ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

PHÂN TÍCH HIỆU QUẢ CHIẾN DỊCH VÀ PHÂN KHÚC KHÁCH HÀNG DỰA VÀO HÀNH VI MUA HÀNG TRÊN THƯƠNG MẠI ĐIỆN TỬ

Nhóm 10			
Sinh viên thực hiện:			
STT	Họ tên	MSSV	Ngành
1	Nguyễn Tấn Dũng	21521977	CNTT
2	Trần Tuyết Minh	21521144	CNTT
3	Nguyễn Minh Duy	21522005	HTTT

TP. HỒ CHÍ MINH – 12/2024

1. GIỚI THIỆU

Đề tài này nhằm mục đích phân tích và đánh giá xu hướng sử dụng trực tiếp các kênh truyền thông đa phương tiện trong thương mại điện tử trong giai đoạn 2021-2023, sử dụng bộ dữ liệu "E-commerce Multichannel Direct Messaging" được cung cấp trên Kaggle. Bộ dữ liệu này chứa thông tin về các chiến dịch truyền thông đa kênh, bao gồm các chiến dịch qua email, SMS, và các phương thức khác, giúp chúng tôi hiểu rõ hơn về tác động của các chiến lược truyền thông đến hành vi của khách hàng và hiệu quả kinh doanh.

Để thực hiện phân tích này, nhóm sử dụng các công cụ và thư viện như Python, Pandas, và Matplotlib trong môi trường Jupyter Notebooks. Dữ liệu được xử lý và làm sạch để loại bỏ các giá trị thiếu, điều chỉnh dữ liệu không hợp lệ, và chuẩn hóa các thông số quan trọng nhằm đảm bảo chất lượng phân tích. Phân tích khám phá dữ liệu (EDA) được áp dụng để tìm ra các mối tương quan giữa các yếu tố trong chiến dịch và kết quả kinh doanh, đồng thời xác định các yếu tố chủ chốt ảnh hưởng đến sự thành công của các chiến dịch truyền thông.

Bộ dữ liệu được tham khảo từ Kaggle [1] và được sử dụng độc lập với các dự án khác. Đề tài này được thực hiện hoàn toàn độc lập, không sao chép từ bất kỳ đồ án nào trước đó. Mã nguồn và dữ liệu minh chứng sẽ được nộp kèm theo để đảm bảo tính minh bạch.

Trong nghiên cứu này, chúng tôi tiến hành phân tích hiệu quả các chiến dịch truyền thông đa kênh và phân khúc khách hàng dựa trên hành vi mua sắm trong thương mại điện tử. Kết quả phân tích xác định được top 10 chiến dịch hiệu quả nhất dựa trên tần suất xuất hiện, số lượt mở, lượt nhấp và số lượng người mua hàng. Đồng thời, khách hàng được phân chia thành 3 nhóm chính gồm Champion, Potential, và Loss, dựa trên các tiêu chí như mở thông báo, nhấp vào liên kết và thực hiện mua sắm. Qua đó, chúng tôi nhận diện sự khác biệt giữa các loại chiến dịch (bulk, trigger, transactional) và đề xuất các chiến lược tối ưu hóa, giúp cải thiện hiệu quả truyền thông và tăng tỷ lệ chuyển đổi khách hàng.

2. MÔ TẢ BỘ DỮ LỆU

Bộ dữ liệu này ghi lại lịch sử các chiến dịch tin nhắn đa kênh trong vòng 2 năm của một công ty bán lẻ vừa và nhỏ. Các chiến dịch được gửi qua các kênh bao gồm email, web push, mobile push và SMS. Mỗi chiến dịch có thể là chiến dịch đại trà (bulk), tự động theo hành động (trigger) hoặc liên quan đến giao dịch (transactional). Dữ liệu này thu thập từ dự án REES46 CDP, một nền tảng quản lý dữ liệu khách hàng. Phiên bản đầy đủ của bộ dữ liệu có chứa 721 triệu tin nhắn, trong khi phiên bản demo chúng tôi sử dụng trong phân tích này có 10 triệu tin nhắn. Bộ dữ liệu có tổng cộng 4 tập:

Bao gồm tập messages-demo có 10000000 dòng và 31 cột. Trong đó bao gồm 3 biến số, 14 biến phân loại và một số cột chứa các trạng thái khác nhau của chiến dịch (ví dụ: mở, nhấp, hủy đăng ký, phản hồi), và có 2 cột không dùng cho mục đích phân tích.

Bên dưới là bảng mô tả các thuộc tính trong bộ dữ liệu.

