DATA ANALYTICS REPORT

Team: LOICHOI

MỤC LỤC

I. XỬ LÝ DỮ LIỆU

1.1. Tổng quan tập dữ liệu

- Bộ dữ liệu chứa thông tin tổng hợp của khách hàng từ 1 sàn thương mại điện tử trong 6 tháng từ (…-…), gồm: Danh sách các gian hàng, Thông tin người dùng, Lịch sử hàng vi người dùng theo thời gian.
- Tập dữ liệu bao gồm 4 bộ dữ liệu:
  - user_info: thông tin cá nhân của người dùng (424171 hàng x 7 cột)
  - user_log: lịch sử hành vi của người dùng (54925330 hàng x 7 cột)
  - test: (52696 hàng x 3 cột)
  - train: (210778 hàng x 3 cột)

2.2. Làm sạch và chuyển đổi dữ liệu

1.2.1. Missing values

- Bảng user_info: có tổng số lượng và tỉ lệ chiếm của giá trị null này trong lần lượt từng cột: age - 95367 (chiếm 22% cột); sex – 5518 (chiếm 1% cột); phone – 21208 (chiếm 5% cột); job – 21208 (chiếm 5% cột); carrier – 21208 (chiếm 5% cột); marital_status – 357 (chiếm 0.008% cột).

→ Chuyển các missing values thành giá trị median (numerical), mode (category)

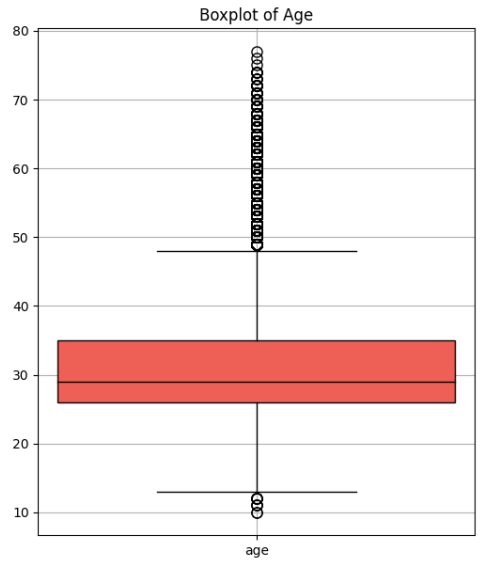
- Bảng user_log: tại cột brand_id có 91015 giá trị null (chiếm 0.0017% cột).

→ Dùng Random Forest để dự đoán các giá trị missing.

- Bảng train: user_id 5590 missing, merchant_id 5498 missing → drop tất cả hàng có missing và duplicates.
- Bảng test: không missing → không cần xử lý.

user_info table

	Variable	Missing Count	Missing Ratio (%)
0	user_id	0	0.0000
1	age	95367	22.4832
2	sex	5518	1.3009
3	phone	21208	4.9999
4	job	21208	4.9999
5	carrier	21208	4.9999
6	marital_status	357	0.0842



1.2.2. Outliers

- Bảng user_info: Tại cột 'age':
  - Nhiều giá trị lỗi như -1, 0, 150, 999 → chuyển các giá trị này thành giá trị 'unknown'.
  - Dùng phân bố quartile trên các giá trị khác 'unknown' thu được p25 = 26, p50 = 29, p75 = 35, min = 10, max = 77, IQR = 9.

→ Các outlier của cột 'age' được xác định là < Lower whisker = 13 và > Upper whisker = 48 được giữ nguyên.

→ Các outlier sẽ được tách riêng để phân tích nhằm tìm hiểu đặc điểm của nhóm cá biệt này.

1.2.3. Normalization

- Bảng user_info:
  - Cột 'sex': Chứa nhiều giá trị không đồng nhất về ngôn ngữ, cách ký hiệu và mã hóa Unicode bất thường

→ chuẩn hóa các giá trị thành 'female', 'male' bằng công cụ unicodedata và unidecode.

  - Cột 'job': Chứa các text không dấu, có dấu và mã hóa Unicode bất thường.

→ chuẩn hóa các giá trị thành không dấu và dạng lowercase bằng công cụ unicodedata, unidecode và lower().

  - Cột 'marital_status': Chứa nhiều giá trị không đồng nhất về ngôn ngữ.

