THIẾT KẾ CHI TIẾT MÔ HÌNH PHÂN KHÚC KHÁCH HÀNG

CHỦ ĐẦU TƯ

NGÂN HÀNG THƯƠNG MẠI CỔ PHẦN SÀI GÒN THƯƠNG TÍN

Hà Nội 09/2022

QUẢN LÝ THAY ĐỔI

Ngày thay đổi	Mục, bảng, sơ đồ được thay đổi	Mô tả thay đổi	Phiên bản
23/09/2022	N/A	Tạo mới	1.0
10/10/2022	Edit	Chỉnh sửa	1.1
11/05/2023	Edit	Chỉnh sửa	1.2

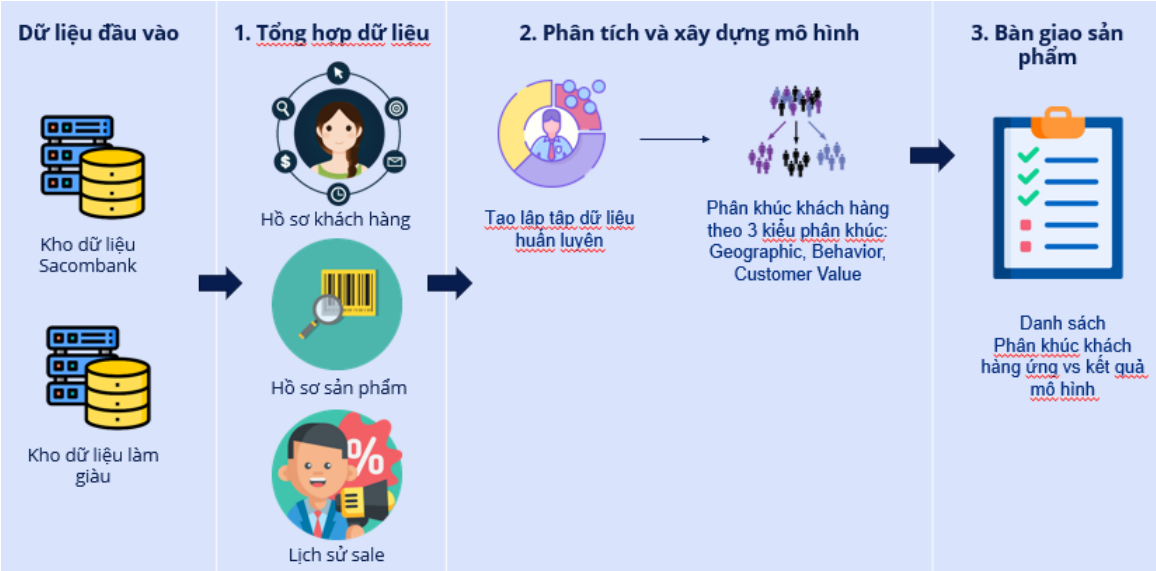
Mục Lục

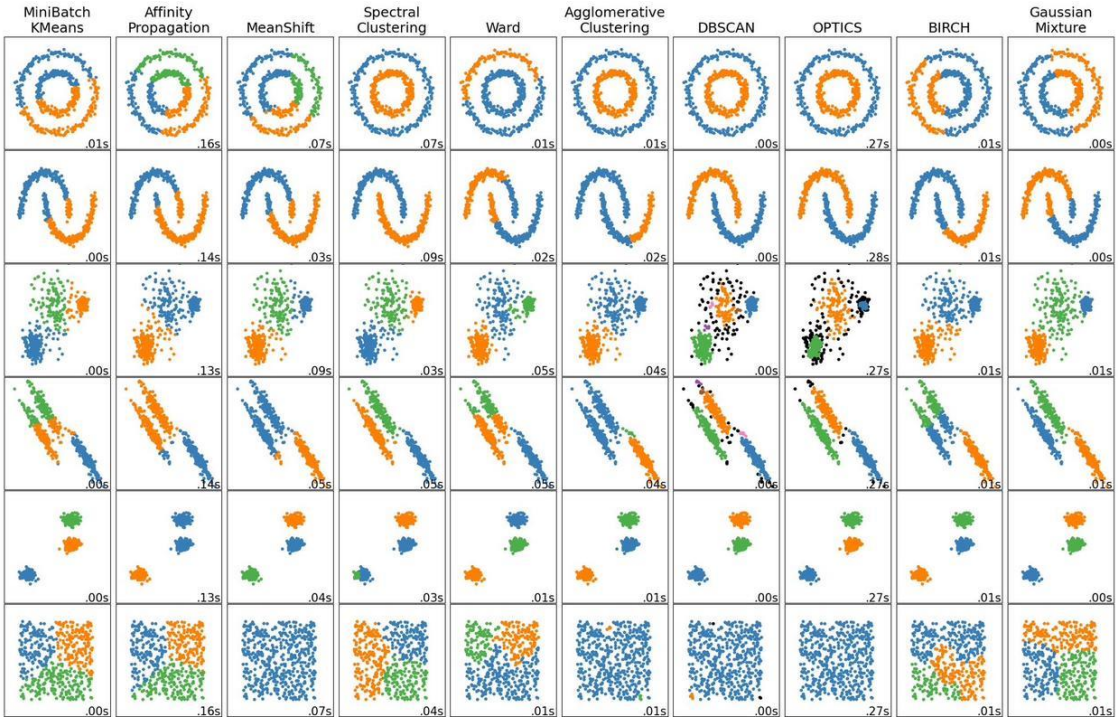
I. TỔNG QUAN

I.1. Đề bài

- Trong bối cảnh hiện tại, thị trường Tài chính – Ngân hàng đang bước vào giai đoạn cạnh tranh khốc liệt trên các mặt trận chính về: Công nghệ - Kinh doanh - Con người. Để giữ chân được khách hàng, giữ vững được thị phần của mình và tiến tới mở rộng thị phần của mình. Sacombank cần đầu tư bàn bản về hệ thống thu thập, xử lý và phân tích dữ liệu để làm giàu dữ liệu của Ngân hàng cũng như để phát triển & nâng cao chất lượng dịch vụ đáp ứng nhu cầu ngày càng cao của khách hàng.
- Hiện tại, Sacombank chưa có hệ thống mô hình AI phân tích dữ liệu để có thể phân chia được nhóm khách hàng nhằm tập trung, đẩy mạnh quan tâm và chăm sóc. Do vậy, để giải quyết bài toán trên, cần thiết phải tìm kiếm các đối tác có kinh nghiệm cũng như năng lực phân tích dữ liệu trong lĩnh vực Tài chính – Ngân hàng nói chung và đặc thù của các Ngân hàng thương mại Việt Nam nói riêng để hỗ trợ Sacombank có thể triển khai phân tích chuyên sâu về mô hình phân khúc khách hàng.
- Phân đoạn khách hàng giúp cho việc hình thành nền tảng khoa học dữ liệu. Bao gồm việc xây dựng các mô hình AI/ML phân đoạn khách hàng theo 3 kịch bản:
    - Geography (địa lý và thông tin chung)
    - Transactional (hành vi giao dịch)
    - Customer Value (giá trị khách hàng)

I.2. Phương pháp luận





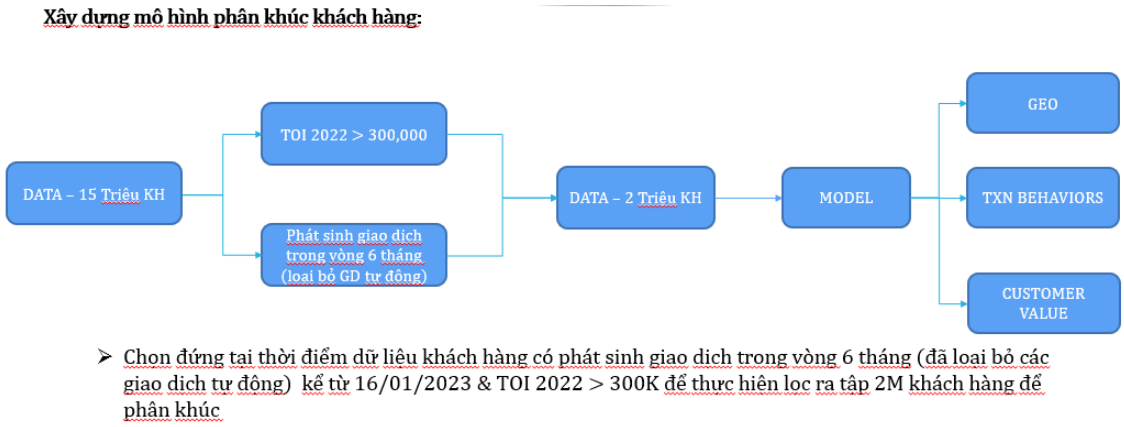
- Để xây dựng mô hình phân khúc khách hàng theo 3 chủ đề: Geographic, TXN Behaviors, Customer Value bao gồm việc thu thập, tổ chức dữ liệu từ 2 nguồn chính:
    - Nguồn dữ liệu nội bộ (Dữ liệu Sacom)
    - Nguồn dữ liệu làm giàu

- Sau đó sẽ thực hiện tổng hợp và xử lý dữ liệu để hình thành bộ dữ liệu huấn luyện, làm đầu vào xây dựng mô hình phân khúc khách hàng. Kết quả mô hình sẽ là danh sách khách hàng được phân cụm.
- Có rất nhiều thuật toán dùng để xây dựng mô hình phân khúc khách hàng. Tuy nhiên, theo đánh giá về độ phân bố dữ liệu cũng như kỹ thuật xử lý từ các thuật toán, tiến hành chọn K-Means và DBSCAN là 2 thuật toán chính dùng để xây dựng mô hình, giúp cho việc phân tách cụm từ tập dữ liệu một cách rõ ràng.

II. QUY TRÌNH XÂY DỰNG MÔ HÌNH

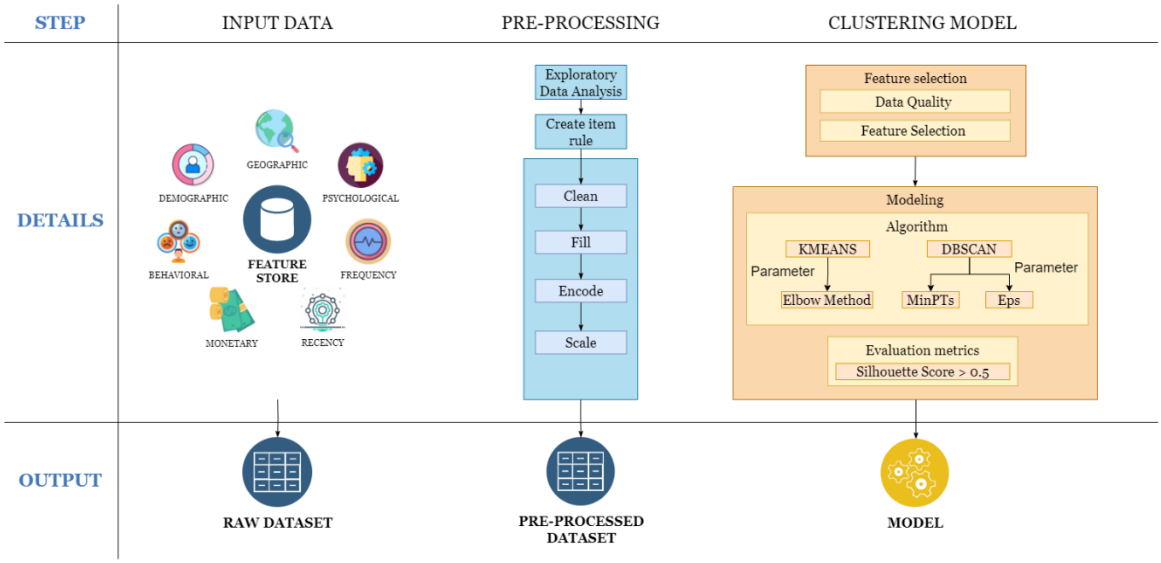
II.1. Đầu vào

- Với 3 kiểu phân khúc : Geographic, Behavior, Customer Value sẽ có những Feature được tổng hợp tương ứng với phân khúc trên để tạo thành bộ dữ liệu huấn luyện xây dựng ba mô hình khác nhau. Tập dữ liệu huấn luyện xây dựng mô hình được tổng hợp theo logic như sau :



- Để xây dựng mô hình, dữ liệu đầu vào bao gồm 15 triệu khách hàng. Tuy nhiên, nhằm tăng độ chính xác trong việc xây dựng mô hình cũng như tiết kiệm thời gian xử lý đối với bộ dữ liệu lớn , chọn đúng tại thời điểm dữ liệu khách hàng có phát sinh giao dịch trong vòng 6 tháng (đã loại bỏ các giao dịch tự động) kể từ 16/01/2023 & TOI 2022 > 300K để thực hiện lọc ra tập 2 triệu khách hàng để phân khúc

II.2. Quy trình xử lý



Các bước xây dựng mô hình phân cụm

Mô hình phân khúc khách hàng được chia làm 3 kiểu phân khúc:

- Geographic
- Behavior
- Customer Value

II.2.1. Data Preparation

- Đây là bước tiền xử lý dữ liệu, bằng cách cleaning data và xử lý những điểm dữ liệu có chứa giá trị null và outliers, bước này giúp cho dữ liệu được sạch hơn và loại bỏ những điểm gây nhiễu để có thể tiến hành xây dựng mô hình với kết quả chính xác cao nhất.

  Sau đó, sẽ thực hiện mã hóa dữ liệu đối với nhóm dữ liệu là Category phục vụ cho việc xử lý của nhiều thuật toán khác nhau và bước Scale dữ liệu giúp cho việc chuẩn hóa các Feature của tập dữ liệu về một khoảng, cùng một đơn vị đo lường giúp cho máy có thể học một cách dễ dàng hơn.

B1: Làm sạch & Làm bù dữ liệu:

- Để làm sạch dữ liệu, sẽ thực hiện các bước như sau:
  - Phân tích insights bằng cách visualize chart

- Xử lý những điểm dữ liệu có chứa giá trị null & outliers với các giá trị như mean, mode, median

B2: Mã hóa dữ liệu:

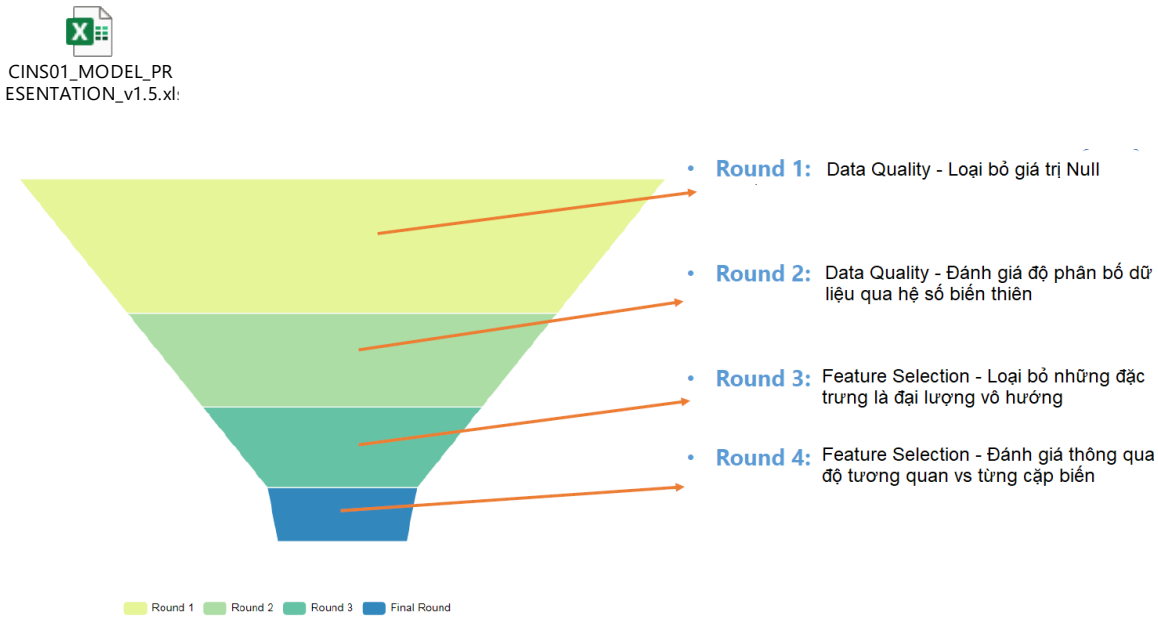
- Sau đó thực hiện mã hóa dữ liệu với những Feature đã được xử lý tại bước trước đó với phương pháp: One hot encoding.