Tên thuộc tính	Mô tả thuộc tính	Loại biến	Kiểu giá trị	Tên thuộc tính	Tên thuộc tính	Tên thuộc tính	Tên thuộc tính
Id	ID chuỗi tin nhắn	Định danh	Int64	**clicked_first_time_at**	Thời gian click lần đầu	Dữ liệu	DateTime
message_id	ID của message	Định danh	Object	**clicked_last_time_at**	Thời gian click lần cuối	Dữ liệu	DateTime
campaign_id	ID Campaign	Định danh	Int64	**is_unsubscribed**	Có hủy đăng ký không?	Boolean	Boolean
message_type	Loại tin nhắn	Dữ liệu	String	**unsubscribed_at**	Thời gian hủy đăng ký	Dữ liệu	DateTime
client_id	ID khách hàng	Định danh	Int64	**is_hard_bounced**	Có bị hard bounce không?	Boolean	Boolean
channel	Kênh gửi tin	Dữ liệu	String	**hard_bounced_at**	Thời gian hard bounce	Dữ liệu	DateTime
category	Danh mục tin nhắn	Dữ liệu	String	**is_soft_bounced**	Có bị soft bounce không?	Boolean	Boolean
platform	Nền tảng tin nhắn	Dữ liệu	String	**soft_bounced_at**	Thời gian soft bounce	Dữ liệu	DateTime
email_provider	Nhà cung cấp email	Dữ liệu	String	**is_complained**	Có phàn nàn không?	Boolean	Boolean
stream	Luồng gửi	Dữ liệu	String	**complained_at**	Thời gian phàn nàn	Dữ liệu	DateTime
date	Ngày gửi tin	Dữ liệu	Date	**is_blocked**	Có bị chặn không?	Boolean	Boolean
sent_at	Thời gian gửi tin	Dữ liệu	DateTime	**blocked_at**	Thời gian bị chặn	Dữ liệu	DateTime

is_opened	Có mở tin không?	Boolean	Boolean	**is_purchased**	Có mua hàng không?	Boolean	Boolean
opened_first_time_at	Thời gian mở lần đầu	Dữ liệu	DateTime	**purchased_at**	Thời gian mua hàng	Dữ liệu	DateTime
opened_last_time_at	Thời gian mở lần cuối	Dữ liệu	DateTime	**created_at**	Thời gian tạo tin nhắn	Dữ liệu	DateTime
is_clicked	Có click vào tin không?	Boolean	Boolean	**updated_at**	Thời gian cập nhật tin nhắn	Dữ liệu	DateTime

Trong bộ dữ liệu gốc, một số thuộc tính có kiểu dữ liệu chưa phù hợp cho phân tích. Chẳng hạn, các cột thời gian như date và sent_at ban đầu được nhận diện là chuỗi (object) thay vì datetime, gây khó khăn cho việc tính toán thời gian hay lọc dữ liệu theo ngày. Tương tự, các cột trạng thái như is_opened, is_clicked, is_unsubscribed, và is_complained được nhận diện là chuỗi do nó ở dạng "t" và "f" thay vì Boolean, trong quá trình xử lý dữ liệu sẽ convert về boolean.

Ngoài ra, các thuộc tính như category cần chuyển sang kiểu số (float) thay vì chuỗi để thực hiện tính toán. Để đảm bảo tính nhất quán và dễ dàng phân tích, nhóm sẽ chuẩn hóa các kiểu dữ liệu này trong bước tiền xử lý, chuyển các cột về định dạng phù hợp như datetime cho thời gian và boolean cho trạng thái. Và loại bỏ những cột thừa, không mang chức năng từ dữ liệu gốc.

Bảng campaigns.csv:

id	ID chiến dịch	Định danh	Int64	**subject_with_emoji**	Chủ đề có emoji hay không	Boolean	Object
campaign_type	Loại chiến dịch (bulk, triggers, transactional)	Phân loại	Object	**subject_with_bonuses**	Chủ đề có bonus hay không	Boolean	Object
channel	Kênh gửi chiến dịch (email, mobile_push,	Phân loại	Object	**subject_with_discount**	Chủ đề có giảm giá hay không	Boolean	Object

	web_push, sms)						
subject_length	Độ dài chủ đề chiến dịch	Định danh	Int64	**subject_with_saleout**	Chủ đề có sale-out hay không	Boolean	Object
subject_with_personalization	Chủ đề có thời hạn hay không	Boolean	Object				

Bảng holiday.csv:

date	Ngày lễ	Thời gian	Object
holiday	Tên lễ hội	Phân loại	Object

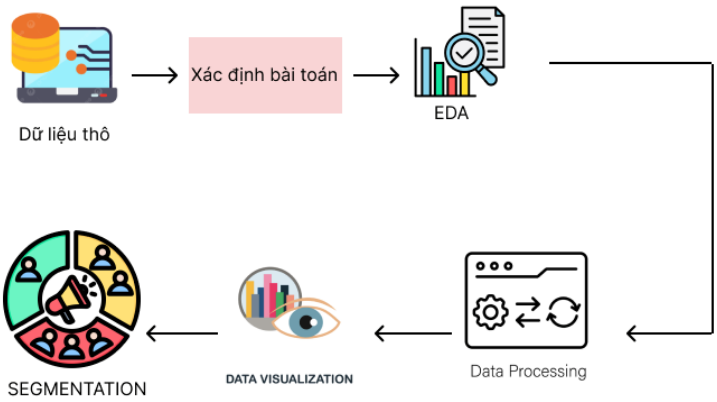
Bộ dữ liệu holiday.csv bao gồm thông tin về các ngày lễ với các thuộc tính như date (Ngày lễ), có kiểu giá trị Object, ghi nhận thời gian của ngày lễ. Thuộc tính holiday chứa tên của lễ hội, có kiểu dữ liệu Object.