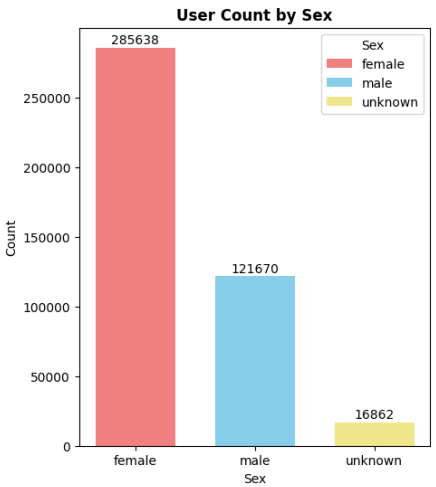
→ chuẩn hóa các giá trị thành 'married', 'single', 'divorced'.

I. XỬ LÝ DỮ LIỆU

1.2.4. Transformation

- Bảng user_info:
  - Chuyển cột 'user_id' thành dạng category.
  - Thêm cột job_sectors để phân 53 công việc ở cột job thành 9 nhóm ngành.
- Bảng user_log:
  - Chuyển các cột 'user_id', 'cat_id', 'brand_id', 'merchant_id', 'action' thành dạng category.
  - Chuyển cột 'datetime' thành dạng datetime.

II. KHÁM PHÁ VÀ GIẢI THÍCH DỮ LIỆU



2.1. Phân tích đơn biến (Univariate analysis)

2.1.1. Giới tính (sex)

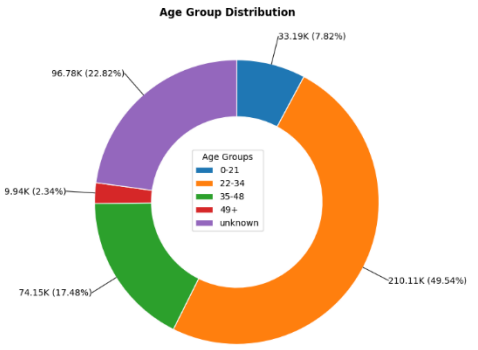
- Nữ chiếm tỷ lệ cao nhất (67.34%), tiếp theo là nam (28.68%).
- Còn lại là unknown (không rõ) chiếm tỉ lệ khá nhỏ (3.98%).

Insight: Nghiên cứu của Nguyễn Anh Tuấn (2025) cũng cho thấy nữ mua sắm trực tuyến nhiều và thường xuyên hơn nam, đặc biệt ở mức độ "mua hàng ngày".

Phụ nữ hiện nay đang là nhóm khách hàng chủ lực trên các sàn thương mại điện tử. Trong khi đó, nam giới thường có xu hướng mua sắm ít thường xuyên hơn.

2.1.2. Tuổi (age)

- Với hơn 420.000 khách hàng:
  - Độ tuổi 22-34 đang là nhóm đối tượng chính của thị trường thương mại điện tử, chiếm hơn 60% người dùng.
  - Các nhóm tuổi tiếp theo lần lượt là 35-48 tuổi (22.6%) và 0-21 tuổi (10.1%),
  - Trong khi nhóm trên 49 tuổi có tỷ lệ sử dụng thấp nhất.

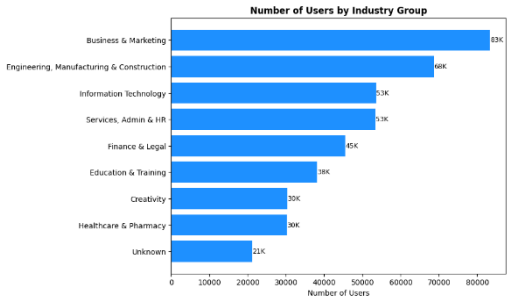


Insight: Nhóm 22-34 tuổi là những người thuộc GenZ và cuối GenY. Là thế hệ lớn lên cùng với Internet và các thiết bị công nghệ thông tin, "GenZ" còn được mệnh danh là những "công dân thời đại kỹ thuật số" (theo Mỹ Hạnh, 2023). Họ nổi bật nhờ sự am hiểu công nghệ, thu nhập ổn định và nhu cầu về sự tiện lợi trong mua sắm.