B3: Scale dữ liệu:

- Thực hiện scale dữ liệu để đưa các giá trị Feature về range (0,1)

II.2.2. Feature Engineering

- Với bài toán này, sẽ thực hiện qua các round như sau để có thể chọn ra được các đặc trưng quan trọng đưa vào mô hình
- File chi tiết 4 round chọn đặc trưng đưa vào mô hình:



Round 1: *Đánh giá chất lượng dữ liệu thông qua tỉ lệ phần trăm Null cho từng đặc trưng.*

- Truyền vào 217 biến, 137 biến không đạt chất lượng và lọc ra được 80 biến chạy cho round tiếp theo

		2,071,398		% NULL
	FEATURE	JAN - DISTINCT		
13	CHILD_PET_SM_AMT_12M	6,672		99.7%
14	CHILD_PET_CT_TXN_12M	6,672		99.7%
15	DEBT_GRP	8,318		99.6%
16	EDUCATION_SM_AMT_1M	12,569		99.4%
17	EDUCATION_CT_TXN_1M	12,569		99.4%
18	HOBBIES_ENTERTAINMENT_CT_TXN_12M	14,719		99.3%
19	HOBBIES_ENTERTAINMENT_SM_AMT_12M	14,719		99.3%
20	BEAUTY_CT_TXN_1M	16,457		99.2%
21	BEAUTY_SM_AMT_1M	16,457		99.2%
22	CARD_CREDIT_UP_SELL_LABEL4_6M	17,491		99.2%
23	VEHICLES_SM_AMT_1M	18,157		99.1%
24	VEHICLES_CT_TXN_1M	18,157		99.1%
25	PUBLIC_SERVICE_HEALTHCARE_SM_AMT_1M	19,860		99.0%
26	PUBLIC_SERVICE_HEALTHCARE_CT_TXN_1M	19,860		99.0%
27	APPLIANCES_SM_AMT_1M	20,064		99.0%
28	APPLIANCES_CT_TXN_1M	20,064		99.0%
29	CARD_CREDIT_CT_TXN_INTER_3M	23,405		98.9%
30	CARD_CREDIT_SUM_TXN_INTER_3M	23,405		98.9%
31	SERVICE_SM_AMT_1M	23,437		98.9%
32	SERVICE_CT_TXN_1M	23,437		98.9%
33	TRAVEL_CT_TXN_1M	28,171		98.6%

Round 2: *Đánh giá sự phân bố dữ liệu thông qua hệ số biến thiên.*

- Truyền vào 80 biến, loại bỏ 13 biến không đạt chất lượng và lọc ra được 67 biến chạy cho round tiếp theo

FTR	Standard Deviation	Averag	Coefficient of Variation	Remove
EB_MBIB_CT_TXN_1M	65.5309	23.6892	2.77	
EB_MBIB_SUM_TXN_AMT_1M	1.48E+09	7E+08	2.11	
EB_CT_TXN_3M	274.0682	59.644	4.60	x
EB_CT_TXN_6M	527.1308	105.3344	5.00	x
EB_SACOMPAY_SUM_TXN_AMT_1M	2.64E+08	96281358	2.74	
EB_SACOMPAY_CT_TXN_1M	2.64E+01	17.871	1.48	
EB_MBIB_DAY_SINCE_LTST_TXN	726.8933	236.316	3.08	x
EB_SACOMPAY_CT_TXN_6M	121.4082	75.9609	1.60	
EB_SACOMPAY_HOLD				
EB_MBIB_HOLD				
CARD_FAV_BRANCH_LOC_3M				
CARD_CT_VAR_BRANCH_3M	5.2504	1.06	4.95	x
CARD_SUM_TXN_AMT_3M	2.16E+08	63746304	3.38	x
CARD_BRANCH_LOC_3M				
CARD_SUM_TXN_AMT_6M	3.66E+08	1.16E+08	3.14	x
CARD_BRANCH_LOC_6M				
CARD_FAV_BRANCH_LOC_6M				
CASA_SUM_TXN_AMT_3M	3.37E+09	1.77E+09	1.91	
CASA_CT_TXN_3M	1.99E+02	96.9031	2.06	
CASA_SUM_TXN_AMT_6M	9.88E+09	3.3E+09	2.99	

Round 3: *Loại bỏ những đặc trưng là đại lượng vô hướng và các đặc trưng được sử dụng làm gốc để sản sinh ra các đặc trưng con đưa vào bài toán phân khúc*.

- Truyền vào 67 biến, loại bỏ 22 biến, 45 biến được sử dụng cho round tiếp theo

FTR	REMOVE
EB_SACOMPAY_HOLD	x
EB_MBIB_HOLD	x
CARD_FAV_BRANCH_LOC_3M	x
CARD_BRANCH_LOC_3M	x
CARD_BRANCH_LOC_6M	x
CARD_FAV_BRANCH_LOC_6M	x
CASA_HOLD	x
GEN_GRP	x
AREA	x
AGE	x
PROFESSION	x
LIFE_STG	x
CARD_TOP5_MERCHANT_6M	x
CASA_CROSS_SELL_LABEL2	x
CARD_TOP4_MERCHANT_6M	x
CARD_TOP3_MERCHANT_6M	x
FAV_POS_6M_CT	x
FAV_POS_6M_SM	x
CARD_TOP2_MERCHANT_6M	x
SACOMBANK_PAY_CROSS_SELL_LABEL1	x

Round 4: *Đánh giá các đặc trưng thông qua độ tương quan với từng cặp biến*

- Truyền vào 45 biến, loại bỏ 7 biến, còn 38 biến làm đầu vào cho mô hình phân khúc

Feature_1	Feature_2	Corr	Remove
EB_MBIB_SUM_TXN_AMT_3M	EB_MBIB_SUM_TXN_AMT_6M	0.463332	
EB_SACOMPAY_CT_TXN_6M	EB_SACOMPAY_CT_TXN_3M	0.960245	EB_SACOMPAY_CT_TXN_6M
EB_MBIB_CT_TXN_1M	EB_MBIB_CT_TXN_3M	0.957353	EB_MBIB_CT_TXN_1M
CASA_CT_TXN_3M	CASA_CT_TXN_1M	0.953188	CASA_CT_TXN_3M
CASA_SUM_TXN_AMT_6M	EB_MBIB_SUM_TXN_AMT_6M	0.945997	CASA_SUM_TXN_AMT_6M
EB_SACOMPAY_SUM_TXN_AMT_3M	EB_SACOMPAY_SUM_TXN_AMT_6M	0.342085	
EB_MBIB_SUM_TXN_AMT_1M	EB_MBIB_SUM_TXN_AMT_3M	0.94031	EB_MBIB_SUM_TXN_AMT_1M
EB_SACOMPAY_CT_TXN_1M	EB_SACOMPAY_CT_TXN_3M	0.939332	EB_SACOMPAY_CT_TXN_1M
CASA_SUM_TXN_AMT_3M	EB_MBIB_SUM_TXN_AMT_3M	0.934998	CASA_SUM_TXN_AMT_3M
EB_MBIB_SUM_TXN_AMT_1M	CASA_SUM_TXN_AMT_1M	0.931121	EB_MBIB_SUM_TXN_AMT_1M
CASA_SUM_TXN_AMT_6M	EB_MBIB_SUM_TXN_AMT_3M	0.921246	CASA_SUM_TXN_AMT_6M
CASA_SUM_TXN_AMT_3M	CASA_SUM_TXN_AMT_1M	0.916707	CASA_SUM_TXN_AMT_3M
CASA_CT_TXN_3M	CASA_CT_TXN_6M	0.907654	CASA_CT_TXN_3M
EB_MBIB_CT_TXN_3M	EB_MBIB_CT_TXN_6M	0.454735	
EB_MBIB_SUM_TXN_AMT_1M	EB_MBIB_SUM_TXN_AMT_6M	0.902072	EB_MBIB_SUM_TXN_AMT_1M
CASA_SUM_TXN_AMT_3M	EB_MBIB_SUM_TXN_AMT_6M	0.893365	CASA_SUM_TXN_AMT_3M
EB_SACOMPAY_SUM_TXN_AMT_1M	EB_SACOMPAY_SUM_TXN_AMT_3M	0.888973	EB_SACOMPAY_SUM_TXN_AMT_1M
EB_MBIB_SUM_TXN_AMT_1M	CASA_SUM_TXN_AMT_3M	0.883282	EB_MBIB_SUM_TXN_AMT_1M
EB_SACOMPAY_CT_TXN_1M	EB_SACOMPAY_CT_TXN_6M	0.879854	EB_SACOMPAY_CT_TXN_1M
EB_MBIB_CT_TXN_1M	EB_MBIB_CT_TXN_6M	0.857045	EB_MBIB_CT_TXN_1M

- Sau khi qua 4 round để có thể lọc chất lượng và làm sạch đặc trưng, thì bộ đặc trưng sử dụng để đưa vào mô hình bao gồm:

  - SEGMENTATION GEOGRAPHIC

FTR_NM	GROUP	SUB_GROUP	DESC
ADDR_POP	GEOGRAPHIC	ENRICH	Dân số theo phường xã khu vực KH sinh sống
DIST_CUST_FAV_POS_TOP1	GEOGRAPHIC	BEHAVIORAL	Khoảng cách nơi ở KH tới top 1 máy POS thường xuyên giao dịch
DIST_CUST_FAV_ATM	GEOGRAPHIC	BEHAVIORAL	Khoảng cách nơi ở KH tới ATM thường xuyên giao dịch
DIST_CUST_FAV_POS_TOP2	GEOGRAPHIC	BEHAVIORAL	Khoảng cách nơi ở KH tới top 2 máy POS thường xuyên giao dịch
ADDR_LAT	GEOGRAPHIC		Vĩ độ của vị trí KH
ADDR_LNG	GEOGRAPHIC		Kinh độ của vị trí KH
FAV_POS_1_LAT	GEOGRAPHIC		Vĩ độ của nơi KH thường xuyên quẹt thẻ
FAV_POS_1_LNG	GEOGRAPHIC		Kinh độ của nơi KH thường xuyên quẹt thẻ
FAV_POS_2_LAT	GEOGRAPHIC		Vĩ độ của nơi KH thường xuyên quẹt thẻ
FAV_POS_2_LNG	GEOGRAPHIC		Kinh độ của nơi KH thường xuyên quẹt thẻ
FAV_ATM_LAT	GEOGRAPHIC		Vĩ độ của ATM KH thường xuyên giao dịch
FAV_ATM_LNG	GEOGRAPHIC		Kinh độ của ATM KH thường xuyên giao dịch

  - SEGMENTATION TXN BEHAVIORS

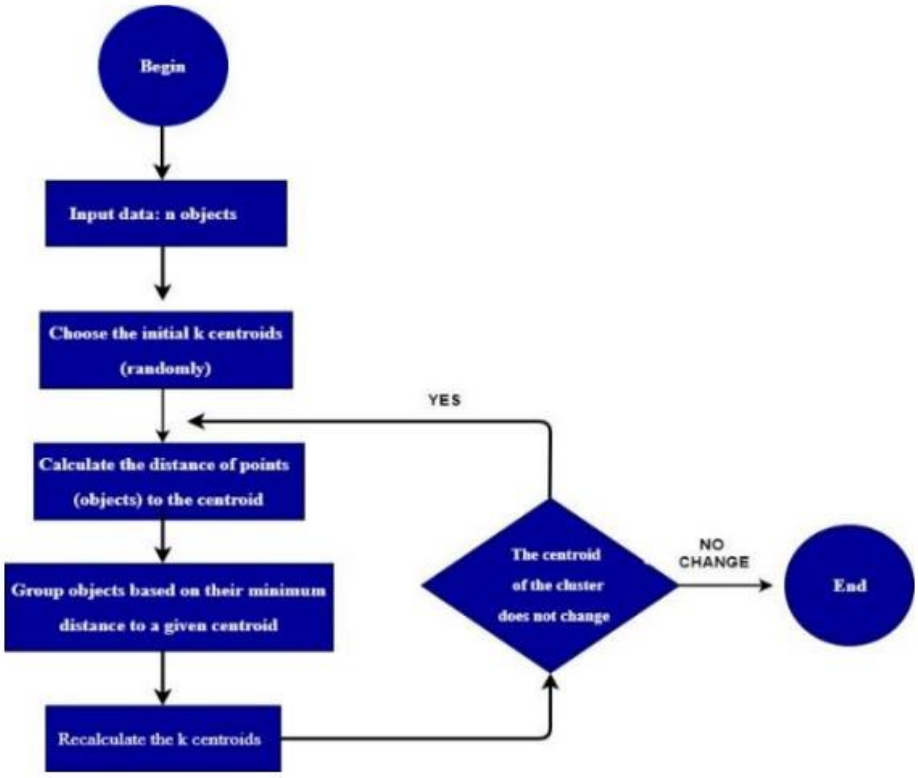
FTR_NM	GROUP	SUB_GROUP	DESC
CASA_SUM_BAL_NOW	PRODUCT USAGE	MONETARY	Số dư CASA tính tới hiện tại
FOOD_GROCERY_CT_TXN_12M	PRODUCT USAGE	FREQUENCY	Số lượng giao dịch FOOD GROCERY Category trong vòng 12 tháng
EB_MBIB_CT_TXN_3M	PRODUCT USAGE	FREQUENCY	Tổng số giao dịch qua MBIB trong quý gần nhất
EB_MBIB_DAY_SINCE_ACTIVE	PRODUCT USAGE	MBIB	Thời gian từ ngày kích hoạt tài khoản MBIB
CASA_CT_TXN_6M	PRODUCT USAGE	CASA	Số giao dịch bằng tài khoản thanh toán trong nửa năm gần nhất
UTILITIES_CT_TXN_12M	PRODUCT USAGE	FREQUENCY	Số lượng giao dịch UTILITIES Category trong vòng 12 tháng
CARD_CREDIT_MAX_LIMIT	PRODUCT USAGE	MONETARY	Hạn mức cao nhất của thẻ tín dụng
EB_SACOMPAY_DAY_SINCE_LTST_TXN	PRODUCT USAGE	SACOMPAY	Thời gian từ lần thực hiện giao dịch qua hệ thống Sacompay gần nhất
CASH_SM_AMT_12M	PRODUCT USAGE	MONETARY	Giá trị giao dịch Cash Category trong vòng 12 tháng
EB_SACOMPAY_CT_TXN_3M	PRODUCT USAGE	FREQUENCY	Tổng số giao dịch qua Sacompay trong quý gần nhất
CASH_CT_TXN_12M	PRODUCT USAGE	FREQUENCY	Số lượng giao dịch CASH Category trong vòng 12 tháng

  - SEGMENTATION CUSTOMER VALUE

FTR_NM	GROUP	SUB_GROUP	DESC
LOR	RELATIONSHIP WITH BANK		Thời gian khách hàng gắn bó với ngân hàng
CASA_SUM_BAL_NOW	PRODUCT USAGE	MONETARY	Số dư CASA cho tới hiện tại
EB_MBIB_CT_TXN_6M	PRODUCT USAGE	MBIB	Tổng số giao dịch qua MBIB trong 6 tháng gần nhất
EB_MBIB_SUM_TXN_AMT_6M	PRODUCT USAGE	MBIB	Tổng giá trị giao dịch bằng MB/IB trong 6 tháng
EB_MBIB_DAY_SINCE_ACTIVE	PRODUCT USAGE	MBIB	Thời gian từ ngày kích hoạt tài khoản MBIB
CASA_CT_TXN_6M	PRODUCT USAGE	CASA	Số giao dịch bằng tài khoản thanh toán trong nửa năm gần nhất
TOI	PRODUCT USAGE	MONETARY	TOI của khách hàng
EB_SACOMPAY_SUM_TXN_AMT_6M	PRODUCT USAGE	SACOMPAY	Tổng giá trị giao dịch bằng Sacompay trong nửa năm

II.2.3. Modeling

- **KMeans**:



Các bước xử lý trong thuật toán K-mean

Input: Số lượng cụm k và các điểm centroid {mj}

Output: Số lượng cụm C[i] ($1 \leq i \leq k$)

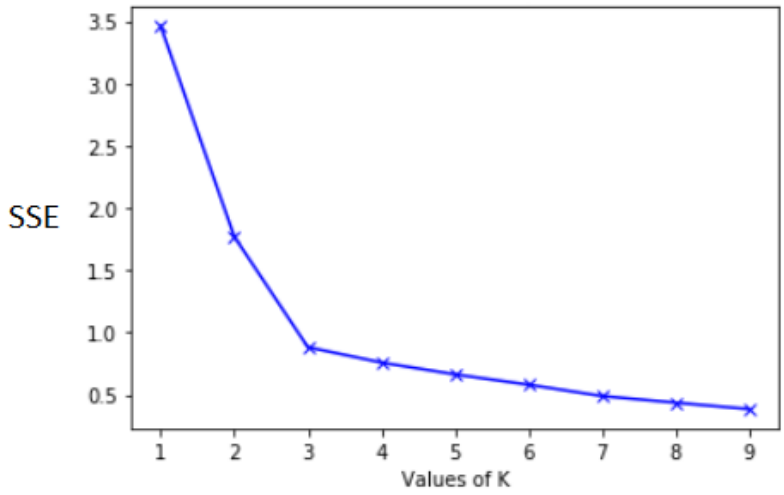
- ***Các bước trong thuật toán K-means:***

