

DATA ANALYTICS REPORT

1.1. Tổng quan tập dữ liệu

- Bộ dữ liệu chứa thông tin đặc điểm và hành vi sử dụng 4G của các khách hàng nước ngoài du lịch đến Việt Nam, bên cạnh đó là thông tin đề xuất các gói Data 4G của Viettel.
- Tập dữ liệu (train set) có 4 bảng:
 - context: **thuộc tính về chuyến đi** (11572 hàng× 12 cột)
 - user: **thuộc tính người dùng** (11572 hàng× 16 cột)
 - mobile_plan_user: **đề xuất gói Data** cho người dùng (45321 hàng× 3 cột)
 - mobile_plan_attr: **thông tin về gói Data** (5 hàng 4 cột)

1.2. Mục tiêu phân tích

Hiểu được thông tin được cung cấp từ bộ dữ liệu

Phân tích và khám phá thông tin từ bộ dữ liệu

Xây dựng mô hình phân loại cho các gói Data 4G

Đề xuất các giải pháp và chiến lược kinh doanh

1.3. Làm sạch và chuyển hóa dữ liệu

1.3.1. Missing values

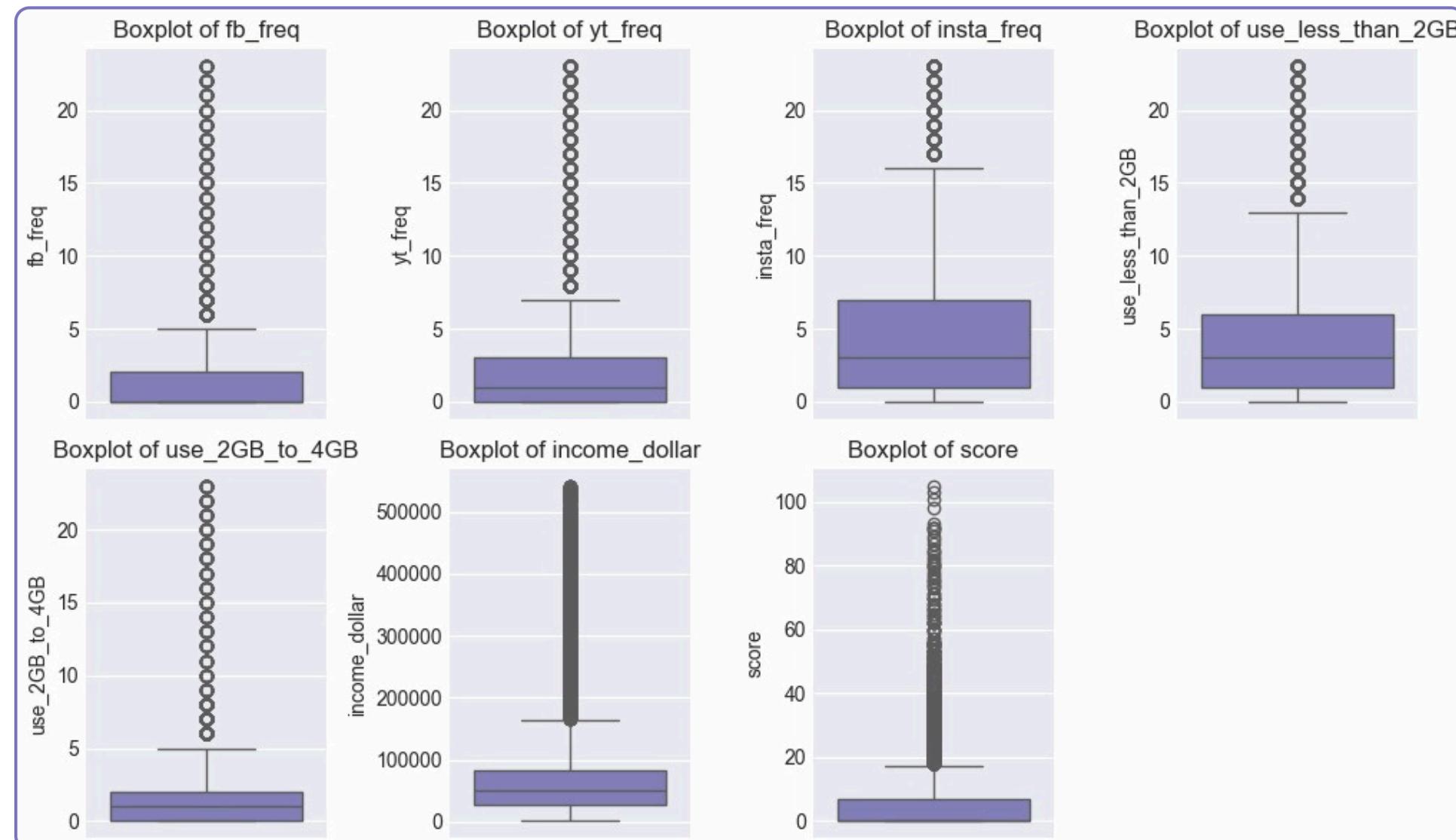
Cột	Vấn đề	Giải pháp
'education'	4003 giá trị null (34.59%)	Chuyển thành “unknown”
'mobile_plan'	2715 giá trị null (5.99%)	Loại bỏ
'accept'	2721 giá trị null (6.00%)	Loại bỏ

1.3.2. Error values

Cột	Vấn đề	Giải pháp
'go_with' & 'weather' & 'living_with'	Chứa ký tự đặc biệt (&%\$!~?)	Loại bỏ bằng công cụ regex
'time'	Format thời gian khác nhau	Quy về một format
'job'	Có nhiều giá trị khác nhau và đa ngôn ngữ (hơn 60%)	Loại bỏ, dùng cột 'profession' biểu thị nghề nghiệp
'income'	Có hai loại tiền tệ USD và VND	Quy về USD (1USD = 25,000VND) thành cột 'income_dollar'

1.3. Làm sạch và chuyển hóa dữ liệu

1.3.3. Outliers



- Các cột chứa outlier đáng kể:**

- 'income_dollar'
- 'fb_freq'
- 'yt_freq'
- 'insta_freq'
- 'score'
- 'use_less_than_2GB'
- 'use_2GB_to_4GB'

- Phương pháp kiểm tra:** Interquartile range (Độ trải giữa)

- Giải pháp:**

- Tạo thêm một cột phụ cho mỗi cột dùng để xét bất thường (bất thường là 1, ngược lại là 0)
- Riêng 'income_dollar' còn được tạo thêm 1 cột chia thu nhập theo level

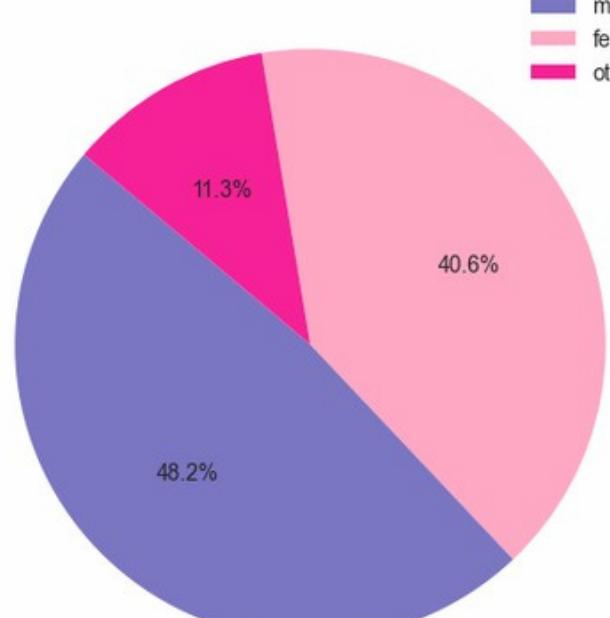
1.3.4. Transformation

Cột	Vấn đề	Giải pháp
'to_hanoi' & 'to_other'	Đã được thể hiện ở cột 'direction'	Loại bỏ
'education' & 'profession'	Có nhiều giá trị khác nhau	Phân lại vào các category
'nation'	Có nhiều quốc gia khác nhau	<ul style="list-style-type: none"> - Phân lại thành các châu lục - Đổi tên thành 'continent'
'living_with'	Nhiều tình trạng hôn nhân khác nhau và có kèm theo số con	<ul style="list-style-type: none"> - Phân lại tình trạng hôn nhân - Tách thành 2 cột 'marital_status' và 'children'

2.1. Phân tích đơn biến

Giới tính (Gender)

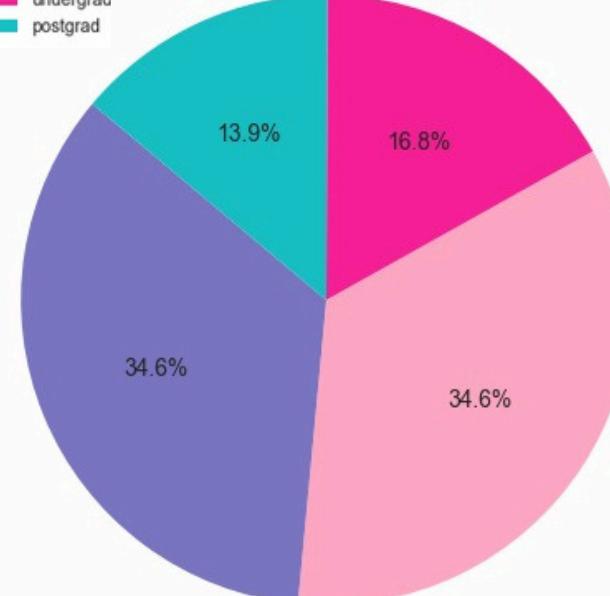
Distribution of Gender



Nam chiếm tỷ lệ cao nhất (48.2%)

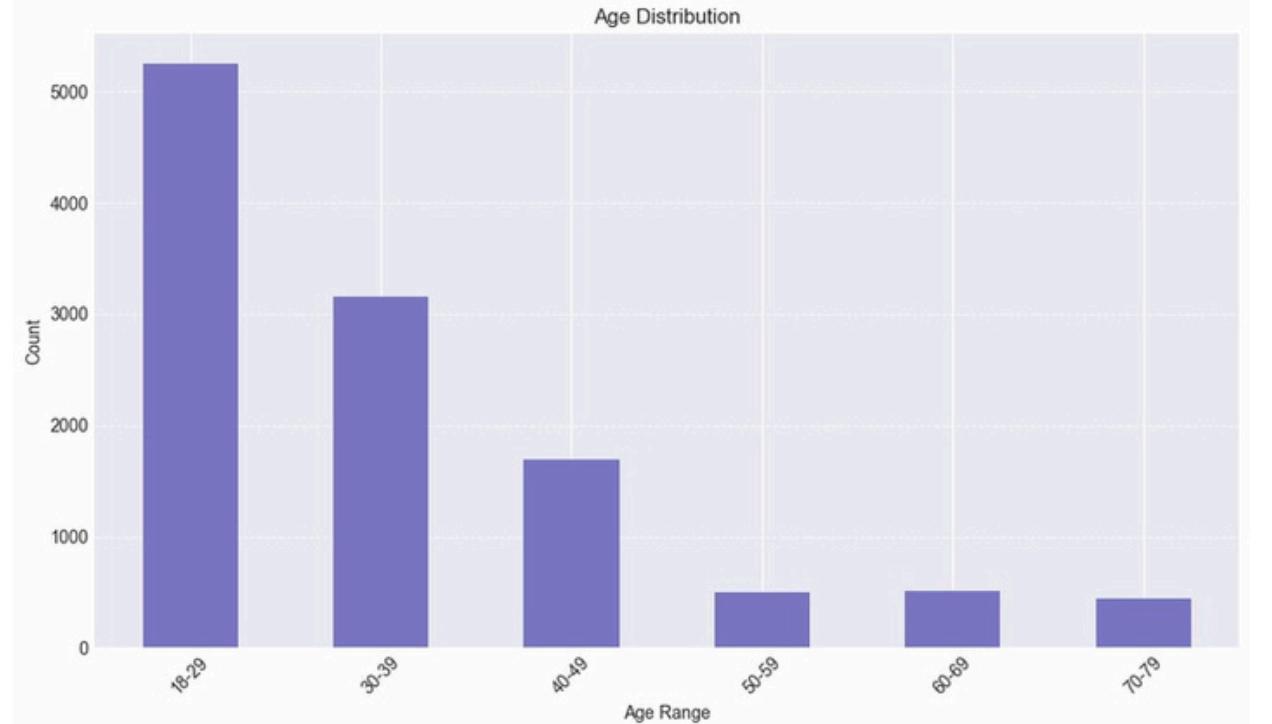
Giới tính (Gender)

Distribution of Education



Tỷ lệ du khách đã tốt nghiệp và không rõ là như nhau. 2 nhóm còn lại chiếm tỷ lệ dưới 20% mỗi nhóm.

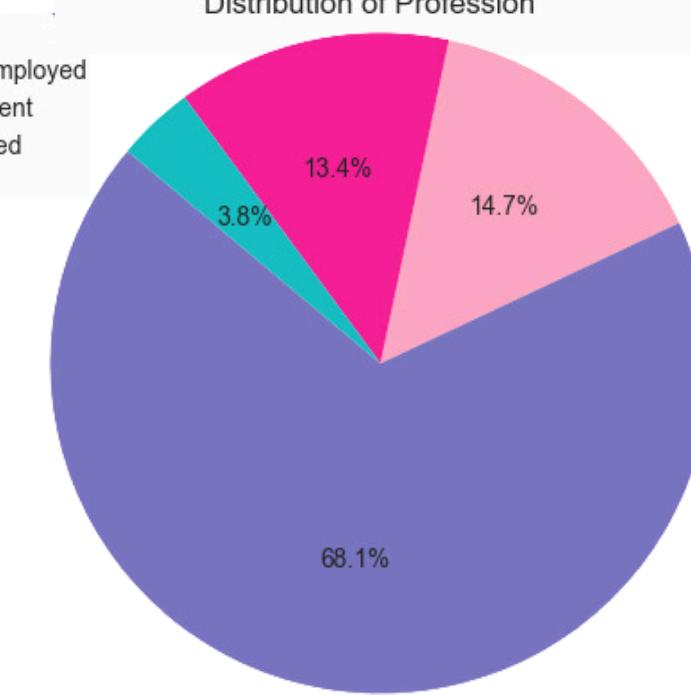
Tuổi tác (Age)



- Tệp tuổi 18-29 chiếm tỷ trọng lớn nhất. Tiếp theo là tệp 30-39 tuổi.
- Các nhóm tuổi còn lại giảm dần về số lượng khi tuổi càng tăng.

Tình trạng việc làm (Profession)

Distribution of Profession

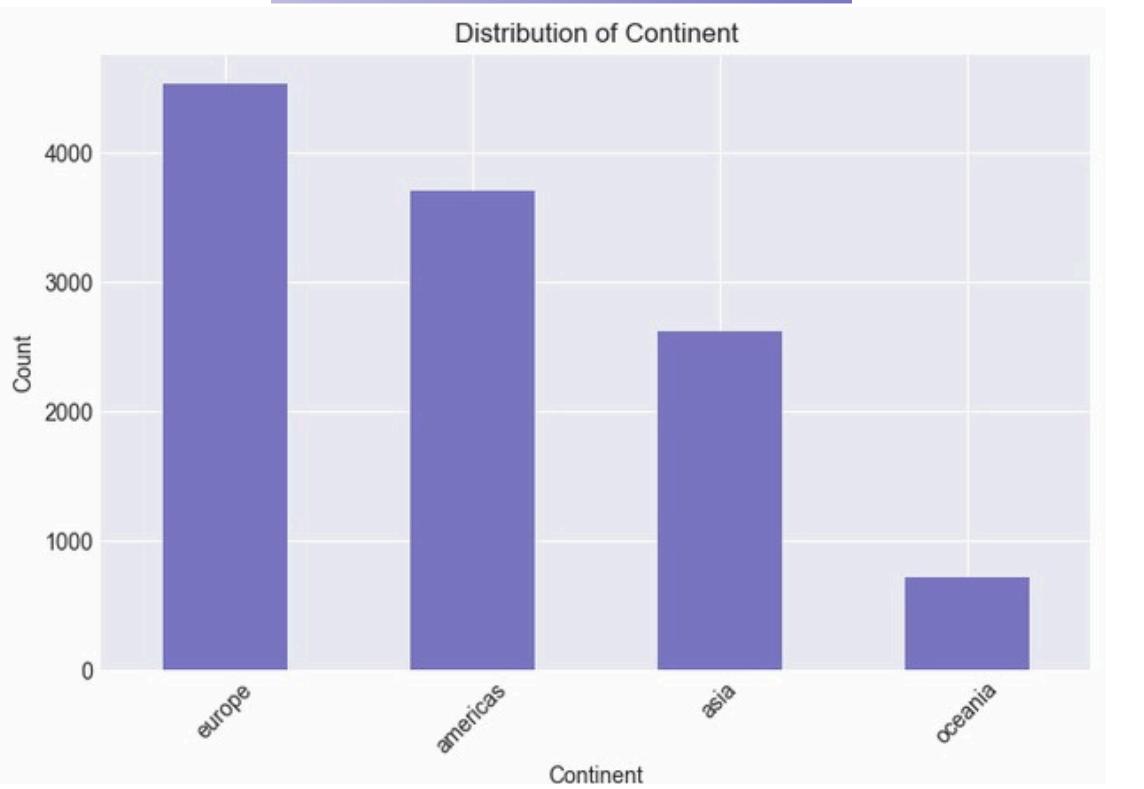


- Phần lớn du khách đều có việc làm, chiếm 68.1%.
- Tệp người nghỉ hưu chiếm tỷ trọng thấp nhất, chỉ khoảng 3.8%.



2.1. Phân tích đơn biến

Châu lục (Continent)



- Đa số là du khách Châu Âu,
- Châu Đại Dương, cụ thể là nước Úc, chiếm số lượng ít nhất.