Bảng client_first_purchase_date.csv:

client_id	ID của khách hàng	Định danh	Float64
first_purchase_date	Ngày mua hàng lần đầu tiên của khách hàng	Thời gian	Object

Bộ dữ liệu client_first_purchase_date.csv chứa thông tin về khách hàng với các thuộc tính như client_id (ID của khách hàng), có kiểu giá trị Float64. Thuộc tính first_purchase_date ghi nhận ngày mua hàng lần đầu tiên của khách hàng, có kiểu giá trị Object.

3. PHƯƠNG PHÁP PHÂN TÍCH



Hình 1 Quy trình PTDL

3.1. Xác định bài toán

Mục tiêu của bài toán là phân tích hiệu quả chiến dịch và phâ phân nhóm khách hàng dựa trên hành vi của họ trong các chiến dịch tiếp thị, đặc biệt là hành vi mở thông báo, nhấp chuột vào liên kết và mua hàng. Câu hỏi chính của bài toán là: Làm thế nào để xác định các nhóm khách hàng có hành vi tương tự và phân loại chúng thành các phân khúc khách hàng.

3.2. Tiền xử lý dữ liệu:

Bộ dữ liệu được sử dụng trong nghiên cứu này được trích xuất từ một bộ dữ liệu lớn hơn, bao gồm hơn 721 triệu dòng, với 10 triệu dòng được chọn từ khoảng thời gian từ ngày 30/4/2021 đến ngày 14/6/2021. Tập dữ liệu chính, messages_demo, lưu trữ lịch sử tương tác của khách hàng với các tin nhắn gửi đi, với hơn 10 triệu dòng dữ liệu liên quan đến các chiến dịch tiếp thị. Ngoài dữ liệu chính này, ba tập dữ liệu phụ liên quan cũng được sử dụng, bao gồm holidays, campaigns, và first_date_purchased.

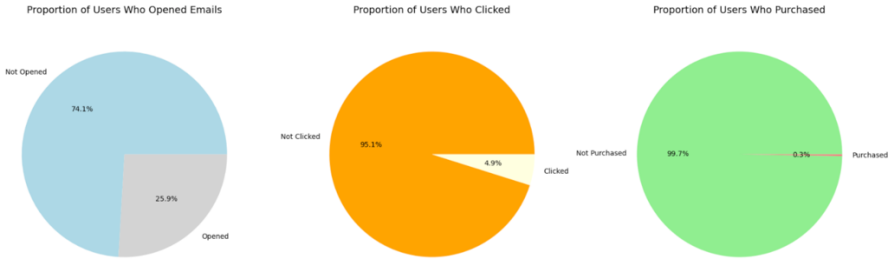
Do thời gian phân tích dữ liệu khá ngắn, tập dữ liệu holidays đã được loại bỏ khỏi quá trình tiền xử lý và phân tích. Các tập dữ liệu còn lại, campaigns và first_date_purchased, tiếp tục được giữ lại để sử dụng trong các phân tích tiếp theo.

Tiền Xử Lý:

- **Giá trị Null: Loại bỏ hai cột Platform (92.62% null) và Email Provider (42.44% null) vì tỷ lệ null cao.**
- **Cột không có giá trị phân tích: Loại bỏ cột stream vì chỉ chứa một giá trị duy nhất là "desktop".**
- **Cột có tỷ lệ bounce thấp: Loại bỏ các dòng có giá trị True trong các cột hard bounce (chiếm 0.3%) và soft bounce(do tỷ lệ thấp và không đáng kể.**

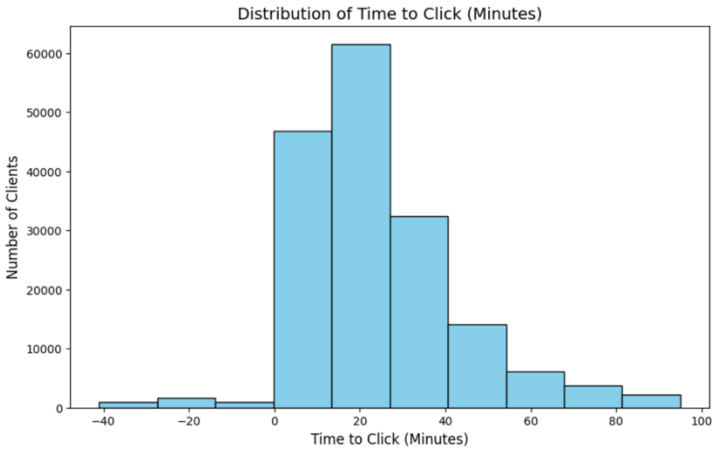
- **Loại bỏ cột không cần thiết: Các cột như id, created_at, updated_at, và category bị loại bỏ vì không có mô tả rõ ràng cho cách sử dụng những cột này.**
- **Loại bỏ dữ liệu bất hợp lý:**
  - Một số khách hàng có lần mua đầu tiên trước 2021, dẫn đến thiếu dữ liệu trong bảng **client_first_purchase**. Dữ liệu tạm thời được đánh dấu khách hàng có ngày mua không hợp lệ.
  - Tồn tại trường hợp khách hàng chưa có hành động mở thông báo (is_opened) mà đã có hành động ấn vào liên kết mua hàng (is_clicked).

