NÂNG CAO CHẤT LƯỢNG PHÂN CỤM CỬA HÀNG THƯƠNG MẠI ĐIỆN TỬ DỰA VÀO TẬP PHỔ BIẾN TỐI ĐẠI

Ứng dụng kết hợp khai thác tập phổ biến tối đại và phân cụm

Nguyễn Văn Đạt - 2186400229

Đại học Công Nghệ Tp. HCM
GVHD: TS. Bùi Danh Hường

Nội Dung

Tổng quan đề tài

Lý do chọn đề tài

- **Hiểu biết thị trường:** Phân cụm hỗ trợ hoạt động kinh doanh sàn TMĐT.
- **Phương pháp mới:** Áp dụng phân cụm kết hợp khai thác tập phổ biến tối đại (Maximal Frequent Itemset). Hỗ trợ nâng cao hiệu quả phân cụm.

Chuẩn bị dữ liệu

Thu thập và tiền xử lý dữ liệu

Thu thập dữ liệu:

- Dữ liệu được thu thập từ sàn thương mại điện tử Tiki.
- Dữ liệu ban đầu gồm 73.000 dòng: Thông tin sản phẩm (id, name, price, quatity sold, ...), thông tin cửa hàng (year joined, rating, followers, chat response rate, ...), các bình luận (feedbacks) của khách hàng về sản phẩm.

Thu thập và tiền xử lý dữ liệu

Tiền xử lý dữ liệu ban đầu:

- Trải qua bước tiền xử lý ban đầu bao gồm:
  - Loại bỏ dữ liệu không hợp lệ.
  - Loại bỏ dữ liệu trùng lặp.
  - Tính toán doanh thu ước tính, phân tích cảm xúc đánh giá.
  - Kết hợp các đặc trưng (Rating & Counter Rating, Positive & Total Feedback).
  - Sử dụng Wilson score đánh giá độ tin cậy các biến giữa các cửa hàng.
- Bảng dữ liệu: 6 đặc trưng và 933 cửa hàng.

Revenue	YearJoined	Followers	ChatResponse	RatingQuality	PositiveQuality
1,326,000	5	982	0	4.621877	0.787091
25,545,495	4	1500	0	4.631574	0.706230
3,148,294	7	479	0	4.350326	0.752689
6,572,000	6	565	1	4.560354	0.792613
529,000	6	137	0	4.648886	0.643643
2,261,000	5	955	0.66	4.545801	0.786455
816,900	8	181	0	4.168226	0.548946
966,000	8	3013	0.5	4.539589	0.770787

Table: Mô tả bảng dữ liệu.

Tiền xử lý dữ liệu

Chuẩn hóa dữ liệu bằng ScalerStandard

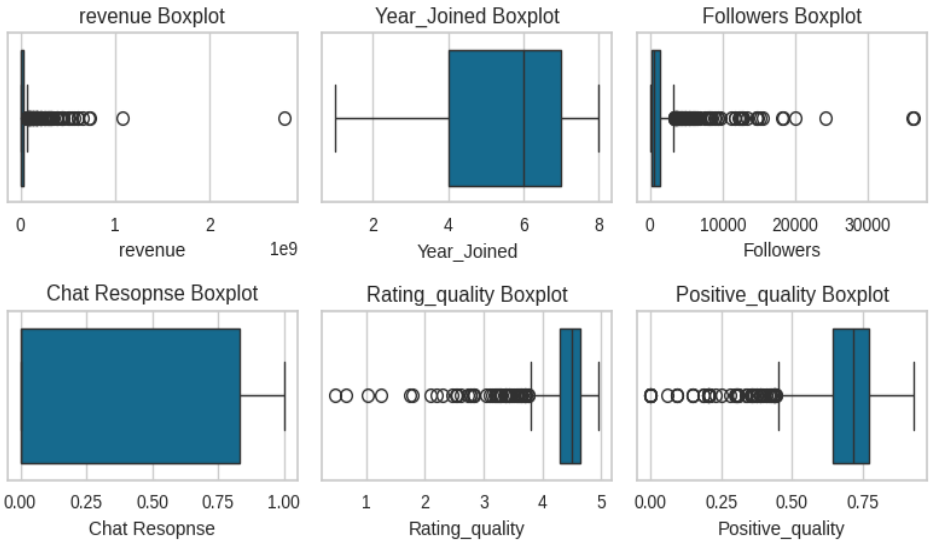
- **Đưa dữ liệu về thang đo chung:** Loại bỏ sự khác biệt về đơn vị đo lường.
- **Cải thiện hiệu suất mô hình:** Đặc biệt quan trọng với các thuật toán nhạy cảm với thang đo.

