DATA ANALYTICS REPORT

I. XỬ LÝ DỮ LIỆU

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 về các gói Data 4G.
- 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. Làm sạch và chuyển hóa dữ liệu

1.2.1. Missing values

- Bảng user: có 4003 giá trị null ở cột 'education' (chiếm 2.16%) → chuyển thành unknown
- Bảng mobile_plan_user: có 2715 giá trị null ở cột 'mobile_plan' và 2721 giá trị ở cột 'accept' (chiếm 4.0%) → drop những giá trị này

1.2.2. Error values

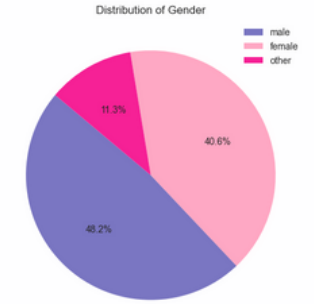
- Bảng context:
  - Nhiều ký tự đặc biệt ở các cột 'go_with' và 'weather' → lọc bỏ những ký tự này bằng công cụ regex
  - Format thời gian ở cột 'time' khác nhau → quy về một format bằng hàm to_datetime
- Bảng user:
  - Nhiều ký tự đặc biệt ở các cột 'living_with' → lọc bỏ những ký tự này bằng công cụ regex
  - Có nhiều giá trị khác nhau và đa ngôn ngữ (hơn 60%) ở cột 'job' → tốn nhiều thời gian để phiên dịch và phân loại lại nên sẽ loại bỏ, chỉ dùng cột 'profession' cho nghề nghiệp
  - Cột income có hai loại tiền tệ USD và VND → quy về chung 1 loại tiền USD (theo tỷ giá 1USD = 25,000VND)

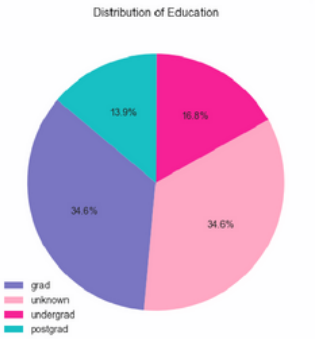
1.2.3. Outliers

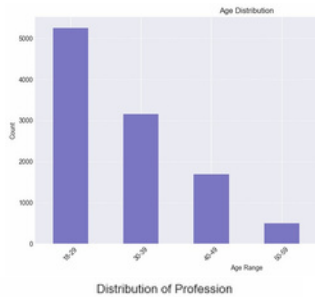
- Bảng context:
  - Nhiều ký tự đặc biệt ở các cột 'go_with' và 'weather' → lọc bỏ những ký tự này bằng công cụ regex
  - Format thời gian ở cột 'time' khác nhau → quy về một format bằng hàm to_datetime
- Bảng user:
  - Nhiều ký tự đặc biệt ở các cột 'living_with' → lọc bỏ những ký tự này bằng công cụ regex
  - Có nhiều giá trị khác nhau và đa ngôn ngữ (hơn 60%) ở cột 'job' → tốn nhiều thời gian để phiên dịch và phân loại lại nên sẽ loại bỏ, chỉ dùng cột 'profession' cho nghề nghiệp
  - Cột income có hai loại tiền tệ USD và VND → quy về chung 1 loại tiền USD (theo tỷ giá 1USD = 23,000VND)
- Nhiều cột chứa nhiều outlier như: 'income', 'fb_freq', 'yt_freq', 'insta_freq', 'score' → được kiểm tra bằng phương pháp IQR, sau đó sẽ tạo thêm cột phụ để xét bất thường (0,1)

1.2.4. Transformation

- Các cột gồm nhiều giá trị khác nhau như 'education', 'profession', 'income', 'living_with', 'nation',... sẽ được quy thành category.
- Bỏ đi cột to_hanoi và to_other vì đã thể hiện trong direction







II. PHÂN TÍCH VÀ TRỰC QUAN HÓA DỮ LIỆU

2.1. Phân tích đơn biến (Univariate analysis)

Giới tính (Gender) Nam chiếm tỷ lệ cao nhất (48.2%), tiếp theo là nữ (40.6%). Giới tính khác chiếm tỉ lệ khá nhỏ.