3.3. Phân tích và trực quan hoá dữ liệu



Hình 2 Tỉ lệ khách hàng mở - click - mua hàng

Ba biểu đồ trên cho thấy sự tương quan giữa tỷ lệ khách hàng mở email, nhấp vào liên kết và mua hàng. Mặc dù có 25.1% khách hàng mở email, chỉ 2.5% trong số đó nhấp vào liên kết, và chỉ 1.5% thực hiện mua hàng. Điều này cho thấy sự sụt giảm đáng kể qua từng giai đoạn, từ mở email đến nhấp vào liên kết và cuối cùng là mua hàng.



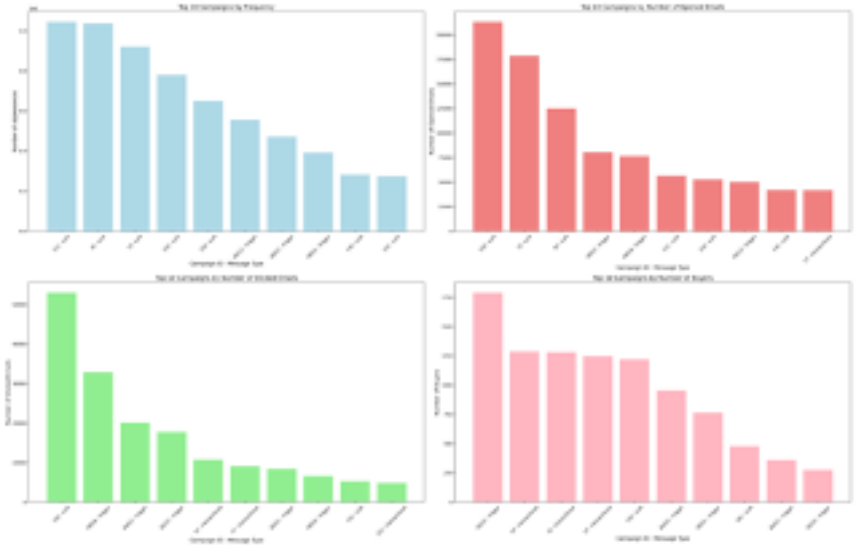
Hình 3 Phân phối thời lượng từ lúc mở đến lúc click

Biểu đồ này thể hiện phân bố thời gian từ khi email được gửi đến lúc khách hàng nhấp vào liên kết, tính theo giây. Dưới đây là nhận xét ngắn gọn:

- **Phần lớn khách hàng nhấp vào liên kết trong khoảng từ 10 đến 40 giây sau khi nhận email, với đỉnh cao nhất vào khoảng 20-30 giây.**

- **Số lượng khách hàng nhấp vào liên kết giảm dần sau khoảng 40 giây.**
- **Rất ít khách hàng nhấp vào liên kết ngay lập tức hoặc sau 60 giây.**

Nhận xét này chỉ ra rằng thời gian phản hồi của khách hàng đối với email là nhanh chóng trong vòng nửa giờ đầu, sau đó giảm dần, cho thấy khoảng thời gian này là quan trọng.



Hình 4 Top những chiến dịch dựa trên tần suất tương tác

Dựa trên 4 biểu đồ trong hình, có thể rút ra các mối tương quan như sau:

- Tương quan giữa tần suất xuất hiện và số email được mở: Các chiến dịch có tần suất xuất hiện cao như "111_bulk" không nhất thiết tương quan với số lượng email được mở cao, cho thấy tần suất xuất hiện không đảm bảo sự quan tâm của người nhận.
- Tương quan giữa số email được mở và số email được nhấp: Các chiến dịch có số email mở cao (như "150_bulk" và "79_bulk") thường có số lượt nhấp cao hơn. Điều này chỉ ra rằng việc mở email có mối quan hệ tích cực với khả năng người nhận thực hiện hành động (nhấp).
- Tương quan giữa số email được nhấp và số người mua: Không có mối tương quan chặt chẽ giữa số lượt nhấp và số người mua. Ví dụ, chiến dịch "150_bulk" có lượt nhấp cao nhất nhưng không dẫn đầu về số người mua, trong khi "1323_trigger" có lượt nhấp thấp nhưng lại đạt được nhiều người mua hơn. Điều này cho thấy chất lượng nhấp chuột (đúng đối tượng) quan trọng hơn số lượng.

Tương quan tổng thể: Sự chuyển đổi từ tần suất xuất hiện -> số email được mở -> số lượt nhấp -> số người mua giảm dần, phản ánh rằng không phải mọi chiến dịch có phạm vi tiếp cận lớn đều mang lại hiệu quả kinh doanh cao. Nội dung và mức độ nhắm đúng đối tượng mới là yếu tố quyết định.