B1: Chọn k centroid $\{mj\}$ ($1 \leq j \leq k$) ban đầu trong Rd (với d là số chiều dữ liệu). Sự lựa chọn này có thể ngẫu nhiên hoặc theo kinh nghiệm

B2: Tính khoảng cách: Với mỗi điểm Xi ($1 \leq i \leq n$), tính khoảng cách từ mỗi điểm dữ liệu tới mỗi centroid $\{mj\}$ ($1 \leq j \leq k$). Sau đó tìm những điểm gần với centroid nhất

B3: Update centroid bằng cách xác định giá trị trung bình các véctơ đặc trưng dữ liệu. Điều kiện dừng thuật toán: Lặp lại bước 2 và bước 3 cho tới khi centroids của clusters không thay đổi

- Thuật toán K-mean sử dụng tốt nhất với bộ dữ liệu nhỏ vì nó sẽ quét toán bộ các điểm dữ liệu => Mất nhiều thời gian để xác định các điểm dữ liệu trong 1 bộ dataset lớn
- Phương pháp để chọn ra số Clusters :
  - **Elbow method**:
- Dựa vào đường cong khuỷu tay, số k thích hợp là vị trí ở khúc quanh (đầu gối) của đường. Tại thời điểm này,giá trị của khoảng cách trung bình không thay đổi đáng kể khi số lượng cụm k tăng lên



Elbow Curve

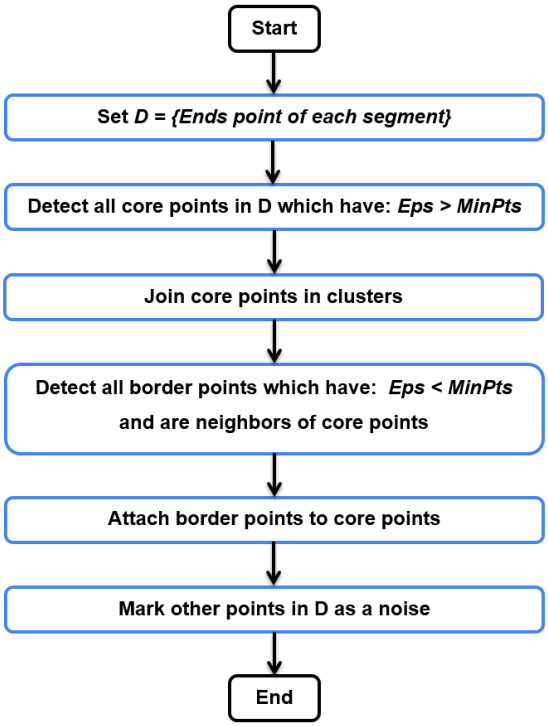
- Với ví dụ minh họa elbow curve tại hình trên, trục tung với tiêu chí đánh giá là SSE (Sum of Errors – đo lường sự khác biệt giữa các điểm trong cluster), có thể thấy với giá trị k = 3, SSE có xu hướng giảm dần. Do đó, số k tối ưu trong trường hợp này sẽ lựa chọn là 3
  - **Silhouette score**:
- Dựa vào Silhouette score để đánh giá chất lượng số cụm được xác định. Với cụm có điểm số cao hơn, khả năng cao số cụm được phân sẽ có chất lượng và kết quả tốt hơn

```
For n_clusters = 3 The average silhouette_score is : 0.5100543174016334
For n_clusters = 4 The average silhouette_score is : 0.47884852105752135
```

Silhouette score với từng k cụm

- Khi chạy mô hình với những thuật toán khác nhau, sẽ cho ra k cụm khác nhau. Trong trường hợp này, sẽ thực hiện chạy đồng thời các methods để cho ra số k bất kì. Với kết quả số k thu được từ từng method, sẽ quyết định chọn số k cụm sao cho kết quả sát với kết quả sát với các method vừa thu được.

  - **DBSCAN** (Density-based spatial clustering of applications with noise)



Các bước trong thuật toán DBSCAN

Hình dạng sau khi phân cụm thuật toán DBSCAN

- **Các bước xử lý trong thuật toán DBSCAN:**

  B1: D ={i} Tập hợp các điểm tại mỗi phân khúc

  B2: Xác định các điểm lõi trong D đáp ứng điều kiện: Eps > MinPts

B3: Kết hợp tất cả các điểm lõi vào chung một cụm

B4: Phát hiện tất cả các điểm biên với điều kiện Eps < MinPts và những điểm này là các điểm hàng xóm với điểm lõi

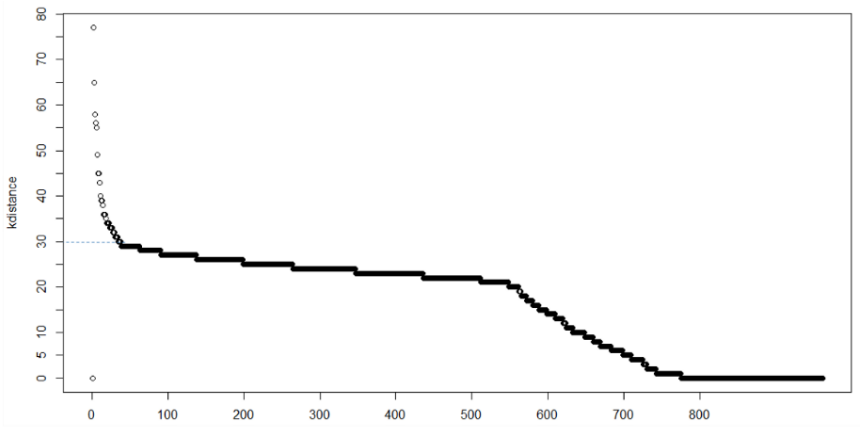
B5: Gán các điểm biên với diểm lõi

B6: Những điểm còn lại được coi là điểm nhiễu

- Density-based algorithm. Thuật toán DBSCAN tốt trong việc phát hiện ra các điểm outliers
- Dựa vào mật độ giữa data points để phân cụm. Không bao gồm các điểm Outliers vào các cụm có mật độ cao.
- **2 parameters để xác định cụm**:
  +) minPTs – Tối thiểu các điểm dữ liệu cần đc phân cụm ở trong khu vực được coi là có mật độ cao
  +) eps – Khoảng cách dùng để xác định một điểm dữ liệu có cùng khu vực với các điểm dữ liệu khác

Cách chọn parameter:

- Đối với tham số minPTs, sẽ phụ thuộc vào chiều của dữ liệu. Với giá trị minPTS = 1 sẽ không có ý nghĩa bởi như vậy tất cả các điểm dữ liệu đều có thể được gom vào một cụm riêng lẻ. Tuy nhiên, các giá trị lớn hơn thường tốt hơn cho các tập dữ liệu có nhiễu và sẽ mang lại nhiều cụm có ý nghĩa hơn. Theo quy tắc chung, MinPts = 2xD có thể được sử dụng, nhưng có thể cần phải chọn các giá trị lớn hơn cho dữ liệu rất lớn, cho dữ liệu nhiễu hoặc cho dữ liệu có nhiều bản sao.
- Với tham số eps, sử dụng k-distance plot để xác định giá trị. Phương pháp này tính toán k-khoảng cách láng giềng gần nhất trong một ma trận điểm. Ý tưởng chính đó là tính khoảng cách trung bình từ mỗi điểm dữ liệu với các điểm hàng xóm xung quanh nó. Giá trị k sẽ được xác định bằng với giá trị minPTs. Tiếp theo, k-khoảng cách này được vẽ theo thứ tự tăng dần. Mục đích là để xác định "knee", tương ứng với tham số epsilon tối ưu. "Knee" tương ứng với ngưỡng xảy ra thay đổi mạnh dọc theo đường cong k-khoảng cách.



Ví dụ minh họa k-distance plot

II.2.4. Evaluate Model

- Sử dụng Silhouette Coefficient để đánh giá mô hình. Silhouette đo lường khoảng cách của một điểm dữ liệu trong cụm đến Centroid - điểm trung tâm của cụm, và khoảng cách của chính điểm đó đến điểm trung tâm của cụm gần nhất. Do vậy, mô hình phân cụm tốt hay không tốt, có thể dùng Silhouette Coefficient để đánh giá.

$$s_i = (b_i - a_i) / \max(a_i, b_i)$$

Trong đó:

- Giả sử có 2 cluster A và B được tìm thấy dựa trên K-means clustering
  - bi là khoảng cách từ điểm i trong cluster A đến điểm trung tâm của cluster B
  - ai là khoảng cách từ điểm i trong cluster A đến điểm trung tâm của cluster A
- Nếu không phải K-means clustering thì:
  - bi là khoảng cách trung bình từ điểm i trong cluster A đến tất cả các điểm trong cluster B với cluster B là cluster láng giềng gần nhất.
  - ai là khoảng cách trung bình từ điểm i trong cluster A đến tất cả các điểm còn lại trong A
  - max(bi, ai) tức lấy chọn giá trị lớn nhất giữa ai và bi

- Để đánh giá liệu một điểm dữ liệu có được phân cụm tốt hay không, sẽ thực hiện như sau:

- Điểm dữ liệu có Silhouette cao, gần bằng 1, chắc chắn nằm đúng trong cluster
- Điểm dữ liệu có Silhouette gần bằng 0, đang nằm giữa 2 cluster
- Điểm dữ liệu có Silhouette thấp, có giá trị âm thì khả năng đã nằm sai cluster.
  - ⇨ Với điểm dữ liệu có giá trị > 0.5, có thể coi là khả năng cao nằm đúng cụm, nhỏ hơn thì có ý nghĩa ngược lại
  - ⇨ Số lượng điểm dữ liệu sau khi chạy mô hình phân khúc được phân chia đồng đều, không quá chênh lệch cũng là 1 yếu tố để có thể đánh giá mô hình

II.2.5. Deploy Model

- Chạy ra danh sách kết quả phân cụm tương ứng và kết quả này được lưu trữ tại bảng CINS_MODEL_RSLT theo RPT_DT – kỳ báo cáo đã chạy, có thể tracking được lịch sử phân cụm khách hàng qua từng kỳ báo cáo đã chạy mô hình

- Trong quá trình tiến hành chạy mô hình, giá trị của mỗi Feature có thể thay đổi theo thời gian, đồng nghĩa với việc chạy thuật toán xử lý xây dựng mô hình sẽ cho ra kết quả phân cụm khách hàng thay đổi, không giống với kết quả ban đầu.

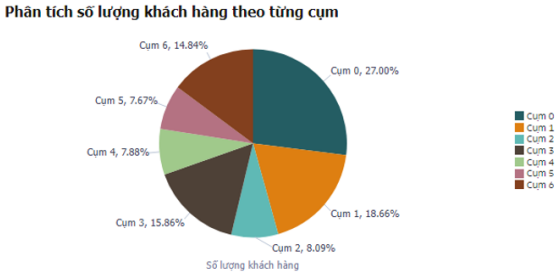
II.3. Đầu ra

- Đầu ra mô hình là danh sách phân khúc khách hàng ứng với kết quả mô hình – lưu trữ tại bảng CINS_MODEL_RSLT, như hình ảnh dưới đây :

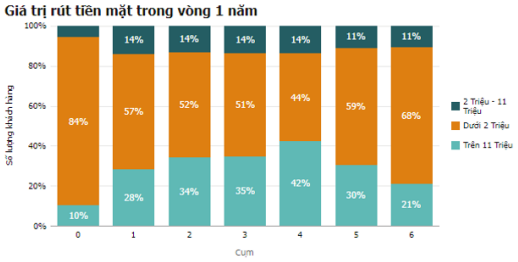
	MODEL_RSLT_KEY	MODEL_NM	PID	CUSTOMER_CDE	RSLT	RPT_DT	CONFIDENCE	ADD_TSTP
1	125936	SEGMENTATION CUSTOMER VALUE	CUST_VAL_01	16175272	UNSTABLE CUST	16-01-2023	(null)	06-APR-23 08.06.58.096277000 AM
2	125941	SEGMENTATION CUSTOMER VALUE	CUST_VAL_01	16175286	NORMAL CUST - LOW TOI	16-01-2023	(null)	06-APR-23 08.06.58.096277000 AM
3	125955	SEGMENTATION CUSTOMER VALUE	CUST_VAL_01	16175431	CONDITIONAL CUST - LOW TOI	16-01-2023	(null)	06-APR-23 08.06.58.096277000 AM
4	125961	SEGMENTATION CUSTOMER VALUE	CUST_VAL_01	16175447	NORMAL CUST - LOW TOI	16-01-2023	(null)	06-APR-23 08.06.58.096277000 AM
5	125963	SEGMENTATION CUSTOMER VALUE	CUST_VAL_01	16175451	NORMAL CUST - LOW TOI	16-01-2023	(null)	06-APR-23 08.06.58.096277000 AM
6	125981	SEGMENTATION CUSTOMER VALUE	CUST_VAL_01	16175527	NORMAL CUST - LOW TOI	16-01-2023	(null)	06-APR-23 08.06.58.096277000 AM
7	125990	SEGMENTATION CUSTOMER VALUE	CUST_VAL_01	16180031	UNSTABLE CUST	16-01-2023	(null)	06-APR-23 08.06.58.096277000 AM
8	125997	SEGMENTATION CUSTOMER VALUE	CUST_VAL_01	1618006	HIGH TXN AMOUNT CUST - LONG TERM - LOW TOI	16-01-2023	(null)	06-APR-23 08.06.58.096277000 AM
9	126778	SEGMENTATION CUSTOMER VALUE	CUST_VAL_01	16233184	UNSTABLE CUST	16-01-2023	(null)	06-APR-23 08.06.58.096277000 AM
10	126788	SEGMENTATION CUSTOMER VALUE	CUST_VAL_01	16233211	NORMAL CUST - LOW TOI	16-01-2023	(null)	06-APR-23 08.06.58.096277000 AM
11	126801	SEGMENTATION CUSTOMER VALUE	CUST_VAL_01	16233441	UNSTABLE CUST	16-01-2023	(null)	06-APR-23 08.06.58.096277000 AM
12	126813	SEGMENTATION CUSTOMER VALUE	CUST_VAL_01	16233472	NORMAL CUST - LOW TOI	16-01-2023	(null)	06-APR-23 08.06.58.096277000 AM
13	126815	SEGMENTATION CUSTOMER VALUE	CUST_VAL_01	16233484	UNSTABLE CUST	16-01-2023	(null)	06-APR-23 08.06.58.096277000 AM
14	126822	SEGMENTATION CUSTOMER VALUE	CUST_VAL_01	16238397	NORMAL CUST - LOW TOI	16-01-2023	(null)	06-APR-23 08.06.58.096277000 AM
15	126824	SEGMENTATION CUSTOMER VALUE	CUST_VAL_01	16238407	UNSTABLE CUST	16-01-2023	(null)	06-APR-23 08.06.58.096277000 AM

- Chân dung từng cụm khách hàng:

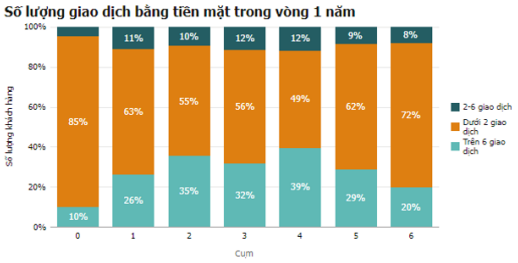
II.3.1. SEGMENTATION TXN BEHAVIORS



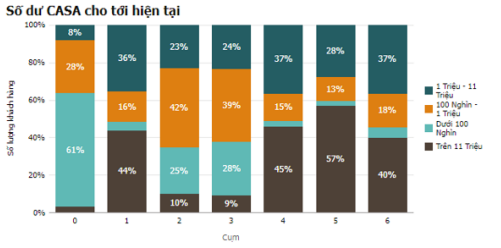
- Cụm 0: Số lượng KH chiếm 27%
- Cụm 1: Số lượng KH chiếm 18.66%
- Cụm 2: Số lượng KH chiếm 8.09%
- Cụm 3: Số lượng KH chiếm 15.86%
- Cụm 4: Số lượng KH chiếm 7.88%
- Cụm 5: Số lượng KH chiếm 7.68%
- Cụm 6: Số lượng KH chiếm 14.84%

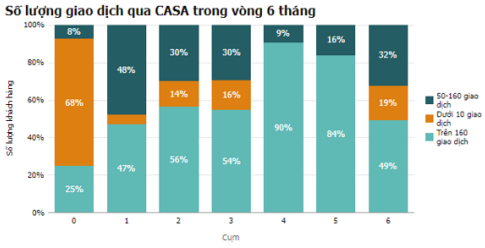


- Cụm 0: Số lượng KH rút tiền mặt có giá trị thấp nhất trong các nhóm (84%) với giá trị dưới 2 tr/năm
- Cụm 4: Số lượng KH có xu hướng rút tiền mặt có giá trị cao nhất trong các nhóm (42%) với giá trị trên 11tr/ năm