Trình độ học vấn (Education)

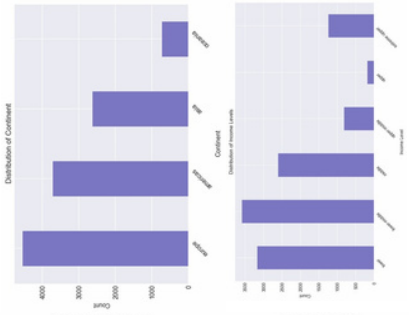
Tỷ lệ trình độ học vấn đã tốt nghiệp và unknown (không rõ) là như nhau, cụ thể 34.6%. 2 nhóm còn lại là chưa tốt nghiệp và học cao học chiếm tỷ lệ dưới 20% mỗi nhóm.

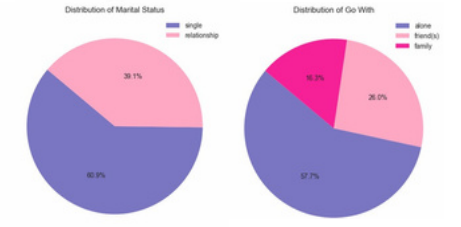
Tuổi tác (Age)

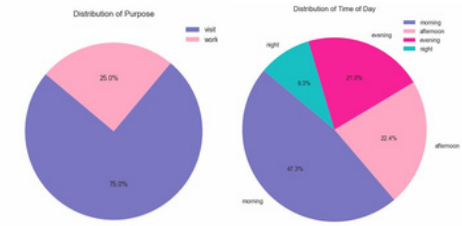
Tệp tuổi 18-29 chiếm tỷ trọng lớn nhất, hơn một nửa bộ dữ liệu. Tiếp theo là nhóm 30-39 tuổi. Các nhóm tuổi còn lại giảm dần về số lượng khi tuổi càng tăng.

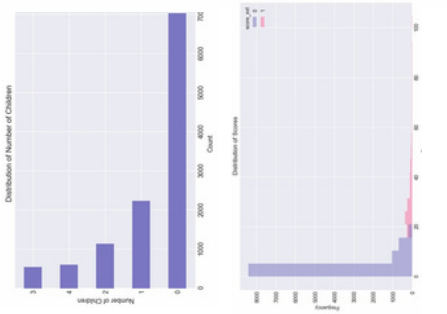
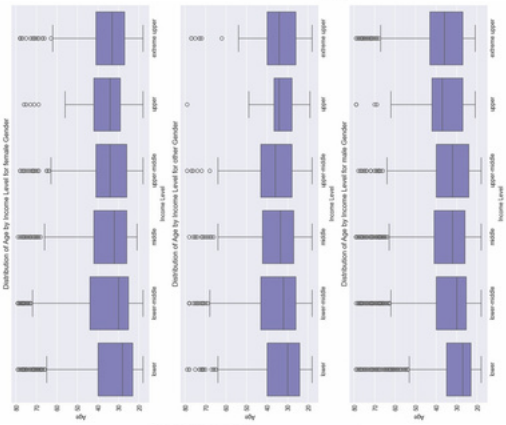
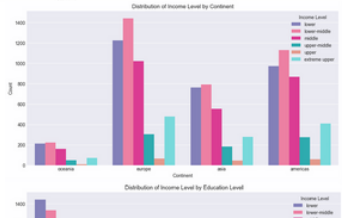
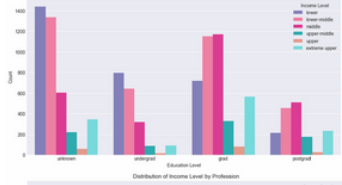
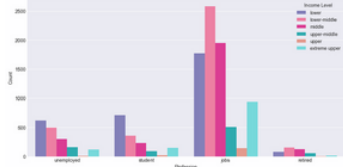
Insight: Việt Nam là một quốc gia Đông Nam Á, thiên về mảng du lịch văn hóa, khám phá thiên nhiên và ẩm thực. Mảng du lịch nghỉ dưỡng ở Việt Nam tuy vẫn đang trên đà phát triển, nhưng chưa thể so sánh được với các nước trong khu vực. Vì vậy, Việt Nam sẽ thu hút khá nhiều nhóm du khách trẻ có nhu cầu khám phá, trải nghiệm cao. Đối với nhóm du khách lớn tuổi, ưu tiên nhu cầu nghỉ dưỡng thì sẽ ít ưu tiên du lịch ở Việt Nam so với nhóm tuổi trẻ hơn. **Tình trạng việc làm (Profession)** Phần lớn bộ dữ liệu được thu thập đến từ những người có việc làm, chiếm 68.1%. Tệp người nghỉ hưu chiếm tỷ trọng thấp nhất, chỉ khoảng 3.8%.