3.4. Customer Segmentation:

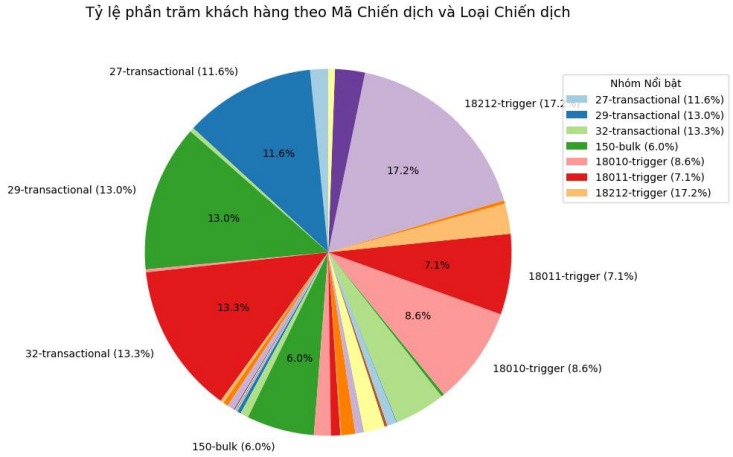
Chúng tôi chia khách hàng thành từng phân khúc với các đặc trưng riêng biệt dựa trên hành vi của khách hàng.

- Champion: Phân khúc khách hàng quan trọng và đáng chú ý.

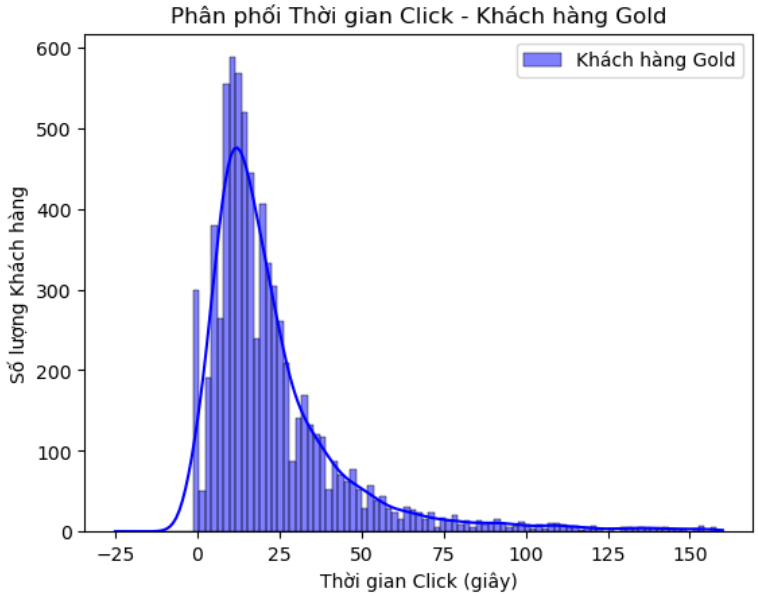
- Potential: Phân khúc khách hàng tiềm năng.
- Loss: Phân khúc khách hàng không có tương tác với sản phẩm.

3.4.1. Phân khúc khách hàng Champion

Chúng tôi dựa trên 4 tiêu chí để phân khúc tập khách hàng này là: có mua hàng lần đầu tiên, có hành động mở thông báo, có hành động mở liên kết mua hàng và có hành động mua hàng.



Hình 5 Tỷ lệ phần trăm khách hàng theo mã chiến dịch và loại chiến dịch



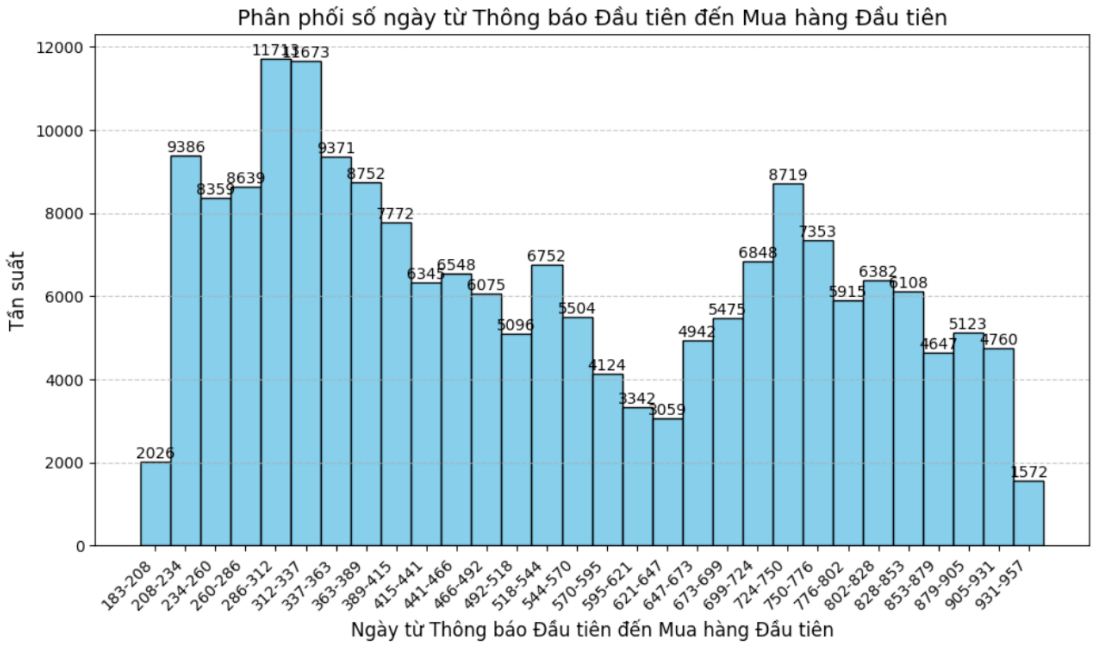
Hình 6 Phân phối thời gian click

Hình 05 thể hiện hành vi của nhóm khách hàng này trên từng loại campaign_id và campain_type. Ngoài ra, Hình 06 thể hiện phân phối về thời gian từ khi khách hàng mở thông báo đến khúc ấn vào liên kết mua hàng.