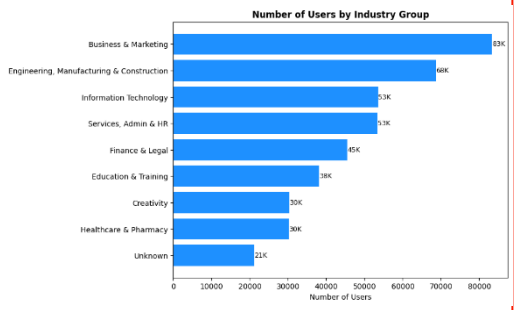
Ngược lại, nhóm 35-48 tuổi dù độc lập tài chính nhưng vẫn còn e ngại công nghệ, nhóm 0-21 tuổi bị hạn chế bởi sự phụ thuộc tài chính, còn người trên 49 tuổi có xu hướng duy trì thói quen mua sắm truyền thống.

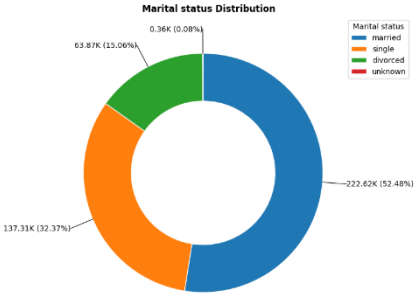
2.1.3. Công việc (job)

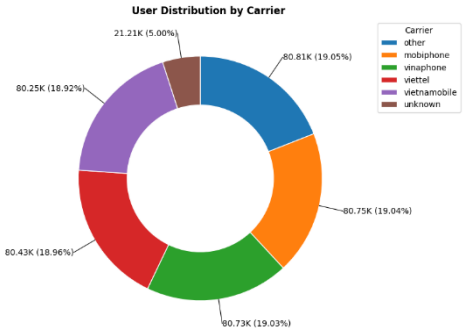
- Sự phân bổ người dùng thương mại điện tử rất rõ rệt giữa các nhóm ngành.
  - Ngành Kinh doanh & Marketing dẫn đầu với 83.000 người dùng, tiếp theo là Kỹ thuật, Sản xuất & Xây dựng với 69.000 người.
    Các ngành CNTT; Dịch vụ, Hành chính & Nhân sự và Tài chính & Pháp lý có số lượng người dùng từ 46.000 đến 54.000, tạo thành nhóm giữa.

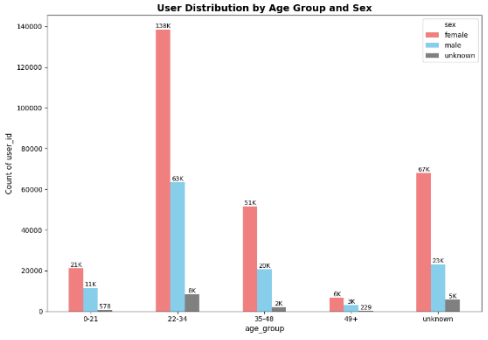


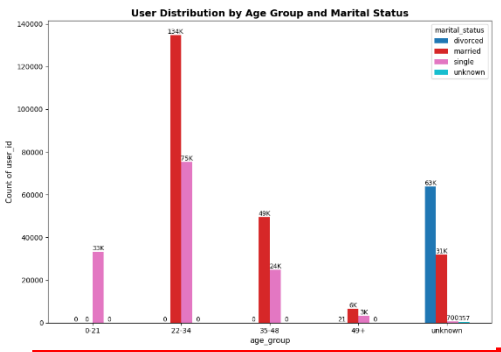
II. KHÁM PHÁ VÀ GIẢI THÍCH DỮ LIỆU











- Các ngành Giáo dục & Đào tạo; Sáng tạo; Y tế & Dược có số lượng người dùng thấp hơn, khoảng 30.000 - 38.000, với một phần không xác định chiếm 21.000 người.