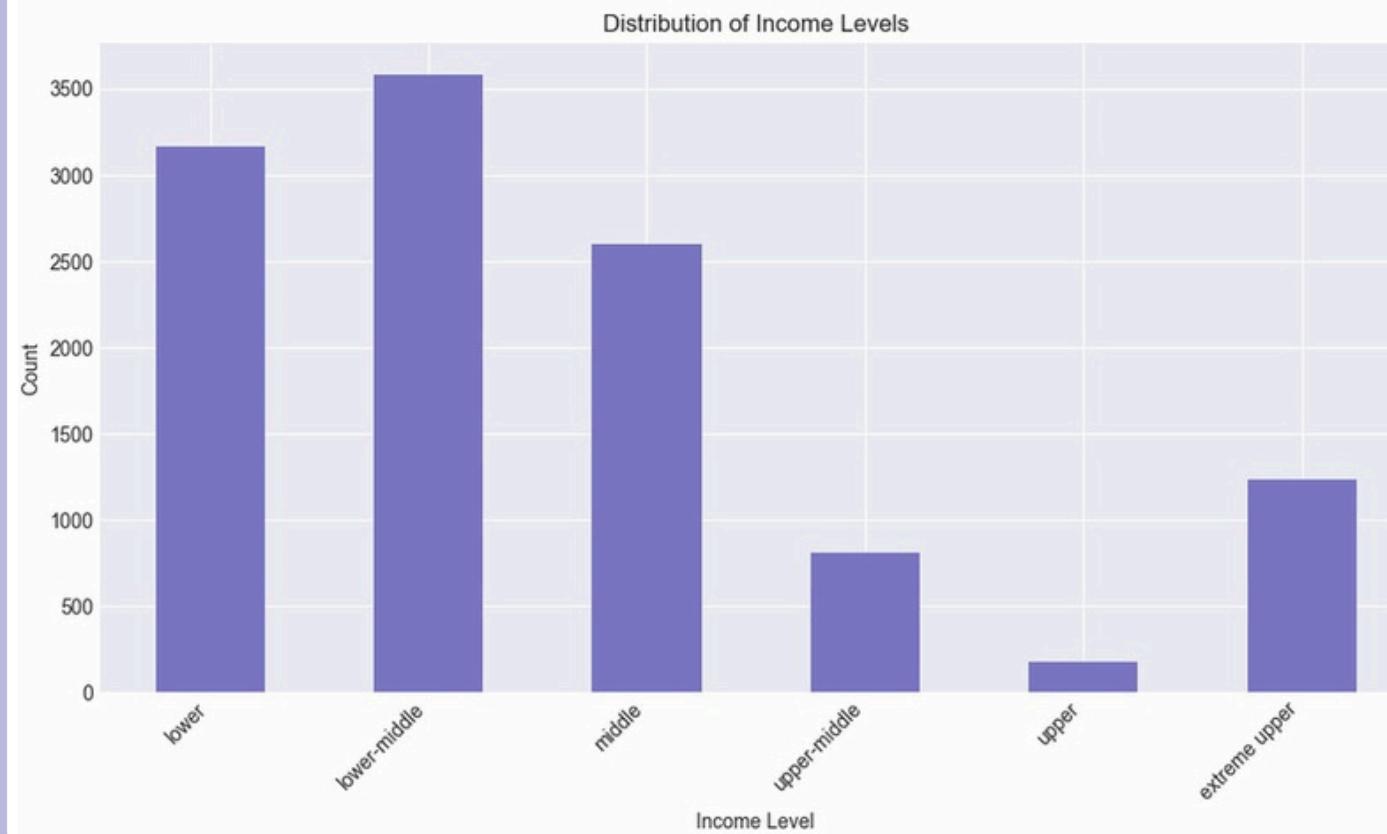
Điểm đến châu Á có mức **tăng trưởng nhanh thứ 3** về lượt tìm kiếm nơi lưu trú **từ du khách châu Âu**.

(Agoda, 2024)

- Thị trường châu Âu được miễn thị thực ngày càng nhiều
- Du khách châu Âu tăng nhu cầu khám phá sự đa dạng văn hóa, ẩm thực, cảnh quan thiên nhiên... qua việc du lịch các nước châu Á.

(Cục du lịch Quốc Gia Việt Nam, 2023)

Nhóm thu nhập (Income Level)



3 nhóm thu nhập hàng đầu gồm:
thấp, cận TB và TB.

Nhóm thu nhập siêu cao chiếm
tỷ trọng đáng kể.

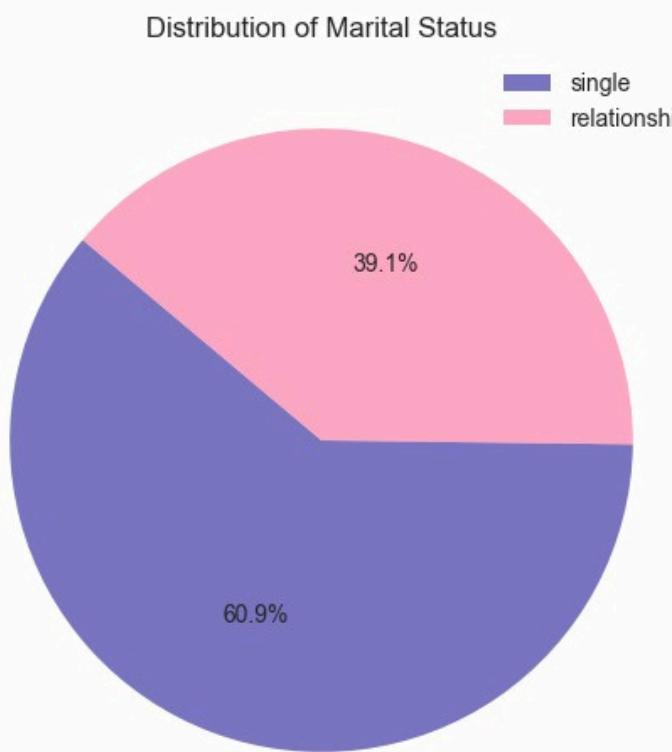
Việt Nam được xem là một trong 10 điểm đến du lịch rẻ nhất châu Á.(Báo Tuổi Trẻ, 2023)

→ **Phù hợp cho người có thu nhập thấp - trung bình.**

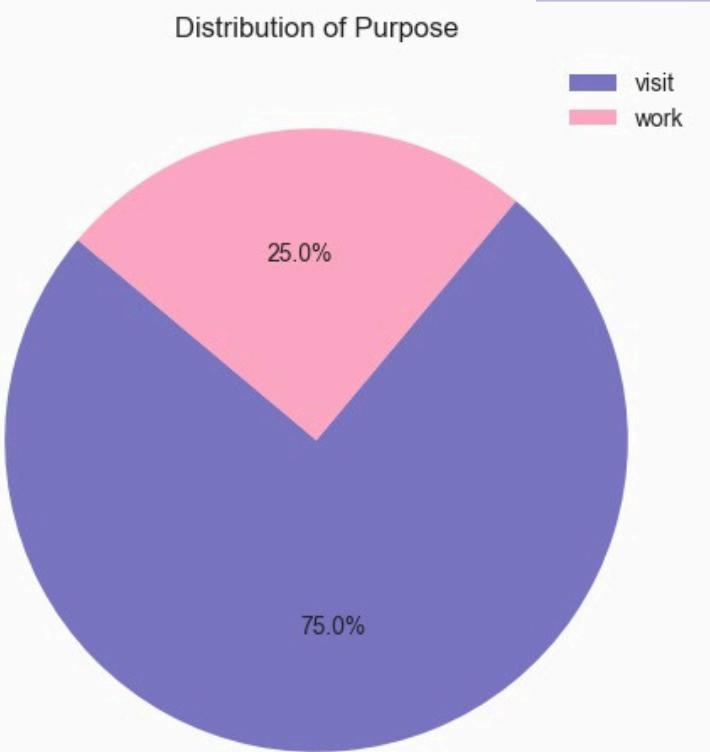
Do **mức lương không đồng đều** giữa các quốc gia, châu lục, điển hình như Mỹ, Đan Mạch luôn nằm trong top 10 quốc gia có lương trung bình cao nhất thế giới.

(CEOWORLD, 2022)

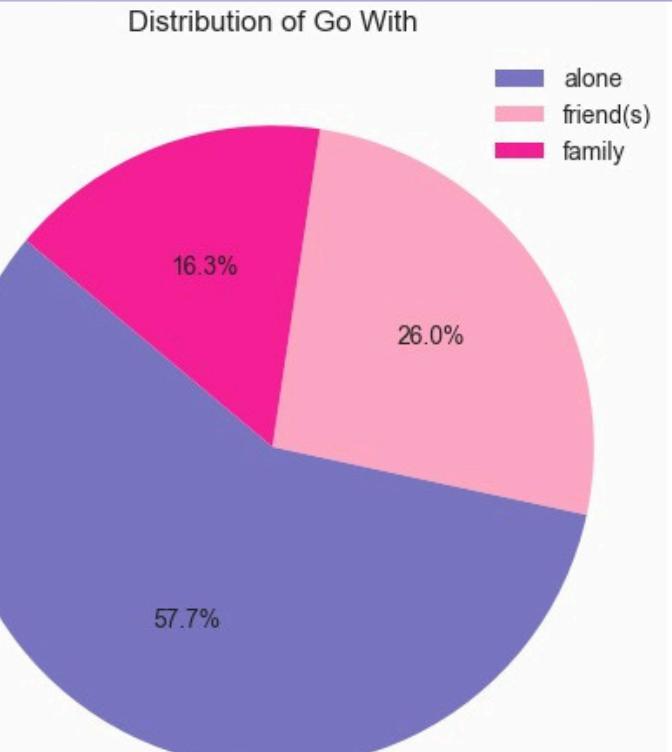
2.1. Phân tích đơn biến (Univariate analysis)



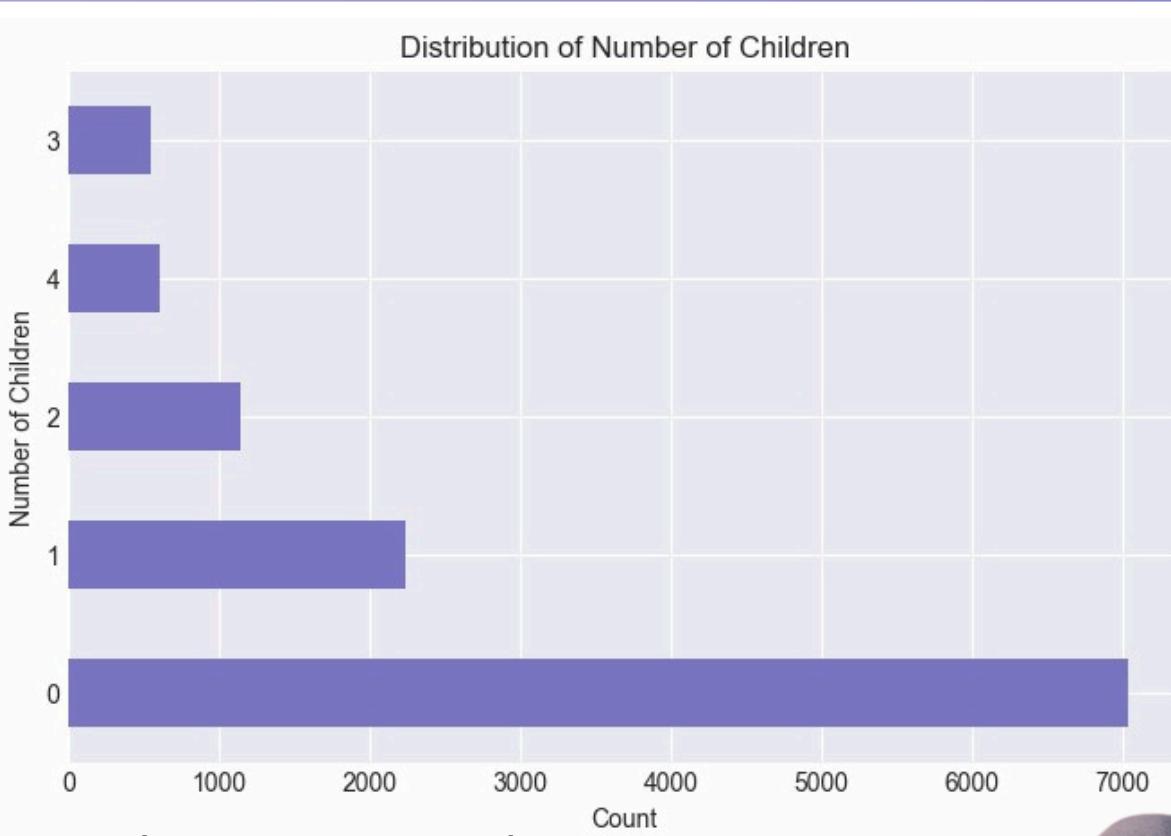
60.9% du khách là đối tượng độc thân.



¾ du khách đến để du lịch
¼ còn lại là mục đích công việc.

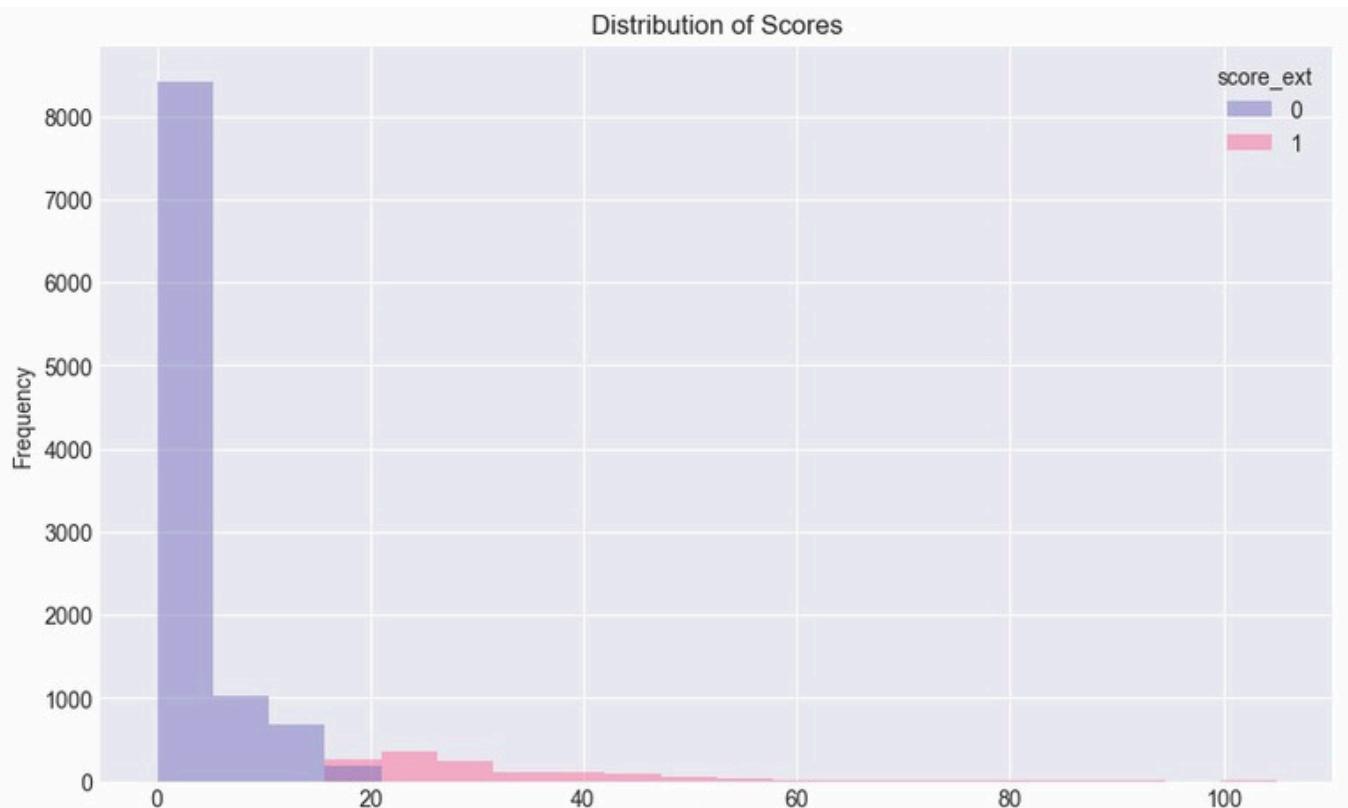


26% trong số họ đi du lịch với bạn bè,
phần còn lại đi với gia đình.



Phần lớn du khách đều không có con.

Điểm Viettel++ (Score)



Đa số khách hàng không có hoặc có rất ít điểm Viettel.

Phần lớn là khách hàng mới, hoặc không phải là khách hàng trung thành đối với Viettel. Vì vậy, họ sẽ có xu hướng so sánh và lựa chọn giữa các dịch vụ khác nhau, quan tâm nhiều đến các yếu tố tiện lợi, giá cả và độ uy tín thương hiệu.

Du khách đến VN sẽ là người nước ngoài, trẻ hoặc trưởng thành (18-39 tuổi), theo chủ nghĩa YOLO, chưa có ràng buộc về quan hệ hay con cái. Họ ưu tiên đi một mình với mục đích trải nghiệm. Họ chưa phải là khách hàng trung thành với Viettel.



2.2. Phân tích đa biến (Multivariate analysis)

2.2.1 Đặc điểm về khách hàng

Age x Income level x Gender

Nam và Nữ: Tuổi trung bình càng tăng thì mức lương càng cao.

- Thu nhập thấp, TB, cận TB:
 - Nữ giới chiếm số lượng lớn hơn so với các nhóm thu nhập còn lại.
 - Nam giới có số lượng khá thấp.
 - Độ tuổi trung bình của Nam và Nữ khá giống nhau (dao động trên dưới 30 tuổi).
- Thu nhập cao và siêu cao: độ tuổi trung bình Nữ trẻ hơn so với Nam.

Other:

- Độ tuổi trung bình ở mức thu nhập cao và siêu cao trẻ hơn so với Nam giới.
- Số lượng người có thu nhập cao lại ít hơn nhiều so với 2 giới còn lại.