3.4.2. Phân khúc khách hàng Loss

Khách hàng thuộc phân khúc này không có hành vi tương tác cũng như mua hàng trong khoảng thời gian của bộ dữ liệu (30/04/2021 – 14/06/2021). Dựa vào hình 2, nhóm khách hàng này chiếm tỉ lệ 74,1%.

Hình 7 Tỉ lệ khách hàng có mua lần đầu và không mua lần đầu trong nhóm khách hàng loss



Hình 8 Phân phối số ngày từ lúc mua hàng lần đầu tiên đến khi nhận được thông báo đầu tiên của nhóm khách hàng loss

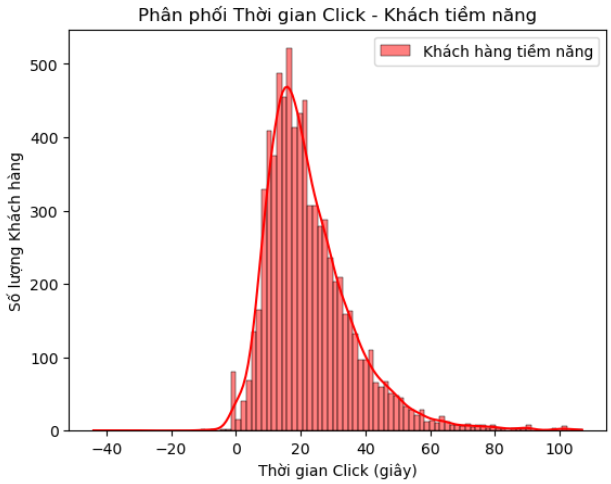
Hình 07 cho thấy nhóm khách hàng chưa từng mua hàng và không tương tác với thông điệp từ kênh thuộc phân khúc "Loss" – nhóm ít quan tâm và tiềm năng thấp. Trong số đó, 7,8% khách hàng từng mua hàng nhưng không tương tác. Tuy nhiên, do thời gian từ lần mua đầu đến thời điểm hiện tại quá lâu (trên 200 ngày), họ có khả năng đã không còn hoạt động. Dữ liệu giới hạn trong khoảng thời gian từ 04/2021 đến 06/2021, nên không đủ thông tin để xác định rõ đặc điểm của những khách hàng này, đặc biệt với

nhóm có lần mua hàng đầu sau năm 2021. Cần bổ sung dữ liệu để đánh giá chính xác hơn.

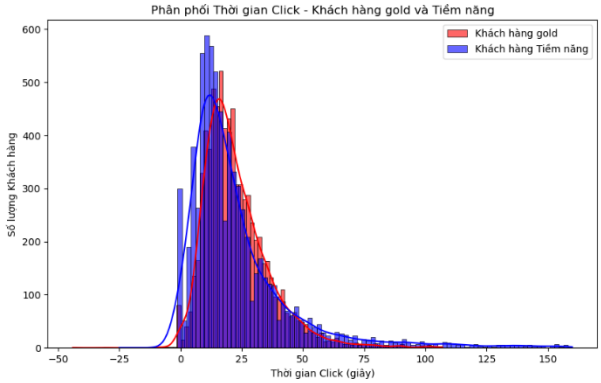
3.4.3. Phân khúc khách hàng Potential

Để xác định tập khách hàng cho phân khúc này, chúng tôi dựa tên hành vi mua hàng và phân phối về thời gian mua hàng của phân khúc Champion theo điều kiện:

- Khách hàng không thuộc phân khúc Champion và Loss.
- Khách hàng có hành vi mở thông báo và ấn vào liên kết mua hàng: Lý do chúng tôi sử dụng bộ lọc này vì muốn đánh mạnh vào tập khách hàng có hành vi mua hàng gần nhất làm điều kiện cần cho phân khúc khách hàng tiềm năng.

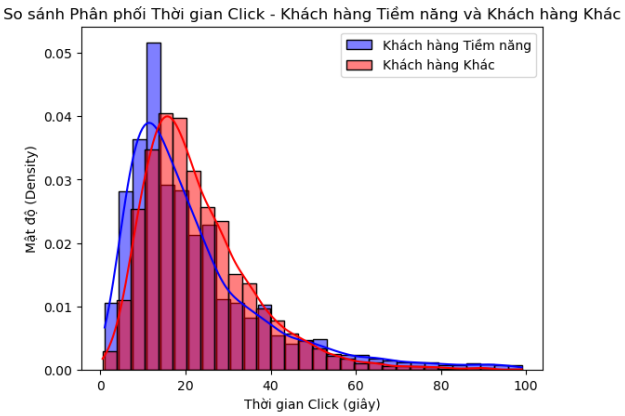


Hình 9 Phân phối thời gian click - khách hàng tiềm năng



Hình 10 Phân phối thời gian click của khách hàng - tiềm năng - chanpion

Sau khi lấy tập khách hàng thoải mãn hai điều kiện trên, chúng tôi thu được tập khách hàng tạm thời. Tập khách hàng này được ánh xạ qua top 7 loại campain_id và campain_type phổ biến của phân khúc Champion và sử dụng KL Divergence để đo lường sự khác biệt giữa hai phân phối xác xuất về thời gian mở thông báo đến thời gian mở liên kết mua hàng của hai phân khúc này. Qua đó lọc ra những khách hàng có hành vi tương tự với nhóm Champion làm phân khúc khách hàng Potential.