Insight: Tương tự, đối tượng nghỉ hưu phần lớn có nhu cầu cao về nghỉ dưỡng. Việt Nam vẫn còn những điểm hạn chế trong dịch vụ công cộng, thông tin chỉ dẫn du lịch, giao thông công cộng... nên sẽ chỉ phù hợp cho những đối tượng linh hoạt, ưu tiên trải nghiệm hơn.







II. PHÂN TÍCH VÀ TRỰC QUAN HÓA DỮ LIỆU

Châu lục (Continent)

Bộ dữ liệu thu thập đa số thông tin từ du khách Châu Âu, kế đến là châu Mỹ, châu Á. Du khách đến từ Châu Đại Dương, cụ thể là nước Úc, chiếm số lượng ít nhất.

Insight: Việt Nam là 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). Điều này xảy ra là do các thị trường châu Âu được miễn thị thực ngày càng nhiều (Cục du lịch Quốc Gia Việt Nam, 2023) và 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.

Nhóm thu nhập (Income Level)

Nhóm thu nhập cận TB chiếm tỷ trọng cao nhất trong 3 nhóm thu nhập hàng đầu gồm: thấp, cận TB và TB. Nhóm thu nhập cận cao và cao chiếm tỷ trọng khá thấp. Tuy vậy, nhóm thu nhập siêu cao ($153000-$200000) lại chiếm tỷ trọng đáng kể.

Insight:

Việt Nam được xem là một trong 10 điểm đến du lịch rẻ nhất châu Á (Báo Tuổi Trẻ, 2023). Do đó khá dễ hiểu khi những người có thu nhập thấp sẽ ưu tiên du lịch tại Việt Nam để phù hợp với mức thu nhập.

Bên cạnh đó, lý do xuất hiện nhóm thu nhập siêu cao có thể là 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). Vì vậy, đây không được xem là các giá trị ngoại lai.

Tình trạng hôn nhân (Marital Status), Đối tượng đi cùng (Go_with), Mục đích (Purpose), Children (Số con)

60.9% du khách là đối tượng độc thân. Phần lớn du khách đều không có con. Hơn phân nửa du khách đến Việt Nam một mình. 26% trong số họ đi du lịch với bạn bè, phần còn lại đi với gia đình. ¾ du khách đến để du lịch, ¼ còn lại là mục đích công việc.

Insight:

Du khách đến VN sẽ là kiểu người 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.

Thời gian đáp (Time of day)

Gần phân nửa 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.

Điểm Viettel++ (Score)

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

Insight: Có thể thấy các du khách 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.

2.2. Phân tích đa biến (Multivariate analysis)

2.2.1 Đặc điểm về khách hàng

Age x Income level x Gender

- Nhìn chung, ở giới tính Nam và Nữ khi tuổi trung bình càng tăng thì mức lương càng cao. Nữ giới với mức thu nhập thấp, cận TB, TB chiếm số lượng lớn hơn so với các nhóm thu nhập còn lại, trong khi nam giới ở mức thu nhập thấp có số lượng khá thấp. Độ tuổi trung bình của Nam và Nữ ở các mức 3 thu nhập đầu tiên khá giống nhau (dao động trên dưới 30 tuổi). Tuy vậy, độ tuổi trung bình Nữ đạt được mức thu nhập cao và siêu cao trẻ hơn so với Nam.
- Đối với giới tính Other, độ tuổi trung bình ở mức thu nhập cao và siêu cao trẻ hơn so với Nam giới. Tuy vậy, số lượng người có thu nhập cao lại nhỏ hơn nhiều so với 2 giới còn lại. Bên cạnh đó, độ tuổi trung bình trong 3 mức thu nhập đầu tiên cũng cao hơn so với 2 giới còn lại.

Insight: Khách hàng trẻ thì có thu nhập thấp hơn, trong đó Nữ và Other sẽ chiếm tỷ trọng cao hơn Nam. Mức thu nhập cao sẽ phân bổ nhiều hơn ở Nam giới. Tuy vậy, Nữ và Other lại có độ tuổi trẻ hơn khi đạt đến những mức thu nhập cao này.

Continent x Income level

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

Insight: Có thể thấy các du khách 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.