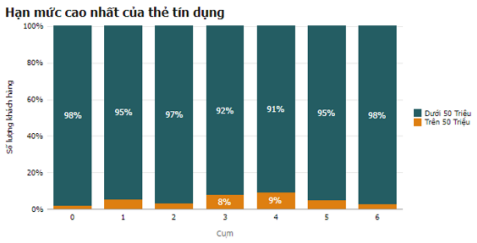
- Cụm 0: Số lượng KH thực hiện giao dịch bằng tiền mặt thấp nhất trong các nhóm (85%) với dưới 2 giao dịch/ năm
- Cụm 4: Số lượng KH thực hiện giao dịch bằng tiền mặt cao (39%) với trên 6 giao dịch/ năm

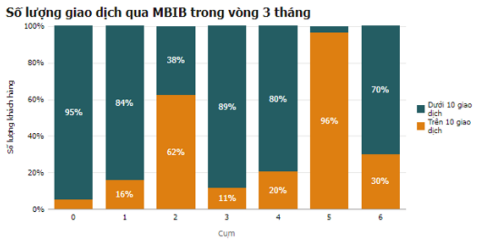




- Cụm 0: Số lượng KH có số dư CASA thấp (61% KH có số dư CASA dưới 100K)
- Cụm 2: Số lượng KH có số dư CASA thấp (42% KH có số dư CASA 100K – 1 Triệu)
- Cụm 4,6: Số lượng KH có số dư CASA mức trung bình (37% KH có số dư CASA 1 Triệu – 11 Triệu)
- Cụm 5: Số lượng KH có số dư CASA cao (57% KH có số dư CASA trên 11 Triệu)

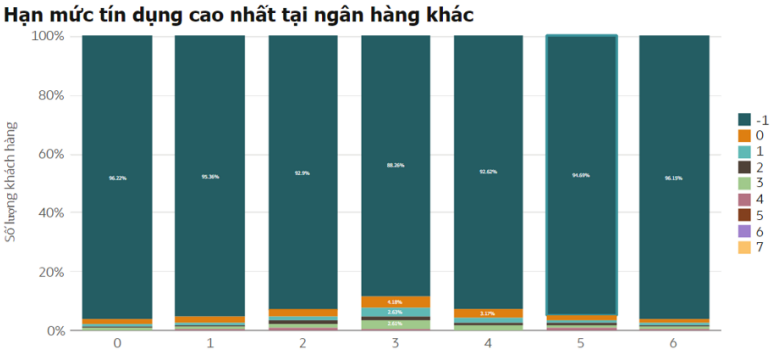
- Cụm 0: Số lượng KH thực hiện giao dịch qua CASA thấp (68% dưới 10 giao dịch)
- Cụm 1: Số lượng KH thực hiện giao dịch qua CASA mức trung bình (48% 50-160 giao dịch)
- Cụm 4: Số lượng KH thực hiện giao dịch qua CASA mức cao (90% trên 160 giao dịch)
- Cụm 5: Số lượng KH thực hiện giao dịch qua CASA mức cao (84% trên 160 giao dịch)



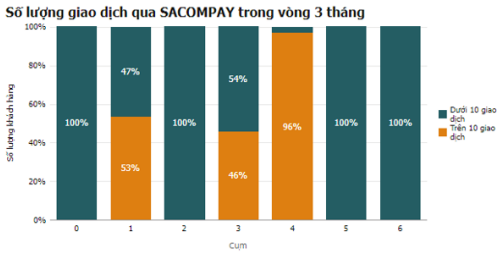


- Cụm 0,6: Số lượng KH có hạn mức cao nhất thẻ tín dụng ở mức thấp (98% KH có hạn mức cao nhất dưới 50 Triệu)
- Cụm 4: Số lượng KH có hạn mức cao nhất thẻ tín dụng ở mức cao (9% KH có hạn mức cao nhất trên 50 Triệu)

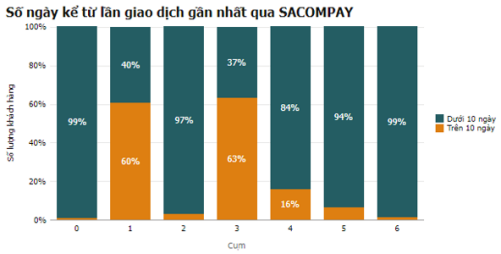
- Cụm 0: Số lượng KH thực hiện giao dịch qua MBIB thấp (95% dưới 10 giao dịch)
- Cụm 5: Số lượng KH thực hiện giao dịch qua MBIB mức trung bình/cao (96% trên 10 giao dịch)



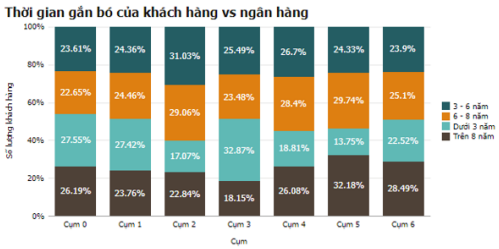
- Đa số khách hàng tại các cụm không có hạn mức tín dụng cao tại các ngân hàng khác
- Cụm 3: Có số lượng nhỏ khách hàng có hạn mức tín dụng cao tại ngân hang khác

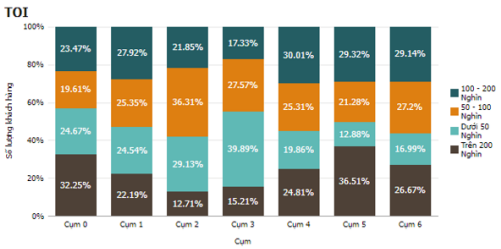


- Cụm 0,2,5,6: Số lượng KH thực hiện giao dịch qua SACOMPAY ở mức thấp (100% dưới 10 giao dịch)
- Cụm 4: Số lượng KH thực hiện giao dịch qua SACOMPAY ở mức trung bình/cao (96% trên 10 giao dịch)



- Cụm 0,6: Số lượng KH có khoảng thời gian ngắn không thực hiện giao dịch qua SACOMPAY (99% dưới 10 ngày)
- Cụm 3: Số lượng KH có khoảng thời gian dài không thực hiện giao dịch qua SACOMPAY (63% trên 10 ngày)





- Cụm 0: Thời gian gắn bó của KH vs ngân hàng đa số là Dưới 3 năm (27.55%), đóng góp TOI ở mức cao
- Cụm 4: Thời gian gắn bó của KH vs ngân hàng đa số là từ 6 – 8 năm (28.4%), đóng góp TOI ở mức thấp
- Cụm 5: Thời gian gắn bó của KH vs ngân hàng đa số là Trên 8 năm (32.18%), đóng góp TOI ở mức cao
- Cụm 6: Thời gian gắn bó của KH vs ngân hàng đa số là Trên 8 năm (28.49%), đóng góp TOI ở mức thấp

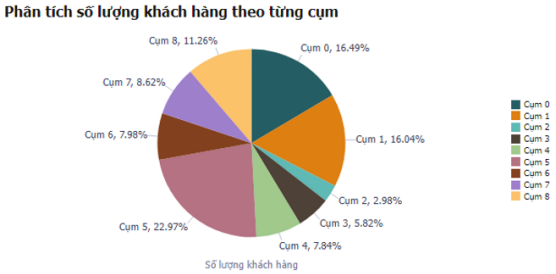
Kết luận cho từng cụm Khách hàng

Cụm	Định nghĩa KH	Hành vi KH	Nhóm KH

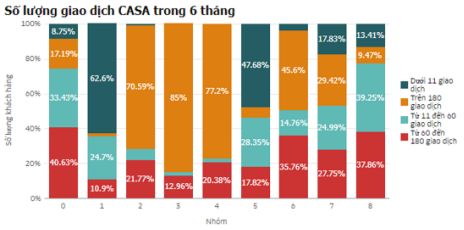
0	- Thuộc thế hệ Gen X - Thời gian gắn bó với ngân hàng < 3 năm - Hạn mức thẻ tín dụng thấp	- Ít giao dịch, rút tiền mặt - Ít giao dịch và có số dư CASA thấp - Ít sử dụng MBIB & Sacompay - Đóng góp TOI cao	Khách hàng ít thực hiện giao dịch
4	- Thuộc thế hệ Gen Y - Thời gian gắn bó với ngân hàng 6 – 8 năm - Hạn mức thẻ tín dụng cao	- Thường xuyên giao dịch, rút tiền mặt - Thường xuyên thực hiện giao dịch qua CASA và có giá trị giao dịch ở mức trung bình/cao - Có xu hướng giao dịch qua Sacompay ở mức trung bình/cao - Đóng góp TOI thấp	Khách hàng chủ động
5	- Thuộc thế hệ Gen Y - Thời gian gắn bó với ngân hàng trên 8 năm - Hạn mức thẻ tín dụng ở mức thấp	- Thường xuyên thực hiện giao dịch qua MBIB - Ít sử dụng Sacompay - Có số dư CASA cao - Đóng góp TOI cao	Khách hàng tiết kiệm gắn bó lâu năm
6	- Thuộc thế hệ Gen Y - Thời gian gắn bó với ngân hàng trên 8 năm - Hạn mức thẻ tín dụng thấp	- Có số dư CASA trung bình - Ít sử dụng Sacompay - Đóng góp TOI thấp	Khách hàng bình thường, lâu năm
1, 2, 3	- Thuộc thế hệ Gen Y - Thời gian gắn bó với ngân hàng phân bố :	- Giao dịch, rút tiền mặt ở mức trung bình - Giao dịch và có số dư CASA	Nhóm KH có hành vi không ổn định

	Dưới 3 năm, 3 – 6 năm	ở mức trung bình - Ít sử dụng MBIB & Sacompay - Đóng góp TOI thấp	

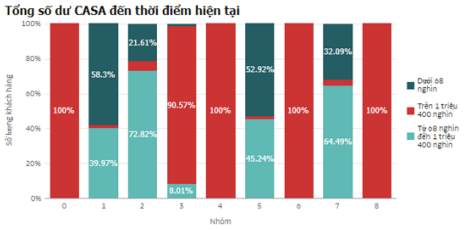
II.3.2. SEGMENTATION CUSTOMER VALUE



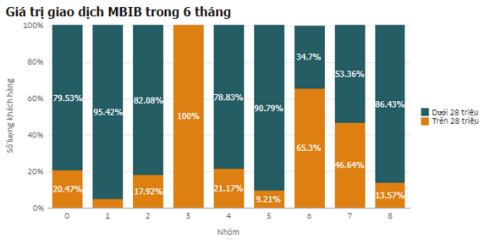
- Cụm 0: Số lượng KH chiếm 16.49%
- Cụm 1: Số lượng KH chiếm 16.04%
- Cụm 2: Số lượng KH chiếm 2.98%
- Cụm 3: Số lượng KH chiếm 5.82%
- Cụm 4: Số lượng KH chiếm 7.84%
- Cụm 5: Số lượng KH chiếm 22.97%
- Cụm 6: Số lượng KH chiếm 7.98%
- Cụm 7: Số lượng KH chiếm 8.62%
- Cụm 8: Số lượng KH chiếm 11.26%



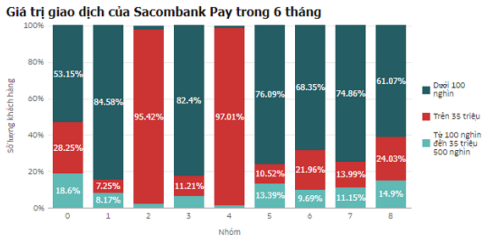
- Cụm 1: Số lượng khách hàng dưới 11 giao dịch qua CASA chiếm tỉ lệ cao nhất trong các cụm (62.6%)
- Cụm 8: Số lượng khách hàng có từ 11 – 60 giao dịch qua CASA chiếm tỉ lệ cao nhất trong các cụm (39.25%)
- Cụm 0: Số lượng khách hàng có từ 60 – 180 giao dịch qua CASA chiếm tỉ lệ cao nhất trong các cụm (40.63%)
- Cụm 3: Số lượng khách hàng có trên 180 giao dịch qua CASA chiếm tỉ lệ cao nhất trong các cụm (85%)



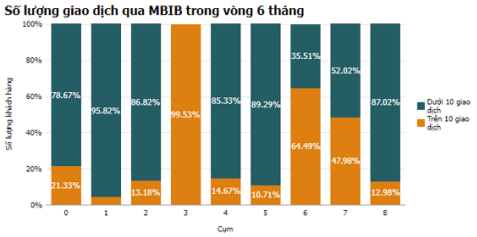
- Cụm 1: Số lượng khách hàng có số dư dưới 68 nghìn chiếm tỉ lệ cao nhất trong các cụm (58.3%)
- Cụm 2: Số lượng khách hàng có số dư 68 nghìn – 1 triệu 400 nghìn chiếm tỉ lệ cao nhất trong các cụm (72.82%)
- Cụm 0,4,6,8: Số lượng khách hàng có số dư trên 1 triệu 400 nghìn chiếm tỉ lệ cao nhất trong các cụm (100%)

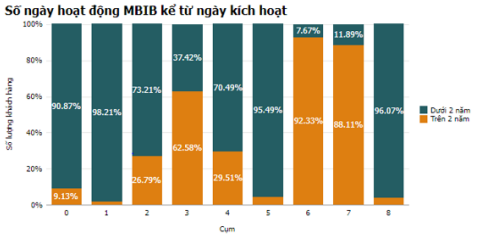


- Cụm 1: Số lượng khách hàng có giá trị giao dịch qua MBIB dưới 28 triệu chiếm tỉ lệ cao nhất trong các cụm (95.42%)
- Cụm 3: Số lượng khách hàng có giá trị giao dịch qua MBIB trên 28 triệu chiếm tỉ lệ cao nhất trong các cụm (100%)



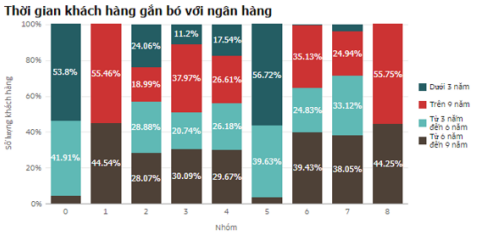
- Cụm 1: Số lượng khách hàng có giá trị giao dịch qua Sacompay dưới 100 nghìn chiếm tỉ lệ cao nhất trong các cụm (84.58%)
- Cụm 0: Số lượng khách hàng có giá trị giao dịch qua Sacompay từ 100 nghìn – 35 triệu 500 nghìn chiếm tỉ lệ cao nhất trong các cụm (18.6%)
- Cụm 4: Số lượng khách hàng có giá trị giao dịch qua Sacompay trên 35 triệu chiếm tỉ lệ cao nhất trong các cụm (97.01%)

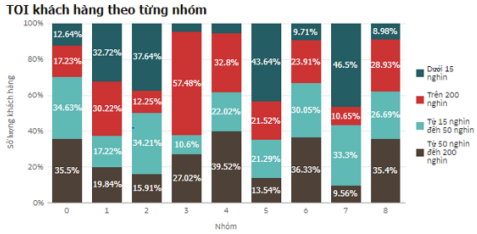




- Cụm 1: Số lượng khách hàng có dưới 10 giao dịch chiếm tỉ lệ cao nhất trong các cụm (95.82%)
- Cụm 3: Số lượng khách hàng có trên 10 giao dịch chiếm tỉ lệ cao nhất trong các cụm (99.53%)

- Cụm 1: Số lượng khách hàng có số ngày hoạt động MBIB kể từ ngày kích hoạt dưới 2 năm chiếm tỉ lệ cao nhất trong các cụm (98.21%)
- Cụm 6: Số lượng khách hàng có số ngày hoạt động MBIB kể từ ngày kích hoạt trên 2 năm chiếm tỉ lệ cao nhất trong các cụm (92.33%)





- Cụm 0: Số lượng khách hàng có thời gian gắn bó vs ngân hàng dưới 3 năm chiếm tỉ lệ cao nhất (53.8%), đóng góp TOI từ 50 nghìn – 200 nghìn (35.5%)
- Cụm 1: Số lượng khách hàng có thời gian gắn bó vs ngân hàng trên 9 năm chiếm tỉ lệ cao nhất (55.46%), đóng góp TOI dưới 15 nghìn (32.72%)
- Cụm 2: Số lượng khách hàng có thời gian gắn bó vs ngân hàng 3 - 6 năm chiếm tỉ lệ cao nhất (28.88%), đóng góp TOI dưới 15 nghìn (37.64%)
- Cụm 3: Số lượng khách hàng có thời gian gắn bó vs ngân hàng trên 9 năm chiếm tỉ lệ cao nhất (37.97%), đóng góp TOI trên 200 nghìn (57.48%)
- Cụm 4: Số lượng khách hàng có thời gian gắn bó vs ngân hàng 6 - 9 năm chiếm tỉ lệ cao nhất (29.67%), đóng góp TOI trên 50 nghìn - 200 nghìn (39.52%)
- Cụm 6: Số lượng khách hàng có thời gian gắn bó vs ngân hàng 6 - 9 năm chiếm tỉ lệ cao nhất (39.43%), đóng góp TOI trên 50 nghìn - 200 nghìn (36.33%)
- Cụm 8: Số lượng khách hàng có thời gian gắn bó vs ngân hàng trên 9 năm chiếm tỉ lệ cao nhất (55.75%), đóng góp TOI trên 50 nghìn - 200 nghìn (35.4%)