Revenue	YearJoined	Followers	ChatResponse	RatingQuality	PositiveQuality
-0.326840	-0.279443	-0.143533	-0.819518	0.517466	0.731572
-0.143226	-0.972848	0.031235	-0.819518	0.539159	0.189270
-0.260906	0.413962	1.358861	1.525984	0.663007	1.056228
-0.313024	1.107368	-0.313240	-0.819518	-0.090066	0.500853
-0.287069	0.413962	-0.284224	1.525984	0.379822	0.768611

Table: Bảng dữ liệu sau khi áp dụng StandardScaler

Xử lý Outliers

Xử lý Outliers bằng Winsorization:



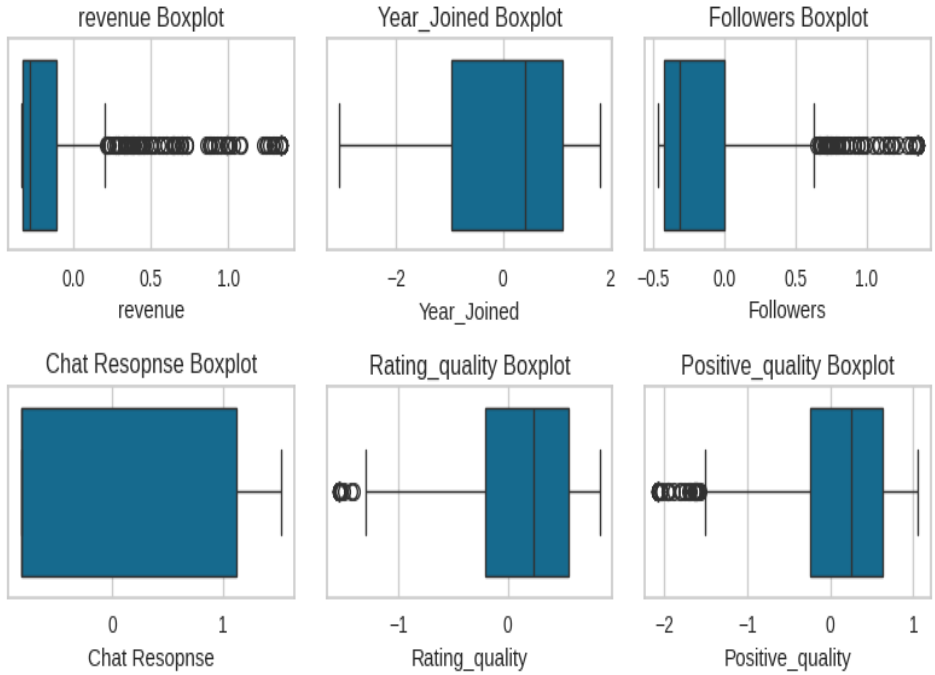
Ảnh 1: Trước khi áp dụng Winsorization.

Nhược điểm:

- Không giải quyết được mọi vấn đề ngoại lệ.

Ưu điểm:

- Giảm tác động của giá trị ngoại lệ.
- Bảo tồn dữ liệu.



Ảnh 2: Sau khi áp dụng Winsorization.

Xử lý Outliers

Xử lý Outliers bằng Isolation Forest:

Isolation Forest

- Phát hiện outliers nhờ tính dễ bị cô lập của chúng.
- Xử lý tốt dữ liệu không tuyến tính hoặc phân phối phức tạp.
- Tiến hành loại bỏ các outliers bị cô lập.