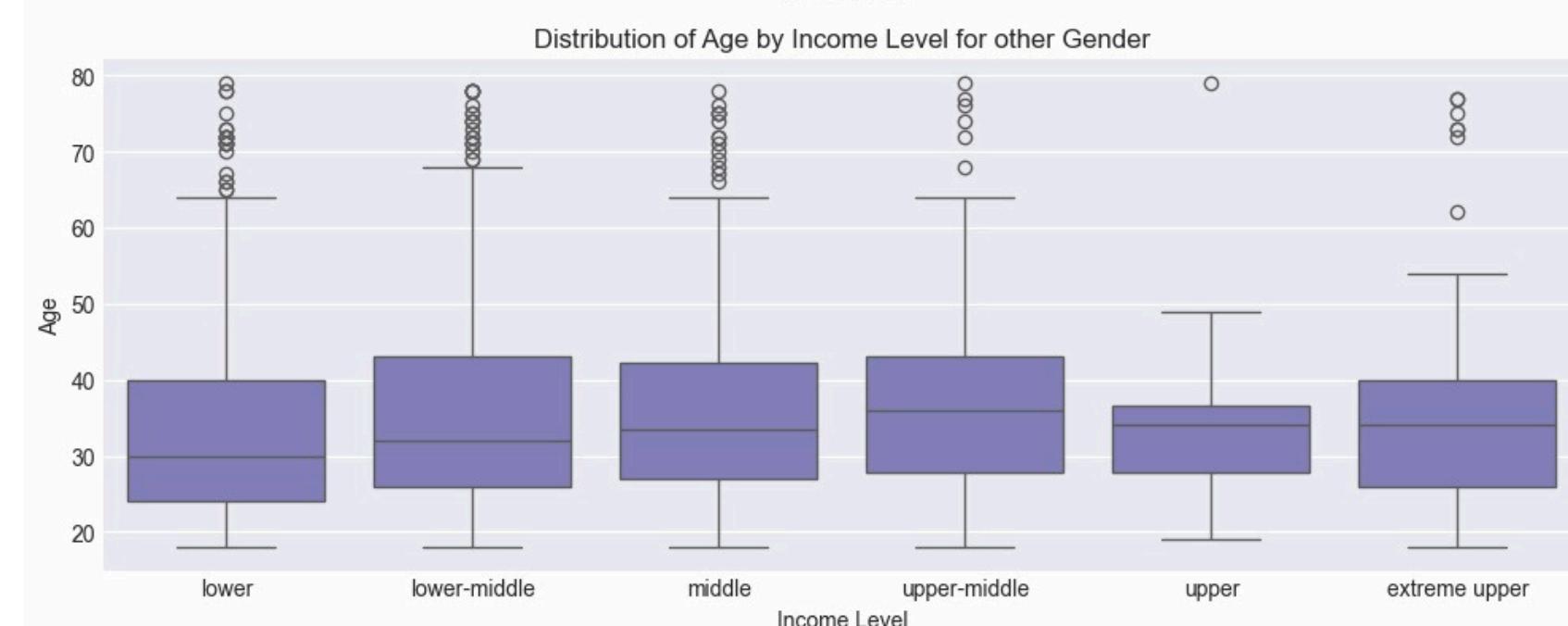
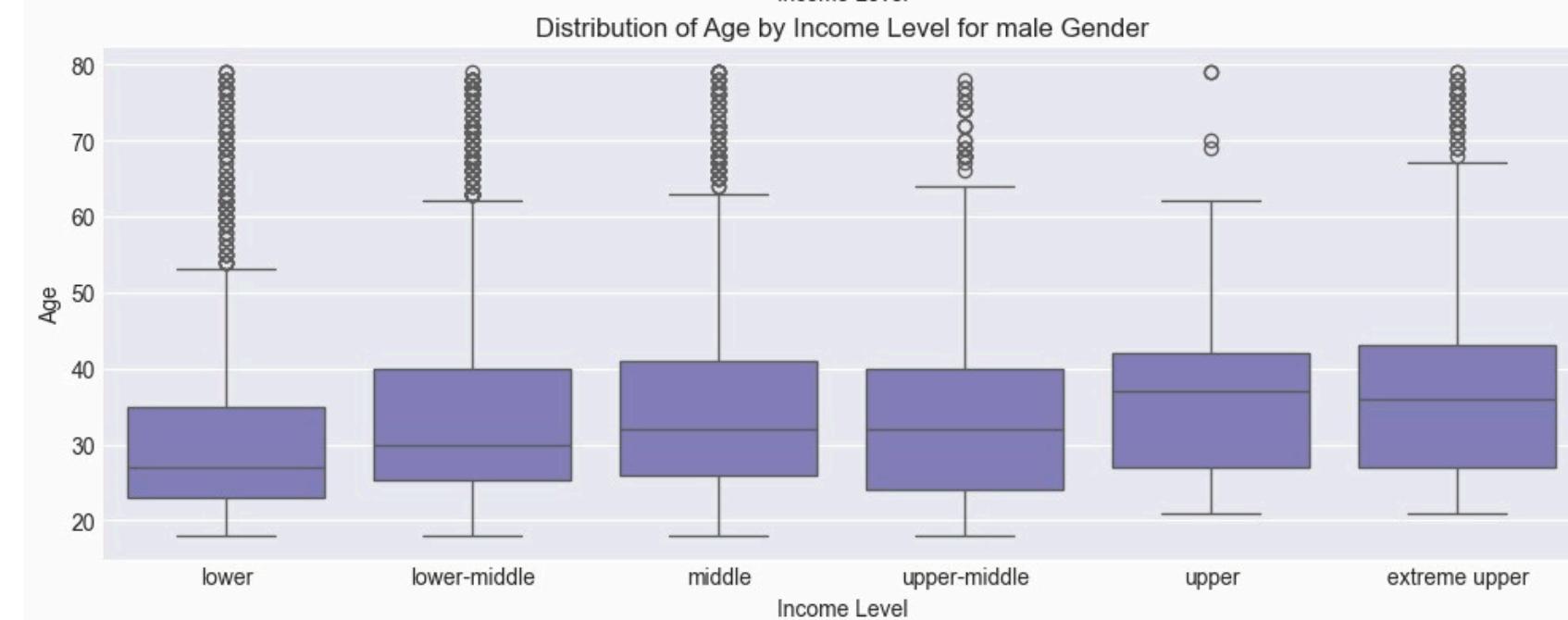
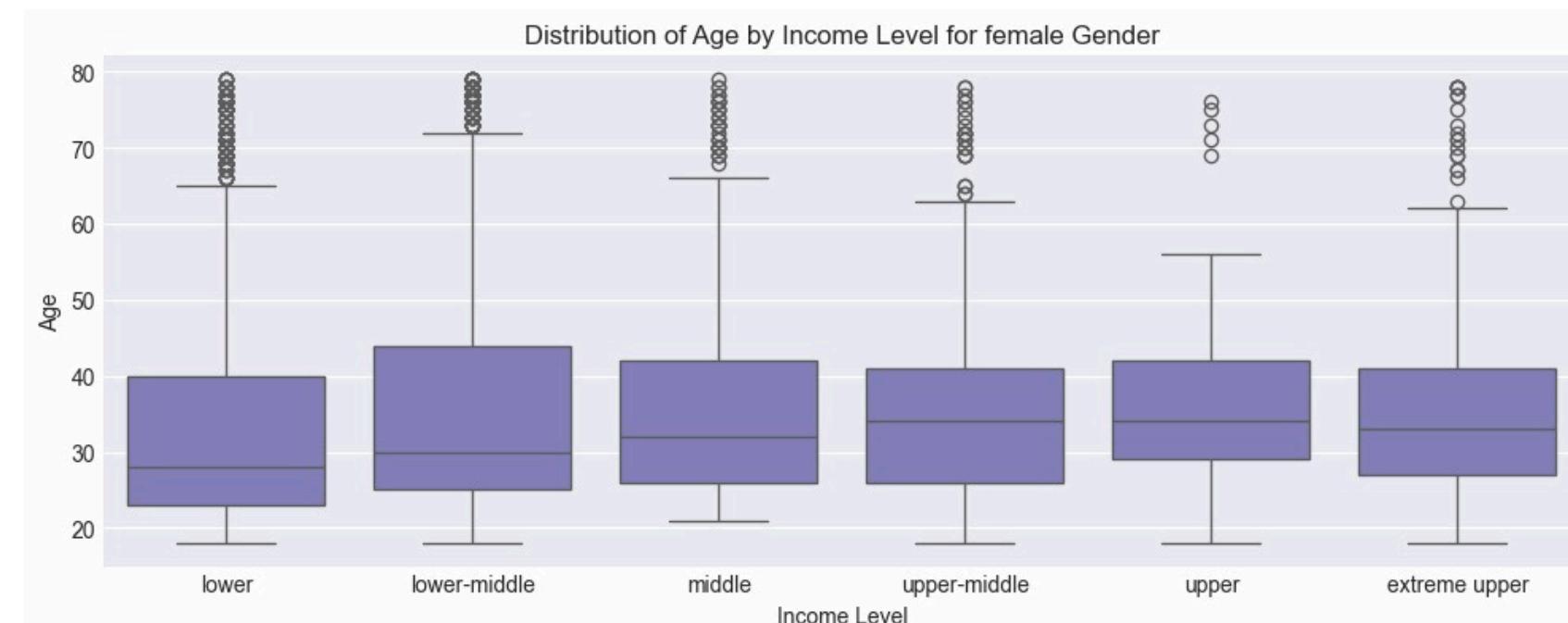


Thu nhập thấp - TB

- Độ tuổi: Khách hàng trẻ thì có thu nhập thấp hơn
- Tỷ trọng: Nữ và Other sẽ chiếm tỷ trọng cao hơn Nam.

Thu nhập cao

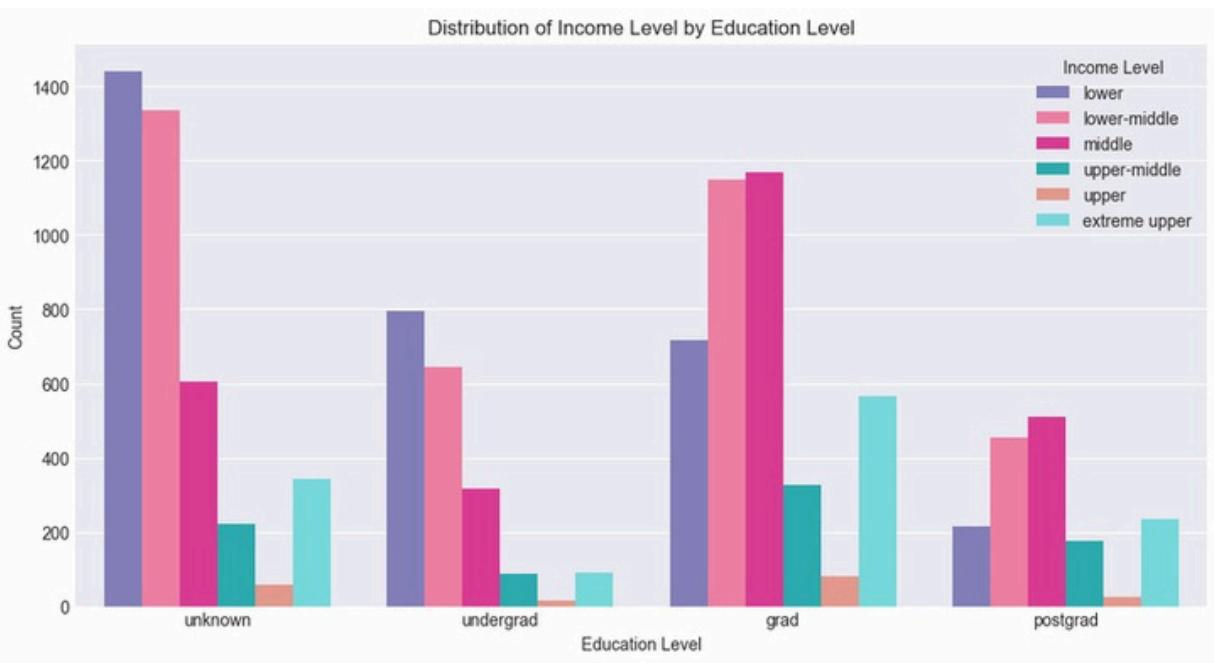
- Tỷ trọng: Phân bổ nhiều hơn ở Nam giới.
- Độ tuổi: Nữ và Other lại có độ tuổi trẻ hơn khi đạt đến những mức thu nhập cao này.



2.2. Phân tích đa biến (Multivariate analysis)

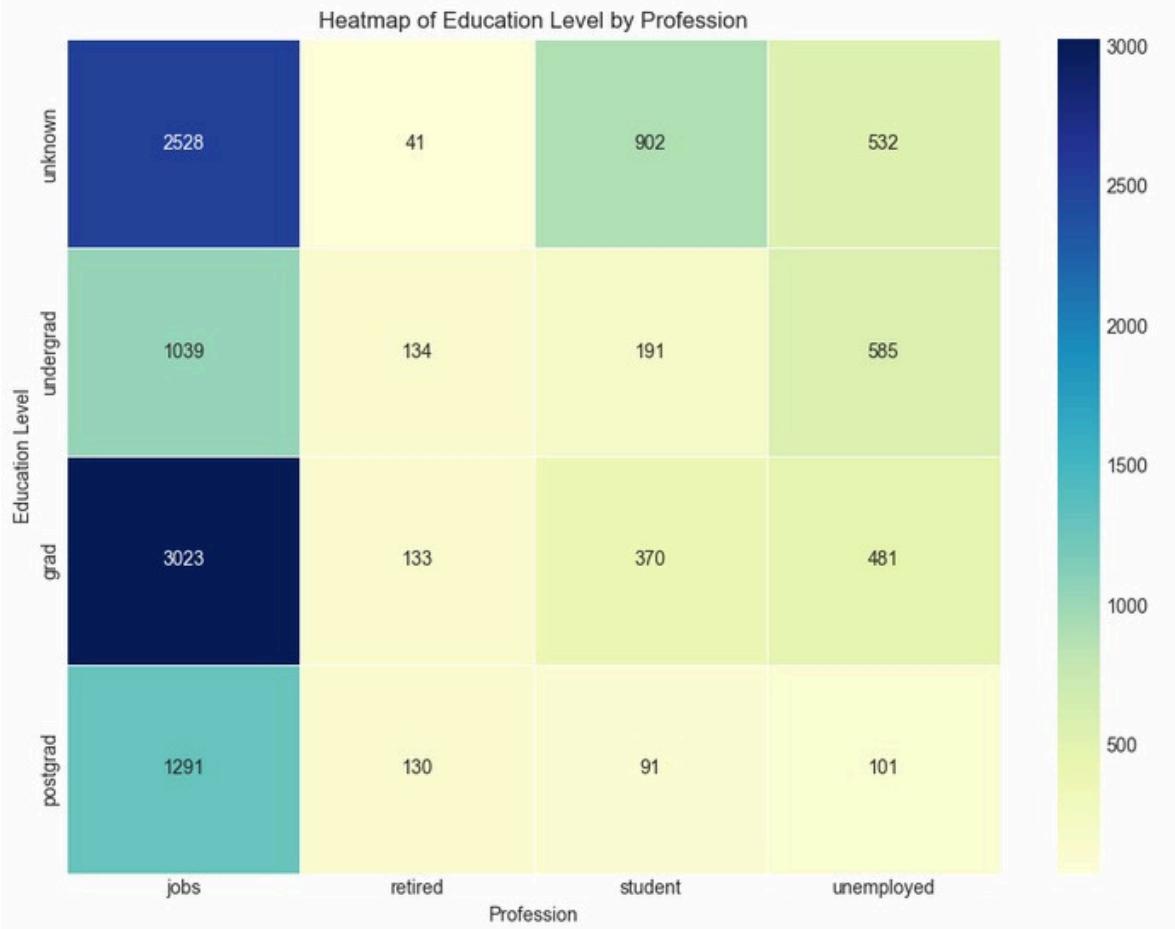
2.2.1 Đặc điểm về khách hàng

Education x Income level



- Nhóm **Undergrad** chủ yếu rơi vào **nhóm thu nhập thấp**.
- Nhóm **Grad** có sự **dịch chuyển về phía thu nhập cao hơn** và có số lượng thu nhập siêu cao nhiều nhất.
- Nhóm **Postgrad** có phân phối thu nhập **khá giống nhóm Grad**.
- Cuối cùng, nhóm **unknown** (không rõ thu nhập) là nhóm có **thu nhập thấp nhiều nhất**.

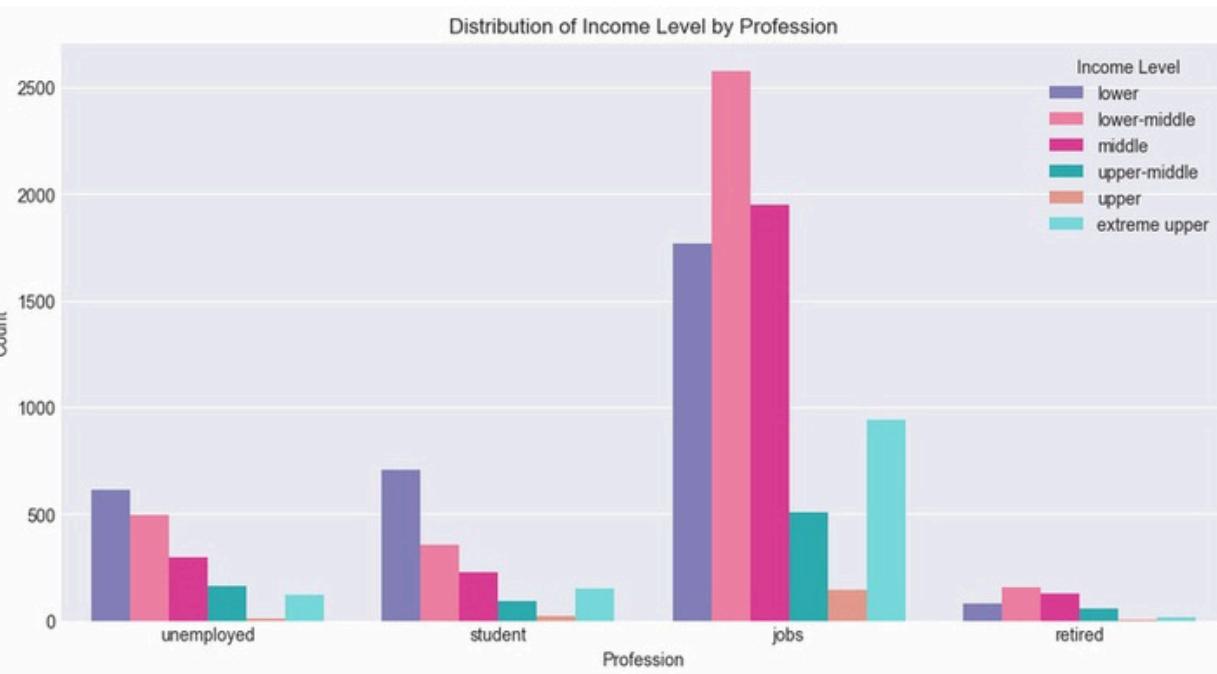
Education x Profession



Phần lớn du khách sẽ thuộc nhóm (**đã tốt nghiệp, có việc làm**) hoặc (**không rõ tình trạng học vấn, có việc làm**). **Điểm chung** của các du khách này là **phần lớn có thu nhập thấp, cận TB và TB**, bên cạnh đó còn có 1 số lượng đáng kể du khách có thu nhập siêu cao.

