viettel

CUỘC THI PHÂN TÍCH DỮ LIỆU DAZONE 2025

Vòng 2.2. PRESENTATION

Team ‘LOICHOI’

Võ Trần Thị Mỹ Vân - Nguyễn Tuấn Huy - Trần Nguyễn Quang Huy

Mục lục

1 Xử lý dữ liệu

2 Tổng quan

3 Mô hình

4 Chân dung khách hàng

5 Đề xuất

Xử lý dữ liệu

Bộ dữ liệu gồm tập dữ liệu với 4 bảng: Bảng user_info, bảng user log, bảng train

Tên cột	Vấn đề	Cách xử lý
Bảng user_info (424171 quan sát) (Phụ lục 1)		
user_id	Không	Không
age	22% missing	Thay thế bằng median
sex	1% missing, chưa chuẩn hóa	Chuẩn hóa thành 'female', 'male', thay thế bằng mode
phone	5% missing	Thay thế bằng mode
job	5% missing, lỗi unicode	Chuẩn hóa unicode, thay thế bằng mode
carrier	5% missing	Thay thế bằng mode
marital_status	0.08% missing, chưa chuẩn hóa	Chuẩn hóa thành 'married', 'single', 'divorced', thay thế bằng mode

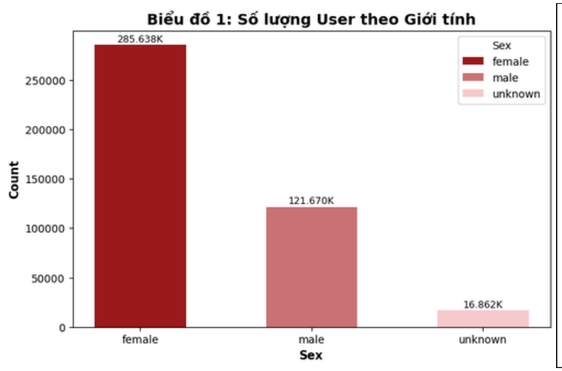
Tên cột	Vấn đề	Cách xử lý
Bảng user_log (54M quan sát)		
user_id item_id cat_id datetime action	Không	Không
brand_id	0.16% missing	Dùng Random Forest để thay thế các giá trị missing

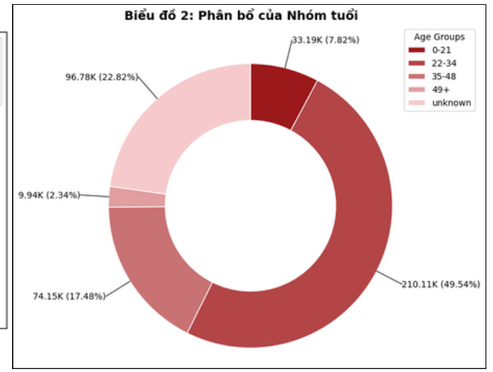
Tên cột	Vấn đề	Cách xử lý
Bảng train (233782 quan sát), test (52695 quan sát)		
user_id	Có missing	Drop hàng
merchant_id	Có missing	Drop hàng
label	Mất cân bằng tỉ lệ 93-7 (Phụ lục 2)	SMOTE, Threshold, Class weighting

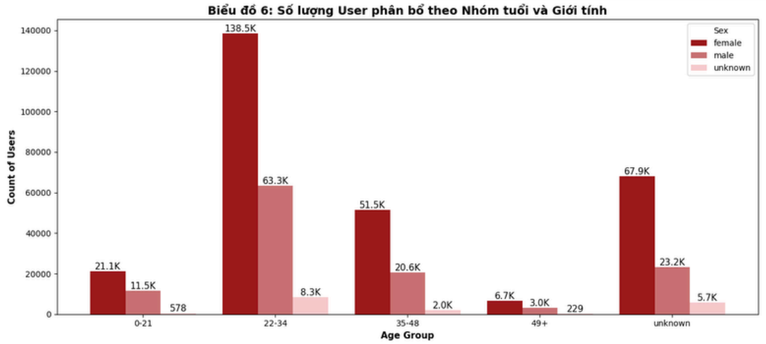
Sau khi xử lý từng bảng dữ liệu, tiến hành kết hợp các bảng thành bảng dữ liệu tổng hợp nhằm phục vụ cho các bước phân tích và mô hình hoá tiếp theo.

Phân tích nhân khẩu học

Theo tuổi và giới tính



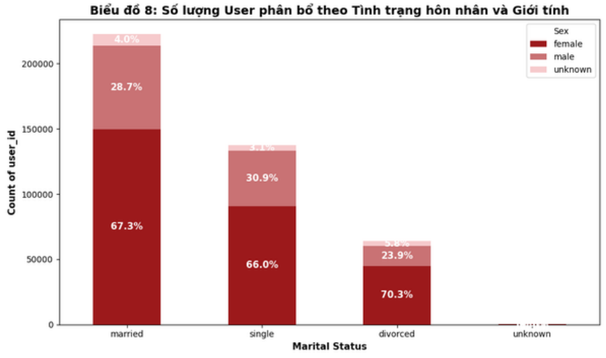


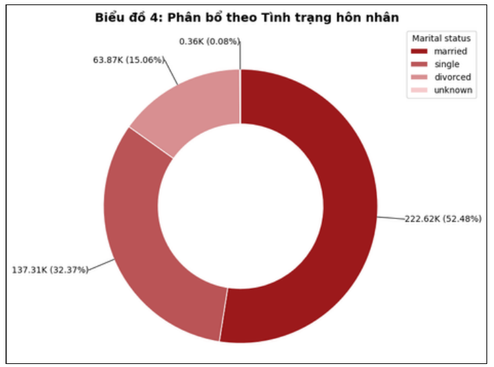


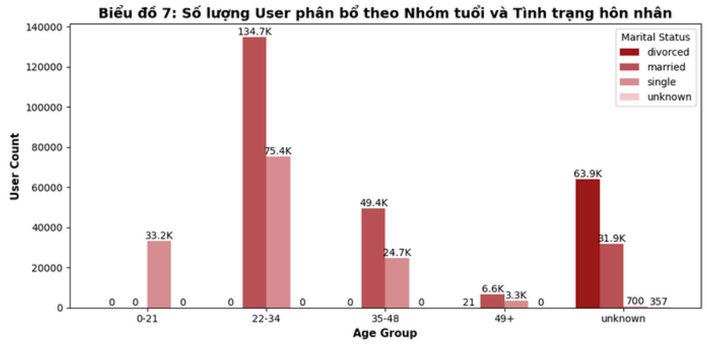
User từ **độ tuổi từ 22-34** chiếm gần **50%** user trên sàn TMĐT.

Nữ giới chiếm chủ yếu hơn **66%** trên tổng số user.
Phân bổ giới tính theo nhóm độ tuổi không có sự khác biệt

Theo tình trạng hôn nhân







Hơn **50% khách hàng đã kết hôn, 32% vẫn độc thân** và còn lại đã ly hôn.