Education x Income level

- Nhóm Undergrad vì là người chưa tốt nghiệp nên họ chủ yếu rơi vào nhóm thu nhập thấp.
- Sang nhóm Grad (đã tốt nghiệp), ta thấy có sự dịch chuyển nhiều hơn về phía thu nhập cao hơn. Đáng chú ý, nhóm Grad có số lượng thu nhập siêu cao nhiều nhất.
- Nhóm Postgrad (học cao học) 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. Nhóm thu nhập siêu cao thu thập được từ nhóm này cũng chiếm số lượng đáng kể.

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.

Education x Profession

II. PHÂN TÍCH VÀ TRỰC QUAN HÓA DỮ LIỆU

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. Ta nên tìm cách thu thập thêm thông tin về trình độ học vấn của nhóm này để hiểu thêm về insight. **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.
Insight: Khách du lịch đến Việt Nam theo dạng gia đình khá hạn chế.

2.2.2. Đặc điểm hành vi

Frequency (fb, yt, ins) x gender
Du khách nam có xu hướng dùng FB và YT nhiều hơn so với 2 tệp còn lại. Cả 3 tệp đều dùng INS khá nhiều và đồng đều.
Insight: Tệp người trẻ hiện nay đang dần dịch chuyển hành vi sử dụng mạng XH, điển hình là số liệu sử dụng FB giảm mạnh vì nền tảng này được xem là nền tảng kết nối giữa gia đình, "già nua" và không mới mẻ. Trong khi đó, 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.
Insight:
- Gói datasilver là gói cơ bản nhất, chỉ gồm dung lượng 4G. Vì tệp dữ liệu phần lớn đến từ khách hàng mới, chưa trung thành với Viettel, nên khả năng cao họ sẽ ngần ngại và chọn sử dụng gói cơ bản nhất để trải nghiệm dịch vụ, trước khi nâng cấp lên các gói đa dạng hơn. Bên cạnh đó, phần lớn du khách đều thuộc nhóm thu nhập thấp nên sẽ nhạy cảm về giá, dẫn đến ưu tiên chọn gói giá rẻ hơn để thử nghiệm.
- Gói socialmedia và socialmediagold là nổi bật ở dung lượng không hạn chế cho các trang mạng XH. Phần lớn du khách là người trẻ hoặc đã trưởng thành, độc thân, du lịch một mình. Vì vậy nhu cầu dùng mạng XH để cập nhật thông tin, giữ liên lạc... là không thể thiếu. Vì vậy, với nhu cầu đó, họ sẵn sàng chi thêm tiền để có thể truy cập mạng XH không giới hạn.
- 2 gói còn lại có tỷ lệ chuyển đổi lại khá thấp. Lý do có thể là gói datacall chú trọng về mảng nghe gọi, nhưng ở thời đại 4.0, có lẽ nhu cầu nghe gọi qua các trang mạng XH có thể thay thế cho nghe gọi trực tiếp. Hơn nữa, họ là du khách nước ngoài độc thân, không đi theo dạng gia đình nhiều nên nhu cầu nghe gọi chưa hẳn là ưu tiên hàng đầu. Gói datagold có giá cao, chưa phù hợp với tệp khách mới muốn dùng thử.

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. Có ...% du khách đến Hà Nội sẽ nghe đề xuất về Viettel sau 30' đáp.
Insight: Các du khách đến Hà Nội sẽ không ở lại sân bay lâu, trong khi đó, du khách chuyển chuyến bay sẽ dành nhiều thời gian hơn ở sân bay, từ đó có khả năng nghe đề xuất nhiều hơn. Vậy nên marketing những gói cơ bản, tiện lợi nhất dành cho các du khách đến Hà Nội. Đối với những khách đến tỉnh khác, họ sẽ có nhiều thời gian "chết" khi di chuyển, vì vậy có thể xem xét kỹ giữa nhiều gói đa tính năng hơn..