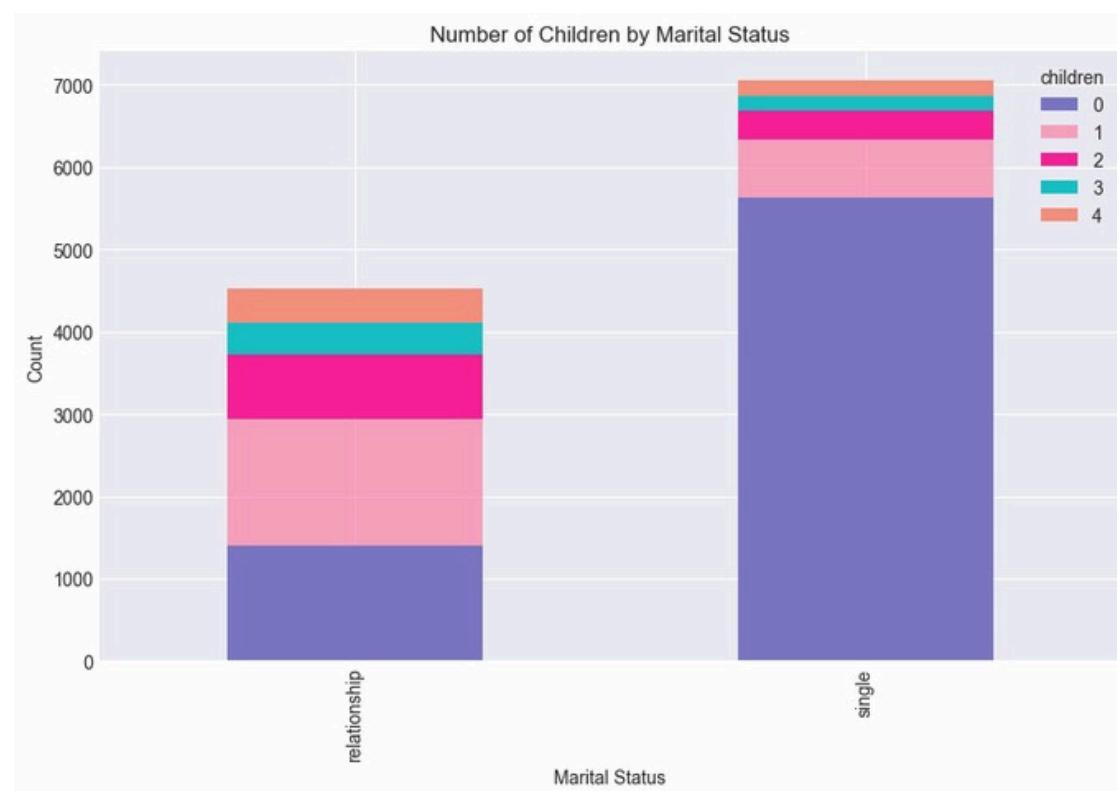
Lưu ý: Nhóm unknown chiếm tỷ trọng lớn và khá quan trọng → nên tìm cách thu thập thêm thông tin.

Profession x Income level



Nhóm du khách **không có việc làm hoặc đang là học sinh** sẽ phân phối nhiều ở mức **thu nhập thấp**. Nhóm **có việc làm** sẽ có phân phối **dịch chuyển về những mức thu nhập cao hơn**.

Marital status x Children



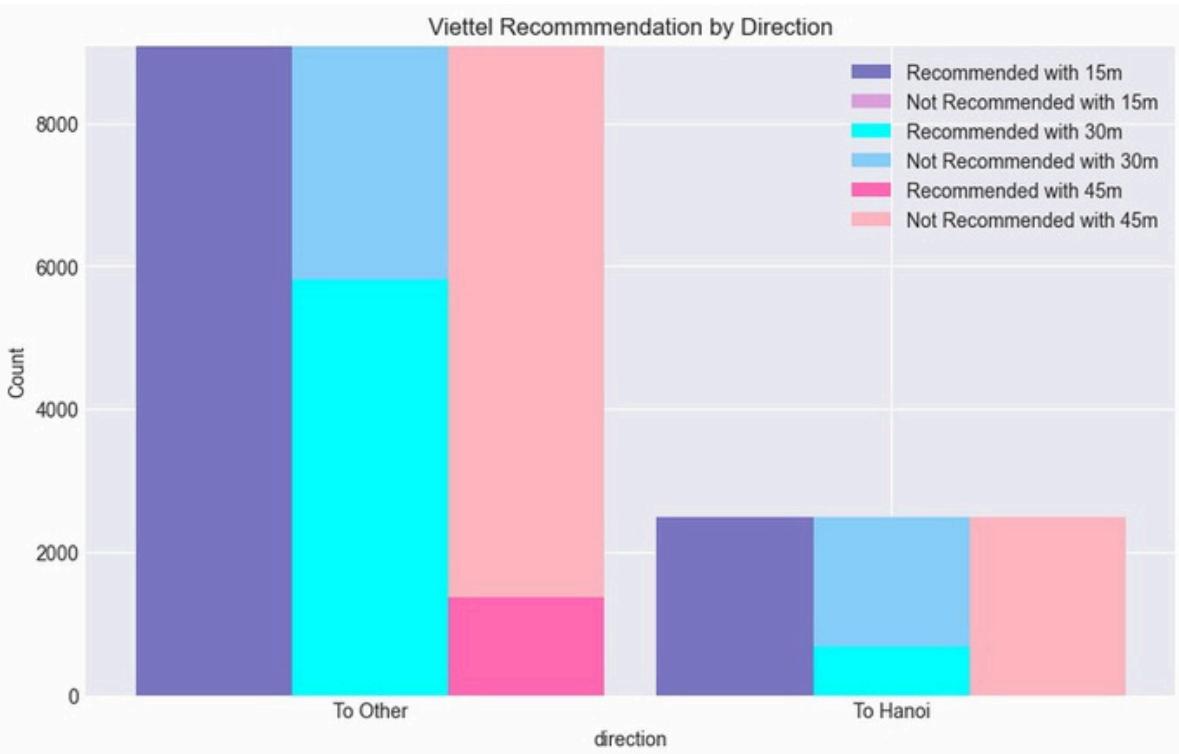
Những du khách **độc thân** hầu hết **không có con**. Tệp du khách **đang trong một mối quan hệ** phần lớn cũng **không có con hoặc có 1-2 con**. Số lượng du khách có trên 2 con là rất ít.

Khách du lịch đến Việt Nam theo dạng gia đình khá hạn chế. Phần lớn đều có việc làm.

2.2. Phân tích đa biến (Multivariate analysis)

2.2.2. Đặc điểm hành vi

Nhận biết gói Viettel x direction

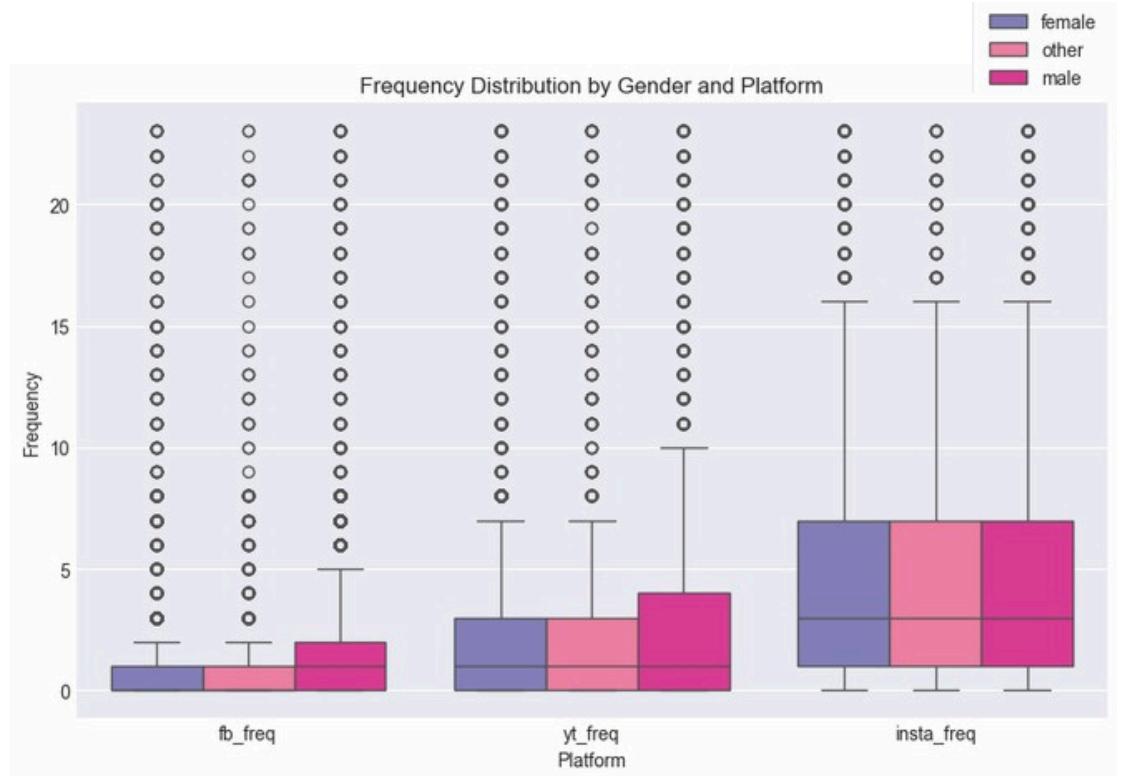


- 100% du khách được đề xuất về Viettel sau 15' đáp xuống sân bay. Nếu du khách được đề xuất về Viettel sau 45' đáp xuống sân bay thì 100% du khách sẽ bay đến tỉnh khác.

→ **Các du khách đến Hà Nội sẽ không ở lại sân bay lâu, ngược lại với du khách đến từ tỉnh khác.**

- Nên marketing những **gói cơ bản, tiện lợi nhất** dành cho các **du khách đến Hà Nội**. Khách đến tỉnh khác sẽ có nhiều thời gian “chết” khi di chuyển, có thể xem xét kỹ giữa nhiều gói đa tính năng hơn.

Frequency (fb, yt, ins) x gender

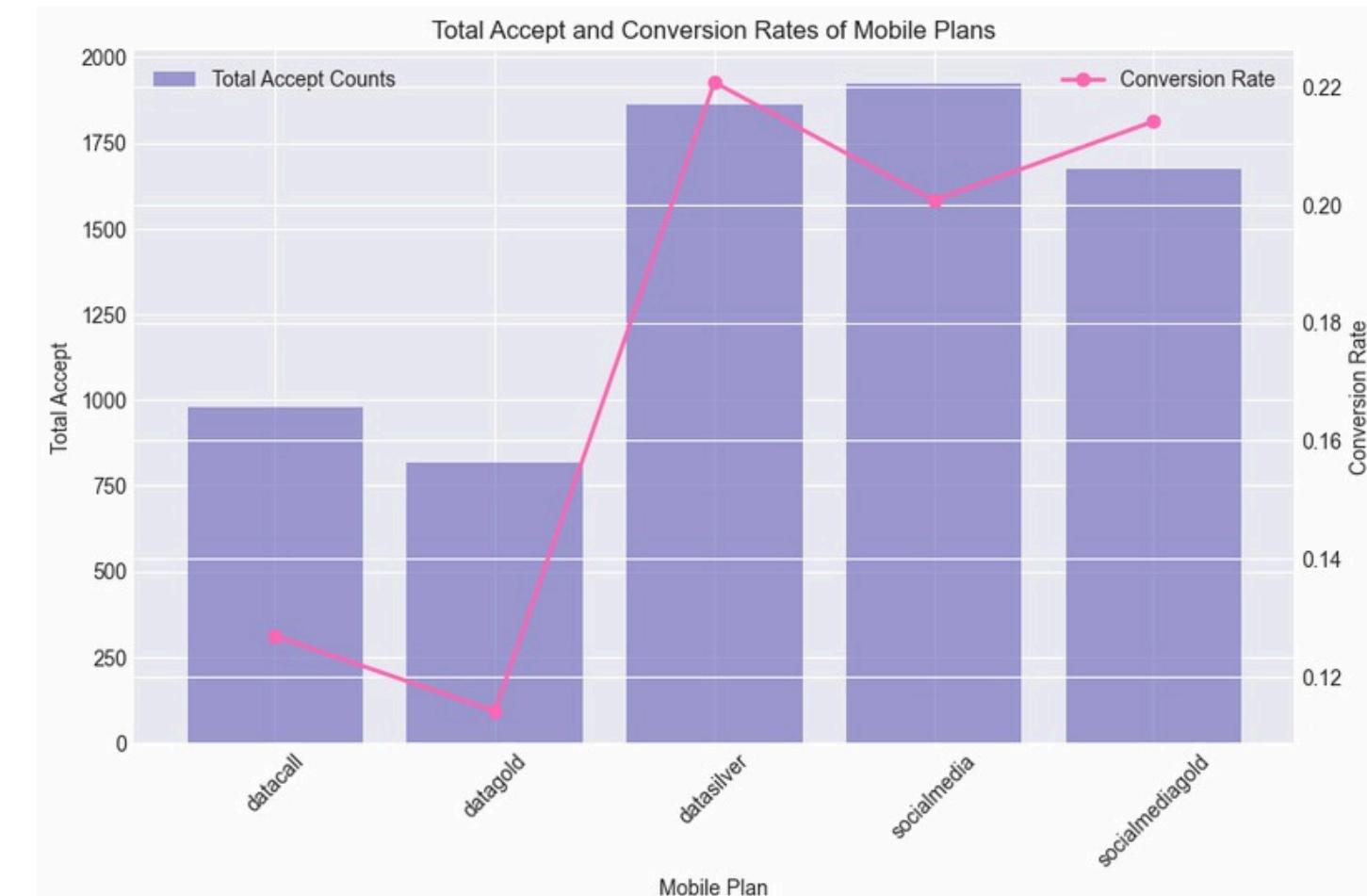


Cả 3 tệp đều dùng INS đồng đều hơn và nhiều hơn 2 trang mạng XH còn lại.

Tệp người trẻ hiện nay đang **dẫn dịch chuyển hành vi sử dụng mạng XH**

- FB: giảm mạnh, nền tảng này được xem là nền tảng kết nối giữa gia đình, “già nua” và không mới mẻ. INS: ngày càng tăng sự thu hút từ giới trẻ vì tính mới mẻ cao, nhiều tính năng và mang tính trẻ trung, riêng tư hơn.

Accept x Mean CVR theo 5 gói



Đa số lượng accept đổ về nhiều từ gói datasilver, socialmedia và socialmediagold.

3 gói này cũng ghi nhận tỷ lệ chuyển đổi cao hơn so với 2 gói còn lại.

Gói datasilver

- Đặc tính: cơ bản nhất, chỉ gồm dung lượng 4G.
- Giá rẻ

Phù hợp khách hàng có thu nhập thấp, chưa trung thành với Viettel để trải nghiệm dịch vụ với giá rẻ

Gói socialmedia, socialmediagold

- Đặc tính: dung lượng không hạn chế truy cập mạng XH

Du khách là người trẻ trưởng thành, độc thân, du lịch một mình, cần dùng mạng XH để cập nhật thông tin, giữ liên lạc...

Gói datacall, datagold

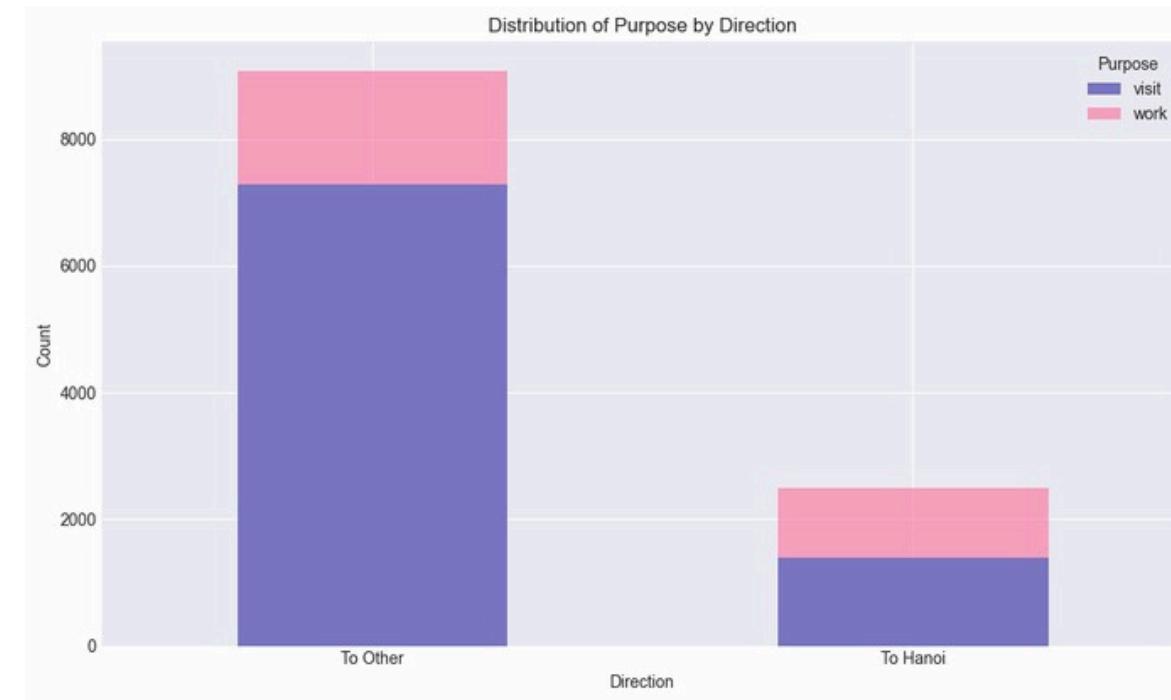
- Đặc tính: chú trọng nghe gọi hoặc dung lượng 4G cao
- Giá cao hơn các gói còn lại

Nhu cầu nghe gọi qua các trang mạng XH có thể thay thế cho nghe gọi trực tiếp.
Giá cao, chưa phù hợp với tệp khách mới muốn dùng thử.

2.2. Phân tích đa biến (Multivariate analysis)

2.2.3. Đặc điểm chuyển đi

Purpose x direction



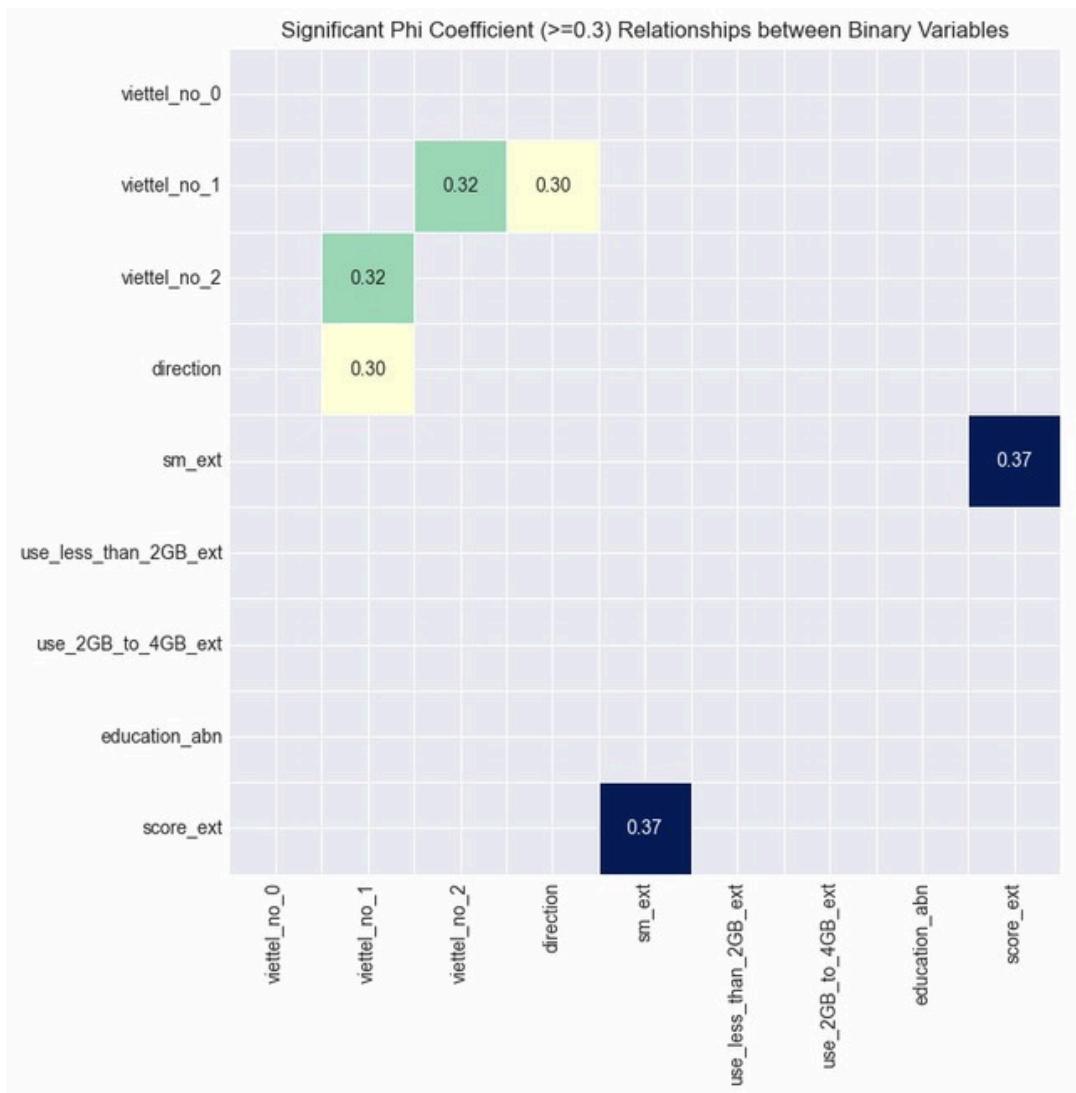
Đa số du khách đến tỉnh khác để du lịch. Trong khi đó, lượng du khách đến Hà Nội để du lịch và làm việc khá tương đồng nhau.