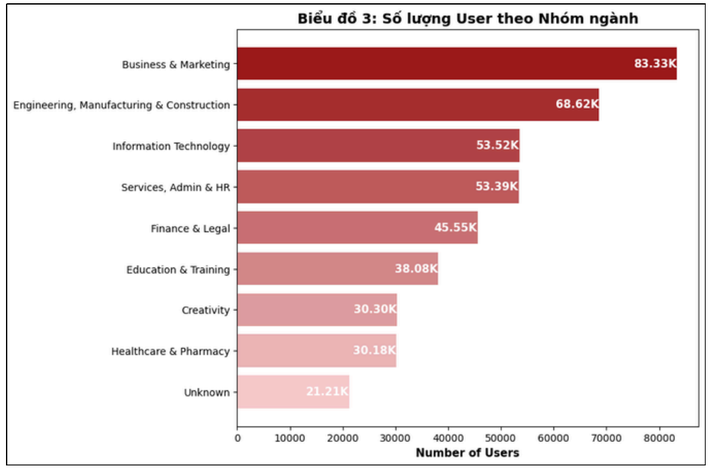
Nữ giới tiếp tục chiếm hơn ⅔ số người trong mỗi tình trạng hôn nhân.
Những người **độc thân xu hướng che giấu độ tuổi** khi mua sắm online.

User nữ thuộc thế hệ Gen Y và Gen Z (từ 22-34 tuổi) là nhóm người chủ yếu quan tâm mua sắm, tương tác nhiều trên sàn thương mại điện tử. Sàn TMĐT này đang phục vụ **> 80% là những người đã kết hôn và độc thân.**

Phân tích nhân khẩu học

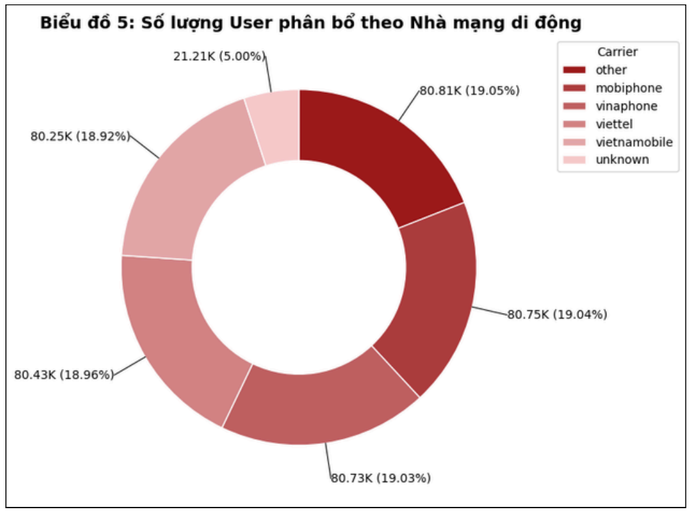
Theo nghề nghiệp

Khách hàng đến từ đa dạng ngành nghề, chủ yếu là **Kinh doanh và marketing, và nhóm ngành kĩ thuật khoa học, công nghệ thông tin**.



Theo nhà mạng

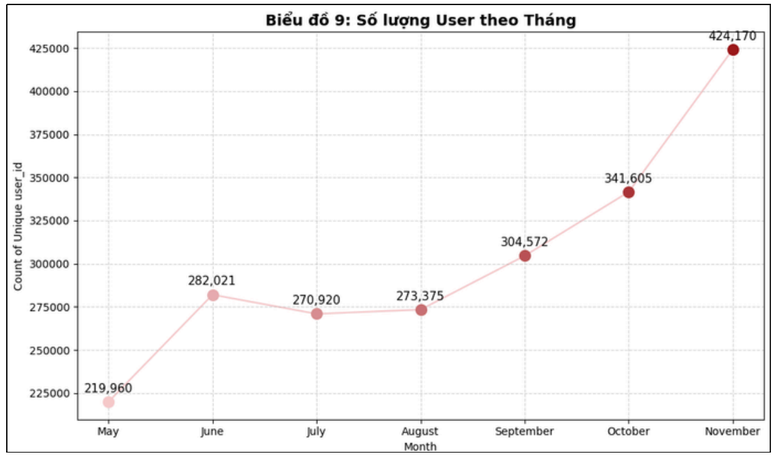
Khách hàng sử dụng các nhà mạng di động không có sự phân hóa rõ ràng.



Phân tích hành vi mua hàng

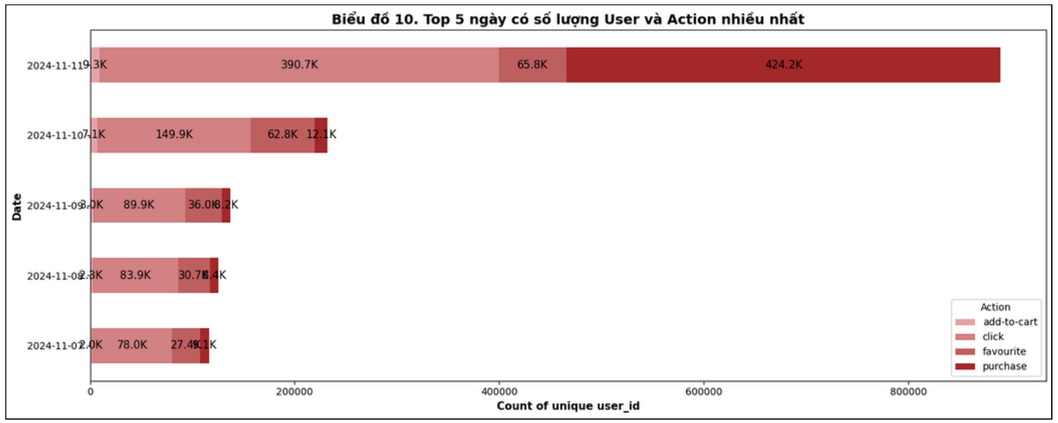
Tổng quan về số lượng user trong 6 tháng qua

Nhìn chung, **6 tháng** qua sàn TMĐT **có tăng trưởng về lượng user**. Sàn này được ghi nhận **đột phá** về số lượng **user trong tháng 11**.



Vậy tại sao lại có sự "đột phá" user trong tháng 11?

Xét **Top 5 ngày** có user **tương tác nhiều nhất** trong **6 tháng** qua, ta thấy tương tác nhiều nhất ngày **07.11 đến ngày 11.11**.



Đặc biệt, ghi nhận **tương tác đột biến ở ngày 11.11** với toàn bộ user đều mua hàng trong ngày này - thậm chí họ không click xem mà mua luôn sản phẩm.

Phân tích hành vi mua hàng

Cụ thể hành vi mua hàng của user trong 6 tháng qua...

Day	add-to-cart	click	favourite	purchase	Total
11	9734	396547	83719	424170	424170
10	7540	221395	79284	43914	233586
9	3518	185979	57808	50975	199010
8	2768	174264	52512	40560	186382
7	2502	165871	49186	39045	178025
6	2372	164615	47039	40647	176470
5	2025	162161	45125	40583	173513
4	1925	158417	42802	40157	169751
28	894	151921	37689	45845	162341
26	786	151553	37896	48407	162002
3	1727	149573	41868	37527	161311
30	1336	150684	38763	39461	160858
22	706	150275	37069	44182	160398
25	818	148989	37743	43403	159336
23	745	148710	37298	43188	159107
29	879	148584	36416	39995	158799
1	1470	147239	40062	37261	158477
27	896	147805	36570	42165	158091
24	738	147758	37084	41675	158052
2	1508	145729	41209	35588	157410
21	750	145433	35662	42002	155522
20	647	137316	33268	37226	147239
17	719	131774	35389	46716	146919
18	649	129300	33777	42747	143988
16	600	126914	35395	41901	142539
15	617	122623	34155	40444	138649
14	585	119837	33392	39094	135935
19	665	117724	32051	36484	132391
12	521	117423	31186	36438	132387
13	577	116115	31679	37920	131986
31	1192	118266	31339	27038	127069
Total	**31044**	**423862**	**233428**	**424170**	**424170**

Trong 1 tháng, **ngày 11** là ngày có lượng hoạt động của user nhiều nhất.