2.2.3. Đặc điểm chuyến đi

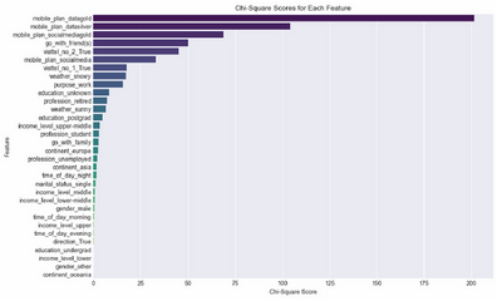
Purpose x direction
Đa số du khách đến tỉnh khác để du lịch. Trong khi đó, lượng du khách đến Hà Nội để du lịch và làm việc khá tương đồng nhau
Insight: Đối với du khách đến tỉnh khác, họ có mục đích du lịch là chính nên có thể mkt những gói cước liên quan đến mạng xã hội. Còn đối với những du khách đến để làm việc, có thể đề xuất những gói cơ bản hoặc có yếu tố nghe gọi để họ phục vụ công việc.
Purpose x go_with100% du khách đến để làm việc thì đi 1 mình.

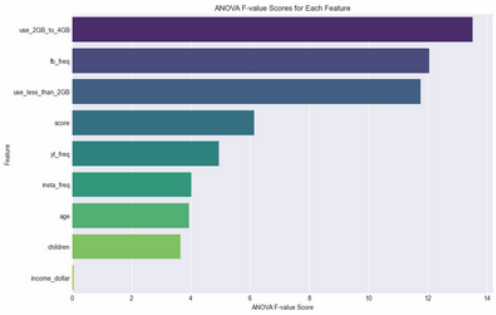
2.3. Mối quan hệ giữa các biến

- Dùng phương pháp Chi-Square cho biến phân loại, p-value càng nhỏ càng thể hiện mối liên kết mạnh → Đa số các cặp biến có mối liên kết mạnh với nhau
- Dùng phương pháp Phi coefficient cho biến nhị phân, hệ số càng lớn thì càng thể hiện mối liên kết mạnh → Trừ các cặp biến về giá trị 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.

2.4. Test set vs. Train set

Giữa test set và train set có những khác biệt tiêu biểu trong các trường:
- Giới tính: Test set gồm 100% dữ liệu ghi nhận từ du khách nữ
- Tuổi: Test set 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 lại đồng đều hơn train set.
- Children: Số lượng con tăng đáng kể, phần lớn mỗi du khách không có con hoặc có từ 1-2 con, so với train set phần lớn du khách sẽ không có con.
- Accept: tỉ lệ chấp nhận đăng ký ở train set mất cân bằng nhiều hơn so với test set





IV. FEATURE SELECTION VÀ ENGINEERING

4.1. Feature selection

- Nhóm chia dữ liệu thành 2 tập, 1 tập chứa các categorical features, 1 tập chứa các continuous features. Trong đó tập categorical features sẽ được one hot encode. Sử dụng
- lớp SelectKBest trong thư viện sklearn để đánh các đặc trưng (features) tốt nhất từ dữ liệu dựa trên phương thức tính điểm cụ thể ứng với từng tập.
  - Đối với tập **categorical features**, phương thức tính điểm là kiểm định Chi-squared. Kết quả cho thấy có 17 đặc trưng có p-value nhỏ hơn hoặc bằng 0.05. Đối với tập
  - **continuous features**, phương thức tính điểm là kiểm định ANOVA. Kết quả cho thấy có 11 đặc trưng có p-value nhỏ hơn 0.05.
- **Tuy nhiên** vẫn cho tất cả đặc trưng vào mô hình. Sau khi chọn được mô hình phù hợp và tunning, nhóm sẽ thực hiện kiểm tra feature importance và kết luận các đặc trưng có thể bỏ ra khỏi mô hình.

4.2. Feature engineering

- Tạo thêm một số nhãn báo bất thường cho các cột như:
  - Nhóm 'fb_freq', 'yt_freq', 'insta_freq': tổng thời gian sử dụng 3 nền tảng này lớn hơn 10h/ngày sẽ được xem là bất thường, dán nhãn 1, ngược lại dán nhãn 0.
  - Nhóm 'use_less_than_2GB', 'use_2GB_to_4GB': sử dụng khoảng 1.5IQR để dán nhãn 1 cho các điểm nằm ngoài khoảng 1.5IQR, nhãn 0 cho các điểm nằm trong.
  - Cột score: áp dụng kỹ thuật tương tự nhóm cột usage.
  - Cột eduacation: các giá trị 'unknown' có trong cột tương đối lớn, dán nhãn 1 cho các giá trị này, các giá trị khác dán nhãn 0.
- 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. Thực hiện oversampling cho cả dataset bằng hàm SMOTE
- trong thư viện imblearn

V. XÂY DỰNG MÔ HÌNH VÀ ĐÁNH GIÁ

5.1. Các phương thức 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. Từ đó, nhóm đề xuất các cách tiếp cận mô hình chung cho tất cả các gói dịch vụ như sau:

- **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, nhóm sẽ 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.