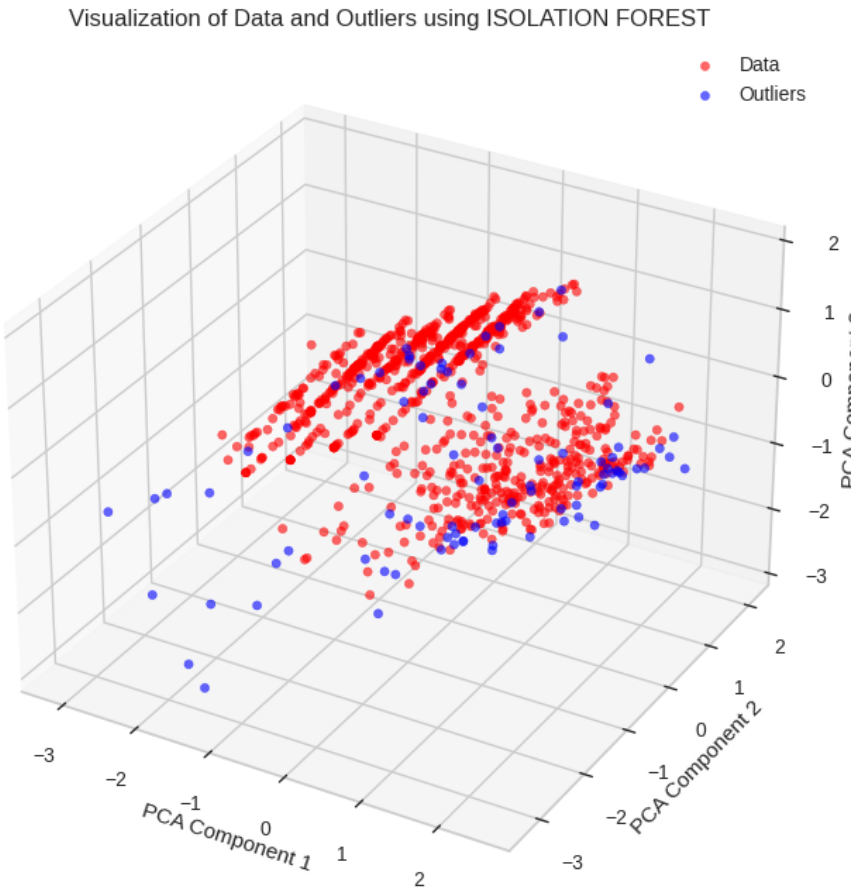


Figure: Isolation Forest.

DBSCAN

Kiểm tra và Xử lý Outliers bằng DBSCAN:

DBSCAN

- Sử dụng DBSCAN để kiểm tra và loại bỏ triệt để các outliers.
- Không yêu cầu phân phối dữ liệu cụ thể.
- Tự động phát hiện điểm mật độ thấp, xa cụm chính (gán nhãn -1).
- Tiến hành loại bỏ triệt để các outliers cuối cùng.