Top 5 ngày có lượng 'click' nhiều nhất

	Datetime	Action 'click'
1	Nov 11, 2024	9188104
2	Nov 10, 2024	2750192
3	Nov 9, 2024	1102507
4	Nov 8, 2024	879656
5	Nov 7, 2024	777904

Top 5 ngày có lượng 'add-to-cart' nhiều nhất

	Datetime	Action 'add-to-cart'
1	Nov 11, 2024	14725
2	Nov 10, 2024	11711
3	Nov 9, 2024	4825
4	Nov 8, 2024	3452
5	Nov 7, 2024	3068

Top 5 ngày có lượng 'favourite' nhiều nhất

	Datetime	Action 'favourite'
1	Nov 10, 2024	161674
2	Nov 11, 2024	156450
3	Nov 9, 2024	88548
4	Nov 8, 2024	73270
5	Nov 7, 2024	64054

Hành động click, add_to_cart và favourite có xu hướng gia tăng cực kỳ mạnh mẽ vào những ngày tháng 11 - cận kề ngày 11.11.

Top 5 ngày có lượng 'purchase' nhiều nhất

	Datetime	Action 'purchase'
1	Nov 11, 2024	1223354
2	Sep 9, 2024	30908
3	Jun 26, 2024	26548
4	Jun 17, 2024	20155
5	Oct 21, 2024	19067

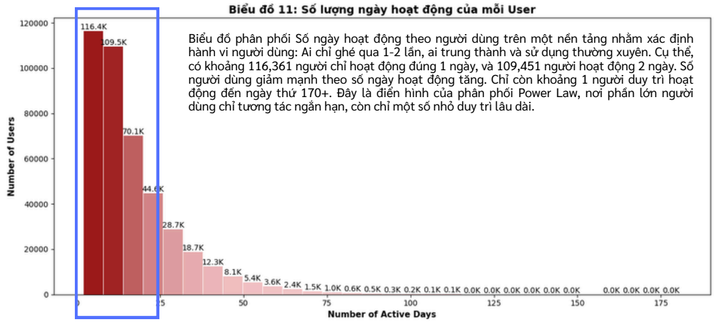
Như 1 thói quen hàng tháng, các sàn TMĐT sẽ có **3 ngày khuyến mãi** lớn trong tháng như **ngày đôi, ngày giữa tháng (15) và ngày lương về (25-28)**
→ tạo ra thói quen tiêu dùng định kỳ của khách hàng.

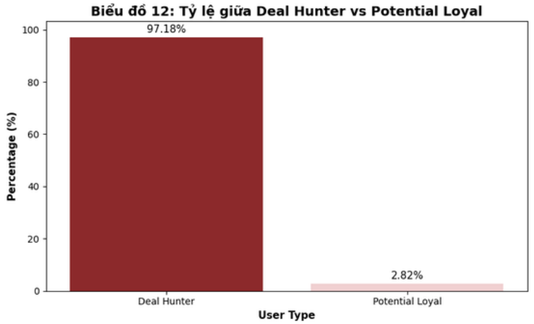
User của sàn tích cực chốt đơn vào 3 ngày sale lớn trong tháng như trên. Đặc biệt, họ quan tâm và mua rất nhiều vào ngày 11.11.

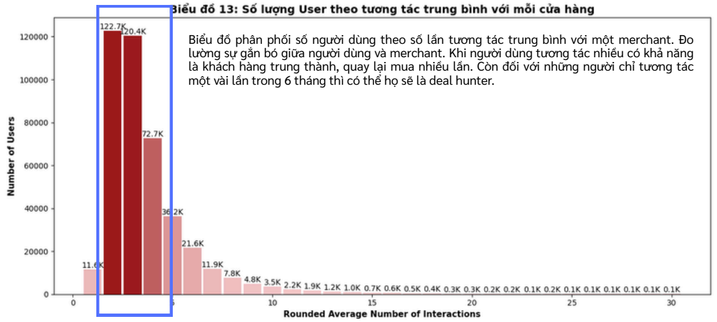
Hầu hết user của sàn đều mua hàng vào ngày 11.11 - ngày các sàn TMĐT hiện nay tung các mã giảm giá cực sâu với user. Họ tận dụng xu hướng mua sắm theo mùa một cách triệt để, **để "hunt" được những "deal" hời mà không phải là người dùng thân thuộc từ các sàn.**

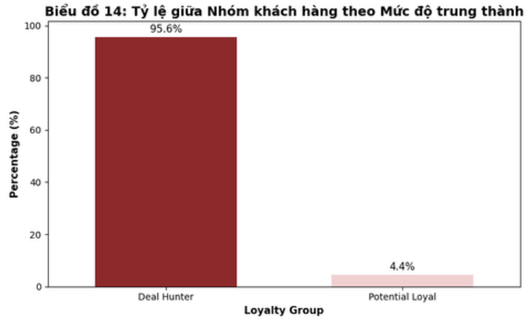
Phân tích hành vi mua hàng

Làm thế nào để sàn TMĐT này có thể tìm "săn" lại được các "Deal hunter"?









Quan điểm cá nhân ban đầu

Cách tiếp cận 1:

