với các nhóm Occasionally và Rarely. Điều này cho thấy rằng các khách hàng thường xuyên quay lại có nhiều khả năng đăng ký hơn.

1.11. Tương quan giữa " Location" và "Subscription Status"



2. Phân tích tương quan và kiểm định giả thuyết.

2.1. Phân tích Tương Quan(Ma trận tương quan)

2.1.1. Cơ sở lý thuyết.

- Khái niệm: Tương quan đo lường mức độ liên quan giữa hai biến. Hệ số tương quan có giá trị từ -1 đến 1, trong đó giá trị gần 1 hoặc -1 cho thấy mối liên hệ mạnh, còn giá trị gần 0 chỉ ra mối liên hệ yếu hoặc không có.

- Ứng dụng: Trong phân tích này, ma trận tương quan được sử dụng để xác định xem các biến số liên tục ‘Age’, ‘Purchase Amount’ và ‘Previous Purchases’, …..có liên hệ gì với Subscription Status hay không. Nếu một yếu tố có hệ số tương quan lớn, yếu tố đó có thể được xem là có ảnh hưởng mạnh mẽ đến Subscription Status.

- Lợi ích: Phân tích tương quan giúp sàng lọc các yếu tố tiềm năng, hỗ trợ xác định những biến có khả năng dự đoán tốt nhất cho việc đăng ký Subscription.

2.1.2. Kết quả thực tế

-> Phân tích kết quả:

Phân tích kết quả

Age: 0.0065 (Mối tương quan gần như không có)

Purchase Amount (USD): -0.007 (Mối tương quan rất yếu, gần như không ảnh hưởng)

Review Rating: -0.0064 (Mối tương quan gần như không có)

Previous Purchases: 0.031 (Mối tương quan rất yếu)

Frequency of Purchases (Encoded): -0.012 (Mối tương quan rất yếu)

2.2. Kiểm định Giả Thuyết:

2.2.1. Cơ sở lý thuyết

- Khái niệm: Kiểm định giả thuyết là phương pháp thống kê để kiểm tra tính hợp lý của một giả thuyết về dữ liệu mẫu. Kết quả kiểm định thường dựa vào giá trị p (p-value) để đánh giá ý nghĩa thống kê: nếu $p < 0.05$, giả thuyết về mối liên hệ giữa các yếu tố được xem là có ý nghĩa thống kê.
- Ứng dụng: Để kiểm tra mối liên hệ giữa Subscription Status và các yếu tố khác, hai loại kiểm định được áp dụng:

+ Kiểm định Chi-Square cho các biến phân loại: Kiểm định này được sử dụng với các biến phân loại như ‘Gender`, `Category`, hoặc `Shipping Type` để xác định liệu có sự khác biệt đáng kể về Subscription Status giữa các nhóm hay không.
+ Kiểm định T-test cho các biến liên tục: Đối với các biến liên tục như `Age`, `Purchase Amount`, và `Previous Purchases`, kiểm định t-test hoặc ANOVA giúp xác định xem có sự khác biệt về trung bình giữa các nhóm Subscription Status hay không.

- Lợi ích: Kiểm định giả thuyết giúp xác nhận các yếu tố có ý nghĩa thống kê và mối quan hệ giữa chúng với Subscription Status, cung cấp bằng chứng rõ ràng về tác động của từng yếu tố.

2.2.1. Kết quả phân tích thực tế:

- **Kiểm định Chi-Square**

+ **Gender:** Chi-square statistic: 676.7944035612919, p-value: 3.3268630006040623e-149

    -> Phân tích kết quả:

    Chi-square statistic: 676.79. Đây là giá trị của thống kê Chi-square cho kiểm định. Giá trị này càng lớn, chứng tỏ rằng sự khác biệt giữa giá trị quan sát và giá trị kỳ vọng càng lớn, có thể dẫn đến kết luận về sự phụ thuộc giữa hai biến.

    Với p-value gần bằng 0, ta có thể bác bỏ giả thuyết không (null hypothesis), và kết luận rằng có mối quan hệ có ý nghĩa thống kê giữa “Gender” và “Subscription Status”. **Điều này nghĩa là “Gender” có khả năng ảnh hưởng đến “Subscription Status”**

+ **Category:** Chi-square statistic: 1.344404607102478, p-value: 0.7186167320249088

    -> Phân tích kết quả:

Chi-square statistic: 1.344. Giá trị Chi-square này khá nhỏ, cho thấy sự khác biệt giữa tần suất quan sát và kỳ vọng không lớn, hàm ý rằng không có mối quan hệ mạnh giữa hai biến.