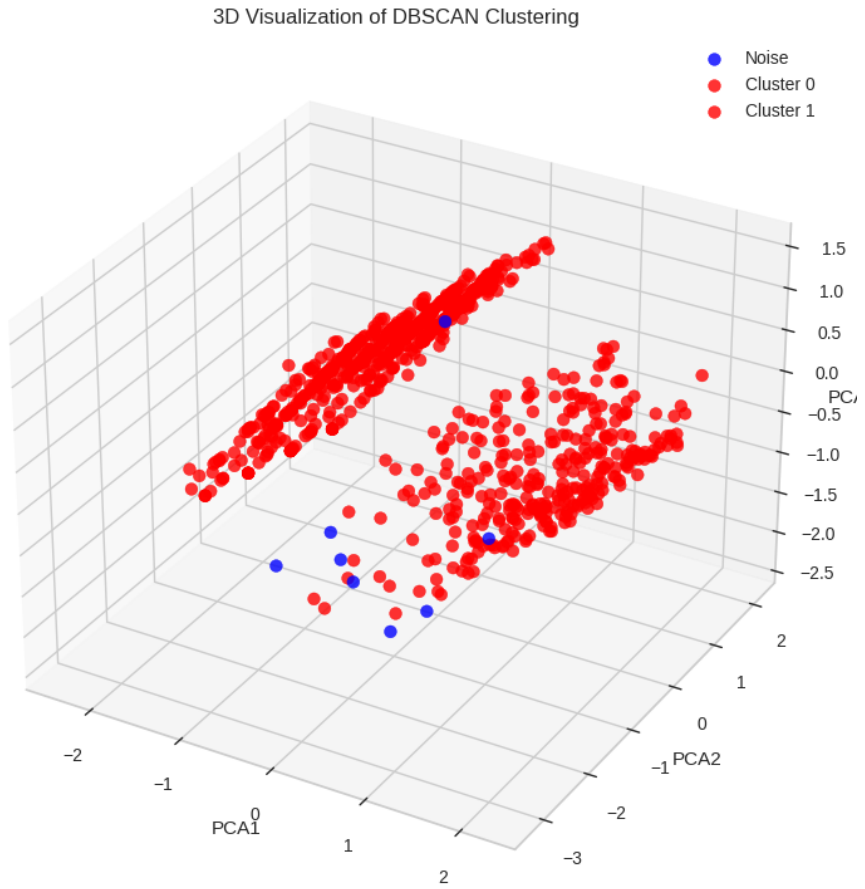


Figure: DBSCAN

Phân cụm trước và sau khi kết hợp tập phổ biến tối đại

Xác định số cụm tối ưu

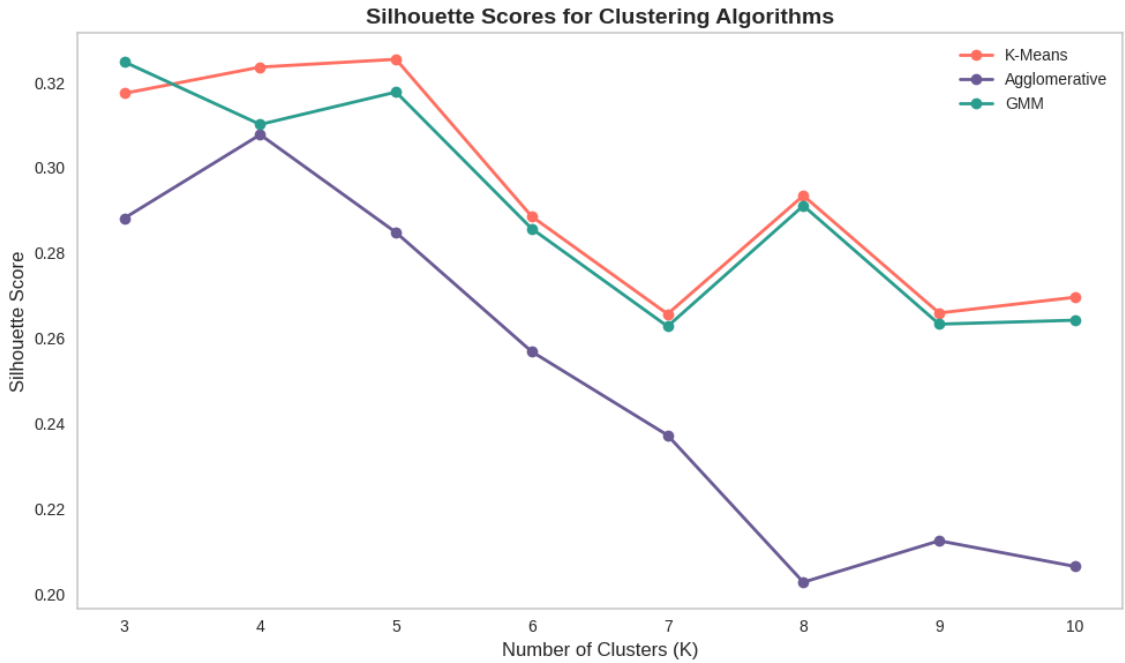


Figure: Chọn số cụm tối ưu

Nhận xét:

- Sử dụng Silhouette để tìm số cụm tối ưu: K-Means (k-means++), Agglomerative (ward), GMM (spherical).
- Số cụm (k) cần lựa chọn để phân cụm cho 3 thuật toán: K-Means, Agglomerative, GMM lần lượt là: 5, 4, 3 cụm.

Phân cụm kết hợp khai thác tập phổ biến tối đại

Tạo ra các đặc trưng mới:

- Sử dụng K-Bins Discretizer (K_Means) để rời rạc hóa dữ liệu về 3 phần: Low, Medium, High.
- Sử dụng FP-Max (min_support = 20%) để tạo ra các tập phổ biến tối đại.
  ==> Tạo ra các đặc trưng nhị phân mới từ FP-Max.

Lựa chọn đặc trưng mới:

- Sử dụng Forward Selection để lựa chọn các đặc trưng nhị phân mới có tác động tích cực đến kết quả phân cum.
- Số lượng đặc trưng nhị phân được thêm vào sau quá trình Forward Selection K-Prototypes (2 biến), Agglomerative (3 biến), GMM (3 biến).

Đánh giá:

- Sử dụng khoảng cách Gower với các chỉ số Silhouette, Davies-Bouldin, Calinski-Harabasz để đánh giá các chỉ số phân cụm với dữ liệu hỗn hợp liên tục và nhị phân.

Phân cụm kết hợp khai thác tập phổ biến tối đại

Các đặc trưng nhị phân mới được thêm vào phân cụm:

1. **K-Prototypes:**
   - ChatLow_YearLow_FollowersLow_revenueLow.
   - YearLow_PositiveHigh.
2. **Agglomerative:**
   - RatingMedium_FollowersLow_ChatLow.
   - RatingHigh_revenueLow_FollowersLow_ChatLow.
   - revenueLow_RatingMedium_ChatLow.
3. **Gaussian Mixture Model:**
   - ChatLow_revenueLow_PositiveHigh.
   - ChatLow_FollowersLow_PositiveHigh.
   - FollowersLow_revenueLow_PositiveHigh.