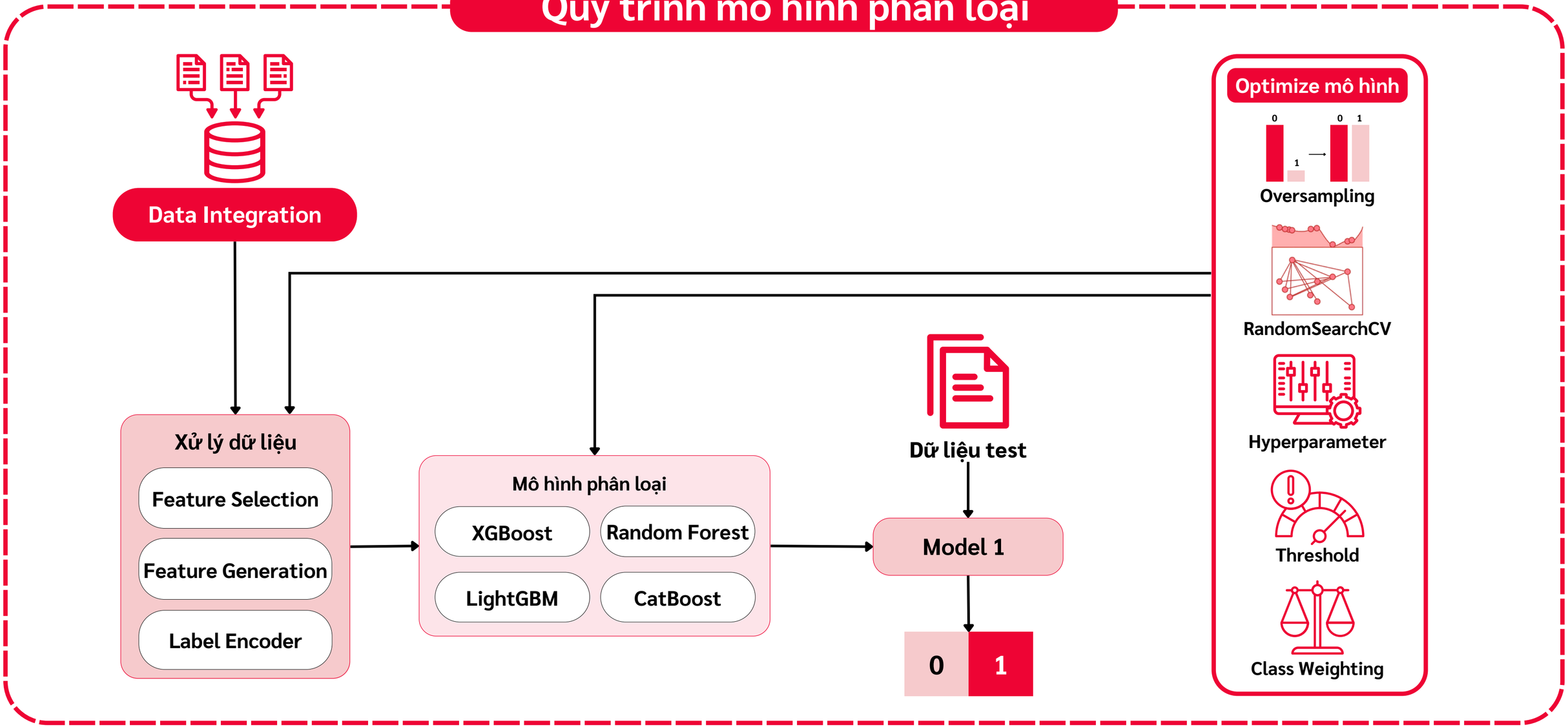
Trung bình 1 tháng có **3 ngày khuyến mãi**, các Deal hunter trung bình cần **2-3 ngày liên tiếp để săn được deal** hời (tìm kiếm trước cửa hàng trên nền tảng, add favourite, add to cart...)
→ Trong 6 tháng, **trung bình Deal hunter vào nền tảng 54 ngày.**
→ Lấy ngưỡng này (54 ngày) vào Active days của mỗi User để phát hiện ra User nào là Deal hunter hay Potential loyal.
→ Từ 0 đến ngày thứ 54 (Biểu đồ 12), ta thấy nền tảng **có 97.18% là Deal hunter.**

Cách tiếp cận 2:

Phần lớn, User có **2-4 lần tương tác trung bình với 1 merchant** (Biểu đồ 13). Từ 4 trở đi, số User giảm nhanh chóng (Power Law). Càng về sau (15, 20 lần...), số User tương tác trung bình cao trở nên rất hiếm.
→ Deal Hunter có thể là những người có **tương tác với cửa hàng dưới 15 lần** trong 6 tháng.
→ Với ngưỡng này (Biểu đồ 14), nền tảng hiện **có 95.6% là Deal hunter.**

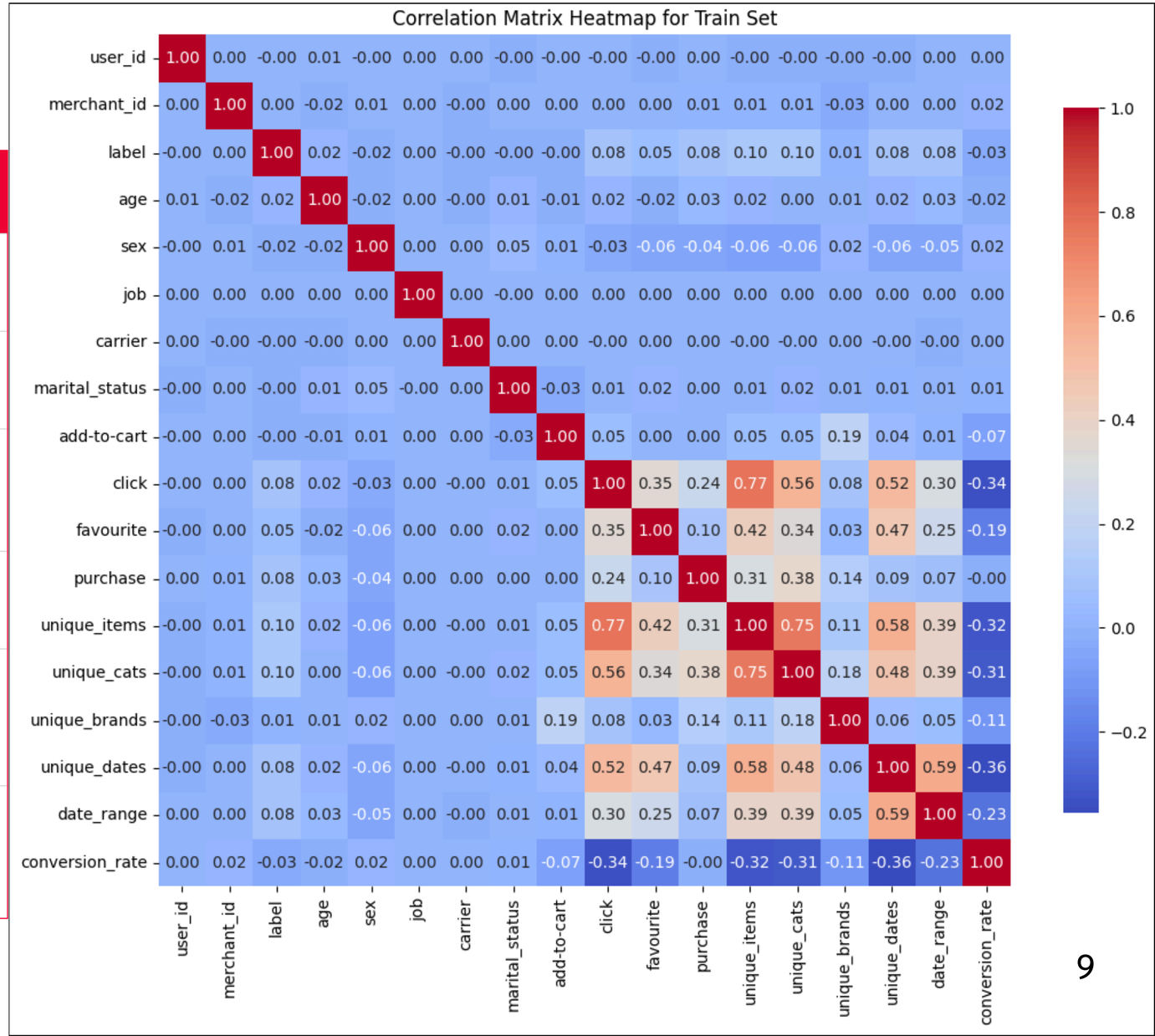
Hai phương thức xác định Deal hunter chỉ ra rằng: có > 95% user hiện đang là deal hunter trong 6 tháng qua

Quy trình mô hình phân loại



Feature Generation

Tên cột	Mô tả	Tên cột	Mô tả
unique_items	Số món hàng tương tác	date_range	Khoảng thời gian tương tác
unique_cats	Số loại hàng tương tác	add-to-cart	Số lần thực hiện add-to-cart
unique_brands	Số nhãn hàng tương tác	click	Số lần thực hiện click
unique_merchants	Số cửa hàng tương tác	favourite	Số lần thực hiện favourite
recency_days	Số ngày kể từ lần tương tác cuối cùng	purchase	Số lần thực hiện purchase
active_days	Số ngày tương tác	conversation_rate	Tỉ lệ mua hàng so với tổng hành vi