Với giá trị p-value cao (0.7186), có thể kết luận rằng không có mối quan hệ có ý nghĩa thống kê giữa "Category" và "Subscription Status" hay "Category" không ảnh hưởng đáng kể đến "Subscription Status"

+ **Shipping Type**: Chi-square statistic: 6.300602351974799, p-value: 0.27805797565678847.

-> Phân tích kết quả:

Chi-square statistic: 6.301. Giá trị Chi-square không quá lớn, cho thấy không có sự khác biệt lớn giữa giá trị quan sát và giá trị kỳ vọng.

Với p-value cao (0.2781), chúng ta không thể kết luận rằng có mối quan hệ có ý nghĩa thống kê giữa "Shipping Type" và "Subcription Status". Điều này cho thấy "Shipping Type" không ảnh hưởng đáng kể đến "Subscription Status"

+ **Discount Applied/ Promo Code Used:** Chi-square statistic: 1908.9213651509538, p-value: 0.0

-> Phân tích kết quả:

Chi-square statistic: 1908.92. Đây là một giá trị rất lớn, cho thấy có sự khác biệt lớn giữa tần suất quan sát và tần suất kỳ vọng, điều này cho thấy khả năng tồn tại mối quan hệ rất mạnh giữa hai biến "Discount Applied", " Promo Code Used" và " Subscription Status".

Với p-value bằng 0, ta có thể bác bỏ giả thuyết không và kết luận rằng có mối quan hệ rất mạnh và có ý nghĩa thống kê giữa việc áp dụng Discount và Promo Code đến " Subscription Status". Do đó cho thấy

" Discount Applied" và " Promo Code Used" **có thể ảnh hưởng đáng kể đến quyết định đăng ký "Subscription Status" của khách hàng.**

+ Frequency of Purchase: Chi-square statistic: 3.8946563289521987, p-value: 0.6909297922870219

-> Tượng tự: Không có ý nghĩa thống kê

+ Payment Method: Chi-square statistic: 2.6055935917437436, p-value: 0.7605152075982343

-> Tương tự: Không có ý nghĩa thống kê

- **Kiểm định T-test**

+ **Age:** T-statistic: 0.4053077788567848, P-value: 0.6852735321521746

-> Phân tích kết quả:

T-statistic: 0.405. Giá trị t-statistic nhỏ cho thấy sự khác biệt về độ tuổi trung bình giữa hai nhóm Subscription Status là "Yes" và "No") là không đáng kể.

Với p-value cao (0.6853), ta không thể kết luận rằng có sự khác biệt có ý nghĩa thống kê về độ tuổi giữa những người có đăng kí và không đăng kí. Điều này cho thấy Age không ảnh hưởng đáng kể đến Subscription Status.

+ **Purchase Amount (USD)**: T-statistic: -0.4395786253862316, p-value: 0.660292253988034

-> Phân tích kết quả:

T-statistic: -0.440. Giá trị t-statistic gần bằng 0 cho thấy không có sự khác biệt lớn giữa trung bình của Purchase Amount và Subcription Status.

Với p-value cao (0.6603), chúng ta không thể kết luận rằng có sự khác biệt có ý nghĩa thống kê về Purchase Amount giữa những người có và không đăng kí. Điều này cho thấy rằng số tiền chi tiêu không bị ảnh hưởng đến khách hàng có đăng kí hay không

+ **Previous Purchases:** T-statistic: 1.9533719191794576, p-value: 0.050919957531247007

-> Phân tích kết quả:

T-statistic: 1.953. Giá trị này cho thấy có sự khác biệt giữa trung bình số lần mua trước “Previous Purchases” của hai nhóm khách hàng đăng kí và không đăng kí

P-value: 0.0509. Đây là giá trị gần với ngưỡng 0.05, cho thấy rằng kết quả có thể có ý nghĩa thống kê.

=Có thể nói rằng số lần mua trước **“ Previous Purchases” có ảnh hưởng đến việc khách hàng có đăng ký hay không**, và sự khác biệt này có thể đáng xem xét trong các phân tích và quyết định tiếp theo về chiến lược kinh doanh hoặc tiếp thị

+ Review Rating: T-statistic: -0.3970059234213549, p-value: 0.6914083059838173

-> Tượng tự: Không có ý nghĩa thống kê

2.3. Kết luận:

Bằng cách sử dụng phân tích tương quan và kiểm định giả thuytô, có thể xác định các yếu tố có ảnh hưởng lớn nhất đến Subscription Status: **“Gender”, “Discount**

Applied", "Promo Code Used", "Previous Purchases" : Những yếu tố có hệ số tương quan cao hoặc giá trị p thấp sẽ được coi là các biến có tác động đáng kể, cung cấp cơ sở cho các chiến lược kinh doanh và các phân tích sâu hơn.