Insight: Những nhóm ngành có tính chất công việc năng động, gắn liền với xu hướng thị trường và có mức độ tiếp xúc công nghệ cao (Kinh doanh & Marketing) hoặc nhóm có nhu cầu mua sắm chuyên biệt (Kỹ thuật, Sản xuất & Xây dựng) đang là những người dùng TMĐT tích cực nhất. Ngược lại, các ngành đòi hỏi sự cẩn trọng cao (Tài chính & Pháp lý) hoặc có tính chất truyền thống hơn (Y tế & Dược, Giáo dục) cho thấy xu hướng sử dụng TMĐT ít hơn, một phần do đặc thù ngành, thói quen hoặc mức độ tiếp cận công nghệ chưa cao bằng các nhóm dẫn đầu.

2.1.4. Tình trạng kết hôn (marital_status)

- 52.48% người sử dụng nền tảng TMĐT này là người đã kết hôn, 32.37% là những người còn độc thân, 15.06% người đã ly hôn và còn lại là unknown.

Insight: Nhóm người đã kết hôn đang là lực lượng chính thúc đẩy thị trường thương mại điện tử. Trong khi đó, người độc thân cũng là phân khúc quan trọng. Người đã ly hôn dù chiếm tỷ lệ nhỏ hơn nhưng vẫn là đối tượng tiềm năng.

2.1.5. Nhà mạng đang sử dụng (carrier)

- Các nền tảng điện thoại họ sử dụng chính là mobiphone, vinaphone, viettel và vietnamobile gần như bằng nhau sấp sỉ 19% thị phần cho mỗi nhà mạng.

Insight: Phản ánh người dùng của sàn TMĐT này không bị chi phối bởi một nhà mạng cụ thể nào.

2.2. Phân tích đa biến (Multivariate analysis)

a) Đặc điểm về khách hàng

2.2.1. Age x Sex

- Nhóm tuổi 13-21 gần như có số lượng user ít nhất trong khoảng độ tuổi được xét đến (13-48 tuổi), nhóm này có lượng user nữ gần gấp đôi số user nam.
- Nhóm từ 22-34 là nhóm user lớn của sàn thương mại điện tử này với 60% user như đã phân tích ở trên. Ta có thể thấy user nữ ở độ tuổi này nhiều gấp 2,2 user nam, có 8K người ở nhóm này chưa rõ về giới tính do missing value.
- Nhóm 35-48, nhóm này có lượng user cao thứ 2 trong khoảng độ tuổi được xét đến và có số user nữ cao gấp 2,4 lần user nam.

Insight: Nhóm nữ giới 22-34, nhóm nam 22-34 và nhóm user nữ 35-40 là nhóm người dùng hoạt động "chủ đạo" trên nền tảng này. Kế đến là nhóm 13-21, nhóm nam 35-40 và 8K unknown user 22-34 cũng là nhóm user tiềm năng.

2.2.2. Age x Marital Status

- Nhóm tuổi 13-21 là những người độc thân với hơn 33K **người.**
- **N**hóm 22-34 là nhóm có lượng người dùng nhiều nhất của sàn.

Họ bắt đầu lập gia đình, nhóm người lập gia đình ở nhóm tuổi này cũng cao nhất với hơn 135K người.

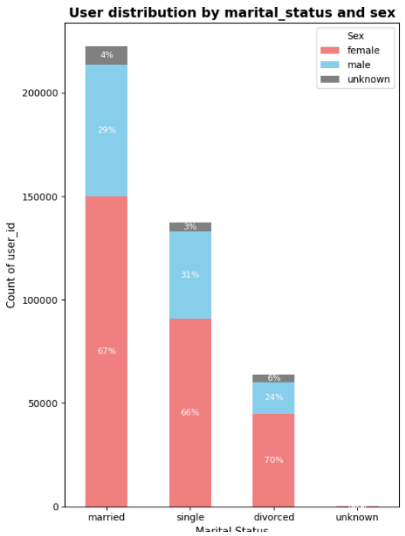
- nhưng số người lại ít hơn với 49K người và vẫn lượng người độc thân vẫn còn khá cao.
- Cả ba nhóm trên gần như ghi nhận không có người ly hôn. Những người dùng ở tình trạng ly hôn thường và hơn 32K người

II. KHÁM PHÁ VÀ GIẢI THÍCH DỮ LIỆU

đã kết hôn không cung cấp tuổi.

Insight: Nhóm tuổi 13-21 sẽ là nhóm khách hàng trung thành tiềm năng trong tương lai. nhóm 22-34 đang là nhóm tiềm năng phát triển nhất. những người độc thân của nhóm 25-48 vẫn là nhóm khách hàng tiềm năng về mua sắm.