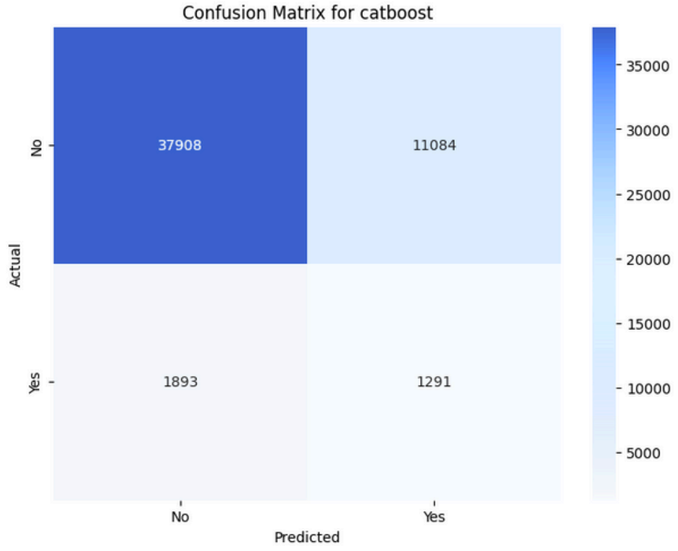
Hiệu suất mô hình phân loại dự đoán với từng phương pháp

SMOTE APPROACH			
	FALSE NEGATIVE RATE	F1 SCORE	ACCURACY
CATBOOST	0.64	0.15	0.75
XGBOOST	0.8	0.12	0.83
LIGHTGBM	0.83	0.11	0.84
RANDOM FOREST	0.84	0.1	0.83

THRESHOLD (=0.1) APPROACH			
	FALSE NEGATIVE RATE	F1 SCORE	ACCURACY
RANDOM FOREST	0.73	0.15	0.81
CATBOOST	0.79	0.16	0.87
XGBOOST	0.81	0.15	0.87
LIGHTGBM	0.83	0.14	0.83

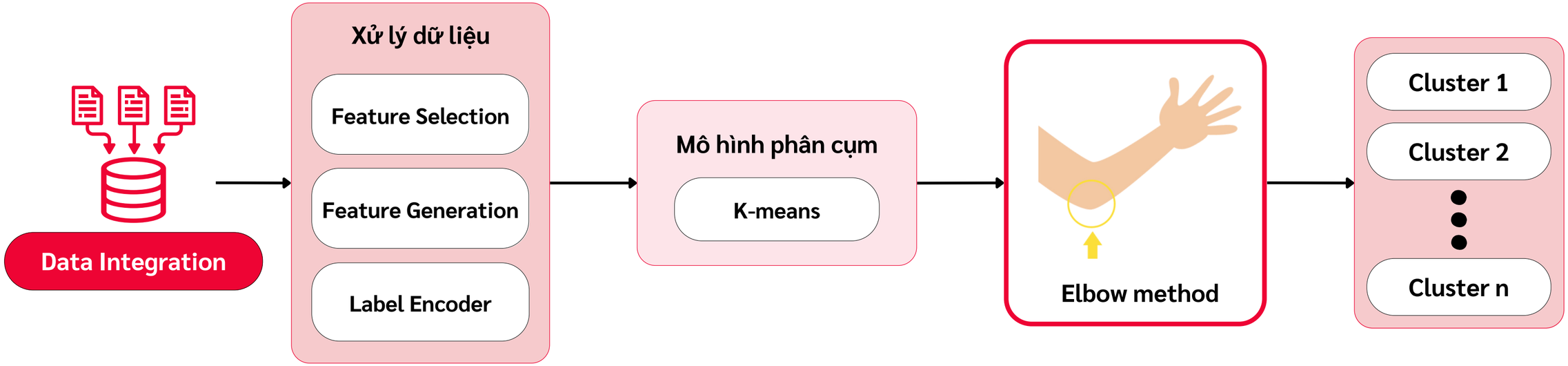
CLASS WEIGHTING APPROACH			
	FALSE NEGATIVE RATE	F1 SCORE	ACCURACY
CATBOOST	0.59	0.17	0.75
RANDOM FOREST	0.61	0.16	0.76
LIGHTGBM	0.61	0.17	0.76
XGBOOST	0.64	0.17	0.79

Mô hình CatBoost với phương pháp Class Weighting đạt FNR thấp nhất và F1 score cao nhất, cho thấy **hiệu suất tốt trong việc tránh bỏ sót những khách hàng có khả năng quay lại** và đảm bảo độ tin cậy của dự đoán.

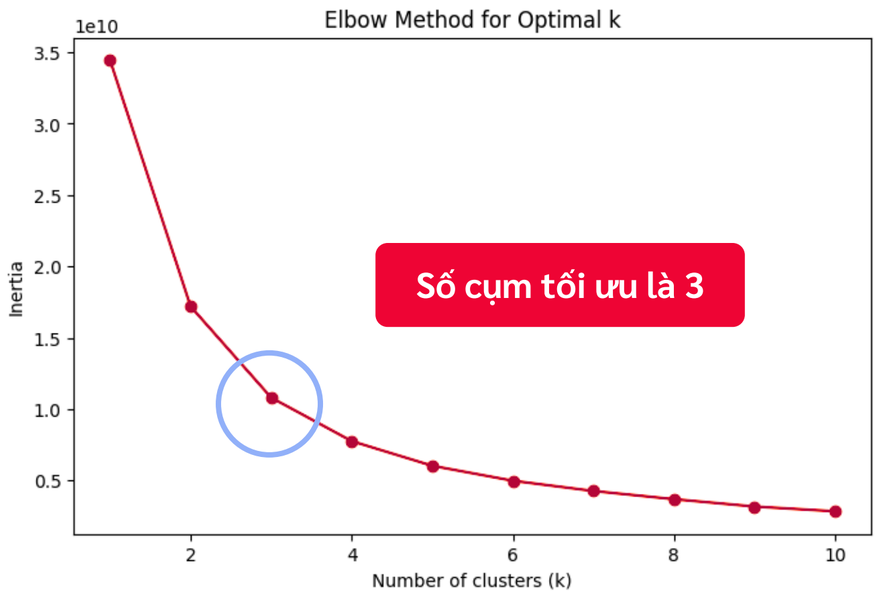


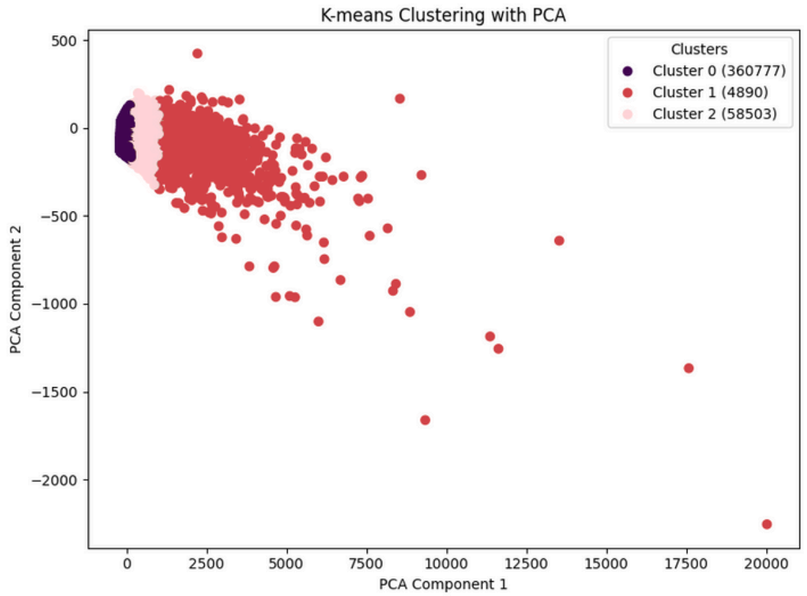
Hình: Ma trận nhầm lẫn của mô hình CatBoost với phương pháp Class weighting