Kết quả

Phân cụm K-Means & K-Prototype

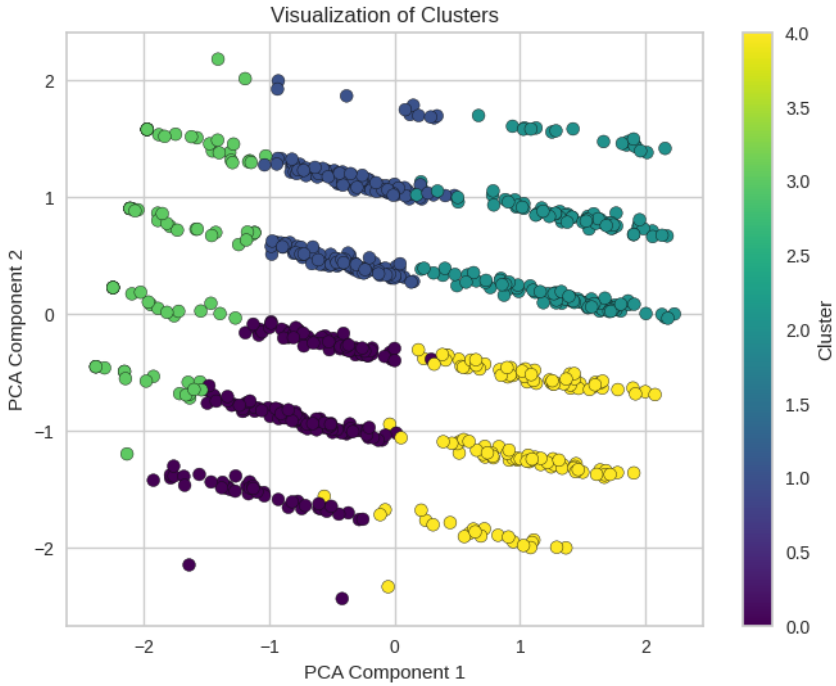


Figure: K-Means trước khi kết hợp đặc trưng mới

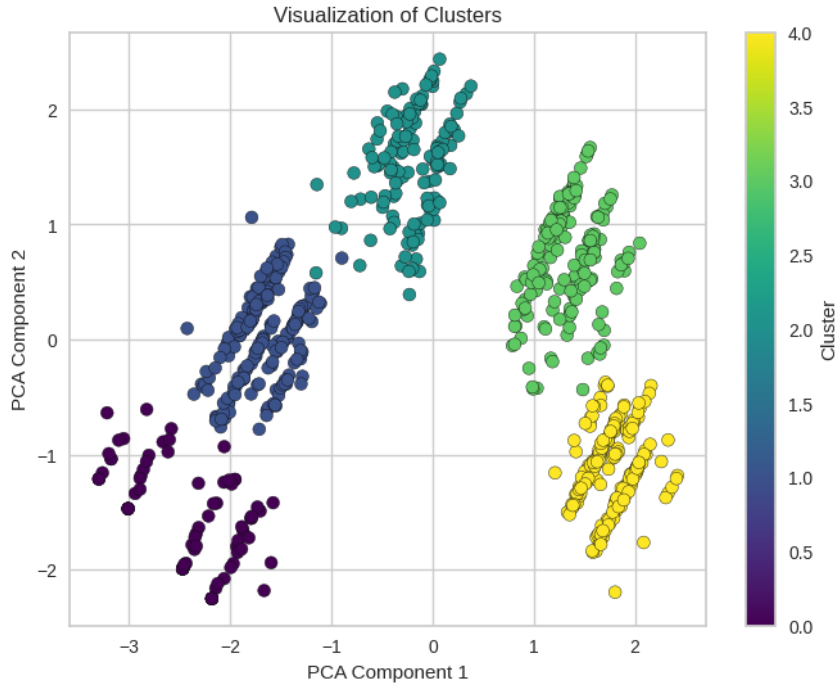


Figure: K-Prototypes sau khi kết hợp đặc trưng mới

- Silhouette: 0.325 -> 0.459 (tăng 0.134 ==> tốt)
- Davies-Bouldin Index: 1.088 -> 0.963 (giảm 0.125 ==> tốt)
- Calinski-Harabasz Index: 415.935 -> 495.393 (tăng 79.458 ==> tốt)