5.2. Độ đo

Nhóm sử dụng các độ đo phổ biến trong bài toán classification như: Accuracy, Precision, Recall, F1-score (Xem thêm tại Appendix 5)

5.3. 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
- Đối với từng cách tiếp cận, nhóm biến đổi dữ liệu test cho phù hợp với định hướng.
- Lần lượt train tập train set với các thuật toán, thuật toán nào có F1-score cao nhất sẽ được lấy làm mô hình cho approach đó.
- Dùng mô hình vừa chọn để test với test set và so sánh chúng dựa trên các độ đo đã chọn.

5.4. Kết quả

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.

- 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.

Approach 1

	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 2

	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
3	...	A	0
3	...	B	0
3	...	C	1
3	...	D	0
3	...	E	0

The result of the validation set using Approach 1.

Model	Accuracy	Precision	Recall	F1 Score
RANDOMFOREST	0.882951	0.889602	0.882951	0.882528
EXTRATREES	0.877794	0.882422	0.877794	0.877492
GRADIENTBOOSTING	0.865162	0.894007	0.865162	0.862838
CATBOOST	0.864639	0.893684	0.864639	0.862287
XGBOOST	0.862097	0.893122	0.862097	0.859605
DECISIONTREE	0.838329	0.838341	0.838329	0.838331
KNN	0.766350	0.766511	0.766350	0.766347
LOGISTIC	0.642948	0.643218	0.642948	0.642901
NAIVEBAYES	0.640706	0.662831	0.640706	0.626749
MLP	0.522760	0.639178	0.522760	0.403005

The result of the testing set using Approach 1.

Model	Accuracy	Precision	Recall	F1 Score
NAIVEBAYES	0.625000	0.625797	0.625000	0.624405
EXTRATREES	0.624302	0.737227	0.624302	0.573552
CATBOOST	0.620810	0.784353	0.620810	0.557132
GRADIENTBOOSTING	0.620810	0.784353	0.620810	0.557132
RANDOMFOREST	0.620112	0.775863	0.620112	0.557678
XGBOOST	0.620112	0.784127	0.620112	0.556042
DECISIONTREE	0.607542	0.626310	0.607542	0.592401
LOGISTIC	0.589385	0.593961	0.589385	0.584325
MLP	0.565642	0.604328	0.565642	0.521263
KNN	0.501397	0.501418	0.501397	0.499507

The result of the validation set using Approach 2.

Model	Accuracy	Precision	Recall	F1 Score
RANDOMFOREST	0.941769	0.947558	0.941769	0.941539
CATBOOST	0.939395	0.945880	0.939395	0.939129
EXTRATREES	0.938863	0.942945	0.938863	0.938685
GRADIENTBOOSTING	0.938572	0.945224	0.938572	0.938295
XGBOOST	0.935956	0.943154	0.935956	0.935644
DECISIONTREE	0.890805	0.891159	0.890805	0.890801
KNN	0.858153	0.864661	0.858153	0.857384
NAIVEBAYES	0.671834	0.701940	0.671834	0.660446
LOGISTIC	0.659384	0.659956	0.659384	0.659278
MLP	0.554743	0.651435	0.554743	0.464169

The result of the testing set using Approach 2.

Model	Accuracy	Precision	Recall	F1 Score
RANDOMFOREST	0.952975	0.956916	0.952975	0.952873
GRADIENTBOOSTING	0.952680	0.956771	0.952680	0.952574
CATBOOST	0.951895	0.956117	0.951895	0.951783
XGBOOST	0.950422	0.954895	0.950422	0.950300
EXTRATREES	0.947281	0.950477	0.947281	0.947187
DECISIONTREE	0.894267	0.894822	0.894267	0.894229
NAIVEBAYES	0.729923	0.748630	0.729923	0.724746
LOGISTIC	0.676713	0.678963	0.676713	0.675694
MLP	0.662478	0.732860	0.662478	0.634889
KNN	0.515806	0.515961	0.515806	0.514627

V. XÂY DỰNG MÔ HÌNH VÀ ĐÁNH GIÁ

5.6. Hyperparameter tuning

Thực hiện tunning bằng phương pháp grid search CV cho mô hình Randomforest Classifier với các tham số như sau (xem thêm ý nghĩa các tham số tại Appendix 7):

- n_estimators: 500
- max_depth: 3
- criterion : gini
- min_samples_split: 2

Tổng quan mô hình sau khi tunning đạt được kết quả tốt hơn, với Accuracy, F1-score, Precision và Recall được cải thiện.