3. Mô hình dự đoán

3.1. Mục tiêu:

Xây dựng và đánh giá các mô hình học máy nhằm dự đoán trạng thái đăng kký dịch vụ (Subscription Status) của người dùng dựa trên các yếu tố như giới tính (Gender) , việc áp dụng giảm giá (Discount Applied), sử dụng mã khuyến mãi (Promo Code Used) và số lần mua hàng trước đây(Previous Purchases).

3.2. Dữ liệu

- Biến phụ thuộc: "Subscription Status" với 2 giá trị:
  - 1: Có đăng ký (Yes)
  - 0: Không đăng ký (No)
- Các biến đọc lập
  - Gender (Male/Female)
  - Discount Applied (Yes/No)
  - Promo Code Used (Yes/No)
  - Previous Purchases

3.3. Cơ sở lý thuyết

- **Logistic Regression**

  Logistic Regression là một phương pháp thống kê dùng để dự đoán xác suất xảy ra của một biến nhị phân (binary variable) dựa trên một hoặc nhiều biến độc lập. Mô hình này sử dụng hàm logistic (hay sigmoid) để biến đổi giá trị đầu ra, đảm bảo rằng nó nằm trong khoảng [0, 1]

  -> Lí do chọn Logistic Regression:

  Logistic Regression là một phương pháp đơn giản nhưng hiệu quả cho các bài toán phân loại nhị phân.

  Mô hình này dễ hiểu và có thể giải thích được, vì các hệ số cho biết ảnh hưởng của từng biến độc lập đến xác suất xảy ra của biến mục tiêu.

  Logistic Regression có thể xử lý các biến độc lập liên tục và phân loại.

- **Decision Tree**

  Decision Tree là một mô hình học máy có cấu trúc dạng cây, trong đó mỗi nút nội bộ đại diện cho một biến kiểm tra, mỗi nhánh đại diện cho kết quả của kiểm

tra đó, và mỗi nút lá đại diện cho một nhãn lớp. Mô hình này sử dụng thuật toán phân chia dữ liệu dựa trên các tiêu chí như độ tinh khiết (impurity) của các lớp, chẳng hạn như Gini impurity hoặc entropy.

-> Lí do chọn Decision Tree:

Decision Tree rất dễ hiểu và trực quan, có thể dễ dàng trực quan hóa mô hình và đưa ra quyết định.

Mô hình này có khả năng xử lý cả biến số liên tục và biến số phân loại.

Decision Tree không yêu cầu các giả định về phân phối của dữ liệu, điều này làm cho nó rất linh hoạt trong các bài toán thực tế.

3.4 Phân tích kết quả:

3.4.1 Kết Quả Mô Hình Hồi Quy Logistic

```
Logistic Regression:
Accuracy: 0.83
Precision: 0.63
Recall: 1.00
F1 Score: 0.77
              precision    recall  f1-score   support

           0       1.00      0.76      0.86       834
           1       0.63      1.00      0.77       336

    accuracy                           0.83      1170
   macro avg       0.81      0.88      0.82      1170
weighted avg       0.89      0.83      0.84      1170
```

- **Độ chính xác (Accuracy): 0.83**
  - Độ chính xác cho biết rằng 83% các dự đoán của mô hình là đúng. Mô hình Hồi quy Logistic đã thực hiện tốt trong việc phân loại tổng thể, cho thấy khả năng phân loại đáng tin cậy.
- **Precision: 0.63**
  - Từ số trường hợp mà mô hình dự đoán là có đăng ký (1), chỉ có 63% thực sự đúng. Điều này cho thấy mô hình có thể gặp khó khăn trong việc phân loại chính xác lớp 1, dẫn đến tỷ lệ giả dương cao. Điều này có thể chỉ ra rằng một số khách hàng không đăng ký đã bị mô hình dự đoán nhầm thành có đăng ký.
- **Recall: 1.00**
  - Mô hình đã phát hiện được 100% số trường hợp thực sự có đăng ký. Đây là một điểm mạnh lớn của mô hình Hồi quy Logistic, cho thấy rằng

không có trường hợp nào bị bỏ sót. Tuy nhiên, điều này cũng có thể cho thấy mô hình thiên về việc phân loại tất cả các trường hợp có đăng ký, có thể gây ra một số dự đoán sai cho lớp 0.