2.2.3. Marital_status x Sex

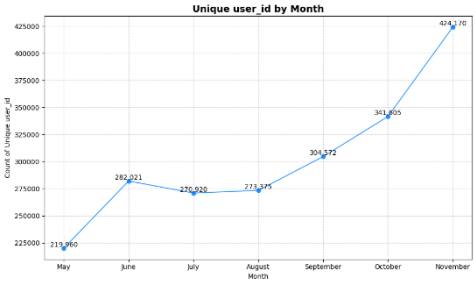


- Số người đã kết hôn chiếm tỷ trọng lớn nhất trong số các người dùng của sàn thương mại điện tử với hơn 222K người, người dùng nữ vẫn chiếm phần lớn trong những người đã kết hôn với 150K gấp hơn 2 lần so với nam giới đã kết hôn.
- Những người độc thân với hơn 137 người, nữ giới ở tình trạng này vẫn tiếp tục chiếm tỷ lệ lớn, gấp 2 lần so với nam. Cuối cùng là người đã ly hôn với 63K người, nữ giới cũng gấp 3 lần nam giới.

Insight: Người đã kết hôn có nhiều lựa chọn phong phú hơn, họ có một "đối tác" san sẻ chi phí (theo VNExpress, 2023) nên sẽ mua sắm và chi tiêu được nhiều hơn 2 nhóm đối tượng còn lại. Phát hiện của Công ty tư vấn Daxue Consulting cũng nhấn mạnh sức tiêu dùng của phụ nữ độc thân là động lực chính của nền kinh tế, khi họ có xu hướng chi tiêu nhiều tiền hơn cho bản thân (theo Báo Phụ Nữ, 2022) so với nam giới.

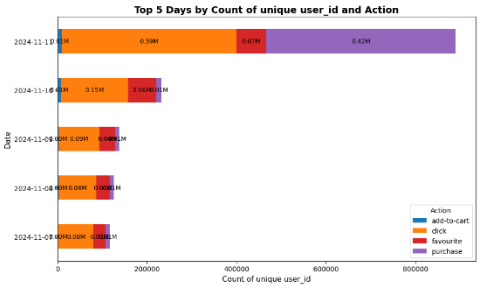
b) Đặc điểm hành vi

2.2.4. User_id x Datetime



- Lượng user tăng trưởng trong 6 tháng qua.
- Tháng 5 bắt đầu với lượng user thấp nhất (220K) và có xu hướng tăng dần về cuối năm.
- Giai đoạn tháng 6 đến tháng 8 có mức tăng trưởng nhẹ và dao động ổn định quanh 270K–282K user. Lượng user tăng mạnh đặc biệt từ tháng 9, tháng 10 (304K và 341K) đến tháng 11 (424K). Tăng đột biến nhất ở tháng 11 với 424K user trên sàn.

***Insight*:** Sàn TMĐT này được ghi nhận đột phá về số lượng user trong tháng 11.



- Top 5 ngày tương tác nhiều nhất ở sàn này từ ngày 07/11 đến ngày 11/11. Ngày đôi 11.11 là ngày được ghi nhận có số lượng tương tác nhiều nhất với 0.42M user mua hàng; 0.39M user click và xem các sản phẩm; 0.07M thêm vào giỏ hàng và cũng là ngày đc người dùng add to cart nhiều nhất.

Insight: Hầu hết user của sàn đều mua hàng vào ngày 11.11 - ngày các sàn TMĐT hiện nay tung các mã giảm giá cực sâu với user.

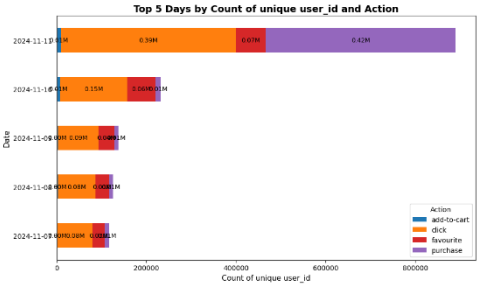
Họ tận dụng xu hướng mua sắm theo mùa một cách triệt để, để "hunt" được những "deal" hời mà không phải là người dùng thân thuộc từ các sàn.