Hình 0911 Phân phối thời gian click của khách hàng gold và khách hàng tiềm năng sau khi áp dụng KL Divergence

Hình 09 Phân phối xác xuất về thời gian mở thông báo đến thời gian mở liên kết sau khi sử dụng KL Divergence để tính sự tương đồng.

4.PHÂN TÍCH THĂM DÒ/SƠ BỘ

Campaign Overview (Tổng quan về chiến dịch):

Phần này cung cấp phân tích chi tiết về loại chiến dịch, kênh sử dụng và chủ đề của các chiến dịch tiếp thị. Các chiến dịch **bulk** chiếm tỷ trọng lớn nhất với 1.8K chiến dịch, cho thấy doanh nghiệp sử dụng nhiều chiến lược gửi thông điệp hàng loạt để tiếp cận khách hàng. Tuy nhiên, các chiến dịch **trigger**, mặc dù có số lượng ít hơn, lại thể hiện hiệu quả cao hơn trong việc cá nhân hóa và tự động hóa. Kênh **email** chiếm ưu thế, với **mobile push** đứng thứ hai, phản ánh tầm quan trọng của email trong tiếp cận khách hàng, đồng thời nhấn mạnh rằng thiết bị di động cũng đang đóng vai trò quan trọng. Về chủ đề, các chiến dịch liên quan đến khuyến mãi, như "Sale out" (779 chiến dịch), chiếm tỷ lệ cao nhất, tiếp theo là các chủ đề mang tính cá nhân hóa, như "Happy birthday" (112 chiến dịch). Điều này nhấn mạnh rằng các thông điệp gắn liền với ưu đãi và cảm xúc cá nhân hóa thường có sức hút cao hơn.

Client First Purchased (Lần mua đầu tiên của khách hàng)

Phân tích này cho thấy rõ xu hướng hành vi mua hàng lần đầu của khách hàng. Dữ liệu chỉ ra rằng số lượng khách hàng mua lần đầu tăng mạnh vào các tháng cuối năm, đặc biệt là tháng 12, có thể liên quan đến hiệu quả của các chiến dịch khuyến mãi cuối năm. Các tháng thấp điểm, như tháng 1 và tháng 7, có số lượng mua lần đầu thấp hơn, cho thấy cần triển khai các chiến dịch bổ sung trong các giai đoạn này để cải thiện doanh thu. Ngoài ra, tỷ lệ khách hàng mua hàng sau khi nhấp chuột vào thông điệp email cao hơn ở các chiến dịch **trigger** và **transactional**, chứng minh rằng các chiến dịch này hiệu quả hơn trong việc dẫn dắt hành động mua hàng.

Message Overview (Tổng quan về thông điệp)

Phân tích này tập trung vào hiệu quả của các thông điệp được gửi qua các loại chiến dịch và kênh khác nhau. Thông điệp từ các chiến dịch **bulk** có số lượng gửi cao nhất với 7.1 triệu, trong khi **trigger** chỉ chiếm 2.1 triệu nhưng cho thấy hiệu quả cao hơn nhờ tính cá nhân hóa và tự động hóa. **Email** tiếp tục là kênh truyền tải thông điệp chính, với **mobile push** là kênh bổ trợ quan trọng, giúp doanh nghiệp tiếp cận khách hàng nhanh chóng qua thiết bị di động. Ngoài ra, tỷ lệ click theo chủ đề chiến dịch nhấn mạnh vai trò của các thông điệp ưu đãi, như "Sale out" và "Offer after purchase," trong việc thúc đẩy sự tương tác của khách hàng.

Kết luận:

Dashboard này cung cấp cái nhìn sâu sắc và toàn diện về hiệu quả chiến dịch và hành vi của khách hàng. Trong **Campaign Overview**, các chiến dịch bulk và email được sử dụng nhiều nhất, nhưng trigger và mobile push lại mang lại hiệu quả cao hơn nhờ tính tự động hóa và khả năng cá nhân hóa. Các chủ đề như "Sale out" và "Offer after purchase" có sức hút lớn, cho thấy khách hàng quan tâm nhiều nhất đến các chương trình ưu đãi. Phân tích về **Client First Purchased** chỉ ra rằng khách hàng thường mua lần đầu vào các tháng cuối năm, đặc biệt là tháng 12, nhấn mạnh vai trò quan trọng của các chiến dịch cuối năm trong việc tăng doanh thu. Các chiến dịch transactional và trigger cũng thể hiện hiệu quả cao trong việc thúc đẩy khách hàng nhấp chuột và mua hàng. **Message Overview** tiếp tục khẳng định email là kênh hiệu quả nhất, nhưng mobile push đang dần trở thành một công cụ bổ sung quan trọng để tiếp cận khách hàng nhanh chóng.