- Du khách đến tỉnh khác:** mục đích du lịch là chính, có thể **chú trọng gói cước liên quan đến mạng xã hội**. Du
- khách đến Hà Nội:** đề xuất những **gói cơ bản, nhanh và tiện**

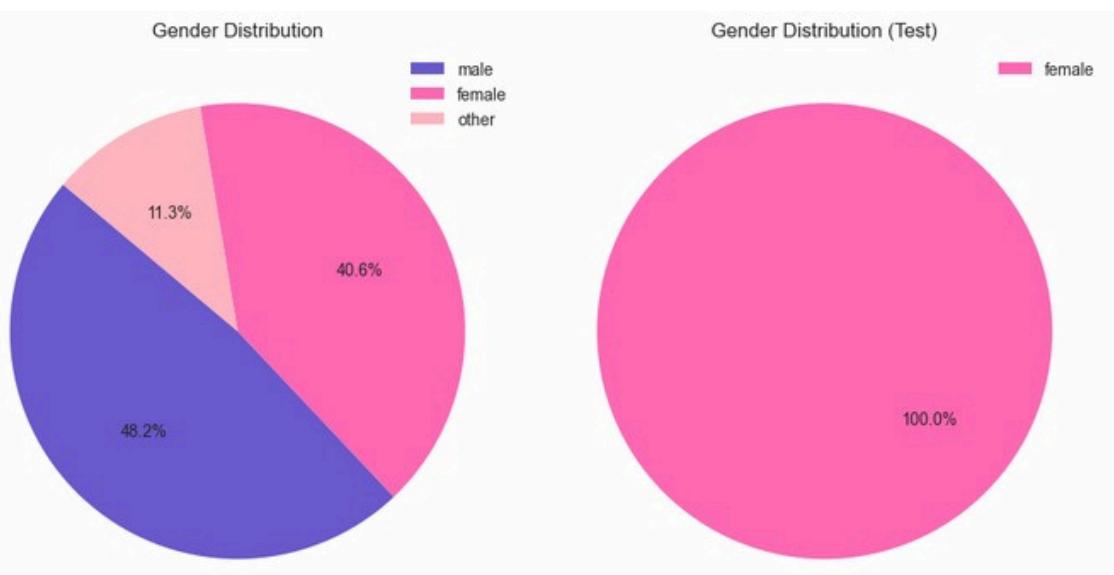
- **Đa số các cặp biến phân loại có mối liên kết mạnh với nhau** - Trừ các cặp biến **extreme value** sẽ độc lập với nhau, các cặp biến còn lại có mối liên kết tương đối tốt - Nhìn chung, các cặp biến liên tục có mối liên kết tương đối.

2.3. Mối quan hệ giữa các biến



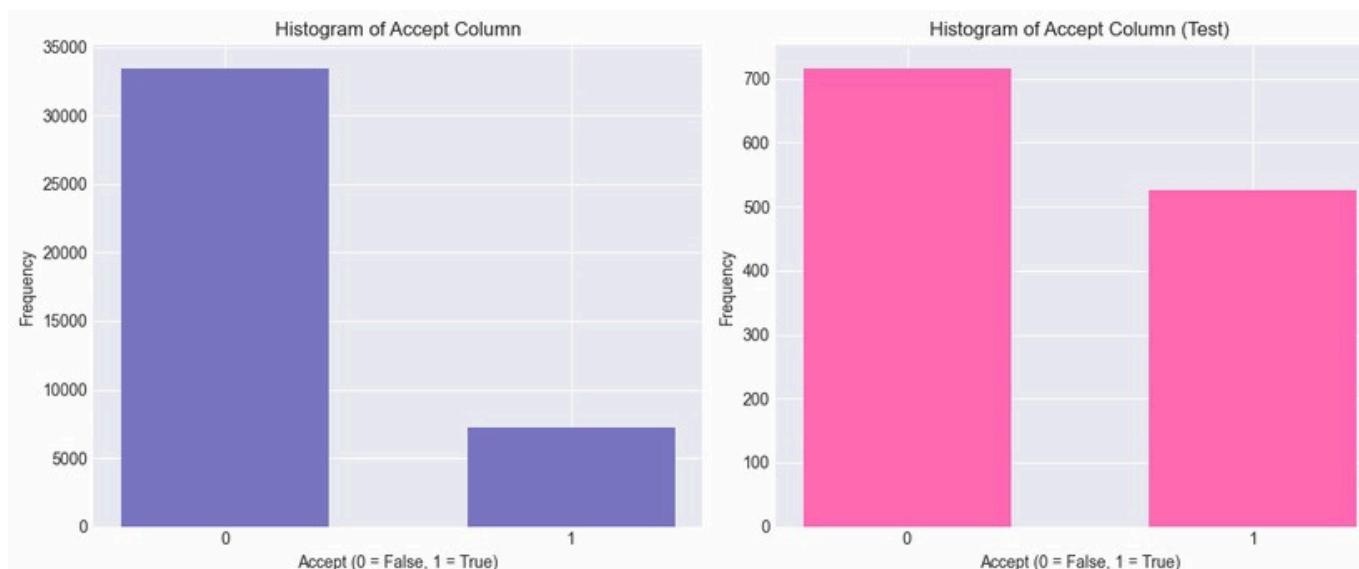
2.4. Test dataset vs. Train dataset

Gender



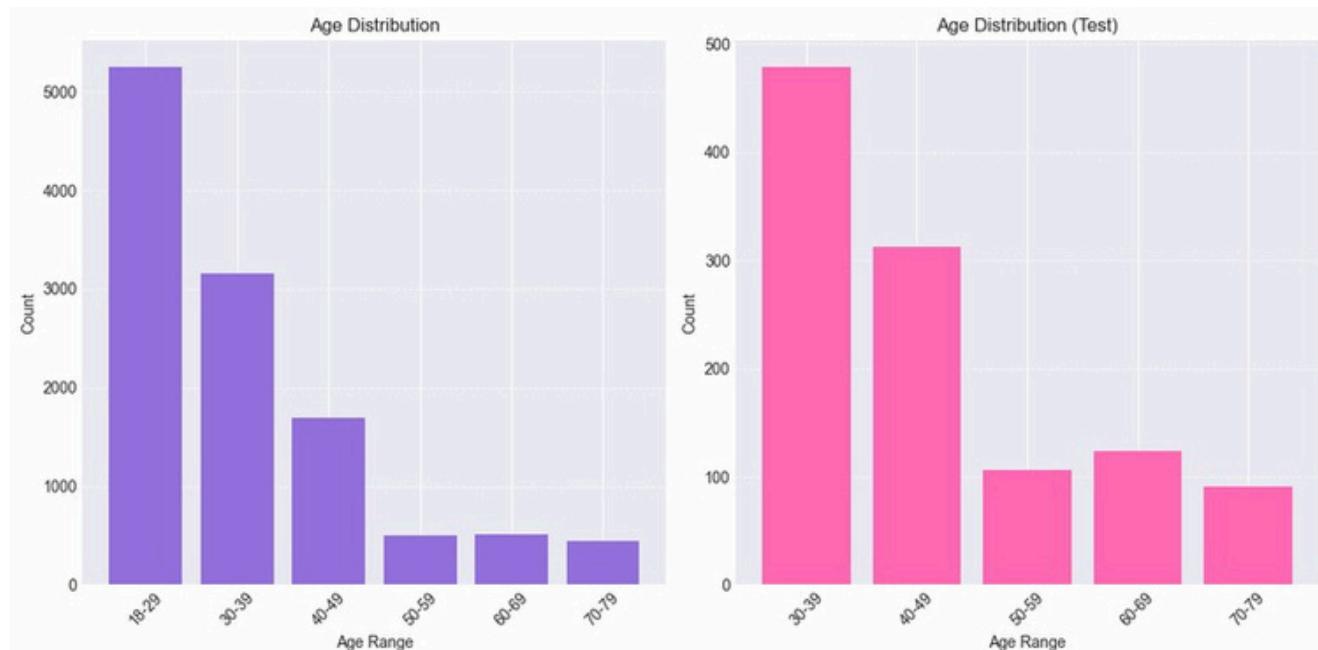
Ở bộ dữ liệu Test thì giới tính được ghi nhận **100% là nữ**

Accept



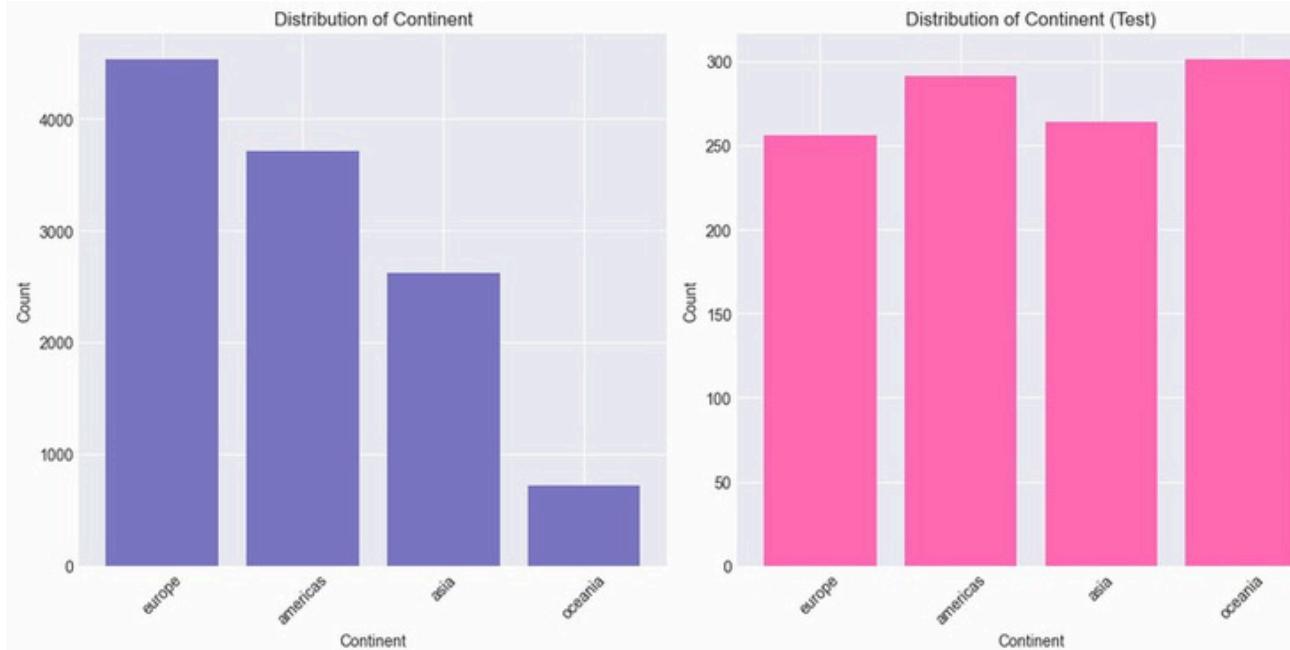
Sự phân bổ 2 giá trị 0,1 trong **tập Train** **mất cân bằng hơn** **tập Test**

Age



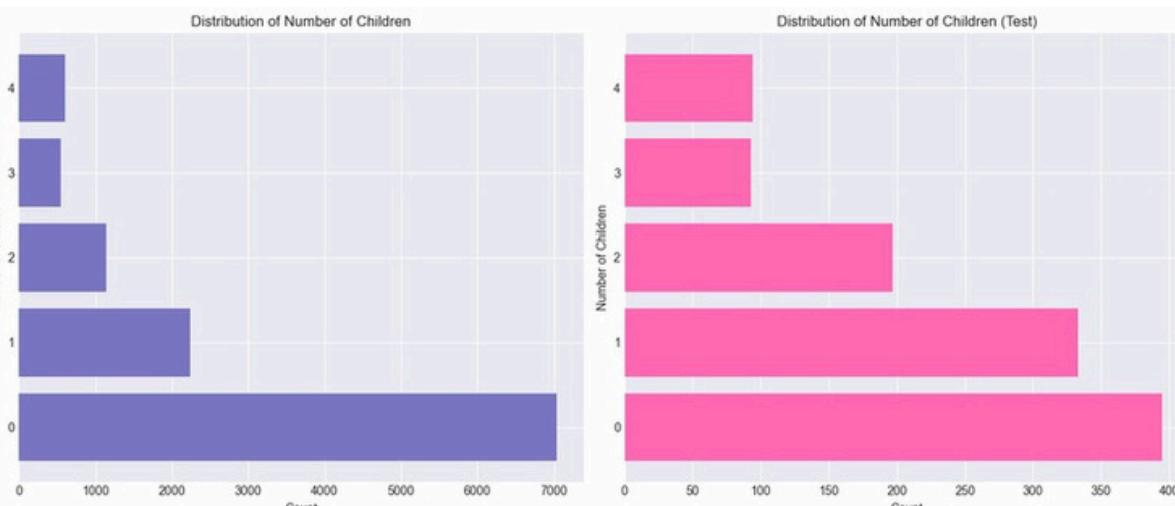
Tập Test không chứa nhóm tuổi 18-29. Bên cạnh đó, nhóm tuổi 60-69 cũng có số lượng nhỉnh hơn phân phối trong train set.

Continent



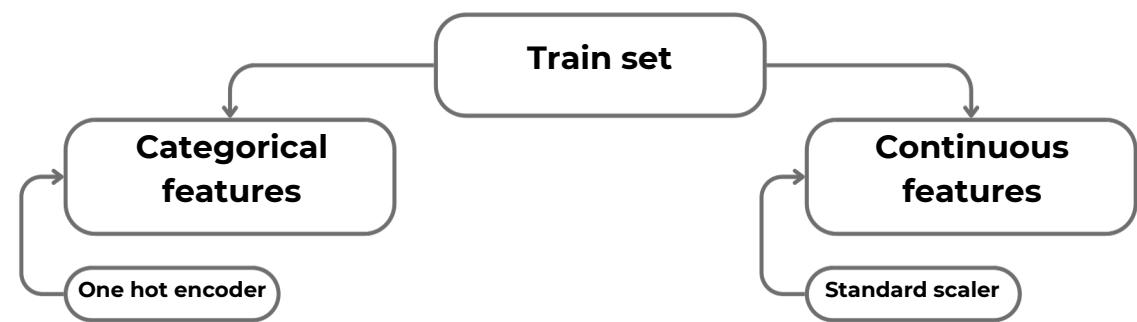
Tuy vẫn có đủ 4 châu lục, **tỷ trọng giữa các châu lục** **tập Test đồng đều hơn**

Children



% số con tăng đáng kể, phần lớn mỗi du khách không có con hoặc có từ 1-2 con, **% du khách không có con** **trong tập train nhiều hơn**

3.1. Feature selection



• Categorical features:

- Phương thức tính điểm: kiểm định Chi-squared.
- **Kết quả:** 17/34 nhãn có p-value <= 0.05 (9/15 đặc trưng).

• Continuous features:

- Phương thức tính điểm: kiểm định ANOVA.
- **Kết quả:** 9/13 đặc trưng có p-value <= 0.05.

Sử dụng toàn bộ các đặc trưng vào các mô hình khả thi. Sau khi chọn được mô hình phù hợp và tuning, thực hiện **kiểm tra feature importance** để **kết luận các đặc trưng có thể bỏ ra khỏi mô hình**.

3.2. Feature engineering

• Tạo nhãn bão bất thường (0;1) cho các cột:

- Nhóm social media '**fb_freq**', '**yt_freq**', '**insta_freq
- Nhóm usage '**use_less_than_2GB**', '**use_2GB_to_4GB
- Cột **score**: áp dụng kỹ thuật tương tự nhóm cột usage.
- Cột **education**: (1) các giá trị 'unknown'.****

• **Xóa cột: 'gender'** do có sự phân phối không đồng đều giữa các giá trị của cột này trong hai tập train set và test set.

• **(Extension) Oversampling** cho train set bằng hàm SMOTE trong thư viện imblearn.

4.1. Cách tiếp cận

Do không có đủ thông tin về bảng mobile_plan_user, ta không biết chắc rằng các gói mobile_plan ứng với từng ID đã được đề xuất một cách đúng đắn hay đề xuất một cách ngẫu nhiên.

Approach 1: Ghép các bảng user left merge với hai bảng context và mobile_plan_user.

Từ đó giữ nguyên tổng số lần các gói mobile_plan được đề xuất cho từng khách hàng cũng như là tình trạng chấp nhận.

Approach 2: Ứng với từng ID trong bảng user, tạo cột 'mobile_plan' gồm đủ 5 loại gói dịch vụ, sau đó left merge với bảng context và mobile_plan_user và fill các giá trị N/A trong cột 'accept' là 0.

X		y	
id	...	mobile_plan	accept
1	...	A	1
1	...	A	0
2	...	A	1
2	...	C	1
3	...	B	0
3	...	B	1
3	...	C	0

Approach 1

X		y	
id	...	mobile_plan	accept
1	...	A	1
1	...	B	0
1	...	C	0
1	...	D	0
1	...	E	0
2	...	A	1
2	...	B	0
2	...	C	1
2	...	D	0
2	...	E	0

Approach 2

4.2. Độ đo

• **Accuracy (Độ chính xác):** Tỉ lệ phần trăm của các dự đoán đúng trên tổng số mẫu kiểm tra.

Precision (Độ chính xác dự đoán dương): Tỉ lệ các dự đoán dương đúng trên tổng số dự đoán dương. Độ chính xác cao cho

• thấy mô hình ít bao động giả.