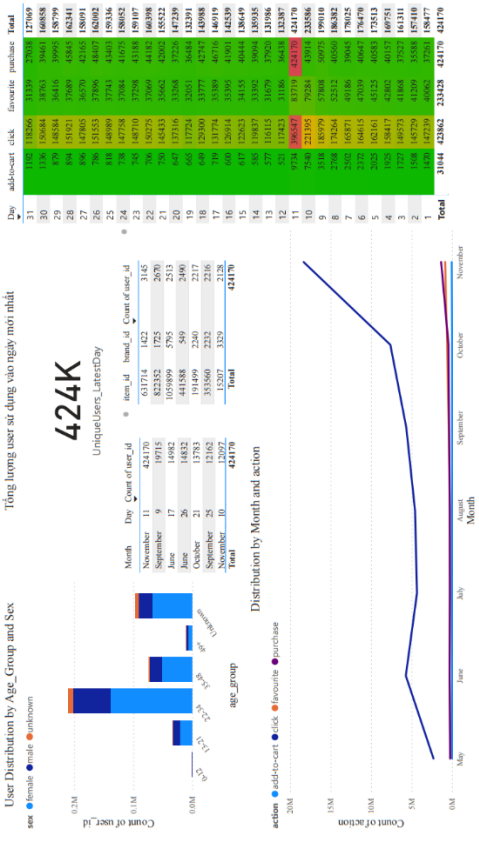
→ Cần có cách tiếp cận để xem user là deal hunter hay là potential loyal.

→ Nhóm đề xuất cách tiếp cận deal hunter bằng 2 cách sau:

- **Cách 1:** Biểu đồ phân phối Số ngày hoạt động theo người dùng trên một nền tảng nhằm xác định hành vi người dùng: Ai chỉ ghé qua 1-2 lần, ai trung thành và sử dụng thường xuyên. Cụ thể, có khoảng 116,361 người chỉ hoạt động đúng 1 ngày, và 109,451 người hoạt động 2 ngày. Số người dùng giảm mạnh theo số ngày hoạt động tăng. Chỉ còn khoảng 1 người duy trì hoạt động đến ngày thứ 170+. Đây là điển hình của phân phối Power Law, nơi phần lớn người dùng chỉ tương tác ngắn hạn, còn chỉ một số nhỏ duy trì lâu dài.

II. KHÁM PHÁ VÀ GIẢI THÍCH DỮ LIỆU





- **Cách xác định Deal hunter:** Trung bình 1 tháng sẽ có 3 lần khuyến mãi, các deal hunter trung bình thường cần 2-3 ngày liền kề để săn được deal hời (tìm kiếm trước cửa hàng trên nền tảng, add favourite, add to cart…)

→ 6 tháng qua trung bình các deal hunter sẽ vào nền tảng 54 ngày → Dùng ngưỡng 54 ngày để phát hiện ra user đó là deal hunter hay là potential loyal → Từ 0 đến ngày thứ 54 như biểu đồ ta thấy nền tảng hiện có 97.18% là deal hunter.

- **Cách 2:** Biểu đồ phân phối số người dùng theo số lần tương tác trung bình với một merchant. Đo lường sự gắn bó giữa người dùng và merchant. Khi người dùng tương tác nhiều có khả năng là khách hàng trung thành, quay lại mua nhiều lần. Còn đối với những người chỉ tương tác một vài lần trong 6 tháng thì có thể họ sẽ là deal hunter.
- Đa số người dùng chỉ có 2-4 lần tương tác trung bình với một merchant. Cụ thể, cao nhất là ở 2 lần (122,670 người) và 3 lần (120,370 người). Từ 4 trở đi, số người dùng giảm nhanh chóng (Power Law). Càng về sau (15, 20 lần...), số lượng người dùng tương tác trung bình cao trở nên rất hiếm.

→ Deal Hunter ở đây được nhóm xác định là những người có tương tác với cửa hàng nhỏ hơn 15 lần trong 6 tháng qua (Bởi có sự giảm mạnh và ít dần đi tương tác trên 15 lần của các user trên nền tảng này)

→ Tương tác dưới 15 lần trong 6 tháng qua như ở biểu đồ ta thấy nền tảng hiện có 95.6% là deal hunter con số cũng xấp xỉ bằng cách phân loại deal hunter bằng cách xem Số ngày hoạt động của user trong 6 tháng qua.