Kết quả

Phân cụm Agglomerative

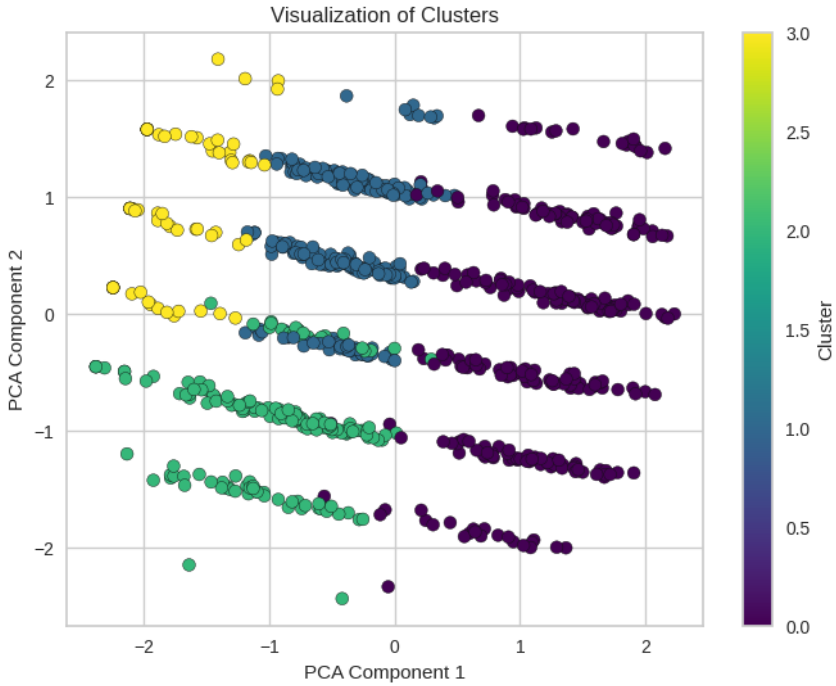


Figure: Agglomerative trước khi kết hợp đặc trưng mới.

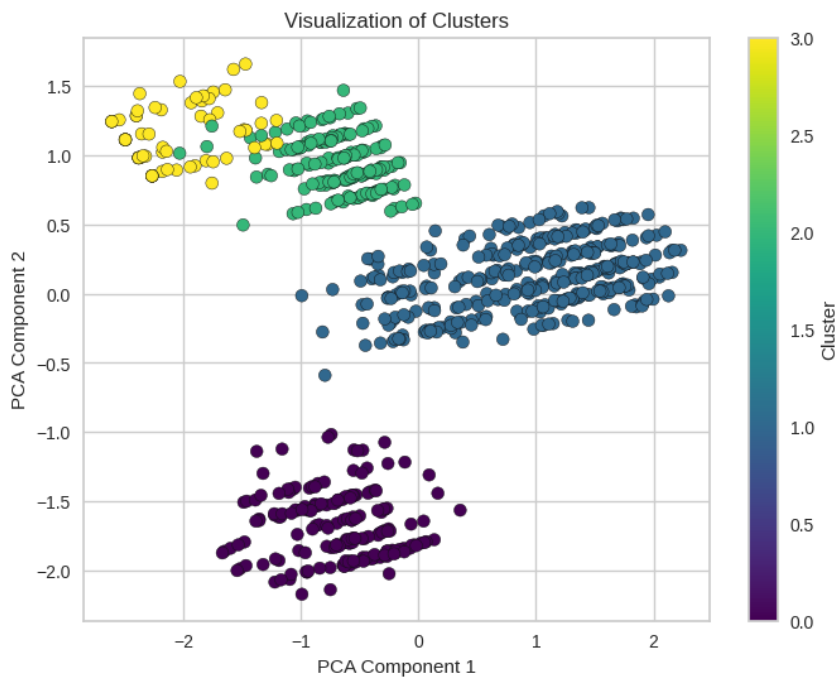


Figure: Agglomerative sau khi kết hợp đặc trưng mới.

- Silhouette: 0.308 -> 0.553 (tăng 0.245 ==> tốt)
- Davies-Bouldin Index: 1.036 -> 0.628 (giảm 0.408 ==> tốt)
- Calinski-Harabasz Index: 361.240 -> 1451.059 (tăng 1089.819 ==> tốt)

Kết quả

Phân cụm GMM (Gaussian Mixture Model)

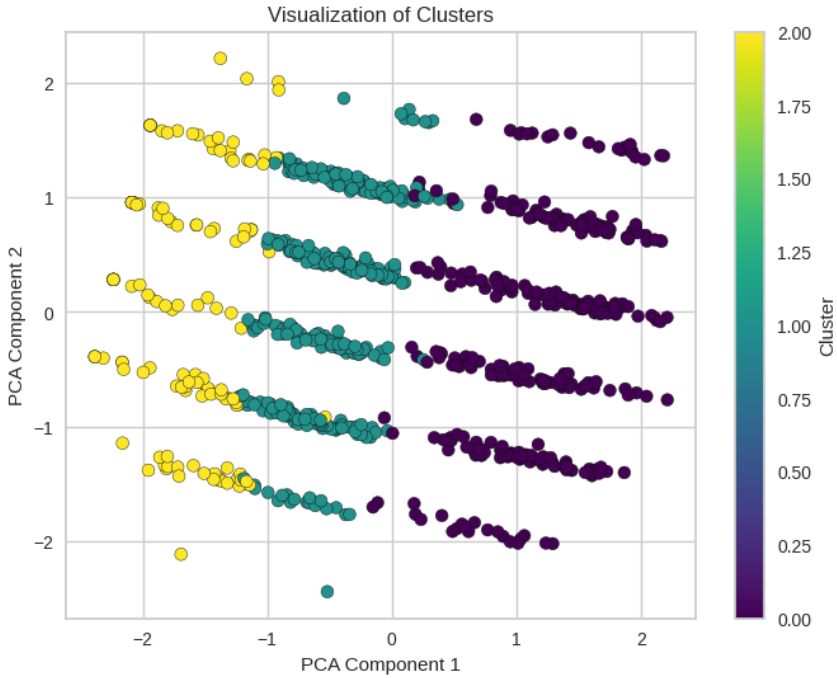


Figure: GMM trước khi kết hợp đặc trưng mới.