- **F1 Score: 0.77**
  - F1 Score là một chỉ số cân bằng giữa precision và recall, cho thấy mô hình có khả năng tổng quát tốt, mặc dù precision không cao.
- **Bảng Chi Tiết Phân Tích**:
  - **Hạng mục 0 (Không có đăng ký)**:
    - Precision: 1.00
    - Recall: 0.76
  - **Hạng mục 1 (Có đăng ký)**:
    - Precision: 0.63
    - Recall: 1.00
- **Tóm tắt**: Hồi quy Logistic cho thấy kết quả khả quan trong việc phát hiện các trường hợp có đăng ký nhưng có thể tạo ra một số dự đoán sai cho lớp không có đăng ký.

3.4.2 Kết Quả Mô Hình Decision

```
Decision Tree:
Accuracy: 0.82
Precision: 0.63
Recall: 0.89
F1 Score: 0.74
              precision    recall  f1-score   support

           0       0.95      0.79      0.86       834
           1       0.63      0.89      0.74       336

    accuracy                           0.82      1170
   macro avg       0.79      0.84      0.80      1170
weighted avg       0.86      0.82      0.83      1170
```

- **Độ chính xác (Accuracy): 0.82**
  - Độ chính xác 82% cho thấy mô hình Cây Quyết Định cũng hoạt động tốt trong việc phân loại nhưng thấp hơn một chút so với hồi quy logistic.
- **Precision: 0.63**
  - Precision ở đây vẫn là 63%, cho thấy rằng tỷ lệ sai lầm trong dự đoán lớp 1 không thay đổi so với hồi quy logistic.
- **Recall: 0.89**
  - Cây Quyết Định phát hiện được 89% số trường hợp có đăng ký, kém hơn hồi quy logistic nhưng vẫn cho thấy khả năng phát hiện tốt. Điều

này chỉ ra rằng mô hình Cây Quyết Định có khả năng phát hiện nhiều trường hợp có đăng ký nhưng không mạnh mẽ như hồi quy logistic.

- **F1 Score: 0.74**
    - F1 Score cho thấy mô hình vẫn duy trì một mức độ chính xác tốt nhưng thấp hơn so với hồi quy logistic, cho thấy mô hình này không đạt được sự cân bằng tốt giữa precision và recall như mô hình trước đó.
- **Bảng Chi Tiết Phân Tích**:
    - **Hạng mục 0 (Không có đăng ký)**:
        - Precision: 0.95
        - Recall: 0.79
    - **Hạng mục 1 (Có đăng ký)**:
        - Precision: 0.63
        - Recall: 0.89
- **Tóm tắt**: Mô hình Cây Quyết Định hoạt động tốt trong việc phân loại không có đăng ký, tuy nhiên, tỷ lệ sai sót trong việc phát hiện các trường hợp có đăng ký vẫn còn.

3.5. So Sánh và Kết Luận

- Cả hai mô hình đều cho thấy khả năng dự đoán đáng kể với độ chính xác cao trên dữ liệu thử nghiệm.
- Hồi quy logistic thể hiện sự vượt trội về recall, giúp phát hiện chính xác tất cả các trường hợp có đăng ký, trong khi cây quyết định có sự cân bằng tốt hơn giữa độ chính xác cho hai lớp.
- Quyết định chọn mô hình nào phụ thuộc vào mục tiêu cụ thể của dự án:
    - Nếu mục tiêu là phát hiện tất cả các trường hợp có đăng ký mà không bỏ sót, Hồi quy Logistic là lựa chọn tốt hơn.
    - Nếu cần một mô hình với khả năng phân loại tốt cho cả hai lớp và giảm thiểu số lượng dự đoán sai cho lớp không có đăng ký, Cây Quyết Định có thể là lựa chọn phù hợp hơn.

4. Tổng kết và đề xuất.

4.1. Tình Trạng Đăng Ký

- **Dữ liệu**: Trong tổng số 3,900 khách hàng, có **2,847 khách hàng không đăng ký** (chiếm 73%) và **1,053 khách hàng đã đăng ký** (chiếm 27%).
- **Nhận định**: Sự chênh lệch lớn giữa số lượng khách hàng không đăng ký và đã đăng ký cho thấy có một lượng lớn khách hàng tiềm năng mà công ty cần tập trung vào để cải thiện tỷ lệ đăng ký.