Quy trình mô hình phân cụm



Mô hình K-means





Kết quả phân cụm

Theo Elbow method, **số cụm k=3 là điểm tối ưu cho mô hình** phân cụm, từ đó, ta có được biểu đồ các điểm 2D được tô màu theo cụm, đồng thời thu được các đặc điểm của nhóm:

- Đặc điểm nhân khẩu của khách hàng trong các nhóm không có sự khác nhau (Phụ lục 3)
- Đặc điểm hành vi của khách hàng thể hiện rõ sự phân biệt ở các nhóm (Phụ lục 4)

→ Mặc dù khách hàng có đặc điểm giống nhau về độ tuổi, giới tính, nghề nghiệp..., họ vẫn có hành vi rất khác nhau trong việc tương tác và mua sắm. Từ đó, tệp khách hàng của sàn thương mại điện tử này được chia thành 3 cụm dựa trên đặc điểm hành vi như sau: **Window Shoppers**, **Potential Loyal** và **Deal Hunter.**

Đặc điểm chung nhóm khách hàng: Chủ yếu là nữ, ~31 tuổi, đã kết hôn, có nghề nghiệp và carrier phân bố đều.

Window Shoppers

1. **Tổng số hoạt động:** >1269 (rất cao)
2. **Thời gian truy cập:** ~66 ngày (dài)
3. **Tỷ lệ chuyển đổi:** ~2.2% (rất thấp)
4. **Đặc điểm tương tác:** click rất nhiều, có add-to-cart, favourite→ purchase (~26 lần)

→ Là nhóm truy cập, xem nhiều nhưng chuyển đổi thành mua rất ít, có thể **đang so sánh sản phẩm/giá cả rất nhiều** giữa các gian hàng.

Deal hunter

1. **Tổng số hoạt động:** ~75 (thấp)
2. **Thời gian truy cập:** ~13 ngày (nhưng chỉ hoạt động gần đây)
3. **Tỷ lệ chuyển đổi:** ~10.79% (cao)
4. **Đặc điểm tương tác:** click là chính, ít add-to-cart, favourite→ purchase (~6.3 lần)

→ Là nhóm có mục đích rõ ràng, ít tương tác nhưng khi **đã truy cập thì mua hàng khá quyết đoán.**

→ Khách hàng này có thể là **"deal-hunter"**.

Potential Loyal

1. **Tổng số hoạt động:** ~368 (trung bình)
2. **Thời gian truy cập:** ~37 ngày (dài)
3. **Tỷ lệ chuyển đổi:** ~4.3% (thấp)
4. **Đặc điểm tương tác:** tương tác khá đều, vừa phải giữa click, add-to-cart, favourite → purchase (~15 lần)

→ Là nhóm **trung thành, tương tác đều đặn,** tỷ lệ chuyển đổi vừa phải và **mua hàng thường xuyên hoặc theo nhu cầu thực tế.**

DASHBOARD

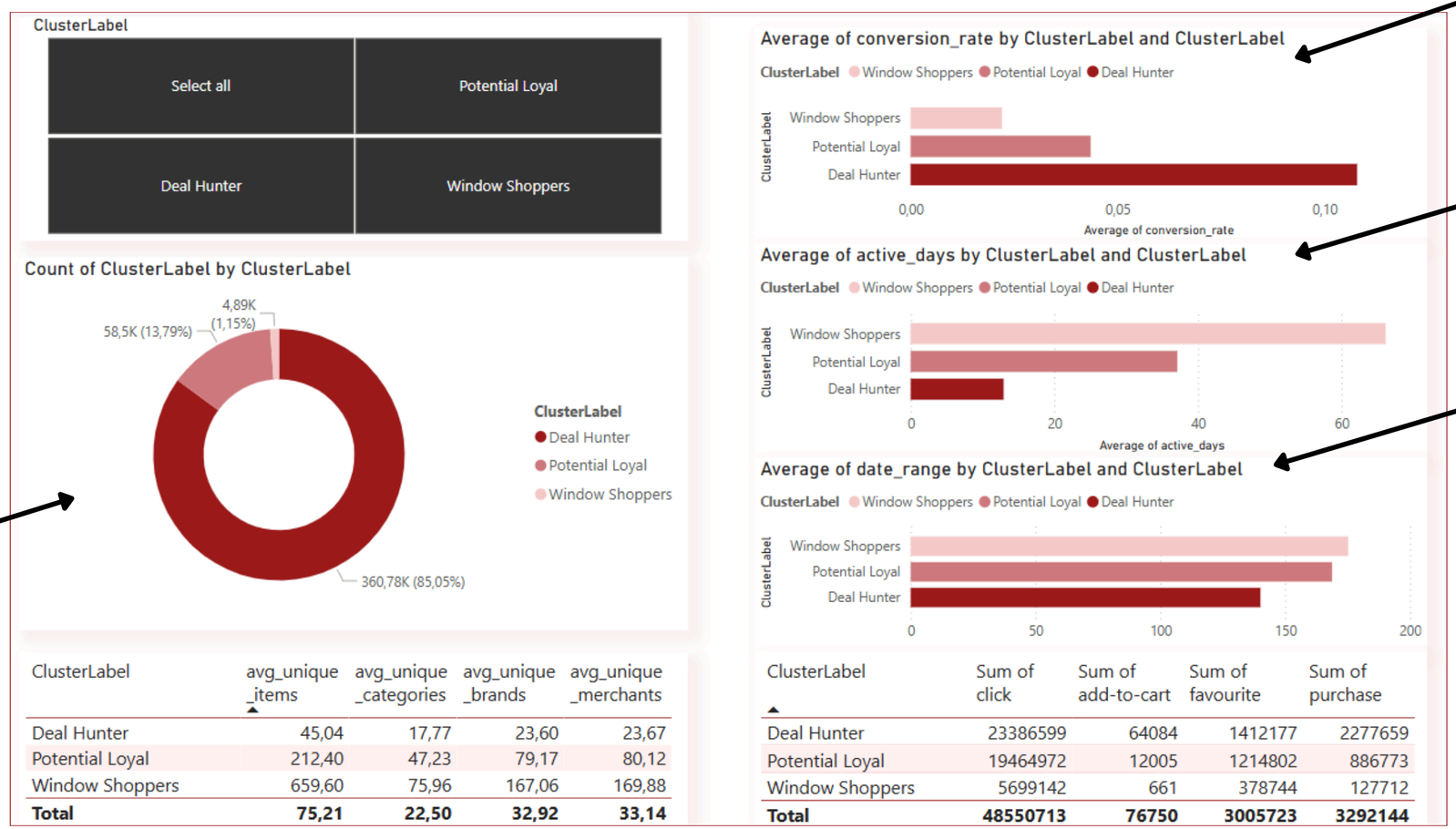
SLICER DÙNG ĐỂ CHỌN VÀ LỌC DỮ LIỆU THEO NHÓM KHÁCH HÀNG, VỚI CÁC TÙY CHỌN.