VI. CHIẾN LƯỢC KINH DOANH

6.1. Cơ sở phát triển chiến lược

Từ các insights có được trong phần EDA và mô hình, ta có những đặc điểm chính như sau: ***Khách hàng là những người trưởng thành, có thu nhập thấp nên nhạy cảm về giá. Với tâm thế YOLO, du lịch trải nghiệm một mình, họ thích dùng mạng XH để cập nhật thông tin và liên lạc. Tuy vậy, đa phần họ chưa biết hoặc chưa trung thành với thương hiệu.*** Nhóm rút ra **cách tiếp cận** khách hàng bằng các **USP** (unique selling points) đáp ứng nhu cầu của du khách như sau:

- Giá rẻ cùng ưu đãi ngập tràn
- Dùng không giới hạn (mạng XH)

6.2. Đề xuất chiến lược

6.2.1. Price

- **Giảm giá trong lần đầu sử dụng**: đối với gói cước có giá cao hơn (datagold, socialmediagold, datacall) → đánh vào **tâm lý price-sensitive** của khách hàng mới và **sự cạnh tranh giá** trong dịch vụ data. **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 giá ưu đãi ở lần tiếp theo → **Thúc đẩy tiêu dùng đối với các gói ít phổ biến hơn & giữ chân khách hàng**.

Vd: ưu đãi datagold và datacall chỉ còn 150,000 VND nếu trước đó dùng gói datasilver.

6.2.2. Place

- **Điểm chạm trực tiếp**: Dựng booth đăng ký sim/gói 4G tự động ở sân bay → người dùng tự tương tác đúng với nhu cầu, dễ thu thập thông tin, nhanh và tiện.
- **Phủ sóng thương hiệu**:
  - Tăng độ phủ quảng cáo tại các địa điểm trong journey của du khách: sân bay, taxi truyền thống & công nghệ, trạm bus, trung tâm thương mại, các địa điểm check-in, các quán ăn trong hoặc gần sân bay... để du khách nhớ về thương hiệu.
  - Tăng độ phủ trên truyền thông, đặc biệt là qua các trang mạng XH mà du khách hay sử dụng (quảng cáo YT, FB, INS).
  - Booking KOL, travel blogger, travel vlogger về mảng du lịch.

6.2.3. Promotion & Product

- Đánh vào 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

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

- Tâm lý 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.

VII. ĐỀ XUẤT MỞ RỘNG

7.1. Thu thập thêm dữ liệu

- Thu thấp đủ thông tin đề xuất và chấp nhận hoặc không đối với cả 5 gói cho mỗi khách
  Mục đích: so sánh được nhu cầu của du khách đối với các gói khác nhau
- Địa điểm du khách hay lui đến
  Mục đích: tăng cường các kênh truyền thông 1 cách hiệu quả nhất
- Nghiên cứu thêm về các platform phổ biến đối với người dùng nước ngoài (vd: Spotify, Apple Music, Twitter...)
  Mục đích: bổ sung các gói cước có các platform này

7.2. Dùng shuffling cho hai tập dữ liệu train và test để cải thiện mô hình

Vì có sự mất cân bằng trong phân phối giữa hai tập dữ liệu train và test, có thể cân nhắc phương pháp shuffling hai bộ với nhau và chia lại để cải thiện kết quả xây dựng mô hình.

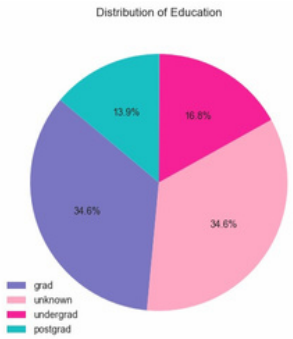
7.3. Dùng resampling cho hai tập dữ liệu train và test để cải thiện mô hình

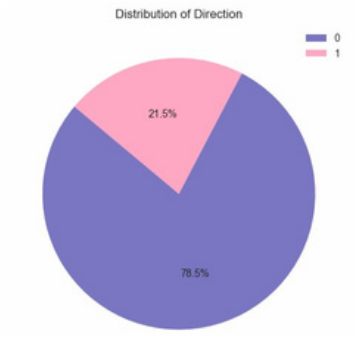
Phương pháp này liên quan đến việc điều chỉnh sự cân bằng giữa các lớp thiểu số và đa số thông qua oversampling hoặc undersampling. Trong tập train set có sự mất cân bằng trong số lượng accept, vì vậy có thể cân nhắc sử dụng resampling cho bộ dataset.