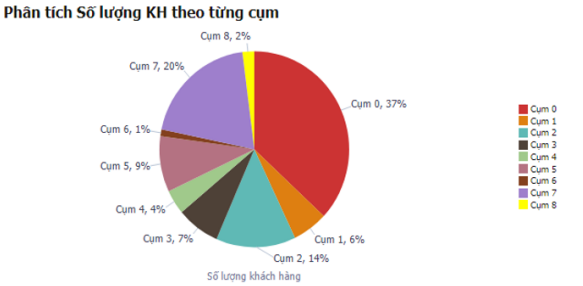
Kết luận cho từng cụm Khách hàng

Cụm	Định nghĩa KH	Tài chính của KH	Nhóm KH
0	- Thuộc thế hệ Gen Y - Thời gian gắn bó với ngân hàng dưới 3 năm	- Thực hiện giao dịch qua CASA ở mức trung bình, có số dư CASA ở mức cao - Giá trị giao dịch qua Sacompay ở mức trung bình - Đóng góp TOI thấp	Khách hàng bình thường, đóng góp TOI thấp
1	- Thuộc thế hệ Gen X - Thời gian gắn bó với ngân hàng trên 9	- Thực hiện giao dịch qua CASA ở mức thấp, có số dư CASA ở mức thấp - Có giá trị giao dịch qua	Khách hàng có giá trị giao dịch thấp, lâu năm , đóng góp TOI thấp

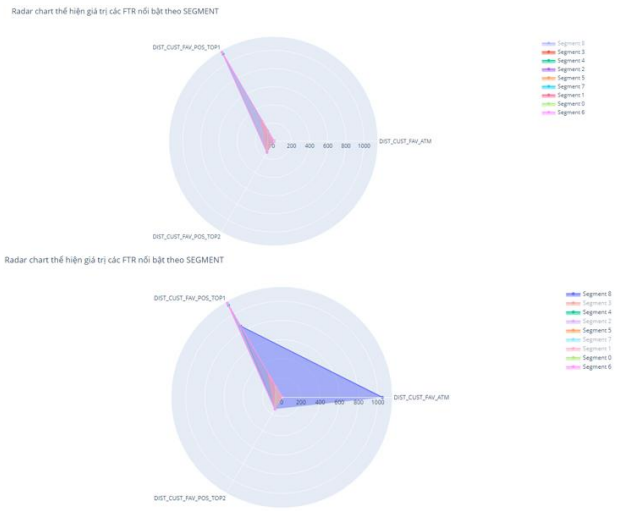
	năm	MBIB, Sacompay ở mức thấp - Đóng góp TOI thấp	
3	- Thuộc thế hệ Gen Y - Thời gian gắn bó với ngân hàng trên 9 năm	- Thường xuyên thực hiện giao dịch qua CASA - Thường xuyên thực hiện giao dịch qua MBIB, có giá trị giao dịch qua MBIB cao - Đóng góp TOI cao	Khách hàng VIP lâu năm, đóng góp TOI cao
4	- Thuộc thế hệ Gen Y - Thời gian gắn bó với ngân hàng dưới 6 – 9 năm	- Có số dư CASA cao - Giá trị giao dịch qua Sacompay ở mức cao - Đóng góp TOI thấp	Khách hàng có điều kiện, đóng góp TOI thấp
8	- Thuộc thế hệ Gen Y - Thời gian gắn bó với ngân hàng trên 9 năm	- Số dư CASA cao - Số lượng giao dịch qua CASA ở mức trung bình - Đóng góp TOI thấp	Khách hàng chủ động lâu năm, đóng góp TOI thấp
6	- Thuộc thế hệ Gen Y - Thời gian gắn bó với ngân hàng 6 – 9 năm	- Thời gian sử dụng MBIB mức trung bình/lâu năm - Số dư CASA cao - Đóng góp TOI thấp	Khách hàng có xu hướng sử dụng công nghệ giao dịch, có điều kiện, đóng góp TOI thấp
2,5,7	- Thuộc thế hệ Gen Y - Thời gian gắn bó với ngân hàng phân bố từ mới – lâu năm	- Số lượng giao dịch qua CASA phân bố ít – nhiều - Gía trị giao dịch qua MBIB thấp - Giá trị giao dịch qua Sacompay phân bố thấp -	Khách hàng có giá trị và hành vi không ổn định

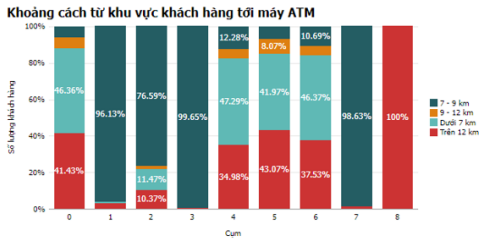
		cao - Đóng góp TOI thấp	

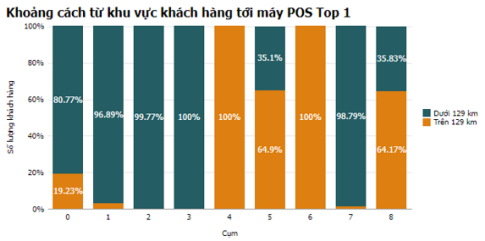
II.3.3. SEGMENTATION GEORAPHIC



- Cụm 0: Số lượng KH chiếm 37%
- Cụm 1: Số lượng KH chiếm 6%
- Cụm 2: Số lượng KH chiếm 14%
- Cụm 3: Số lượng KH chiếm 7%
- Cụm 4: Số lượng KH chiếm 4%
- Cụm 5: Số lượng KH chiếm 9%
- Cụm 6: Số lượng KH chiếm 1%
- Cụm 7: Số lượng KH chiếm 20%
- Cụm 8: Số lượng KH chiếm 2%

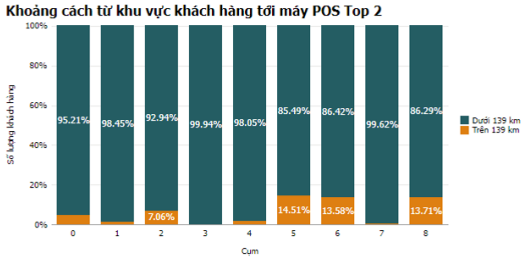




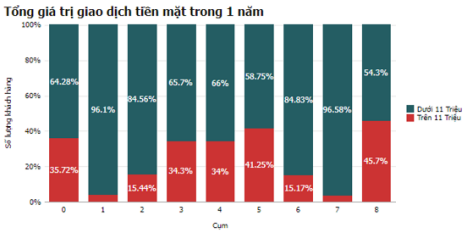


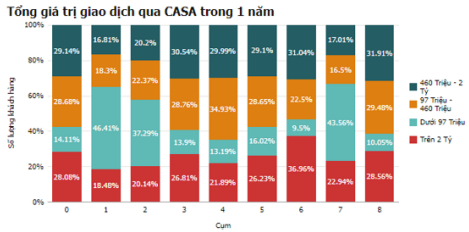
- Cụm 3: Số lượng KH có xu hướng thực hiện giao dịch qua ATM vs khoảng cách từ 7 – 9 km chiếm tỉ lệ cao nhất trong các nhóm (99.65%)
- Cụm 4: Số lượng KH có xu hướng thực hiện giao dịch qua ATM gần nhà chiếm tỉ lệ cao nhất trong các nhóm (47.29%) với khoảng cách dưới 7 km
- Cụm 5: Số lượng KH có xu hướng thực hiện giao dịch qua ATM vs khoảng cách từ 9 – 12 km chiếm tỉ lệ cao nhất trong các nhóm (8.07%)
- Cụm 8: Số lượng KH có xu hướng thực hiện giao dịch qua ATM xa nhà vs khoảng cách trên 12 km chiếm tỉ lệ cao nhất trong các nhóm (100%)

- Cụm 3: Số lượng KH có xu hướng thực hiện giao dịch, mua sắm tại máy POS ưa thích Top 1 chiếm tỉ lệ cao nhất trong các nhóm vs khoảng cách dưới 129 km (100%)
- Cụm 4, 6: Số lượng KH có xu hướng thực hiện giao dịch, mua sắm tại máy POS ưa thích Top 1 chiếm tỉ lệ cao nhất trong các nhóm vs khoảng cách trên 129 km (100%)



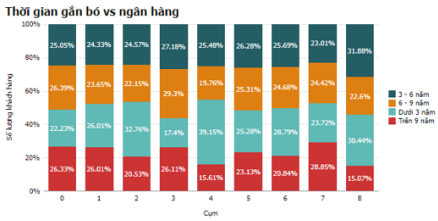
- Cụm 3: Số lượng KH có xu hướng thực hiện giao dịch, mua sắm tại máy POS ưa thích Top 2 chiếm tỉ lệ cao nhất trong các nhóm vs khoảng cách dưới 139 km (99.94%)
- Cụm 5: Số lượng KH có xu hướng thực hiện giao dịch, mua sắm tại máy POS ưa thích Top 2 chiếm tỉ lệ cao nhất trong các nhóm vs khoảng cách trên 139 km (14.51%)

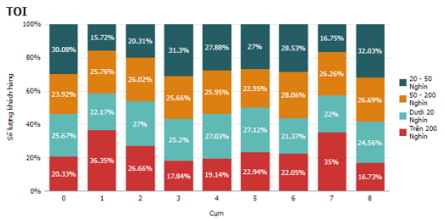




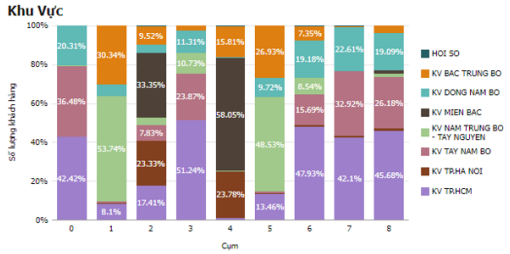
- Cụm 7: Số lượng KH có giá trị giao dịch bằng tiền mặt dưới 11 Triệu chiếm tỉ lệ cao nhất trong các nhóm (96.58%)
- Cụm 8: Số lượng KH có giá trị giao dịch bằng tiền mặt trên 11 Triệu chiếm tỉ lệ cao nhất trong các nhóm (45.7%)

- Cụm 1: Số lượng KH có giá trị giao dịch qua CASA dưới 97 triệu chiếm tỉ lệ cao nhất trong các nhóm (46.41%)
- Cụm 4: Số lượng KH có giá trị giao dịch qua CASA 97 - 460 triệu chiếm tỉ lệ cao nhất trong các nhóm (34.93%)
- Cụm 8: Số lượng KH có giá trị giao dịch qua CASA 460 Triệu – 2 Tỷ chiếm tỉ lệ cao nhất trong các nhóm (31.91%)
- Cụm 6: Số lượng KH có giá trị giao dịch qua CASA Trên 2 Tỷ chiếm tỉ lệ cao nhất trong các nhóm (36.96%)





- Cụm 1: Số lượng KH có thời gian gắn bó vs ngân hàng dưới 3 năm và trên 9 năm chiếm tỉ lệ cao nhất (26.01%), đóng góp TOI ở mức cao trên 200 nghìn (36.35%)
- Cụm 3: Số lượng KH có thời gian gắn bó vs ngân hàng 6 – 9 năm chiếm tỉ lệ cao nhất (29.3%), đóng góp TOI ở mức thấp từ 20 – 50 nghìn (31.3%)
- Cụm 4: Số lượng KH có thời gian gắn bó vs ngân hàng dưới 3 năm chiếm tỉ lệ cao nhất (39.15%), đóng góp TOI ở mức thấp từ 20 – 50 nghìn (27.88%)
- Cụm 5: Số lượng KH có thời gian gắn bó vs ngân hàng 3 – 6 năm chiếm tỉ lệ cao nhất (26.28%), đóng góp TOI ở mức thấp dưới 20 nghìn (27.12%)
- Cụm 6: Số lượng KH có thời gian gắn bó vs ngân hàng dưới 3 năm chiếm tỉ lệ cao nhất (28.79%), đóng góp TOI ở mức thấp 20 - 50 nghìn (28.53%)
- Cụm 8: Số lượng KH có thời gian gắn bó vs ngân hàng 3 – 6 năm chiếm tỉ lệ cao nhất (31.88%), đóng góp TOI ở mức thấp 20 - 50 nghìn (32.03%)



- Cụm 1: Chủ yếu KH đến từ khu vực Nam Trung Bộ - Tây Nguyên (chiếm 53.74%)
- Cụm 3: Chủ yếu KH đến từ khu vực TP Hồ Chí Minh (chiếm 51.24%)
- Cụm 4: Chủ yếu KH đến từ khu vực Miền Bắc (chiếm 58.05%)
- Cụm 5: Chủ yếu KH đến từ khu vực Nam Trung Bộ - Tây Nguyên (chiếm 48.53%)
- Cụm 6: Chủ yếu KH đến từ khu vực TP Hồ Chí Minh (chiếm 47.93%)
- Cụm 8: Chủ yếu KH đến từ khu vực TP Hồ Chí Minh (chiếm 45.68%)

Kết luận cho từng cụm Khách hàng

Cụm	Định nghĩa KH	Phân bố địa lý – đặc trưng KH	Nhóm KH
1	- Thuộc thế hệ Gen Y - Thời gian gắn bó với ngân hàng phân bố bao gồm dưới 3 năm và trên 9 năm	- Có giá trị giao dịch CASA ở mức thấp trong vòng 1 năm - Đóng góp TOI cao - Giao dịch. mua sắm gần khu vực mình sống - Tập trung chủ yếu khu vực Nam Trung Bộ - Tây Nguyên	Khách hàng có xu hướng thích giao dịch, mua sắm gần nhà, tập trung chủ yếu khu vực Nam Trung Bộ - Tây Nguyên
3	- Thuộc thế hệ Gen Y - Thời gian gắn bó với ngân hàng 6 – 9 năm	- Có xu hướng thực hiện giao dịch qua ATM, máy POS gần khu vực mình sống - Đóng góp TOI thấp - Tập trung chủ yếu khu vực TP Hồ Chí Minh	Khách hàng bình thường xu hướng thích giao dịch, mua sắm gần nhà, tập trung chủ yếu tại TP HCM
4	- Thuộc thế hệ Gen Y	- Có xu hướng giao dịch gần nhà, sẵn sang đi	Khách hàng bình thường xu hướng thích giao dịch gần nhà,

	- Thời gian gắn bó với ngân hàng dưới 3 năm	xa để mua sắm - Đóng góp TOI thấp - Có giá trị giao dịch qua CASA ở mức Trung bình - Tập trung chủ yếu khu vực Miền Bắc	sẵn sàng đi xa mua sắm, tập trung chủ yếu khu vực Miền Bắc
5	- Thuộc thế hệ Gen Y - Thời gian gắn bó với ngân hàng 3 – 6 năm	- Có xu hướng giao dịch gần nhà, sẵn sang đi xa để mua sắm - Đóng góp TOI thấp - Có giá trị giao dịch qua CASA ở mức cao - Tập trung chủ yếu khu vực Nam Trung Bộ - Tây Nguyên	Khách hàng giao dịch với giá trị cao, xu hướng giao dịch gần nhà, đi xa mua sắm, tập trung chủ yếu khu vực Nam Trung Bộ - Tây Nguyên
6	- Thuộc thế hệ Gen Y - Thời gian gắn bó với ngân hàng dưới 3 năm	- Có xu hướng giao dịch xa nhà - Đóng góp TOI thấp - Có giá trị giao dịch qua CASA ở mức cao - Tập trung chủ yếu khu vực TP HCM	Khách hàng giao dịch với giá trị cao, xu hướng giao dịch xa nhà, tập trung chủ yếu TP HCM
8	- Thuộc thế hệ Gen Y - Thời gian gắn bó với ngân hàng 3 – 6 năm	- Có xu hướng giao dịch xa nhà - Đóng góp TOI thấp - Có giá trị giao dịch qua CASA ở mức cao - Có giá trị giao dịch bằng tiền mặt ở mức	Khách hàng có điều kiện, xu hướng giao dịch xa nhà, tập trung chủ yếu TP HCM

		cao - Tập trung chủ yếu khu vực TP HCM	
0,2,7	- Thuộc thế hệ Gen X, Gen Y - Thời gian gắn bó phân bố dưới 3 năm, 6 – 9 năm, trên 9 năm	- Bao gồm cả xu hướng giao dịch, mua sắm gần lẫn xa nhà - Ít thực hiện giao dịch bằng tiền mặt - Đóng góp TOI không ổn định - Tập trung chủ yếu khu vực TP HCM, Miền Bắc	Khách hàng phân bố rải rác tại các vị trí khác nhau như TP HCM & Miền Bắc

III. THIẾT KẾ CSDL PHỤC VỤ CHO MÔ HÌNH

Nhằm quản lý và tổng hợp các rule dùng để tổng hợp Feature cho bài toán – Bảng CINS_FTR_DIM sẽ được lưu trữ tại tầng SMY. Bộ dữ liệu huấn luyện được tổng hợp từ các nguồn dữ liệu đã được làm sạch và theo những quy tắc cụ thể bằng cách khai báo rule dưới dạng item – value. Người dùng thực hiện khai báo rule theo quy tắc bao gồm các trường thông tin như sau:

- FTR_CD: Mã Feature theo các dạng Basic, Aggregated, Specific, Derivative-basic...
- FTR_NM: Tên Feature
- GRP: Tên nhóm dữ liệu
- SUB_GRP: Tên nhóm dữ liệu lvl2
- DESC: Mô tả Feature
- EXPS: Câu lệnh SQL tổng hợp dữ liệu từ các nguồn ứng với từng Feature type
- COND: Điều kiện lấy dữ liệu
- FLTR_CD: Điều kiện sửa, lọc dữ liệu sau khi đã tổng hợp từ nguồn SMY
- FILL_CD: Làm bù dữ liệu trong trường hợp Feature bị null. Một số giá trị dùng để làm bù như 'mean', 'median', 'mode', 'max', 'min'
- LABEL: Là khóa đánh dấu nếu nhãn bài toán cần tổng hợp dữ liệu từ nguồn
- UNIT : Đơn vị của Feature
- FTR_TP: Đánh dấu loại Feature
- MODEL_NM: Tên mô hình
- ADD_TSTP: Là trường ghi lại thời gian rule được khai báo vào bảng
- ACTIVE: Là khóa đánh dấu nếu người dùng muốn vô hiệu hóa việc tổng hợp 1 Feature nào đó.