Recall (Độ nhạy): Tỉ lệ các dự đoán dương đúng trên tổng số mẫu dương thực sự. Độ nhạy cao cho thấy mô hình ít bỏ sót các

• trường hợp dương.

F1-score: Trung bình điều hòa của Precision và Recall, cung cấp một thước đo cân bằng giữa hai chỉ số này. F1-score cao cho thấy mô hình cân bằng tốt giữa độ chính xác và độ nhạy.

4.3. Xây dựng mô hình

Thuật toán và các kỹ thuật xây dựng mô hình

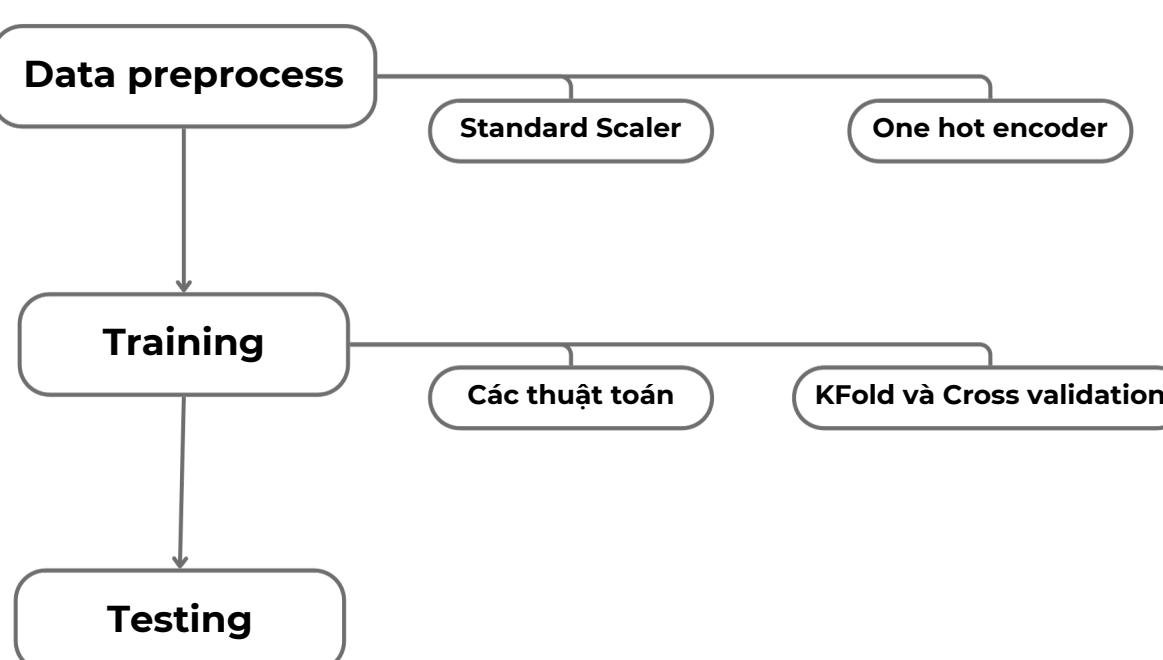
Thuật toán sử dụng:

- KNeighbors Classifier
- Logistic Regression
- Random Forest Classifier
- Extratrees Classifier
- MLP
- Gradient Boosting Classifier
- Decision Tree
- XGBoost Classifier
- Naive Bayes Classifier
- Catboost Classifier

Các kỹ thuật khác:

- Standard Scaler
- One hot encoder
- KFold và Cross validation

Các bước thực hiện:



4.4. Kết quả

Training and Validation

Khi thực hiện train và cross validate, các mô hình tỏ ra khá hiệu quả với bản thân bộ valid set, tuy nhiên khi test các mô hình với test set thì performance có giảm, lí do là do sự bất cân xứng giữa 2 bộ dữ liệu.

Approach	Model	Accuracy	Precision	Recall	F1 Score
AP1	Logistics Regression	0.818305	0.669623	0.818305	0.736535
AP2	Logistics Regression	0.892931	0.797326	0.892931	0.842425

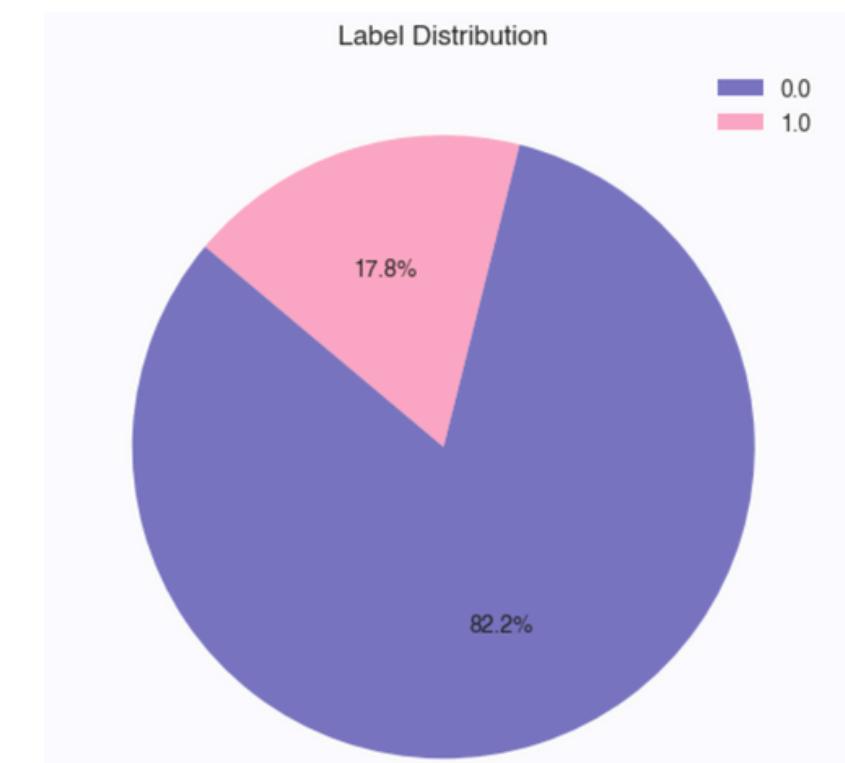
Testing

Tuy nhiên khi test các mô hình với test set thì performance có giảm, lí do là do sự bất cân xứng giữa 2 bộ dữ liệu và sự bất cân xứng nhãn của đặc trưng “accept”.

Approach	Model	Accuracy	Precision	Recall	F1 Score
AP1	Extratrees Classifier	0.577295	0.756118	0.577295	0.423443
AP2	Extratrees Classifier	0.916727	0.889064	0.916727	0.881829

- Tuy approach 1 tỏ ra khá hiệu quả với valid set, tuy nhiên trên tập test set kết quả khá tệ và ở đây, ta xem mobile_plan là một đặc trưng của mô hình. Điều này không hợp lý trong thực tế, khi mà ta chưa có thông tin về gói phù hợp với khách hàng.
- Approach 2 cũng xem mobile_plan là một đặc trưng của mô hình nhưng bằng việc lặp lại hàng theo từng gói sẽ giúp ta có thể ứng dụng mô hình trong thực tế. Vậy nên nhóm quyết định chọn approach 2 làm mô hình để phục vụ bài toán classification.
- Tổng quan mô hình của approach 2 đạt được kết quả tương đối ổn định, với độ chính xác và F1-score khá cao.

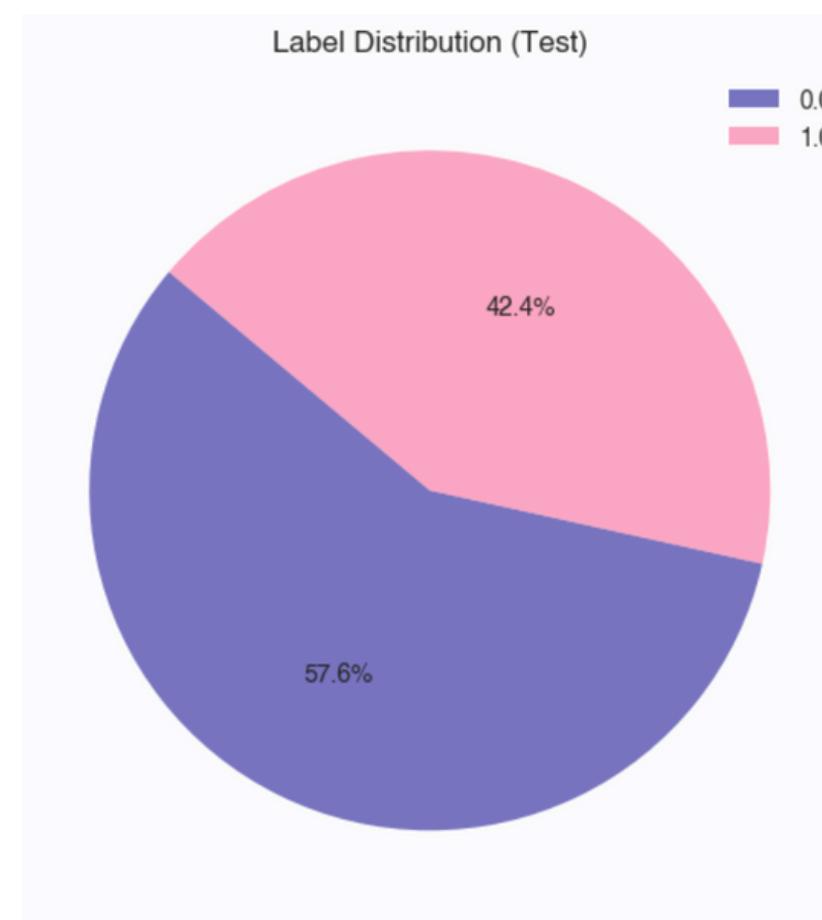
4.5. Resampling



Sự mất cân đối trong bộ dữ liệu train và bộ dữ liệu test khả năng cao ảnh hưởng tới hiệu quả của mô hình, vì vậy ta có thể sử dụng kỹ thuật resampling cho bộ dữ liệu train để cân bằng lại nhãn “accept”, từ đó giảm thiểu thiên vị trong quá trình training.

Approach	Model	Accuracy	Precision	Recall	F1 Score
AP1	Random Forest Classifier	0.883325	0.889936	0.883325	0.882906
AP2	Random Forest Classifier	0.941721	0.947539	0.941721	0.941489

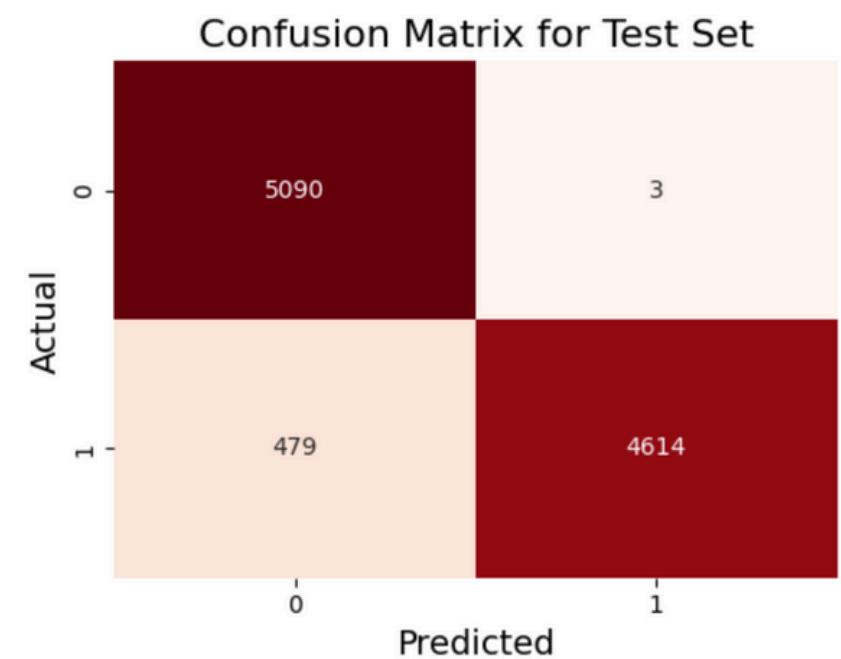
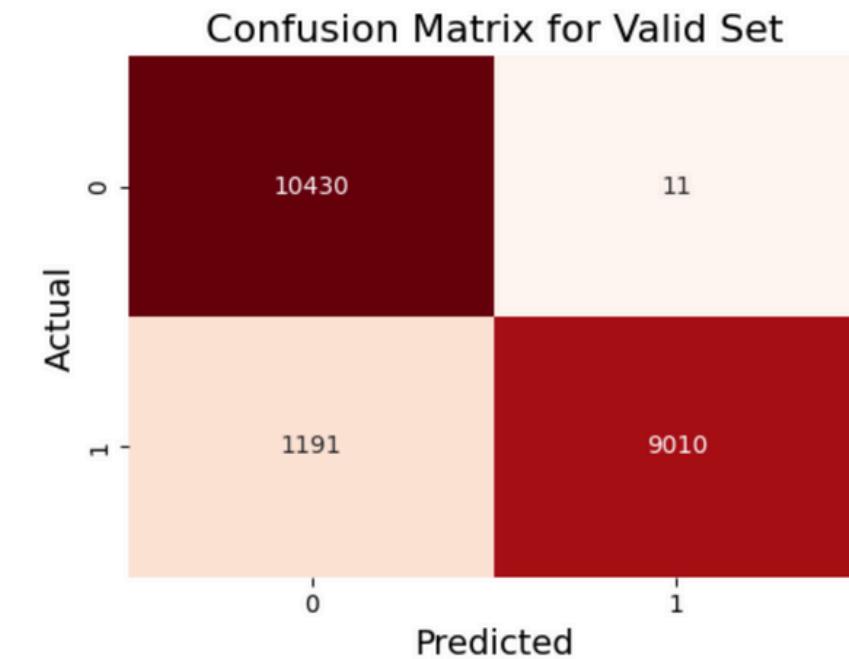
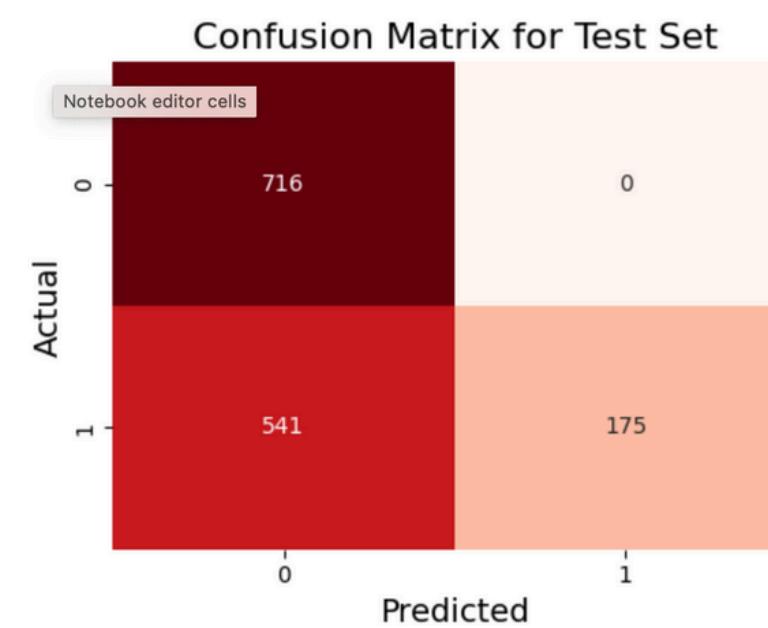
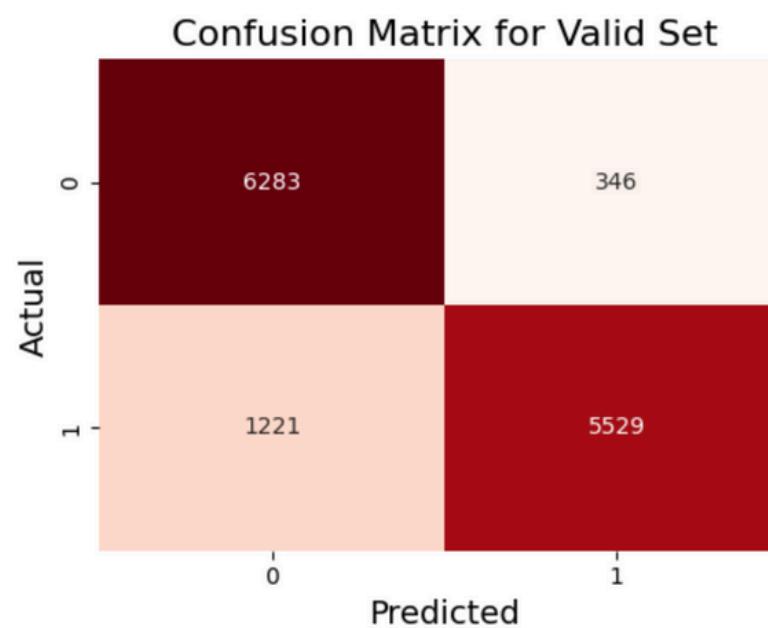
Kết quả tập valid sau khi resampling



Approach	Model	Accuracy	Precision	Recall	F1 Score
AP1	Random Forest Classifier	0.625000	0.625797	0.625000	0.624405
AP2	Random Forest Classifier	0.952680	0.956602	0.956602	0.952578

Kết quả tập test sau khi resampling

4.6. Confusion Matrix



Confusion Matrix của Approach 1

Confusion Matrix của Approach 2

4.7. Hyperparameter tuning

Sử dụng phương pháp **grid search CV** với các tham số như sau:

n_estimators: [100, 200, 300]

max_depth: [None, 10, 20, 30]

min_samples_split: [2, 5, 10]

min_samples_leaf: [1, 2, 4]

Bộ tham số tối ưu cho mô hình của approach 2 là:

n_estimators: 100

max_depth: 30

min_samples_split: 2

min_samples_leaf: 2

Metrics	Based	Tunned
Accuracy	0.952680	0.953290
Precision	0.956602	0.958742
Recall	0.956602	0.952900
F1 Score	0.952578	0.955812

5.1. Về Viettel

Tập đoàn Công nghiệp – Viễn thông Quân đội, hay Viettel, có lịch sử hình thành 35 năm với nhiều cột mốc giá trị:

- Là doanh nghiệp viễn thông, công nghiệp, công nghệ số 1 Việt Nam, với giá trị thương hiệu số 1 tại Đông Nam Á và thứ 9 tại Châu Á
- Top 15 doanh nghiệp viễn thông phát triển nhanh nhất thế giới
- Vươn tầm thương hiệu đến với 10 thị trường nước ngoài ở 3 châu lục gồm Châu Á, Châu Mỹ và Châu Phi

TÂM NHÌN

Sáng tạo vì con người

SỨ MỆNH

Tiên phong, chủ lực kiến tạo
xã hội số

5.2. Cơ sở phát triển

Từ các insights có được trong phần EDA và mô hình, ta có những đặc điểm chính như sau:

Tôi là người trưởng thành, có thu nhập thấp nên nhạy cảm về giá. Với tâm thế YOLO, du lịch trải nghiệm một mình, tôi thích dùng mạng XH để cập nhật thông tin và liên lạc.