VIII. 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 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.

Đích đến (Direction)

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

APPENDIX 3 Các phân tích EDA đa biến khác

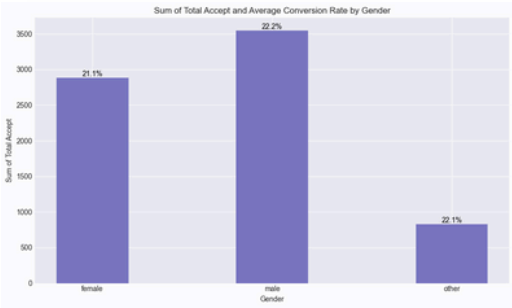
Gender x accept x Mean CVR Nam và nữ giới có mức độ chấp nhận gói cước khá cao so với tệp khách hàng còn lại. Tuy nhiên, dù có số lượng accept thấp, tỉ lệ CVR của tệp này lại khá đồng đều so với 2 tệp nam, nữ. → Vẫn nên tập trung thúc đẩy đều cho 3 tệp

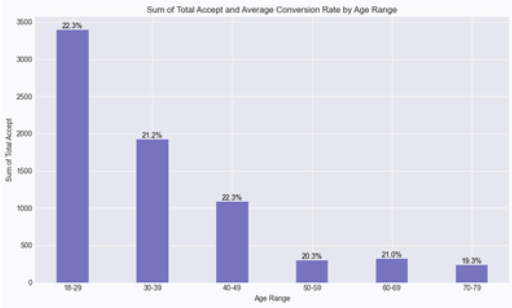
Age x accept x Mean CVR

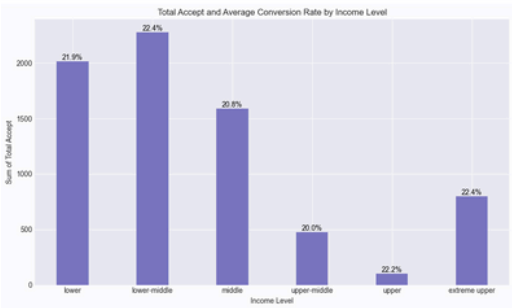
3 nhóm tuổi trẻ nhất đều chiếm số lượng accept và CVR cao nhất.

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 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.

VIII. 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)

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

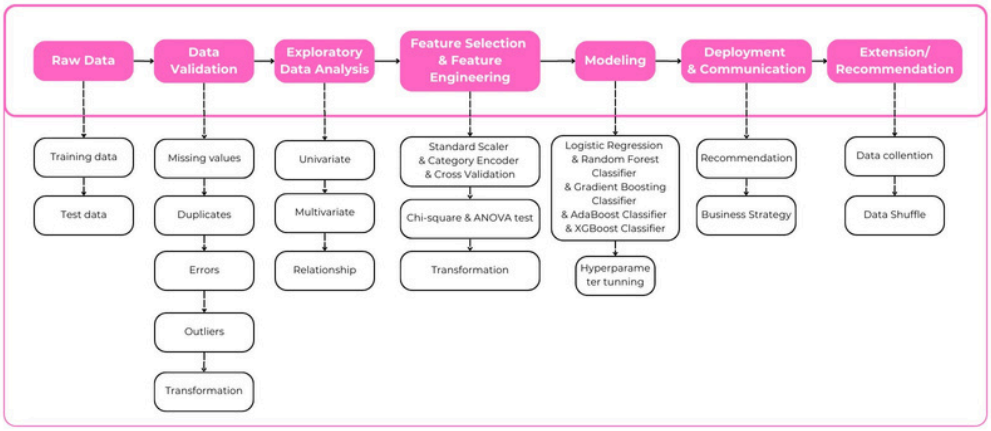
Đây là tỉ lệ các dự đoán dương đúng trên tổng số mẫu dương thực sự. Độ nhạy cao cho thấy mô hình ít bỏ sót các trường hợp dương.

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 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