Ngoài ra, các bảng còn lại trong hệ thống được mô tả chi tiết dưới đây:

III.1. CINS_FEATURE_STORE

- Bảng CINS_FEATURE_STORE sẽ lưu trữ bộ dữ liệu (tổng hợp được từ nguồn) phục vụ cho việc chạy mô hình.

TBL_NM	TBL DESC	FIELD	FIELD DESC	DATA TYPE	KEY	NOT NULL	DEFAULT	NOTE
CINS_FEATURE_STORE	Lưu trữ bộ dữ liệu (tổng hợp được từ nguồn) phục vụ cho việc chạy mô hình.	CUSTOMER_CDE	Mã Khách hàng	VARCHAR2(20)		X		
CINS_FEATURE_STORE		FTR_NM	Tên Feature	VARCHAR2(200)				
CINS_FEATURE_STORE		FTR_VAL	Giá trị Feature	VARCHAR2(2000)				
CINS_FEATURE_STORE		RPT_DT	Kỳ chạy	VARCHAR2(20)			Định dạng: 01-12-2022	
CINS_FEATURE_STORE		ADD_TSTP	Thời điểm nhấn vào bảng	TIMESTAMP				

III.2. CINS_FEATURE_STORE_DERIVED

- Bảng CINS_FEATURE_STORE_DERIVED sẽ lưu trữ bộ dữ liệu – phái sinh từ các Feature gốc đã được tổng hợp từ bảng CINS_FEATURE_STORE phục vụ cho việc chạy mô hình

TBL_NM	TBL DESC	FIELD	FIELD DESC	DATA TYPE	KEY	NOT NULL	DEFAULT	NOTE
CINS_FEATURE_STORE_DERIVED	Lưu trữ bộ dữ liệu – phái sinh từ các Feature gốc đã được tổng hợp từ bảng CINS_FEATURE_STORE phục vụ cho việc chạy mô hình	CUSTOMER_CDE	Mã Khách hàng	VARCHAR2(20)		X		
CINS_FEATURE_STORE_DERIVED		FTR_NM	Tên Feature	VARCHAR2(200)				
CINS_FEATURE_STORE_DERIVED		FTR_VAL	Giá trị Feature	VARCHAR2(2000)				
CINS_FEATURE_STORE_DERIVED		RPT_DT	Kỳ chạy	VARCHAR2(20)			Định dạng: 01-12-2022	
CINS_FEATURE_STORE_DERIVED		ADD_TSTP	Thời điểm nhấn vào bảng	TIMESTAMP				

III.3. CINS_MODEL_RSLT

- Bảng CINS_MODEL_RSLT sẽ lưu trữ toàn bộ kết quả của mô hình

TBL_NM	TBL DESC	FIELD	FIELD DESC	DATA TYPE	KEY	NOT NULL	DEFAULT	NOTE
CINS_MODEL_RSLT	Lưu trữ kết quả mô hình	MODEL_RSLT_KEY	Id tự tăng	INTERGER	X			
CINS_MODEL_RSLT		MODEL_NM	Tên mô hình khái quát	VARCHAR2(200)			VD: SEGMENTATION - GEO, REACTIVATE…	
CINS_MODEL_RSLT		PID	Process id lấy từ lần chạy python	VARCHAR2(100)				
CINS_MODEL_RSLT		CUSTOMER_CDE	Mã Khách hàng	VARCHAR2(20)		X		
CINS_MODEL_RSLT		RSLT	Kết quả mô hình	VARCHAR2(200)				
CINS_MODEL_RSLT		RPT_DT	Kỳ chạy	VARCHAR2(20)			Định dạng: 01-12-2022	
CINS_MODEL_RSLT		CONFIDENCE	Độ tự tin	NUMBER(20,6)			Giữ giá trị float nếu là dạng %, không *100	
CINS_MODEL_RSLT		ADD_TSTP	Thời điểm nhấn vào bảng	TIMESTAMP				

III.4. CINS_MODEL_EVAL

- Bảng CINS_MODEL_EVAL lưu trữ kết quả đánh giá mô hình theo metric

TBL_NM	TBL DESC	FIELD	FIELD DESC	DATA TYPE	KEY	NOT NULL	DEFAULT	NOTE
CINS_MODEL_EVAL	Lưu trữ kết quả đánh giá mô hình theo metric	MODEL_EVAL_KEY	Id tự tăng	INTERGER	X			

CINS_MODEL_EVAL	MODEL_NM	Tên mô hình khái quát	VARCHAR2(200)			VD: SEGMENTATION - GEO, REACTIVATE…	
CINS_MODEL_EVAL	PID	Process id lấy từ lần chạy python	VARCHAR2(100)				
CINS_MODEL_EVAL	SUBSET		VARCHAR2(200)			Nhằm lưu giá trị metric đối với nhóm nhỏ như silhouette score riêng từng cluster , nếu là metric chung thì để trống	
CINS_MODEL_EVAL	METRIC	Tên metric theo chuẩn quốc tế, viết hoa	VARCHAR2(200)			VD: ACCURACY, RMSE…	
CINS_MODEL_EVAL	METRIC_VAL	Giá trị metric đo được	NUMBER(20,6)			Giữ giá trị float nếu là dạng %, không *100	
CINS_MODEL_EVAL	RPT_DT	Kỳ chạy	VARCHAR2(20)			Định dạng: 01-12-2022	
CINS_MODEL_EVAL	ADD_TSTP	Thời điểm nhấn vào bảng	TIMESTAMP				

III.5. CINS_FTR_DIM

- Lưu trữ thông tin bộ Feature theo từng bài toán

TBL_NM	TBL DESC	FIELD	FIELD DESC	DATA TYPE	KEY	NOT NULL	DEFAULT	NOTE
CINS_FTR_DIM	Lưu trữ thông tin bộ Feature theo từng bài	FTR_DIM_KEY	Id tự tăng	INTERGER	X			
CINS_FTR_DIM		FTR_CD	Mã Feature theo quy chuẩn	VARCHAR2(20)				
CINS_FTR_DIM		FTR_NM	Tên Feature	VARCHAR2(200)				

CINS_FTR_DIM	toán	GRP	Nhóm Feature	VARCHAR2(200)				
CINS_FTR_DIM		SUB_GRP	Nhóm nhỏ	VARCHAR2(200)				
CINS_FTR_DIM		DESC	Mô tả Feature	NVARCHAR2(20000)				
CINS_FTR_DIM		EXPS	Code tổng hợp ra Feature	NVARCHAR2(20000)				
CINS_FTR_DIM		COND	Điều kiện lấy	NVARCHAR2(20000)				
CINS_FTR_DIM		FLTR_CD	Điều kiện làm sạch	NVARCHAR2(20000)				
CINS_FTR_DIM		FILL_CD	Giá trị làm bù	VARCHAR2(2000)				
CINS_FTR_DIM		LABEL	Là Feature nhãn hay không?	INTERGER			0	
CINS_FTR_DIM		UNIT	Đơn vị của Feature	NVARCHAR2(2000)				Nếu có unit là dạng số, không unit là dạng chữ
CINS_FTR_DIM		FTR_TP	DERIVED/ROOT	VARCHAR2(20)				
CINS_FTR_DIM		MODEL_NM	Tên mô hình khái quát	VARCHAR2(200)				Phục vụ trường hợp phát sinh cách xử lý khác nhau khi dùng chung Feature, nếu không gặp trường hợp vậy thì không cần khai
CINS_FTR_DIM		ACTIVE	Có còn được dùng cho bài toán không?	INTERGER			1	
CINS_FTR_DIM		ADD_TSTP	Thời điểm nhấn vào bảng	TIMESTAMP				

CINS_FTR_DIM		CHART	Sử dụng biểu thị chart hay không?	VARCHAR2(2)				

III.6. CINS_SPLITTED_TBL

- Lưu trữ danh sách khách hàng cần chạy cho mỗi tập dữ liệu của mỗi bài toán

TBL_NM	TBL DESC	FIELD	FIELD DESC	DATA TYPE	KEY	NOT NULL	DEFAULT	NOTE
CINS_SPLITTED_TBL	Lưu trữ danh sách khách hàng cần chạy cho mỗi tập dữ liệu của mỗi bài toán	MODEL_NM	Tên mô hình khái quát	VARCHAR2(200)	X			
CINS_SPLITTED_TBL		SPLIT_KEY	Điều kiện chia bộ dữ liệu, nếu nhiều điều kiện thì phải lấy bộ Customer_cde thõa mãn tất cả tiêu chí	VARCHAR2(200)				VD: TOI
CINS_SPLITTED_TBL		SPLIT_EXPS	Điều kiện chia bộ dữ liệu, nếu nhiều điều kiện thì phải lấy bộ Customer_cde thõa mãn tất cả tiêu chí	NVARCHAR2(20)				VD: >
CINS_SPLITTED_TBL		SPLIT_VAL	Điều kiện chia bộ dữ liệu, nếu nhiều điều kiện thì phải lấy bộ Customer_cde thõa mãn tất cả tiêu chí	NVARCHAR2(2000)				VD: 100000
CINS_SPLITTED_TBL		CUSTOMER_CDE	Mã Khách hàng	VARCHAR2(20)				
CINS_SPLITTED_TBL		RPT_DT	Kỳ chạy	VARCHAR2(20)				Định dạng 01-11-2022
CINS_SPLITTED_TBL		ADD_TSTP	Thời điểm nhấn vào bảng	TIMESTAMP				

III.7. CINS_CLEANED_TBL

- Lưu trữ bộ Feature đã làm sạch

TBL_NM	TBL DESC	FIELD	FIELD DESC	DATA TYPE	KEY	NOT NULL	DEFAULT	NOTE
CINS_CLEANED_TBL	Lưu trữ bộ Feature đã làm sạch	CUSTOMER_CDE	Mã Khách hàng	VARCHAR2(20)		X		
CINS_CLEANED_TBL		FTR_NM	Tên Feature	VARCHAR2(200)				Phục vụ trường hợp phát sinh cách xử lý khác nhau khi dùng chung Feature, nếu không gặp trường hợp vậy thì không cần
CINS_CLEANED_TBL		FTR_VAL	Giá trị Feature	VARCHAR2(2000)				
CINS_CLEANED_TBL		RPT_DT	Kỳ chạy	VARCHAR2(20)				Định dạng 01-12-2022
CINS_CLEANED_TBL		ADD_TSTP	Thời điểm nhấn vào bảng	TIMESTAMP				

III.8. CINS_FILLED_TBL

- Lưu trữ bộ Feature đã làm bù

TBL_NM	TBL DESC	FIELD	FIELD DESC	DATA TYPE	KEY	NOT NULL	DEFAULT	NOTE
CINS_FILLED_TBL	Lưu trữ bộ Feature đã làm bù	CUSTOMER_CDE	Mã Khách hàng	VARCHAR2(20)	X			
CINS_FILLED_TBL		FTR_NM	Tên Feature	VARCHAR2(200)				
CINS_FILLED_TBL		FTR_VAL	Giá trị Feature	VARCHAR2(2000)				
CINS_FILLED_TBL		RPT_DT	Kỳ chạy	VARCHAR2(20)				Định dạng 01-12-2022
CINS_FILLED_TBL		ADD_TSTP	Thời điểm nhấn vào bảng	TIMESTAMP				

III.9. CINS_ENCODED_TBL

- Lưu trữ bộ Feature đã mã hóa

TBL_NM	TBL DESC	FIELD	FIELD DESC	DATA TYPE	KEY	NOT NULL	DEFAULT	NOTE
CINS_ENCODED_TBL	Lưu trữ bộ Feature đã mã hóa	CUSTOMER_CDE	Mã Khách hàng	VARCHAR2(20)	X			
CINS_ENCODED_TBL		FTR_NM	Tên Feature	VARCHAR2(200)				
CINS_ENCODED_TBL		FTR_VAL	Giá trị Feature	VARCHAR2(2000)				
CINS_ENCODED_TBL		RPT_DT	Kỳ chạy	VARCHAR2(20)				Định dạng 01-12-2022
CINS_ENCODED_TBL		ADD_TSTP	Thời điểm nhấn vào bảng	TIMESTAMP				

III.10. CINS_SCALED_TBL

- Lưu trữ bộ Feature đã co giãn

TBL_NM	TBL DESC	FIELD	FIELD DESC	DATA TYPE	KEY	NOT NULL	DEFAULT	NOTE
CINS_SCALED_TBL	Lưu trữ bộ Feature đã co giãn	CUSTOMER_CDE	Mã Khách hàng	VARCHAR2(20)	X			
CINS_SCALED_TBL		FTR_NM	Tên Feature	VARCHAR2(200)				
CINS_SCALED_TBL		FTR_VAL	Giá trị Feature	VARCHAR2(2000)				
CINS_SCALED_TBL		RPT_DT	Kỳ chạy	VARCHAR2(20)				Định dạng 01-12-2022
CINS_SCALED_TBL		ADD_TSTP	Thời điểm nhấn vào bảng	TIMESTAMP				

III.11. CINS_MODEL_COMBINE

- Bảng quy định cách gộp cụm mô hình

TBL_NM	TBL DESC	FIELD	FIELD DESC	DATA TYPE	KEY	NOT NULL	DEFAULT	NOTE
CINS_MODEL_COMBINE	Bảng quy định cách gộp cụm mô hình	MODEL_COMBINE_KEY	Id tự tăng	INTERGER	X			
CINS_MODEL_COMBINE		PID	Process id lấy từ lần chạy python	VARCHAR2(100)				

CINS_MODEL_COMBINE		RSLT_BEFORE	Kết quả mô hình	VARCHAR2(200)				
CINS_MODEL_COMBINE		PID_LIVE	Mã mô hình gộp	VARCHAR2(100)				
CINS_MODEL_COMBINE		RSLT_AFTER	Kết quả mô hình gộp	VARCHAR2(200)				
CINS_MODEL_COMBINE		ACTIVE	Có còn được dùng cho bài toán không?	INTERGER			1	
CINS_MODEL_COMBINE		ADD_TSTP	Thời điểm nhấn vào bảng	TIMESTAMP				