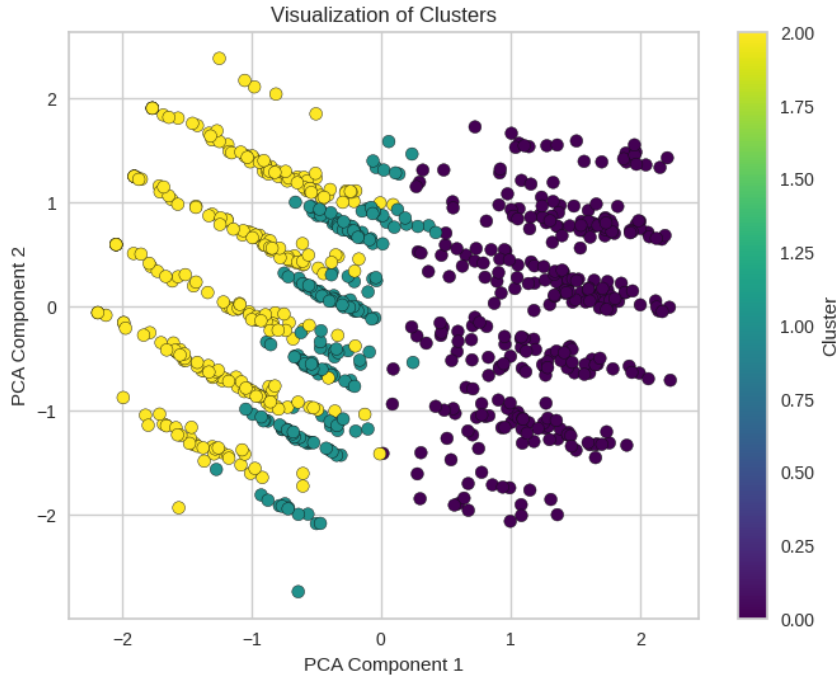


Figure: GMM sau khi kết hợp đặc trưng mới.

- Silhouette: 0.325 -> 0.341 (tăng 0.016 ==> tốt)
- Davies-Bouldin Index: 1.249 -> 1.234 (giảm 0.015 ==> tốt)
- Calinski-Harabasz Index: 378.246 -> 373.191 (giảm 5.055 ==> xấu)