BIỂU ĐỒ CỘT DÙNG ĐỂ SO SÁNH TỶ LỆ CHUYỂN ĐỔI TRUNG BÌNH GIỮA CÁC NHÓM.

BIỂU ĐỒ CỘT DÙNG ĐỂ XEM SỐ NGÀY HOẠT ĐỘNG TRUNG BÌNH CỦA TỪNG NHÓM.

BIỂU ĐỒ CỘT DÙNG ĐỂ ĐO KHOẢNG THỜI GIAN TRUNG BÌNH CỦA TỪNG NHÓM.

BIỂU ĐỒ TRÒN DÙNG ĐỂ XEM PHÂN BỐ SỐ LƯỢNG KHÁCH HÀNG THEO NHÓM



ClusterLabel	avg_unique_items	avg_unique_categories	avg_unique_brands	avg_unique_merchants
Deal Hunter	45,04	17,77	23,60	23,67
Potential Loyal	212,40	47,23	79,17	80,12
Window Shoppers	659,60	75,96	167,06	169,88
Total	**75,21**	**22,50**	**32,92**	**33,14**

ClusterLabel	Sum of click	Sum of add-to-cart	Sum of favourite	Sum of purchase
Deal Hunter	23386599	64084	1412177	2277659
Potential Loyal	19464972	12005	1214802	886773
Window Shoppers	5699142	661	378744	127712
Total	**48550713**	**76750**	**3005723**	**3292144**

BẢNG DÙNG ĐỂ HIỂN THỊ TRUNG BÌNH SỐ LƯỢNG MẶT HÀNG, DANH MỤC, THƯƠNG HIỆU, VÀ NHÀ BÁN LẺ ĐỘC NHẤT THEO TỪNG NHÓM

BẢNG DÙNG ĐỂ CUNG CẤP SỐ LIỆU CHI TIẾT NHƯ SỐ LẦN NHẤP, THÊM VÀO GIỎ, YÊU THÍCH, VÀ MUA HÀNG.

	Window Shoppers	Deal hunter	Potential Loyal
INSIGHT	Họ dành nhiều thời gian để lướt và so sánh giá trên sàn TMĐT và họ cảm thấy hứng thú với việc “shopping giải trí”.	Họ cực kỳ nhạy cảm với giá, các deal chưa đủ hấp dẫn để họ “xuống tiền”. Họ phải chờ thời cơ có deal cực hời mới đi “săn”, mòn mỏi “check” khi nào sàn tung voucher “đậm” vào một ngày nào đó.	Sàn TMĐT này đang khiến họ có trải nghiệm tốt nhưng họ vẫn lăn tăn hoặc chưa quyết định trở thành khách hàng trung thành, do còn quá nhiều sàn TMĐT để họ lựa chọn.
PROBLEM	Không rõ họ thật sự cần gì, họ không bị ảnh hưởng bởi các quảng cáo push trực tiếp.	Sàn TMĐT đang “đốt tiền” quảng cáo cho những ngày khuyến mãi “đậm” nhưng không giữ chân họ được dài hạn.	Hiện tại, tỷ lệ chuyển đổi action (tỷ lệ mua hàng) thành purchase của nhóm này vẫn còn thấp.
QUESTION	Làm sao nuôi dưỡng họ dần dần để tăng xác suất mua lúc xem hàng trong tương lai và cá nhân hóa được các mặt hàng họ cần?	Làm sao để tăng tỷ lệ chuyển đổi với khách hàng có hành vi "săn deal" để họ thành những user tiềm tăng thậm chí là trung thành?	Làm sao khiến họ quay lại mua đều đặn hơn và luôn giữ chân dược nhóm khách hàng này?
PROPOSE	Xác định sản phẩm mà User này hay ngắm nghía và xem nhiều nhất lên hàng đầu. Dùng mô hình học máy xem xét đến các yếu tố khác (giá/1hóa đơn, chi tiêu...) để cá nhân hóa trải nghiệm	Gắn nhãn những đối tượng này và tạo ra Deal chọn lọc riêng cho họ, Flash sale cá nhân hóa, tạo ra chuỗi streak mua sắm càng nhiều càng nhiều voucher giảm sâu.	Gia tăng thêm các chương trình chăm sóc khách hàng, hậu mãi đối với nhóm đối tượng này. Gợi ý những mặt hàng có liên quan bằng học máy cho nhóm khách hàng này.

NHÓM ĐỐI TƯỢNG CHÍNH CẦN HƯỚNG TỚI

BIG IDEA

GIỮ STREAK MUA

Càng mua càng hời

Thông điệp muốn lan tỏa qua Big idea: Với sự đồng hành của sàn TMĐT, người dùng luôn sẵn sàng mua những thứ mà mình thích, tận hưởng những deal hời hằng ngày mà chẳng cần đến dịp deal khủng.

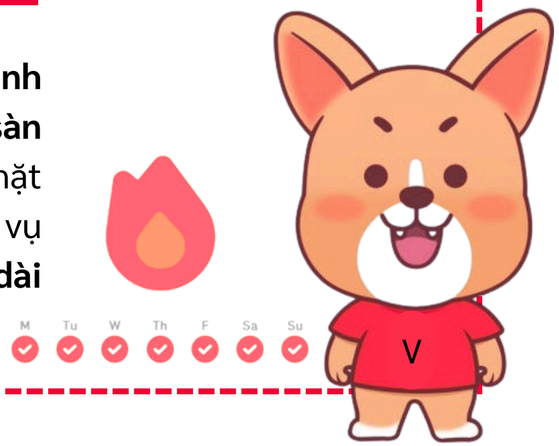
KEY MESSAGE

Không chờ DEAL - DEAL chờ mình

PHASE 01

Mascot giữ chuỗi

User được thu hút bằng cách **điểm danh - giữ chuỗi ngày hoạt động dài trên sàn TMĐT**. Nó đại diện cho sự liên kết chặt chẽ giữa người tiêu dùng và các dịch vụ mà sàn TMĐT cung cấp. **Chuỗi càng dài càng nhận được nhiều deal hời.**



TVC: KHÔNG CHỜ DEAL, DEAL CHỜ MÌNH

Video content

Một người phụ nữ đứng, tay cầm điện thoại trong tháng máy đông người. Cô lướt qua một món đồ yêu thích thì màn hình hiện lên dòng chữ quen thuộc: "Flash Sale Thứ Sáu – còn 2 ngày nữa". Cô thở nhẹ, ánh mắt đầy vẻ chán nản - “Lại phải chờ...”

Bất ngờ, màn hình hiện lên: "Deal cá nhân – chỉ hôm nay". Ánh mắt cô sáng lên, lập tức bấm “Mua ngay”. Thang máy mở cửa, cô bước ra đầy tự tin. Màn hình hiện dòng chốt: "Không chờ deal – deal chờ mình."

Logo app xuất hiện, cùng thông điệp chốt: "Ưu đãi cá nhân – mở app là có.".