Tuy có nhu cầu tiêu dùng gói cước 4G khi du lịch, tôi chưa biết hoặc chưa trung thành với thương hiệu Viettel.

Painpoints

Giá cả

Trải nghiệm dùng dịch vụ có mạng xã hội (YT, FB, đặc biệt là INS)

Strategic direction

Cạnh tranh về giá: Giá rẻ

Chuyển đổi khách hàng trung thành:
Ưu đãi ngập tràn

Đáp ứng nhu cầu: đa dạng dịch vụ, trong đó có gói không giới hạn dung lượng truy cập mạng xã hội



5.3. Đề xuất chiến lược

PRICE

Giảm giá đối với lần đầu sử dụng: cho gói cước có giá cao hơn (datagold, socialmediagold, datacall)

Ưu đãi cho lần đăng ký tiếp theo:

Nếu trước đó khách dùng gói datasilver, có thể thúc đẩy chuyển đổi lên gói cao hơn (datacall, socialmediagold, datagold) với ưu đãi ở lần tiếp theo

Đánh vào **tâm lý price-sensitive** của khách hàng mới và sự cạnh tranh giá trong dịch vụ data

Thúc đẩy tiêu dùng đối với các gói ít phổ biến hơn & **giữ chân khách hàng.**

PLACE (ĐỊA ĐIỂM)

Dựng booth đăng ký 4G tự động ở sân bay

Người dùng tự tương tác đúng với nhu cầu, dễ thu thập thông tin, nhanh và tiện

Tăng độ phủ quảng cáo trong hành trình của du khách mới

Sân bay, taxi truyền thống & công nghệ, trạm xe, trung tâm thương mại...

Tăng độ phủ trên truyền thông

Qua các trang mạng XH mà du khách hay sử dụng (YT, FB, INS).
Thậm chí có thể tiếp cận qua các KOL về mảng du lịch.
Các địa điểm du lịch hay được check-in, các quán ăn trong hoặc gần sân bay

5.3. Đề xuất chiến lược

PROMOTION & PRODUCT

Tâm lý du khách thích dùng mạng XH & đi du lịch một mình sẽ cần truy cập Internet. Trong hành trình đó, wifi free tuy tồn tại nhưng còn gặp nhiều bất cập

- tắc nghẽn
- trải nghiệm không mượt mà
- không phủ sóng dày đặc

Du khách trẻ sẽ hay **xem review, vlog...** về du lịch Việt Nam
 → **tăng độ phủ sóng qua các kênh KOL, travel blogger và vlogger** để tăng sức nhận biết đối với các khách hàng mới và độ trung thành với các khách hàng hiện có.

Gói cước 4G SOCIALMEDIA là Key Product của Viettel, đáp ứng được hầu hết các painpoint của du khách.

6.1. Thu thập thêm dữ liệu

Thu thập đầy đủ thông tin để xuất các gói Data 4G

Để có thể đưa ra được insight chính xác hơn thì việc thu thập đầy đủ thông tin về chấp nhận đối với cả **5 gói cho mỗi khách hàng** là cần thiết

Thu thập thông tin về mục đích sử dụng 4G

Ngoài thông tin về thời gian sử dụng 4G chỉ trên 3 nền tảng (Facebook, Instagram, Youtube) thì có thể thu thập thêm thông tin **gọi điện, nghe nhạc, đọc tin tức,...**

Thu thập thông tin về địa điểm du lịch

Bên cạnh thông tin về đi đến Hà Nội hay các tỉnh khác, việc thu thập thêm **địa điểm du lịch cụ thể** của khách hàng có thể giúp cho việc quảng cáo/tiếp thị thực hiện tốt hơn

6.2. Xử lý mất cân bằng dữ liệu (imbalanced dataset)

Undersampling

Kỹ thuật giảm số lượng mẫu của lớp chiếm đa số để cân bằng với lớp thiểu số.

SMOTE

Là kỹ thuật tạo ra các mẫu giả (synthetic samples) cho lớp thiểu số bằng cách sử dụng phương pháp nội suy giữa các mẫu hiện có.

Thu thập thêm dữ liệu

Thu thập đầy đủ dữ liệu cho các yếu tố bị thiếu, mất cân xứng như về đề xuất đầy đủ các gói cho từng khách hàng và sự chấp nhận của họ đối với gói

APPENDIX

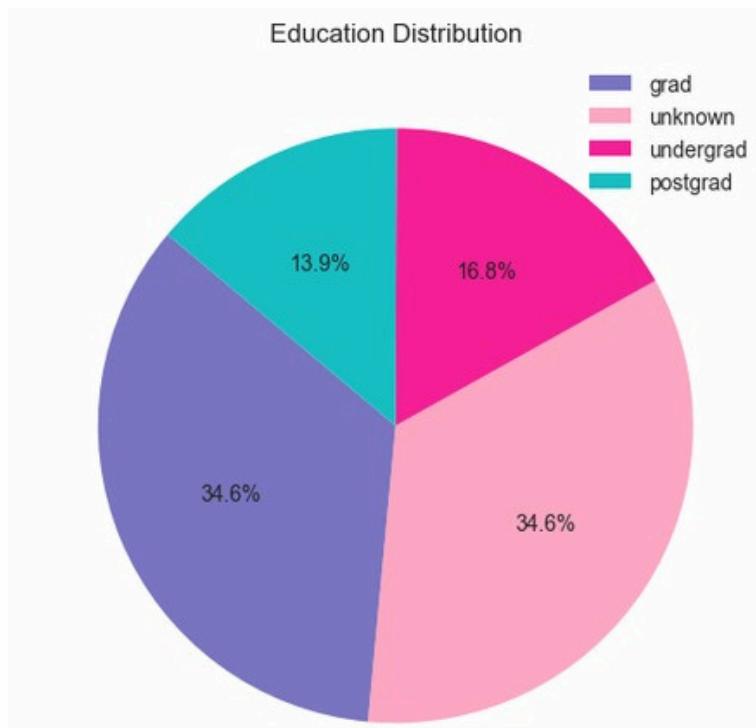
APPENDIX 1. Danh mục tài liệu tham khảo

- 1.Thu Phương. (2023). “Làm gì để hút khách du lịch quốc tế đến Việt Nam?”. Báo Điện Tử Đảng Cộng Sản Việt Nam.
2. T. Đ. (2019). “Những trải nghiệm mà du khách Tây thích thú khi đến thăm Việt Nam”. Chuyên trang của Báo Lao động du lịch.
3. Minh Huyền. (2023). “Khách quốc tế: 'Du lịch Việt Nam quá rẻ, nhiều món ăn giá trên dưới 1 đô la'”. Tuổi Trẻ Online.
4. CEOWORLD. (2022). “These are the countries with the highest average salary, 2022”.
5. Huệ Anh (2022). “Facebook dần trở nên vô giá trị với những người dùng trẻ”. GenK, Trang thông tin điện tử tổng hợp.
6. Hải Nam. (2024). “Du khách châu Âu ưa chuộng du lịch Việt Nam trong dịp hè 2024”. VOV.
7. Linh Phương (2023). “Khách châu Âu đến Việt Nam tăng mạnh”. Pháp luật Thành phố Hồ Chí Minh.
- 8.The U.S. Census Bureau. (2021).

APPENDIX

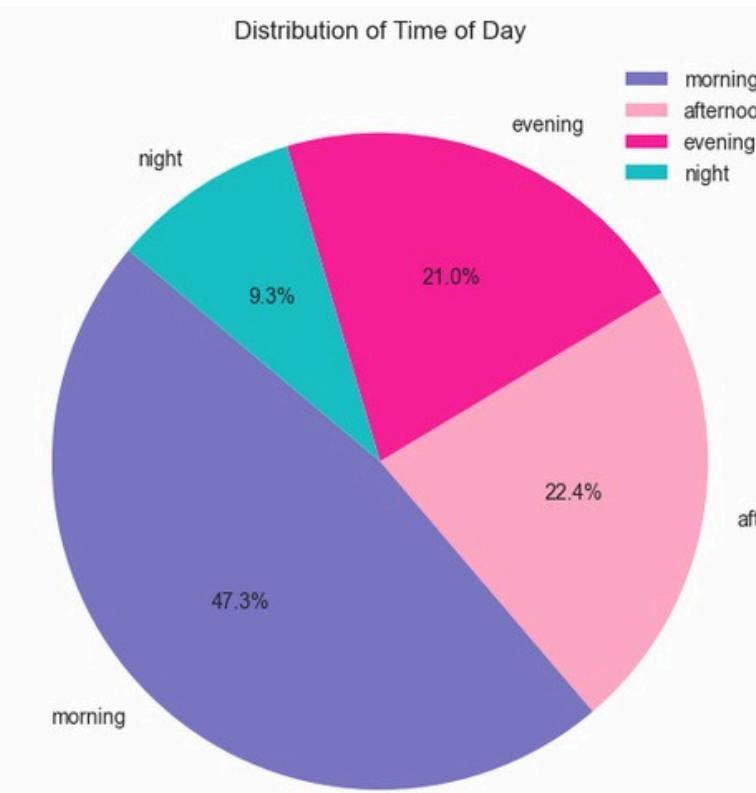
APPENDIX 2. Các phân tích EDA đơn biến khác

Trình độ học vấn (Education)



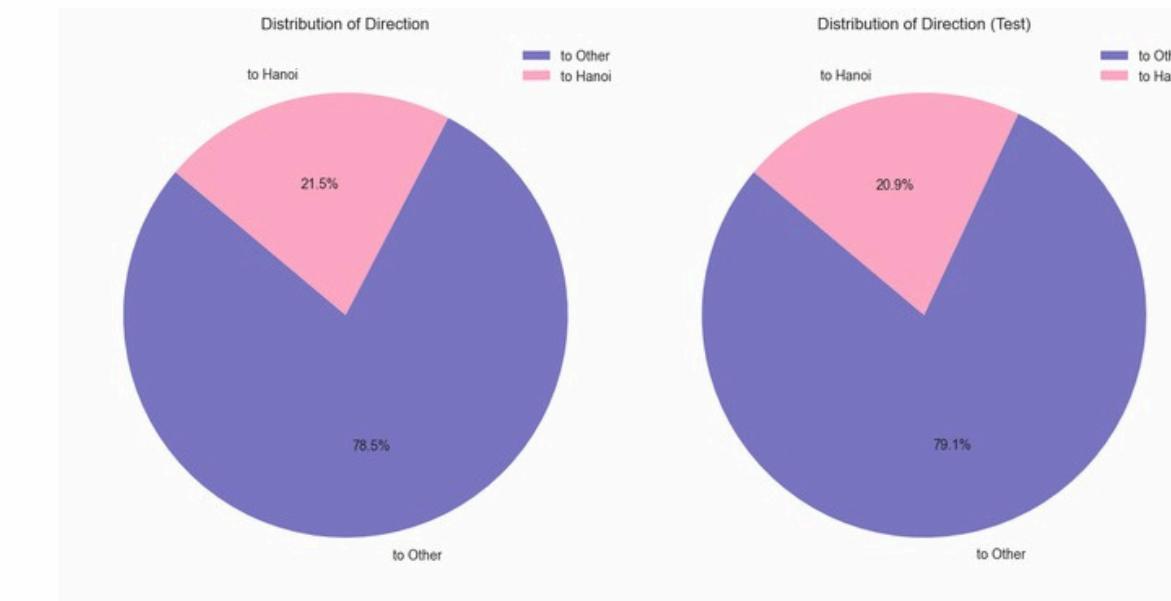
Tỷ lệ trình độ học vấn **graduate (đã tốt nghiệp)** và **unknown (không rõ)** là như nhau, cụ thể **34.6%**. 2 nhóm còn lại là **undergraduate (chưa tốt nghiệp)** và **postgraduate (học cao học)** chiếm tỷ lệ dưới 20% mỗi nhóm.

Thời gian đáp (Time of day)



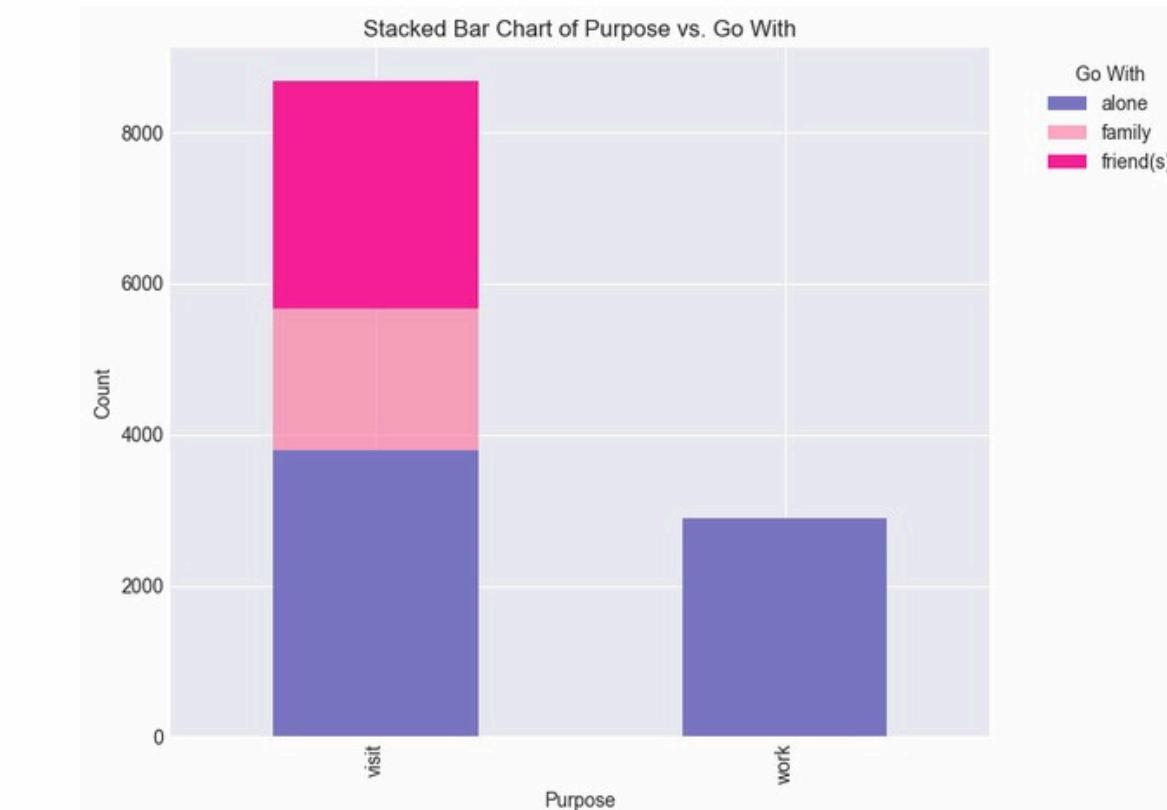
Gần **1/2** du khách hạ cánh vào **buổi sáng**. Chỉ **khoảng 10%** lượng du khách được thu thập thông tin vào **buổi tối**. Tuy vậy, điều này **có thể do giữa các buổi khác nhau về số lượng chuyến bay** hay phân bổ nguồn lực để thu thập thông tin.

Đích đến (Direction)



78.5% du khách **đến các tỉnh thành khác** không phải Hà Nội.

Mục đích chuyến đi và đi với ai (Purpose x go_with)

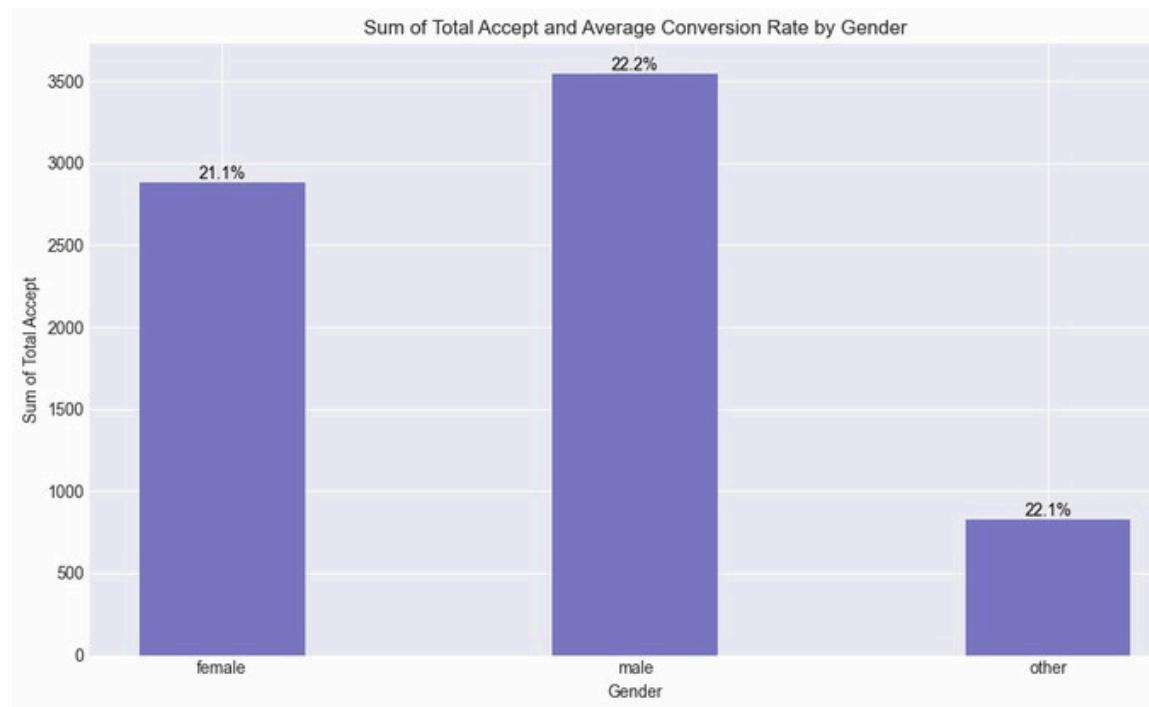


100% du khách đến để **làm việc** thì **đi 1 mình**.