III.12. CINS_JOB_REGISTRY

- Bảng đăng ký và theo dõi lịch chạy Job mô hình

TBL_NM	TBL DESC	FIELD	FIELD DESC	DATA TYPE	KEY	NOT NULL	DEFAULT	NOTE
CINS_JOB_REGISTRY	Bảng đăng ký và theo dõi lịch chạy Job mô hình	JOB_REGISTRY_KEY	Id tự tăng	INTERGER	X			
CINS_JOB_REGISTRY		PID_LIVE	Mã mô hình gộp	VARCHAR2(100)				
CINS_JOB_REGISTRY		SCHED	Thời gian đăng ký chạy Job	TIMESTAMP		X		
CINS_JOB_REGISTRY		ACT_PID	Process id lấy từ lần chạy python	INTERGER				
CINS_JOB_REGISTRY		ACT_START	Thời gian Job thực tế bắt đầu	TIMESTAMP				
CINS_JOB_REGISTRY		ACT_END	Thời gian Job thực tế kết thúc	TIMESTAMP				
CINS_JOB_REGISTRY		PARAMS	Tham số truyền vào Job	VARCHAR2(2000)				VD: Truyền vào '05-04-2023' cho RPT_DT của lần chạy Job tới
CINS_JOB_REGISTRY		ACTIVE	Có còn được dùng cho bài toán không?	INTERGER		X	1	
CINS_JOB_REGISTRY		STATUS	Trạng thái Job	VARCHAR2(20)		X	'QUEUE'	Default: QUEUE, RUNNING, DONE, ERROR
CINS_JOB_REGISTRY		FILE_NM	Tên file Python cần	VARCHAR2(200)				

			chạy					

III.13. CINS_LOC_DIM_POP

- Bảng dim ghi nhận dân số Việt Nam tới cấp huyện

TBL_NM	TBL DESC	FIELD	FIELD DESC	DATA TYPE	KEY	NOT NULL	DEFAULT	NOTE
CINS_LOC_DIM_POP	Bảng dim ghi nhận dân số Việt Nam tới cấp huyện	LOC_POP_KEY	Id tự tăng	INTERGER	X			
CINS_LOC_DIM_POP		PROVINCE	Tỉnh	VARCHAR2(200)				
CINS_LOC_DIM_POP		DISTRICT	Quận	VARCHAR2(200)				
CINS_LOC_DIM_POP		WARD	Huyện	VARCHAR2(200)				
CINS_LOC_DIM_POP		POPULATION	Dân số	INTERGER				
CINS_LOC_DIM_POP		ACTIVE	Có còn được dùng cho bài toán không?	INTERGER		1		
CINS_LOC_DIM_POP		ADD_TSTP	Thời điểm nhấn vào bảng	TIMESTAMP				

III.14. CINS_LOC_DIM_LNGLAT

- Bảng dim lưu tọa độ các vị trí địa lý trên hệ thống

TBL_NM	TBL DESC	FIELD	FIELD DESC	DATA TYPE	KEY	NOT NULL	DEFAULT	NOTE
CINS_LOC_DIM_LNGLAT	Bảng dim lưu tọa độ các vị trí địa lý trên hệ thống	LOC_LNGLAT_KEY	Id tự tăng	INTEGER	X			
CINS_LOC_DIM_LNGLAT		SRC_ID1	Id lấy từ hệ thống	VARCHAR2(50)		X		
CINS_LOC_DIM_LNGLAT		SRC_ID2	Id lấy từ hệ thống (lvl2)	VARCHAR2(50)				

CINS_LOC_DIM_LNGLAT	SUBJECT	Loại vị trí	VARCHAR2(50)				DEFAULT: CIF, ATM, BRANCH, POS
CINS_LOC_DIM_LNGLAT	ADDR1	Chi tiết địa chỉ	VARCHAR2(2000)				
CINS_LOC_DIM_LNGLAT	ADDR2	Chi tiết địa chỉ (cont)	VARCHAR2(2000)		1		
CINS_LOC_DIM_LNGLAT	LNG	Kinh độ	NUMBER(10,7)				
CINS_LOC_DIM_LNGLAT	LAT	Vĩ độ	NUMBER(10,7)				
CINS_LOC_DIM_LNGLAT	ADD_TSTP	Thời điểm nhấn vào bảng	TIMESTAMP				

III.15. TMP_SEGMENT_RSLT

- Lưu trữ kết quả mô hình được biến đổi phục vụ Dashboard theo dõi mô hình

TBL_NM	TBL DESC	FIELD	FIELD DESC	DATA TYPE	KEY	NOT NULL	DEFAULT	NOTE
TMP_SEGMENT_RSLT	Lưu trữ kết quả mô hình được biến đổi phục vụ Dashboard theo dõi mô hình	MODEL_RSLT_KEY	Id tự tăng	INTEGER	X			
TMP_SEGMENT_RSLT		MODEL_NM	Tên mô hình khái quát	VARCHAR2(200)				
TMP_SEGMENT_RSLT		PID	Process id lấy từ lần chạy Python	VARCHAR2(100)				
TMP_SEGMENT_RSLT		CUSTOMER_CDE	Mã Khách hàng	VARCHAR2(20)		X		
TMP_SEGMENT_RSLT		RSLT	Kết quả mô hình	VARCHAR2(200)				
TMP_SEGMENT_RSLT		RPT_DT	Kỳ báo cáo	VARCHAR2(20)				
TMP_SEGMENT_RSLT		CONFIDENCE	Độ tự tin	NUMBER(20,6)				Giữ giá trị float nếu là %, không *100
TMP_SEGMENT_RSLT		ADD_TSTP	Thời điểm nhấn vào bảng	TIMESTAMP				

III.16. SEGMENT_CNT_CST

- Bảng phân bố kết quả mô hình Segmentation theo nhóm tuổi phục vụ Dashboard

TBL_NM	TBL DESC	FIELD	FIELD DESC	DATA TYPE	KEY	NOT NULL	DEFAULT	NOTE
SEGMENT_CNT_CST	Bảng phân bố kết quả mô hình Segmentation theo nhóm tuổi phục vụ Dashboard	PID	Process id lấy từ lần chạy Python	VARCHAR2(100)				
SEGMENT_CNT_CST		RPT_DT	Kỳ báo cáo	VARCHAR(20)				
SEGMENT_CNT_CST		CLUSTER_NM	Tên cụm	VARCHAR(100)				
SEGMENT_CNT_CST		GEN_GRP	Nhóm tuổi	VARCHAR(200)				
SEGMENT_CNT_CST		CNT_CST	Số lượng Khách hàng	NUMBER(38,0)				

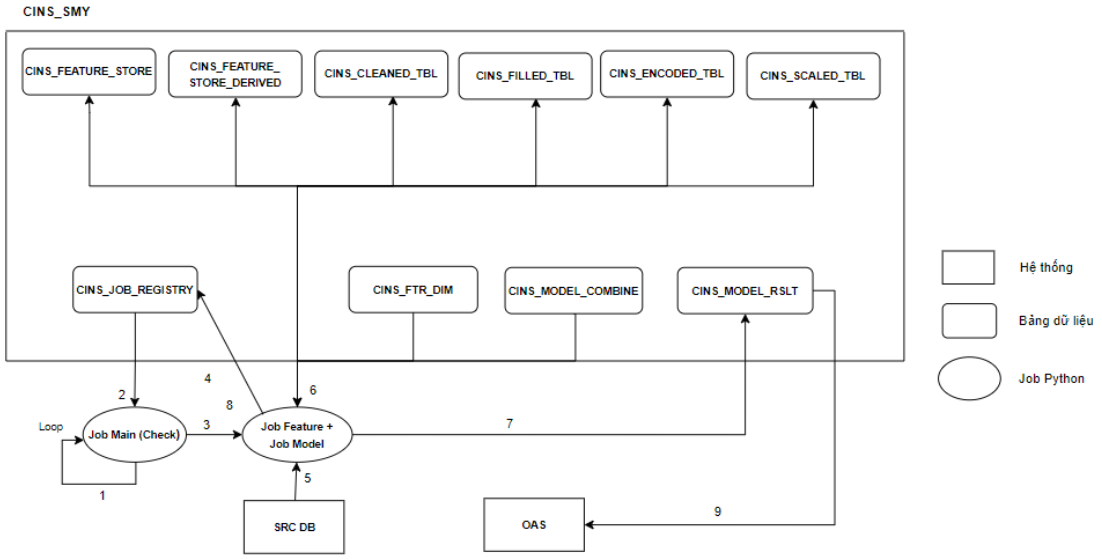
III.17. SEGMENT_FTR

- Bảng phân tích mô hình phục vụ Radar Chart trên Dashboard theo dõi mô hình

TBL_NM	TBL DESC	FIELD	FIELD DESC	DATA TYPE	KEY	NOT NULL	DEFAULT	NOTE
SEGMENT_FTR	Bảng phân bố kết quả mô hình Segmentation theo nhóm tuổi phục vụ Dashboard	PID	Process id lấy từ lần chạy Python	VARCHAR2(100)				
SEGMENT_FTR		MODEL_NM	Tên mô hình khái quát	VARCHAR(200)				
SEGMENT_FTR		RPT_DT	Kỳ chạy	VARCHAR(20)				
SEGMENT_FTR		CLUSTER_NM	Tên cụm	VARCHAR(100)				
SEGMENT_FTR		AGG_TP	Cách thức tổng hợp	VARCHAR(200)				
SEGMENT_FTR		FTR_NM	Tên Feature	VARCHAR(200)				
SEGMENT_FTR		AGG_VAL	Giá trị được tổng hợp	NUMBER(38,0)				

IV. LUỒNG XỬ LÝ

IV.1. Thiết kế tổng quan



B1 : Job Main được lặp đi lặp lại với tần suất 6 tiếng/lần – để thực hiện bước số 2 dưới đây. (*Được mô tả chi tiết tại mục IV.2.1.1. Job Main (Check)*)

B2 : Job Main (check) nhằm kiểm tra dữ liệu từ bảng CINS_JOB_REGISTRY trong trường hợp có đặt lịch chạy Job hay không (*Được mô tả chi tiết tại mục IV.2.1.1. Job Main (Check)*)

B3 : Nếu bước 2 kiểm tra có Job được đăng ký sẽ gọi tới Job Feature và Job Model (*Được mô tả chi tiết tại mục IV.2.1.1. Job Main (Check)*)

B4 : Sau khi Job Main tiến hành chạy, bảng CINS_JOB_REGISTRY sẽ được cập nhật trạng thái của Job (*Được mô tả chi tiết tại mục IV.2.1.1. Job Main (Check)*)

B5 : Trong quá trình chạy Job Main - Job Feature, dữ liệu từ Source Database sẽ được xử lý rồi đẩy vào các bảng dữ liệu đã được tạo sẵn và được lưu trữ phục vụ cho các bước tiếp theo

B6 : Lúc này, Job Feature thực hiện tổng hợp và lưu dữ liệu Feature tại kì báo cáo chạy Job tại bảng :

- CINS_FEATURE_STORE
- CINS_FEATURE_STORE_DERIVED

Đọc bảng CINS_FTR_DIM nhằm lấy cấu hình các Feature phục vụ cho mô hình, sau đó dữ liệu được đẩy vào lần lượt các bảng tương ứng :

- CINS_CLEANED_TBL

- CINS_FILLED_TBL
- CINS_ENCODED_TBL
- CINS_SCALED_TBL

Tiếp đến Job Model sẽ được tiến hành, bảng CINS_MODEL_COMBINE lưu cấu hình được sử dụng để gộp kết quả chi tiết của từng cụm trong trường hợp mô hình phân cụm quá sâu. (*Được mô tả chi tiết tại mục IV.2.1.1. Job Feature, IV.2.1.2. Job Feature – Preprocessing Data, IV.2.1.3. Job Model*)

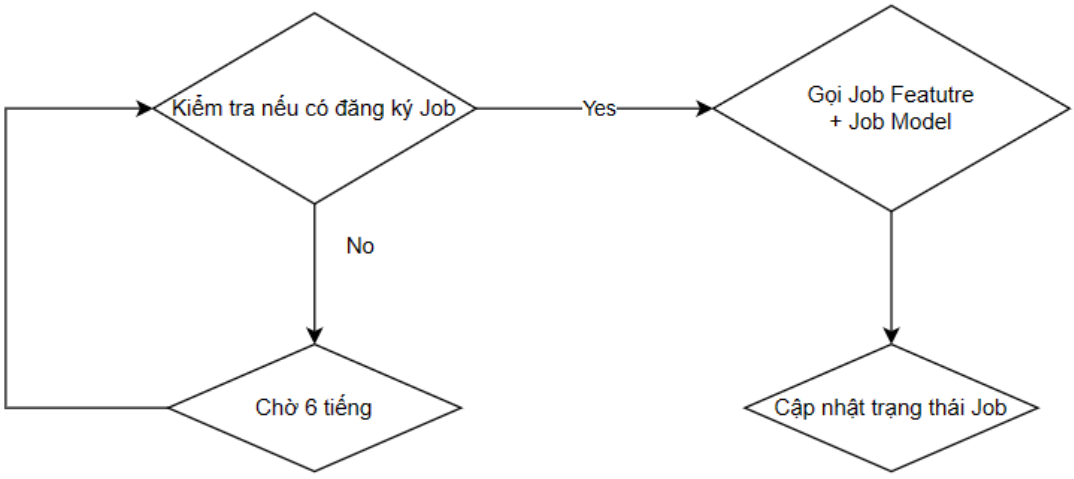
B7 : Sau khi chạy Job Model kết quả dự đoán của mô hình sẽ được đẩy vào và lưu trữ tại bảng CINS_MODEL_RSLT. (*Được mô tả chi tiết tại mục IV.2.1.3. Job Model*)

B8 : Trạng thái Job được cập nhật sau khi đã hoàn thành (*Được mô tả chi tiết tại mục IV.2.1.1. Job Main (Check)*)

B9 : Sau khi hoàn thành việc đẩy dữ liệu vào bảng CINS_MODEL_RSLT tại bước 7, dữ liệu sẽ tiếp tục được đẩy vào các bảng dữ liệu phục vụ cho Dashboard và hiển thị trên OAS (*Được mô tả chi tiết tại mục IV.2.1.3. Job Model*)

IV.2. Quy trình chạy mô hình trên Live

IV.2.1. Job Main (Check)



Là Job bao ngoài để điều khiển luồng của toàn bộ chương trình chạy dự đoán trên live của mô hình Segmentation bao gồm các bước như sau :

B1 : Thực hiện loop lặp lại với tần suất 6 tiếng/lần để call lại chính nó

B2 : Kiểm tra dữ liệu bảng CINS_JOB_REGISTRY kết hợp các điều kiện ACTIVE = 1, STATUS = 'QUEUE', SCHED <= CURRENT_TIMESTAMP. Trong trường hợp tìm được Job đã đăng ký nhưng chưa chạy, sẽ tiến hành tiến hành thực hiện Bước 4 và chuyển trạng thái Job đó từ 'QUEUE' thành 'RUNNING'.

B3 : Gọi lần lượt tới Job Feature và Job Model

B4 : Sau khi Job Feature và Job Model chạy hoàn tất, cập nhật trạng thái của Job tại bảng CINS_JOB_REGISTRY thành 'DONE'. Trong trường hợp Job chạy lỗi trạng thái c sẽ được chuyển đổi thành 'ERROR'.

IV.2.2. Job Feature

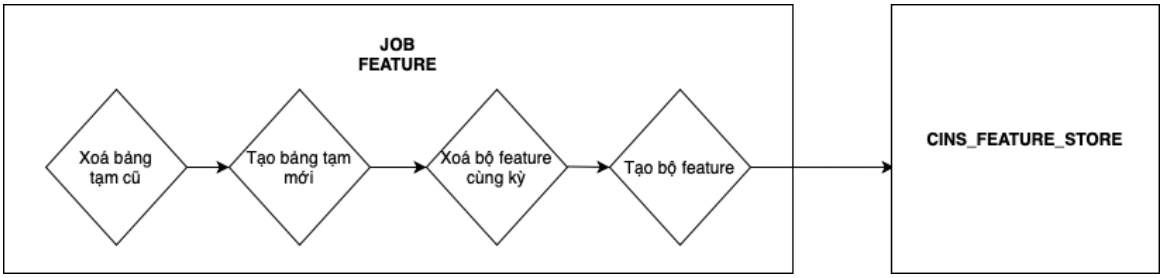
Job Feature là bước đầu tiên trong quy trình chạy mô hình, mục đích nhằm tổng hợp các Feature đã được chọn lọc để đưa vào mô hình Segmentation.

Các Feature sau khi được tổng hợp bằng Job Python qua việc gọi đến các câu lệnh SQL đã được tạo thành script sẵn, sẽ được đưa vào bảng CINS_FEATURE_STORE để lưu trữ, và sử dụng ở các bước tiếp theo trong quá trình dự đoán của mô hình.