2.3. Dashboard

- Dashboard này tổng quan hóa hành vi người dùng sàn TMĐT, nhấn mạnh nhóm nữ 22–34 tuổi chiếm đa số — mục tiêu lý tưởng cho các chiến dịch thời trang, mỹ phẩm. Tháng 11 chứng kiến lượng người dùng mới và sản phẩm tăng vọt, gợi ý sức hút từ sự kiện 11/11. Dù lượt "add-to-cart" lên đến ~20M, "purchase" chỉ ~1M, cho thấy tỷ lệ chuyển đổi thấp. Dữ liệu gần nhất tiếp tục phản ánh tình trạng này, với hàng triệu lượt click nhưng giao dịch thực tế còn hạn chế. Nguyên nhân có thể đến từ phí ship hoặc quy trình thanh toán chưa tối ưu. Sàn cần cải thiện chuyển đổi và tận dụng các dịp cao điểm để thúc đẩy doanh thu.

2.4. Objective

- Trong bối cảnh thương mại điện tử phát triển mạnh mẽ tại Việt Nam, việc phân tích dữ liệu giúp doanh nghiệp xác định khách hàng có khả năng quay lại và tiềm năng trung thành để tối ưu hóa chi phí tiếp thị. Nghiên cứu cho thấy chất lượng dịch vụ ảnh hưởng trực tiếp đến lòng trung thành của khách hàng (Jescae, 2023). Bên cạnh đó, sự tương tác trên mạng xã hội cũng góp phần thúc đẩy sự gắn kết với thương hiệu (ResearchGate, 2024). Niềm tin và sự hài lòng được xem là yếu tố then chốt trong việc giữ chân khách hàng lâu dài (MPRA, 2023). Do đó, các mô hình phân tích hành vi khách hàng giúp doanh nghiệp xây dựng chiến lược hiệu quả trong thị trường cạnh tranh hiện nay.

PHỤ LỤC

Anh, T., Yen, T.T. and Trang, N.T.T. (2022) Impact of E-commerce service quality on customer loyalty: A case of Vietnam, Journal of Social Sciences and Management Studies. Available at: https://www.jescae.com/index.php/jssms/article/view/73 (Accessed: 26 May 2025).

Dang, S.-H. and Nguyen, L.-T. (2023) What drives customer loyalty in Social Commerce Sector? ... Available at: https://mpra.ub.uni-muenchen.de/119509/1/MPRA_paper_119509.pdf (Accessed: 26 May 2025).

Long, H.C.L., Tra, N.H.X. and Quan, P.N.A. (2024) Full article: Factors affecting customer engagement and brand loyalty in Vietnam FMCG: The moderation of artificial intelligence. Available at: https://www.tandfonline.com/doi/full/10.1080/23311975.2024.2428778 (Accessed: 26 May 2025).

Nguyễn Anh Tuấn (2025) Thực trạng mua Sắm Trực Tuyến Trong Thanh Niên, Tạp chí điện tử TRUNG ƯƠNG ĐOÀN TNCS HỒ CHÍ MINH. Available at: https://thanhnienviet.vn/thuc-trang-mua-sam-truc-tuyen-trong-thanh-nien-209250204105939831.htm (Accessed: 25 May 2025).

Báo Phụ Nữ (2022) Thời đại của những khách hàng độc Thân. Available at: https://www.phunuonline.com.vn/thoi-dai-cua-nhung-khach-hang-doc-than-a1465908.html (Accessed: 25 May 2025).

VnExpress (2023) Sống độc Thân tốn kém Hơn 5 triệu so Với Khi Lấy Chồng, vnexpress.net. Available at: https://vnexpress.net/song-doc-than-ton-kem-hon-5-trieu-so-voi-khi-lay-chong-4661648.html (Accessed: 25 May 2025).

Mỹ Hạnh (2023)‘Genz’ làm chủ công Nghệ Số, Báo An Giang Online. Available at: https://baoangiang.com.vn/-genz-lam-chu-cong-nghe-so-a354060.html (Accessed: 25 May 2025).