4.2. Các Nhân Tố Ảnh Hưởng

Trong nghiên cứu này, có bốn yếu tố chính:

- **Giới tính (Gender)**
- **Giảm giá (Discount)**
- **Mã khuyến mãi (Promo Code)**
- **Số lần mua hàng trước đó (Previous Purchases)**

4.2.1 Giới Tính (Gender)

- **Dữ liệu**: Số lượng khách hàng đăng ký chủ yếu là nam. Cụ thể, nếu chỉ có nam giới đăng ký trong số những khách hàng đã thực hiện việc đăng ký, điều này cho thấy sự thiên lệch về giới tính trong việc tiếp cận và tham gia dịch vụ.
- **Phân tích**: Việc chủ yếu có nam giới đăng ký có thể liên quan đến những yếu tố văn hóa và hành vi tiêu dùng. Nam giới có thể cảm thấy sản phẩm hoặc dịch vụ phù hợp với nhu cầu của họ hơn so với nữ giới.
- **Đề xuất**: Công ty cần xem xét lại chiến lược tiếp thị để tạo ra các chương trình thu hút cả nam và nữ. Cần phát triển nội dung quảng cáo và chương trình khuyến mãi phù hợp với nhu cầu và sở thích của cả hai giới nhằm tăng cường mức độ tham gia.

4.2.2. Giảm Giá (Discount)

- **Dữ liệu**: Khách hàng đã đăng ký có tỷ lệ sử dụng giảm giá cao hơn so với khách hàng không đăng ký.
- **Phân tích**: Giảm giá tạo ra cảm giác giá trị cao hơn cho khách hàng và khuyến khích họ thực hiện hành động đăng ký. Tuy nhiên, nhiều khách hàng không đăng ký có thể không cảm thấy được khuyến khích hoặc không có đủ thông tin về các chương trình giảm giá.
- **Đề xuất**: Tăng cường quảng bá về các chương trình giảm giá, đặc biệt cho nhóm khách hàng không đăng ký, có thể giúp tăng cường khả năng đăng ký. Cần thực hiện các chiến dịch truyền thông mạnh mẽ hơn để giới thiệu rõ ràng các lợi ích của việc đăng ký, bao gồm các chương trình giảm giá hấp dẫn.

4.2.3 Mã Khuyến Mãi (Promo Code)

- **Dữ liệu**: Việc sử dụng mã khuyến mãi được ghi nhận là một trong những yếu tố quan trọng ảnh hưởng đến quyết định đăng ký của khách hàng.
- **Phân tích**: Khách hàng có mã khuyến mãi có xu hướng cảm thấy được hưởng lợi nhiều hơn, từ đó kích thích họ tham gia vào việc đăng ký. Ngược lại, những khách hàng không có mã khuyến mãi có thể cảm thấy thiếu động lực.

- **Đề xuất**: Công ty cần triển khai các chương trình mã khuyến mãi hấp dẫn cho cả nhóm khách hàng không đăng ký để khuyến khích họ thực hiện hành động đăng ký.

4.2.4. Số Lần Mua Hàng Trước Đó (Previous Purchases)

- **Dữ liệu**: Những khách hàng có số lần mua hàng trước đó cao có xu hướng đăng ký dịch vụ cao hơn.
- **Phân tích**: Sự hài lòng từ những lần mua hàng trước có thể dẫn đến lòng trung thành, nhưng với tỷ lệ không đăng ký cao, có thể có nhiều khách hàng chưa trải nghiệm đủ tích cực để chuyển sang đăng ký.
- **Đề xuất**: Cần có các chương trình chăm sóc khách hàng và ưu đãi dành riêng cho những khách hàng đã mua hàng nhiều lần để khuyến khích họ đăng ký dịch vụ.

4.3. Kết Luận

Phân tích cho thấy rằng có một lượng lớn khách hàng không đăng ký, với 2,847 khách hàng không đăng ký trong tổng số 3,900. Những khách hàng đã đăng ký chủ yếu là nam giới, cho thấy sự cần thiết phải điều chỉnh chiến lược tiếp thị và chăm sóc khách hàng để nâng cao tỷ lệ đăng ký, đặc biệt là đối với khách hàng nữ.

Bằng cách áp dụng các đề xuất như tăng cường quảng bá các chương trình khuyến mãi, phát triển nội dung quảng cáo đa dạng và nhắm đến cả hai giới, cũng như cải thiện dịch vụ khách hàng cho những khách hàng đã mua hàng nhiều lần, công ty có thể nâng cao trải nghiệm của khách hàng và thúc đẩy tỷ lệ đăng ký cao hơn trong tương lai.