Dựa trên các phân tích, doanh nghiệp có thể tối ưu hóa chiến lược bằng cách: đầu tư nhiều hơn vào các chiến dịch cuối năm với nội dung ưu đãi hấp dẫn, tăng cường cá nhân hóa thông điệp, và tận dụng kênh mobile push để tiếp cận khách hàng hiệu quả hơn. Đồng thời, việc tối ưu hóa nội dung dòng tiêu đề và lời kêu gọi hành động (CTA) có thể giúp tăng tỷ lệ mở email, click, và chuyển đổi thành mua hàng. Nhìn chung, dashboard này không chỉ cung cấp thông tin chi tiết mà còn đưa ra những gợi ý chiến lược để doanh nghiệp nâng cao hiệu quả tiếp thị và trải nghiệm khách hàng.

5.TRIỂN KHAI MÔ HÌNH VÀ KẾT QUẢ:

5.1. Huấn luyện mô hình máy học:

Chúng tôi áp dụng mô hình học máy để dự đoán hành vi mua hàng của khách hàng, với biến mục tiêu là is_purchased. Các biến đầu vào được xử lý qua kỹ thuật feature engineering, trong khi các vấn đề mất cân bằng dữ liệu được điều chỉnh bằng undersampling. Để tối ưu mô hình, chúng tôi sử dụng grid search và xác định các biến quan trọng nhất.

5.2. Thang đo đánh giá:

Chúng tôi sử dụng thang đo accuracy để ung cấp cái nhìn tổng quát về hiệu suất mô hình.

Mô hình Logistic Regression với tham số tối ưu (C: 0.001, max_iter: 100, penalty: ‘l2’, solver: ‘liblinear’) đạt độ chính xác 80.32% trên tập kiểm tra. Các thước đo khác bao gồm F1-Score: 0.149, ROC-AUC: 0.585, và MSE: 0.197. Các biến đầu vào quan trọng được lựa chọn bao gồm các đặc trưng về nội dung, thời gian gửi email, hành vi mở/click trước đó, và mã hóa chủ đề, kênh.

6.KẾT QUẢ PHÂN TÍCH

Sau quá trình tiền xử lý, dữ liệu ban đầu đã được cải thiện đáng kể về chất lượng. Các giá trị null, nhiễu và không hợp lý đã được loại bỏ, giúp đảm bảo tính chính xác và đáng tin cậy khi đưa vào phân tích và trực quan hóa.

Kết quả phân tích cho thấy hiệu quả của các chiến dịch marketing vẫn còn nhiều hạn chế. Tỷ lệ người mua hàng sau các chiến dịch chỉ đạt mức thấp (1.5%). Đáng chú ý, ngay cả nhóm chiến dịch xuất hiện với tần suất cao cũng không tạo ra sự gia tăng đáng kể về số lượng khách hàng mua hàng.

Nghiên cứu đã thành công trong việc phân loại khách hàng thành ba phân khúc dựa trên hành vi mua sắm. Đặc biệt, nhóm khách hàng tiềm năng (potential) đã được xác định thông qua việc ánh xạ với phân khúc khách hàng "champion". Điều này giúp phát hiện ra các khách hàng có điểm chung về các chiến dịch và phân phối xác suất liên quan đến thời gian mua hàng, mở ra cơ hội tối ưu hóa chiến lược kinh doanh trong tương lai.

7. KẾT LUẬN

Bài nghiên cứu đã ứng dụng các kỹ thuật tiền xử lý dữ liệu, phân tích và trực quan hóa để đánh giá hiệu quả của các chiến dịch marketing trong bộ dữ liệu E-commerce Multichannel Direct Messaging. Đồng thời, nghiên cứu cũng thực hiện phân khúc khách hàng dựa trên thói quen, nhu cầu và hành vi thông qua việc sử dụng thuật toán KL Divergence để phân tích hành vi người dùng.

Tuy nhiên, do hạn chế về thời gian và nguồn lực, nghiên cứu chưa khai thác toàn diện các khía cạnh tiềm năng của bộ dữ liệu. Cụ thể, các yếu tố quan trọng như thời gian mua hàng lần đầu hay tần suất mua hàng vẫn chưa được phân tích sâu. Đây sẽ là những

hướng phát triển quan trọng trong tương lai để tối ưu hóa giá trị của bộ dữ liệu này và nâng cao hiệu quả chiến lược marketing.

TÀI LIỆU THAM KHẢO

[1] Dataset: E-commerce multichannel direct messaging 2021-2023. Link: https://www.kaggle.com/datasets/mkechinov/direct-messaging/data (01/12/2024)

[2] Tên pandas documentation — pandas 1.5.2 documentation (pydata.org). Link: https://pandas.pydata.org/docs/. Link: https://pandas.pydata.org/docs/ (01/12/2024)

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Trần Tuyết Minh	Visualization, Customer Segmentation, Preprocessing, viết model
2	Nguyễn Tấn Dũng	Viết docs, làm slide, Preprocessing, Xây dựng dashboard PowerBI
3	Nguyễn Minh Duy	Viết docs , Xây dựng dashboard PowerBI, Preprocessing, Viết model