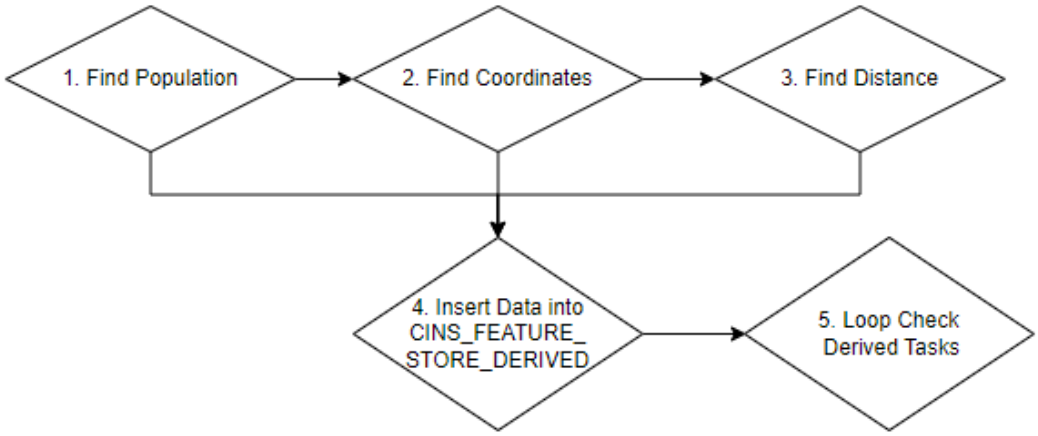
Job Feature sẽ được chạy theo kỳ (1 tháng 1 lần), và tương tự mỗi bộ Feature tổng hợp sẽ được lưu trữ theo kỳ chạy đó – Report Date (trường RPT_DT trong bảng CINS_FEATURE_STORE). Khi bắt đầu chạy Job, Report Date sẽ cần được nhập dưới định dạng: 'DD-MM-YYYY'

Job Feature bao gồm 4 bước chính:

1. Xoá các bảng tạm hỗ trợ tổng hợp Feature đã được tạo ra từ lượt chạy Job trước
2. Tạo các bảng tạm hỗ trợ tổng hợp Feature cho kỳ chạy này
3. Xoá bộ Feature trong bảng CINS_FEATURE_STORE với Report Date của lần chạy Job này (trong trường hợp cần chạy lại)
4. Tổng hợp và đẩy bộ Feature vào bảng CINS_FEATURE_STORE với Report Date được nhập khi chạy Job



Ngoài ra, các Feature phái sinh được tổng hợp từ Feature gốc tại bảng CINS_FEATURE_STORE được tổng hợp và lưu trữ tại bảng CINS_FEATURE_STORE_DERIVED với Reported Date được truyền vào. Job Feature Derived bao gồm các bước sau:



B1: Tổng hợp Feature ADDR_POP tại kỳ báo cáo được chạy, dữ liệu thay đổi so với kỳ báo cáo trước đó, trong trường hợp dữ liệu tại kỳ báo cáo chạy Job chưa có sẽ thực hiện tổng hợp từ các bảng CINS_LOC_DIM_LNGLAT và bảng CINS_FEATURE_STORE nhằm lấy dữ liệu liên quan tới Population và Address

B2: Tìm tọa độ các Feature phái sinh từ các Feature:

- CARD_TOP1_MERCHANT_6M
- FAV_POS_6M_SM
- CARD_FAV_BRANCH_LOC_6M
- ADDR_TOWN

Từ đó sinh ra các Feature sau:

- ADDR_LNG, ADDR_LAT
- FAV_POS_1_LNG, FAV_POS_1_LAT
- FAV_POS_2_LNG, FAV_POS_2_LAT
- FAV_ATM_LNG, FAV_ATM_LAT

Được tổng hợp từ các bảng CINS_LOC_DIM_LNGLAT và bảng CINS_FEATURE_STORE với kỳ báo cáo tương ứng được truyền vào

B3: Tìm khoảng cách của các Feature:

- CARD_TOP1_MERCHANT_6M
- FAV_POS_6M_SM
- CARD_FAV_BRANCH_LOC_6M

Dựa vào các tọa độ đã được tổng hợp trước đó, từ đó sinh ra các Feature sau:

- DIST_CUST_FAV_ATM
- DIST_CUST_FAV_POS_TOP1
- DIST_CUST_FAV_POS_TOP2

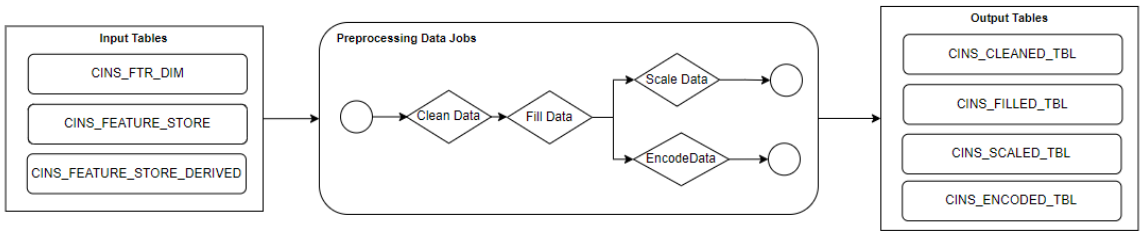
B4: Sau khi tổng hợp và lưu trữ dữ liệu các Feature từ bước 1, bước 2, bước 3, sẽ tiến hành thực hiện đẩy dữ liệu bộ Feature phái sinh vào bảng CINS_FEATURE_STORE_DERIVED

B5: Feature Derived được tính toán song song. Nhằm quản lý được tiến độ , đảm bảo tất cả các Feature phái sinh đã hoàn thành tính toán, bước này chạy vòng lặp với tần suất 30 phút/ lần để kiểm tra dữ liệu tại bảng CINS_FEATURE_STORE_DERIVED.

IV.2.3. Job Feature – Preprocessing Data

Luồng tiền xử lí dữ liệu bao gồm bốn bước clean data - làm sạch dữ liệu, fill data – làm bù dữ liệu, encode data – mã hoá dữ liệu và scale data - biến đổi khoảng giá trị dữ liệu. Bốn bước xử lý dữ liệu được thực hiện tuần tự nhằm tạo ra các dữ liệu sạch theo cấu hình được lưu trữ trong bảng CINS_FTR_DIM phục vụ cho việc huấn luyện mô hình và sử dụng mô hình dự báo. Cấu hình giúp chương trình xác định các Feature cần tiền xử lý, cách thức tiền xử lý và các bảng nguồn chứa các Feature đó, cụ thể là các bảng dữ liệu tổng hợp CINS_FEATURE_STORE và CINS_FEATURE_STORE_DERIVED.

Luồng tiền xử lý dữ liệu được lập trình bằng ngôn ngữ Python. Chương trình được chạy định kì một tháng một lần với tham số đầu vào là giá trị của kì chạy tương ứng với giá trị trong trường report date (RPT_DT) với định dạng 'DD-MM-YYYY' trong các bảng nguồn CINS_FEATURE_STORE và CINS_FEATURE_STORE_DERIVED. Chi tiết các bước được mô tả dưới đây:



Hình 1: Mô tả luồng tiền xử lý dữ liệu cùng các bảng đầu vào / đầu ra

Clean data - làm sạch dữ liệu từ các bảng dữ liệu nguồn và lưu trữ vào bảng CINS_CLEANED_TBL. Các bước thực hiện như sau:

1. Đọc cấu hình làm sạch dữ liệu từ bảng CINS_FTR_DIM để lấy tên Feature trường FTR_NM, cách thức làm sạch dữ liệu trường FLTD_CD, tên bảng chứa Feature tương ứng trường SRC_TBL với điều kiện các luật đang được sử dụng ACTIVE = 1.
2. Chương trình thực hiện xoá các Feature theo tham số kì report date RPT_DT trong bảng CINS_FILLED_TBL nếu có (trong trường hợp chạy lại kỳ).
3. Chương trình thực hiện làm sạch dữ liệu từ các bảng nguồn được khai báo trong bảng CINS_FTR_DIM với tham số là kì report date RPT_DT được truyền vào.
4. Dữ liệu sau khi làm sạch được lưu trữ vào bảng đích CINS_CLEANED_TBL sẵn sàng cho bước làm bù dữ liệu.

Fill data – làm bù dữ liệu từ bảng dữ liệu được làm sạch CINS_CLEANED_TBL và lưu trữ vào bảng CINS_FILLED_TBL. Các bước thực hiện như sau:

1. Đọc cấu hình làm sạch dữ liệu từ bảng CINS_FTR_DIM để lấy tên Feature trường FTR_NM, giá trị làm bù dữ liệu cho từng Feature theo trường FILL_CD với điều kiện các luật đang được sử dụng ACTIVE = 1.
2. Chương trình thực hiện xoá các Feature theo tham số kì report date RPT_DT trong bảng CINS_FILLED_TBL nếu có (trong trường hợp chạy lại kỳ).
3. Chương trình thực hiện làm bù dữ liệu từ bảng CINS_CLEANED_TBL được khai báo trong bảng CINS_FTR_DIM với tham số là kì report date RPT_DT được truyền vào.
4. Dữ liệu sau khi làm bù được lưu trữ vào bảng đích CINS_FILLED_TBL sẵn sàng cho bước làm encode và scale dữ liệu.

Encode data – mã hoá dữ liệu từ bảng dữ liệu được làm bù CINS_FILLED_TBL và lưu trữ vào bảng CINS_ENCODED_TBL. Các bước thực hiện như sau:

1. Đọc cấu hình từ bảng CINS_FTR_DIM để lấy tên Feature trường FTR_NM cần mã hoá với điều kiện các luật đang được sử dụng ACTIVE = 1.
2. Chương trình thực hiện xoá các Feature theo tham số kì report date RPT_DT trong bảng CINS_ENCODED_TBL nếu có (trong trường hợp chạy lại kỳ).
3. Chương trình thực hiện mã hoá dữ liệu từ bảng CINS_FILLED_TBL được khai báo trong bảng CINS_FTR_DIM với tham số là kì report date RPT_DT được truyền vào.
4. Dữ liệu sau khi mã hoá được lưu trữ vào bảng đích CINS_ENCODED_TBL sẵn sàng cho bước mô hình dự báo.

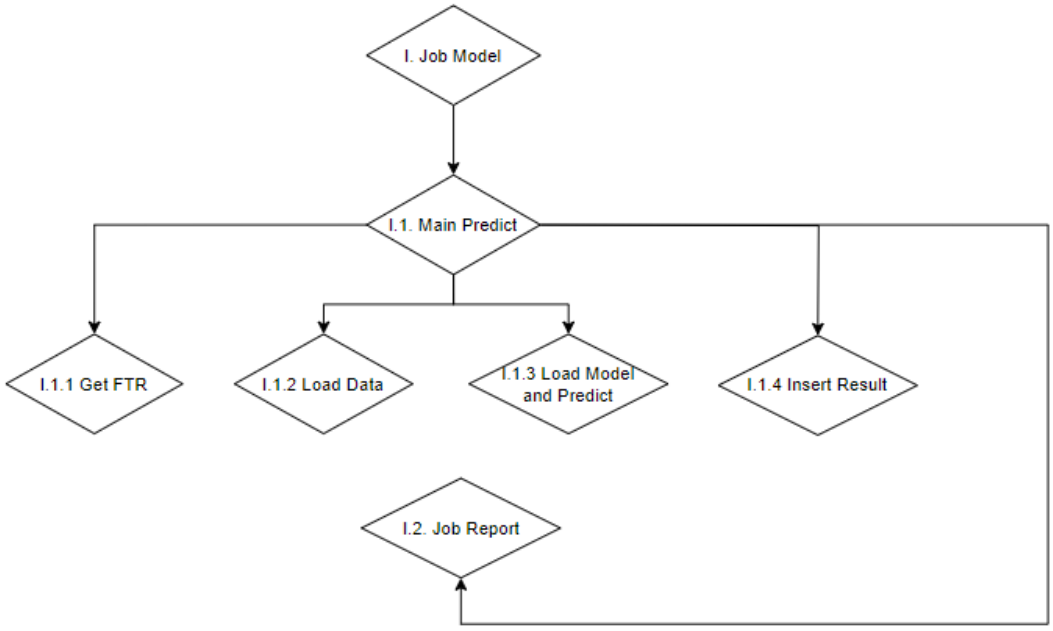
Scale data – mở rộng dữ liệu, đưa giá trị từng Feature về chung một khoảng giá trị [0,1] từ bảng dữ liệu được làm bù CINS_FILLED_TBL và lưu trữ vào bảng CINS_SCALED_TBL. Các bước thực hiện như sau:

1. Đọc cấu hình từ bảng CINS_FTR_DIM để lấy tên Feature trường FTR_NM cần mở rộng với điều kiện các luật đang được sử dụng ACTIVE = 1.
2. Chương trình thực hiện xoá các Feature theo tham số kì report date RPT_DT trong bảng CINS_SCALED_TBL nếu có (trong trường hợp chạy lại kỳ).
3. Chương trình thực hiện mở rộng dữ liệu từ bảng CINS_FILLED_TBL được khai báo trong bảng CINS_FTR_DIM với tham số là kì report date RPT_DT được truyền vào.
4. Dữ liệu sau khi mở rộng được lưu trữ vào bảng đích CINS_SCALED_TBL sẵn sàng cho bước mô hình dự báo.

IV.2.4. Job Model

Job Model bản chất là Job Python nhằm sử dụng 3 mô hình đã được Train trước đó với bộ dữ liệu 2 triệu khách hàng: Segmentation Geographic, Segmentation TXN Behaviors, Segmentation Customer Value để thực hiện phân cụm tập khách hàng hoạt động với tần suất chạy là 1 tháng 1 lần tại kỳ báo cáo RPT_DT (Reported Date) được truyền vào.

Job Model gồm các bước xử lý chính như sau:



I.1. Main Predict:

I.1.1. Get FTR: Hàm getFTR(pid_live,m,v) dùng để tổng hợp các Feature sử dụng cho mô hình. Đọc cấu hình từ bảng CINS_FTR_DIM với điều kiện Active = 1 và tên mô hình tương ứng muốn chạy

- Đầu vào: Sử dụng dữ liệu được lưu trữ tại bảng CINS_FTR_DIM để tổng hợp Feature
- Đầu ra: Bộ Feature sử dụng cho mô hình để thực hiện phân cụm

Trong đó các tham số:

- pid_live: Mã mô hình (VD: CUST_VAL_01, TXN_BEHAVIORS_01, GEO_01)
- m: Tên mô hình (VD: SEGMENTATION CUSTOMER VALUE,…)
- v: Nhóm dữ liệu (VD: ALL,…)

I.1.2. Load Data: Hàm loadData(ftr,rpt_dt,x1,x2) được sử dụng nhằm lấy bộ dữ liệu đầu vào để thực hiện phân cụm khách hàng. Nhằm lấy dữ liệu khách hàng hoạt động tại bảng CINS_TMP_CIF sau đó join với 2 bảng CINS_ENCODED_TBL (nếu Feature là dạng Category) và bảng CINS_SCALED_TBL nhằm lấy dữ liệu các đặc trưng đã được mã hóa và co giãn với kỳ báo cáo RPT_DT được truyền vào

- Đầu vào: Feature được tổng hợp từ bước 1 , đã thực hiện qua các bước preprocessing data, dữ liệu tại các bảng: CINS_ENCODED_TBL, CINS_SCALED_TBL, danh sách khách hàng hoạt động tại kỳ báo cáo chạy Job
- Đầu ra: Bộ dữ liệu chứa khách hàng hoạt động tại kỳ báo cáo được truyền vào

Trong đó các tham số:

- ftr: Feature sử dụng cho mô hình - sau khi tổng hợp được từ bước 1
- rpt_dt: Reported Date – kỳ báo cáo chạy Job
- x1,x2: Tham số dùng để phân tách bộ dữ liệu lớn thành các tập dữ liệu nhỏ hơn

I.1.3. Load Model: Hàm loadModel(pid_live,df) nhằm load mô hình đã được Train trước đó với 2 triệu khách hàng và sau đó thực hiện phân cụm tập khách hàng mới tại kỳ báo cáo truyền vào

- Đầu vào: Bộ dữ liệu chứa khách hàng hoạt động tại kỳ báo cáo được truyền vào
- Đầu ra: Kết quả phân cụm khách hàng

Trong đó các tham số:

- pid_live: Mã mô hình

- df: Bộ dữ liệu chứa khách hàng hoạt động tại kỳ báo cáo được truyền vào

I.1.4. Insert Result: Hàm insertResult(rslt,m,pid_live,rpt_dt) nhằm nhấn kết quả sau khi mô hình thực hiện phân cụm khách hàng và lưu trữ tại bảng dữ liệu CINS_MODEL_RSLT, kết quả phân cụm sẽ được reference tại bảng CINS_MODEL_COMBINE

- Đầu vào: Kết quả sau khi chạy mô hình
- Đầu ra: Dữ liệu được insert vào bảng CINS_MODEL_RSLT

Trong đó các tham số:

- rslt: Kết quả phân cụm
- m: Tên mô hình
- pid_live: Mã mô hình
- rpt_dt: Kỳ báo cáo

I.2. Job Report:

Cuối cùng, sau khi chạy ra kết quả phân cụm khách hàng, Job Report sẽ được chạy ngay sau đó với đầu vào dữ liệu là bảng CINS_MODEL_RSLT và CINS_MODEL_EVAL, sau đó dữ liệu sẽ được đẩy vào lần lượt các bảng: TMP_SEGMENT_RSLT, SEGMENT_CNT_CST, SEGMENT_FTR nhằm trực quan hóa kết quả mô hình, phục vụ báo cáo 3 mô hình: Segmentation Geographic, Segmentation Txn Behaviors, Segmentation Customer Value với tham số là RPT_DT – kỳ báo cáo chạy Job được truyền vào.