APPENDIX

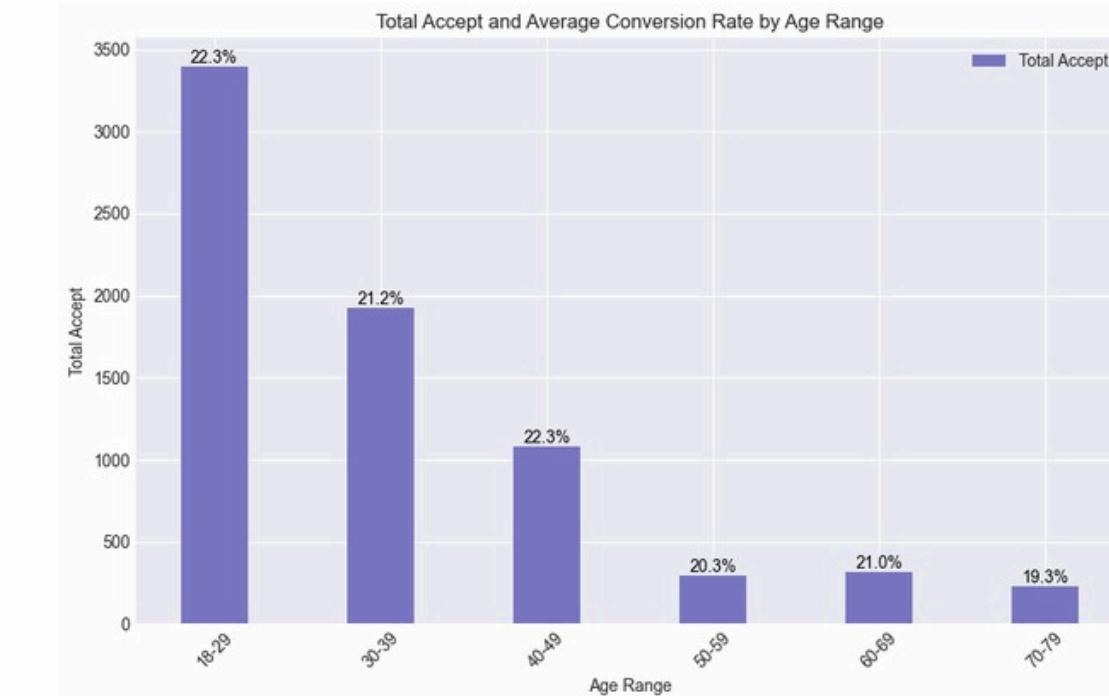
APPENDIX 3. Các phân tích EDA đa biến khác

Gender x Accept x Mean CVR



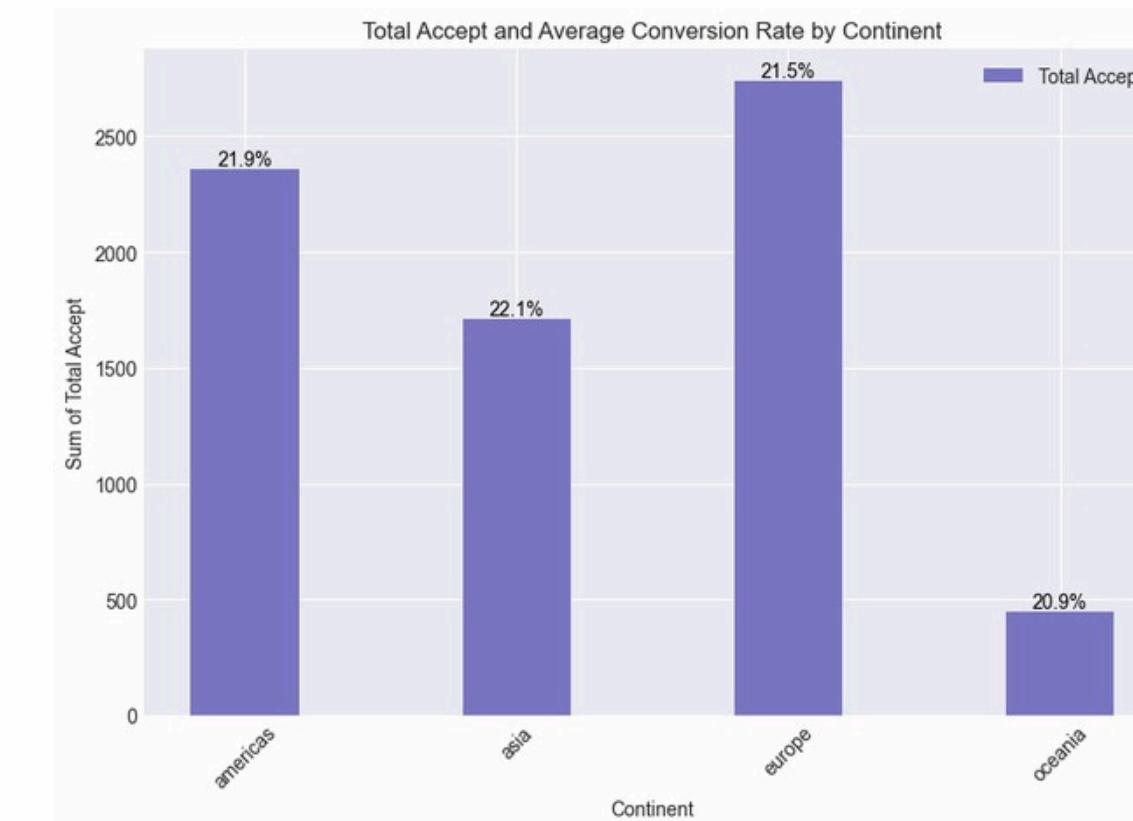
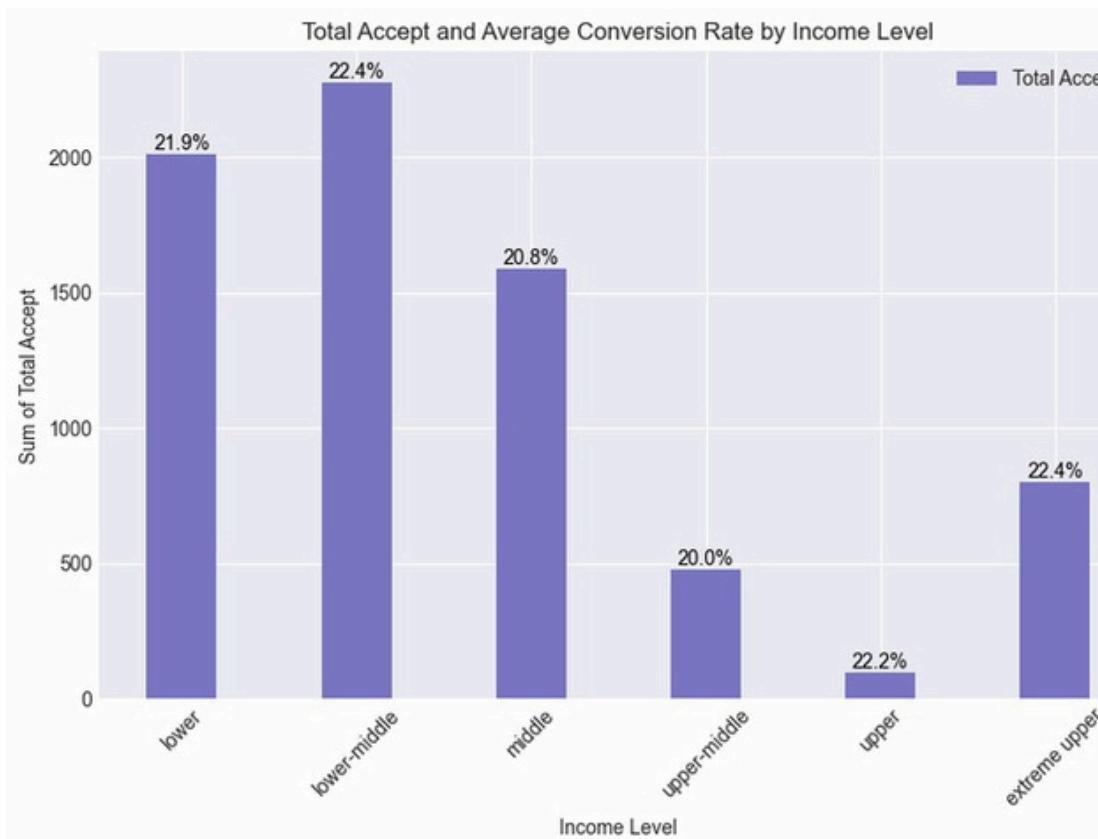
Nam và nữ giới có mức độ chấp nhận gói cước khá cao so với tệp khách hàng còn lại. Tuy nhiên, dù có số lượng accept thấp, tỉ lệ CVR của tệp này lại khá đồng đều so với 2 tệp nam, nữ. → Vẫn nên tập trung thúc đẩy đều cho 3 tệp

Age x Accept x Mean CVR



3 nhóm tuổi trẻ nhất đều chiếm số lượng accept và CVR cao nhất.

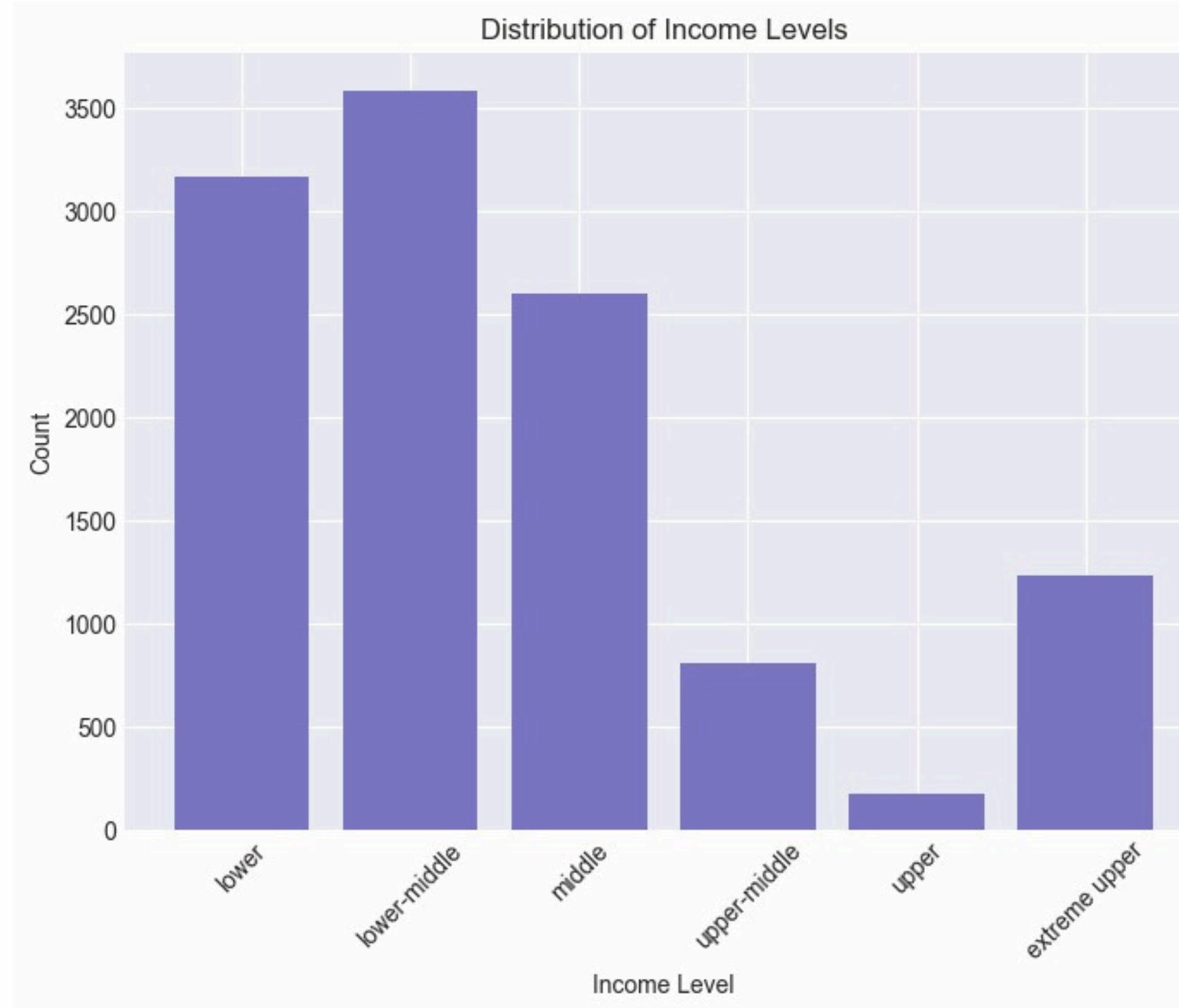
Income level x Accept x Mean CVR và Continent x accept x mean CVR



Phân phối accept giữa các nhóm thu nhập và châu lục khá tương đồng với phân phối số lượng du khách. Trong đó, tỉ lệ CVR cũng đồng đều giữa các nhóm.

APPENDIX

APPENDIX 4. Cơ sở chia nhóm thu nhập



Dựa theo thống kê của U.S. census data vào năm 2021, nhóm phân chia các nhóm thu nhập (Income Class) như sau:

- Lower: thu nhập thấp, nhỏ hơn hoặc bằng \$30000
- Lower-middle: thu nhập cận TB, từ \$30000-\$58000
- Middle: thu nhập TB, từ \$58000-\$94000
- Upper-middle: thu nhập cận cao, từ \$94000-\$153000
- Upper: thu nhập cao, \$94000-\$153000
- Extreme Upper: thu nhập siêu cao, từ \$153000-\$200000

Trong đó, nhóm thu nhập siêu cao dành cho những du khách có thu nhập được xem là extreme value trong trường thu nhập.

APPENDIX

APPENDIX 5. Các độ đo đối với mô hình

1. Accuracy (Độ chính xác)

$$\text{Accuracy} = \frac{\text{số dự đoán đúng}}{\text{tổng số mẫu}}$$

Đây là tỉ lệ phần trăm của các dự đoán đúng trên tổng số mẫu kiểm tra. Độ chính xác cao cho thấy mô hình dự đoán đúng nhiều.

2. Hamming Loss

$$\text{Hamming Loss} = \frac{1}{N \times L} \sum_{i=1}^N \sum_{j=1}^L [y_{ij} \neq \hat{y}_{ij}]$$

(với N là số mẫu và L là số nhãn)

Đây là tỉ lệ lỗi Hamming, đo lường tỉ lệ các nhãn bị phân loại sai. Chỉ số này càng thấp càng tốt.

3. Precision (Độ chính xác dự đoán dương)

$$\text{Precision} = \frac{\text{số true positive}}{\text{số true positive} + \text{false positive}}$$

Đây là tỉ lệ các dự đoán dương đúng trên tổng số dự đoán dương. Độ chính xác cao cho thấy mô hình ít báo động giả.

4. Recall (Độ nhạy)

$$\text{Recall} = \frac{\text{số true positive}}{\text{số true positive} + \text{false negative}}$$

Đây là tỉ lệ các dự đoán dương đúng trên tổng số mẫu dương thực sự. Độ nhạy cao cho thấy mô hình ít bỏ sót các trường hợp dương.

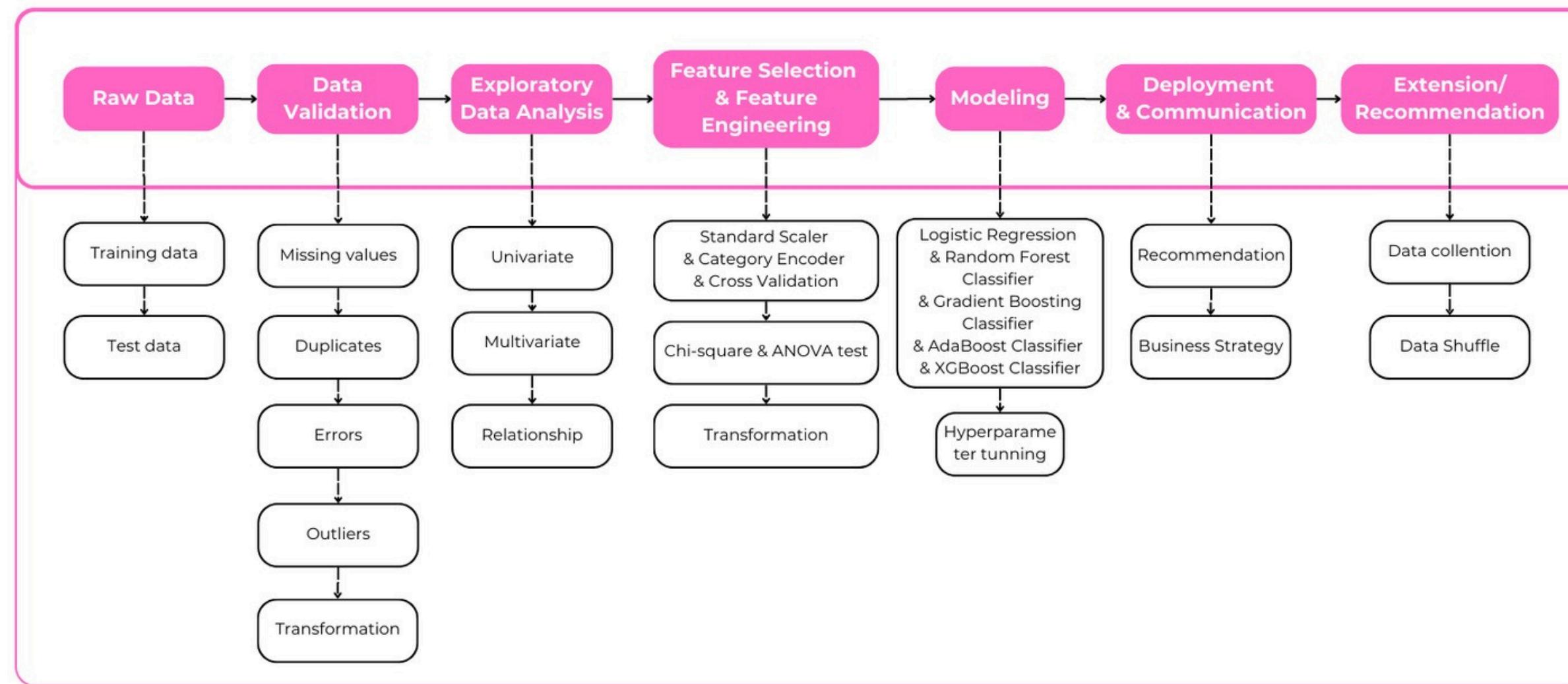
5. F1-score

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

Đây là trung bình điều hòa của Precision và Recall, cung cấp một thước đo cân bằng giữa hai chỉ số này. F1-score cao cho thấy mô hình cân bằng tốt giữa độ chính xác và độ nhạy.

APPENDIX

APPENDIX 6. Quy trình tiếp cận và giải quyết bài toán



- **Raw Data:** lấy bộ dữ liệu train và test được cung cấp
 - **Data Validation:** làm sạch dữ liệu và biến đổi dữ liệu để phục vụ phân tích tốt hơn
 - **EDA:** phân tích đặc điểm các biến, mối quan hệ giữa các biến, insight từ dữ liệu
 - **Feature Selection & Feature Engineering:** lựa chọn và biến đổi các biến trong bộ dữ liệu để phục vụ xây dựng mô hình tốt hơn
 - **Modeling:** xây dựng và tối ưu mô hình phân loại các gói Data 4G để phục vụ việc dự đoán xu hướng chọn gói 4G của khách hàng
 - **Deployment & Communication:** từ những dữ kiện phân tích được đưa ra các chiến lược kinh doanh, đề xuất cho doanh nghiệp để hoạt động tối ưu hơn
 - **Extension/Recommendation:** đề xuất các giải pháp để nâng cao hiệu quả việc khai thác bộ dữ liệu