Phân tích

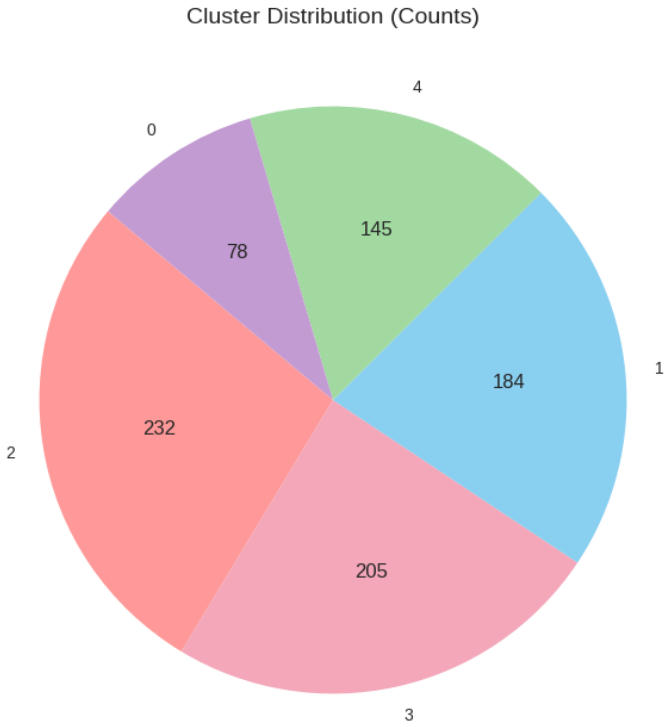


Figure: Pie Chart các cụm

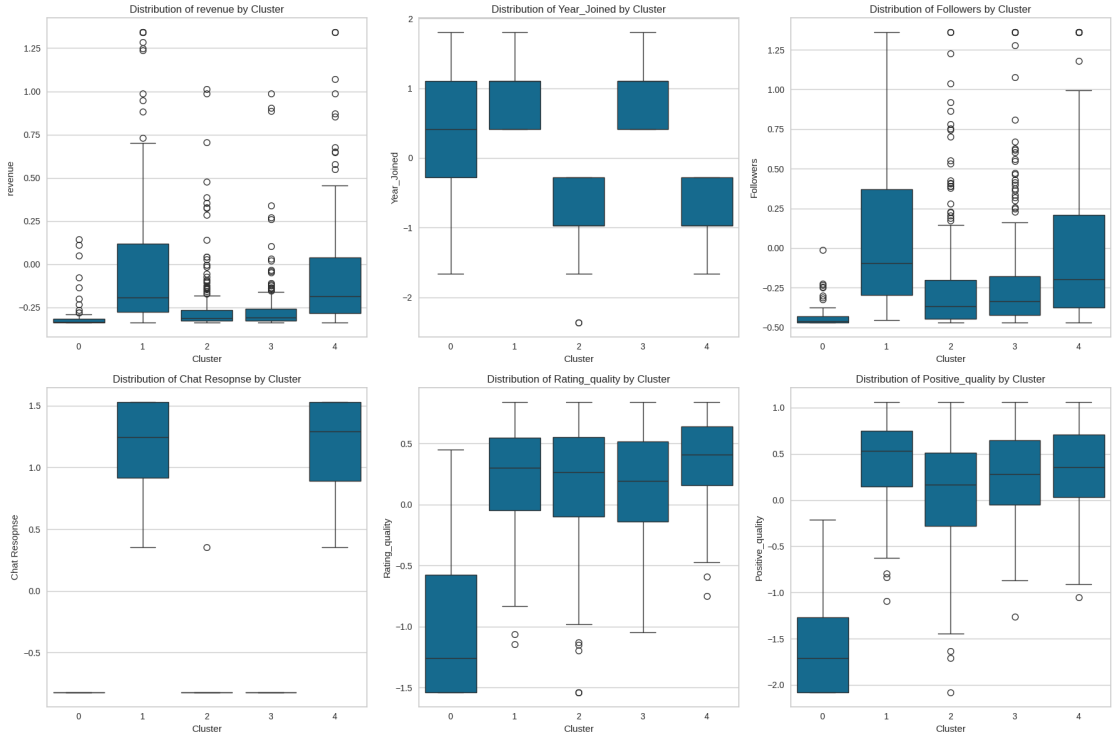


Figure: Boxplot các cụm

Phân tích

Cluster	Analyze
1	Thời gian hoạt động khá lâu nhưng hiệu quả kinh doanh thấp. Doanh thu, lượng người theo dõi, và phản hồi tích cực đều ở mức rất thấp. Chất lượng dịch vụ cũng chưa được đánh giá cao. Đây là nhóm cửa hàng cần cải thiện toàn diện về mọi khía cạnh.
2	Đây là các cửa hàng mới gia nhập thị trường, với hiệu quả kinh doanh và chất lượng dịch vụ tương đối tốt. Doanh thu và các chỉ số khác dao động lớn giữa các cửa hàng, cho thấy sự không đồng đều về hiệu quả hoạt động.
3	Nhóm này chủ yếu là các cửa hàng lâu năm nhưng hoạt động không hiệu quả. Doanh thu và các chỉ số liên quan đều thấp, tuy nhiên một số cửa hàng vẫn giữ được điểm tích cực. Đây là nhóm cần cải thiện hiệu quả kinh doanh nhưng vẫn có tiềm năng để phát triển.
4	Các cửa hàng trong cụm này tương đối mới và có hiệu quả trung bình. Điểm phản hồi tích cực và chất lượng dịch vụ không cao nhưng ổn định. Nhóm này có tiềm năng phát triển nếu cải thiện thêm chất lượng và trải nghiệm khách hàng.
5	Đây là nhóm cửa hàng lâu năm hoạt động hiệu quả nhất. Doanh thu dao động lớn nhưng điểm chất lượng và phản hồi tích cực cao hơn các cụm khác. Nhóm này có thể được coi là các cửa hàng tiêu biểu về hiệu quả kinh doanh và dịch vụ.

Table: Phân tích chung cho từng cụm

Kết luận

Kết luận

Đóng góp:
==> **Về mặt lý thuyết**: Phương pháp kết hợp giữa phân cụm và khai thác tập phổ biến tối đại giúp nâng cao đáng kể kết quả phân cụm.
==> **Về mặt thực tế**: Phân cụm các cửa hàng, hỗ trợ hoạt động kinh doanh của sàn Tiki.

Cảm ơn mọi người đã lắng nghe!