PHASE 02

Promotion Campaign

Miễn phí freeship 1 tuần khi đạt chuỗi 1 tháng

Voucher Freeship

Voucher giảm giá cực khủng cho người có chuỗi lớn

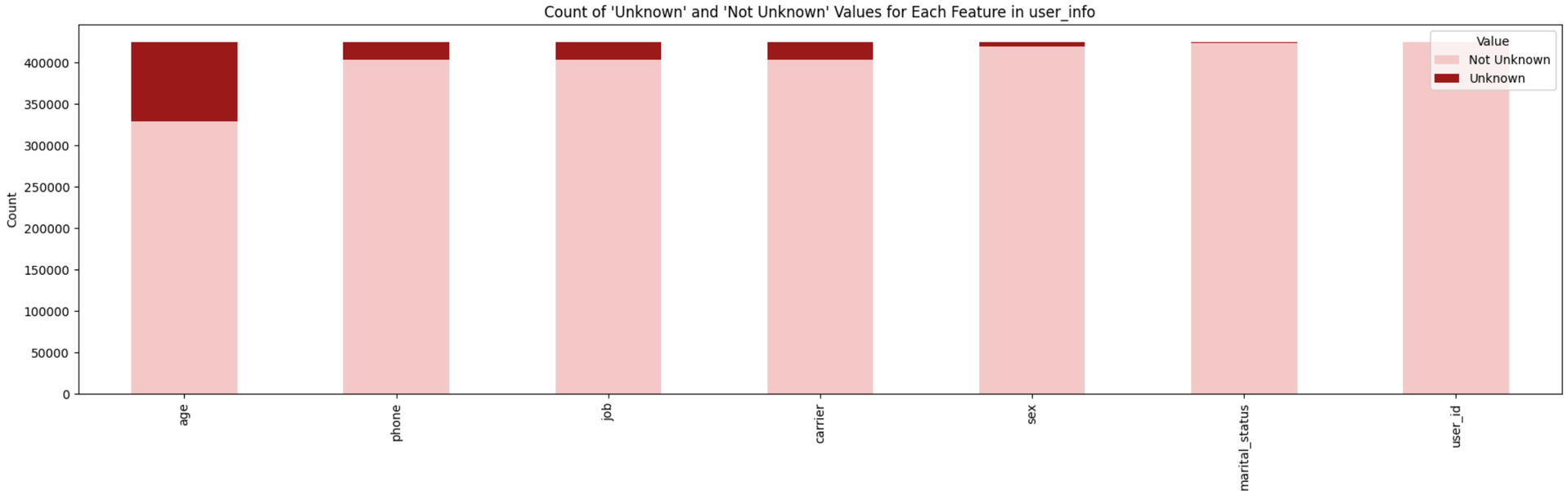
Gamification: My Streak Streak

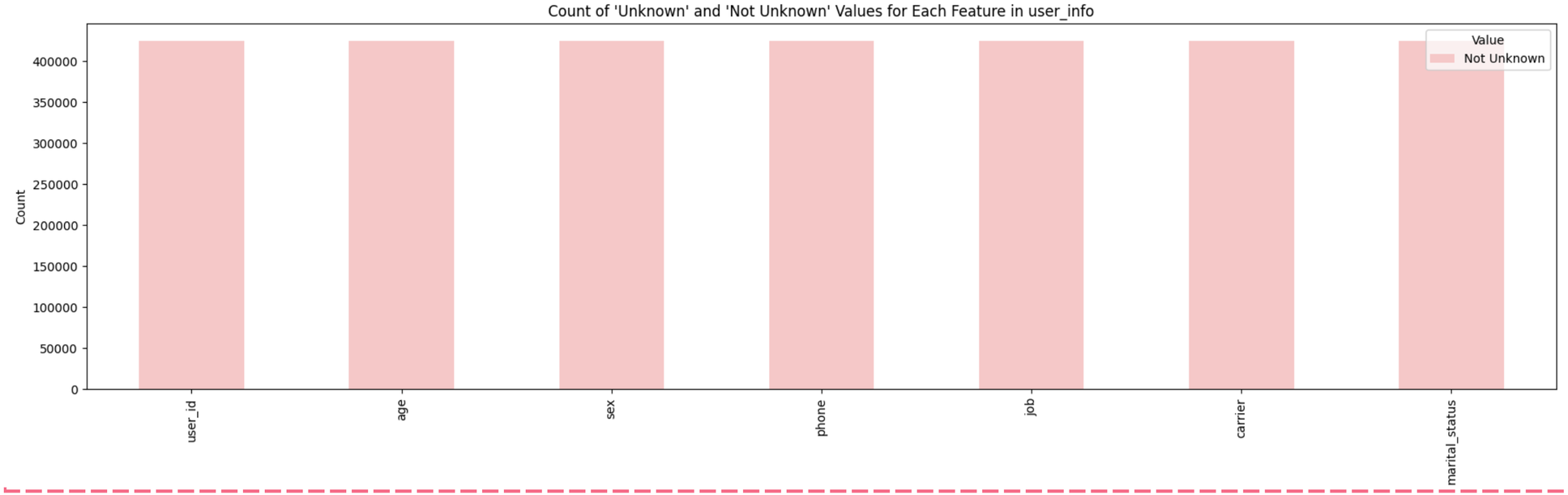
SUPPORT ACTIVITIES

- Social post: về hoạt động của Promotion Campaign
- KOLs
- PR articles
- Paid channels
- OOH, TVC

Phụ lục

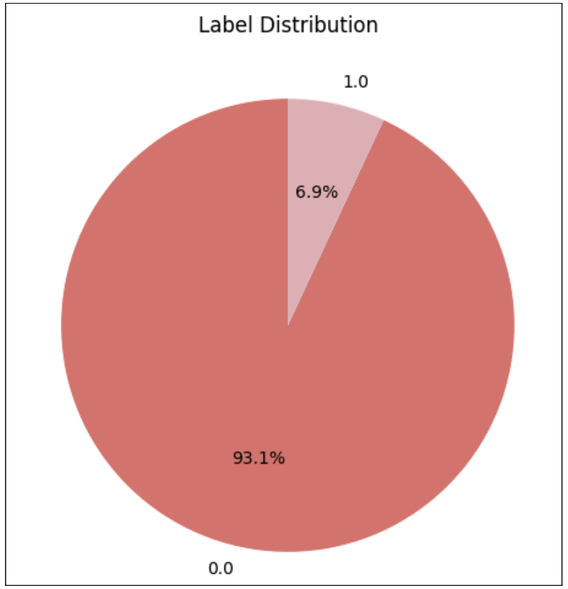
Phụ lục 1:
Số lượng missing của từng cột trong bảng user_info trước và sau khi làm sạch dữ liệu





Phụ lục

Phụ lục 2: Phân phối label trong dữ liệu



Phụ lục 3: Đặc điểm nhân khẩu khách hàng của từng nhóm

Cluster	#age	#sex	#job	#carrier	#marital_status
1	30	0	27	2	1
2	32	0	27	2	1
3	31	0	27	2	1

Phụ lục 4: Đặc điểm hành vi khách hàng của từng nhóm

cluster	unique _items	unique _categories	unique _brands	unique _merchants	recency _days	active _days	conversion _rate	date _range	total _actions	add-to-cart	click	favourite	purchase
0	45.03	17.77	23.59	23.67	1	12.94	0.11	140.12	75.22	0.18	64.81	3.91	6.31
1	659.6	75.96	167.06	169.88	1	66.13	0.02	175.23	1269.17	0.14	1165.47	77.45	26.12
2	212.36	47.23	79.15	80.11	1	37.13	0.04	168.81	368.78	0.21	332.66	20.76	15.16

viettel

CUỘC THI PHÂN TÍCH DỮ LIỆU DAZONE 2025

THANK YOU
FOR YOUR LISTENING!

Team ‘LOICHOI’
Võ Trần Thị Mỹ Vân - Nguyễn Tuấn Huy - Trần Nguyễn Quang Huy