NGHIÊN CỨU PHƯƠNG PHÁP CẢI THIỆN PHÂN CỤM DỮ LIỆU CÁC CỬA HÀNG TRÊN SÀN THƯƠNG MẠI ĐIỆN TỬ TIKI

Ứng dụng kết hợp khai thác mẫu thường xuyên và học máy

Nguyễn Văn Đạt - 2186400229

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

Nội Dung

Tổng quan đề tài

Lý do chọn đề tài

- Tầm quan trọng của hiểu biết thị trường: Giúp đưa ra quyết định chiến lược hiệu quả.
- Phân tích phân cụm: Nhận diện nhóm cửa hàng có đặc điểm tương đồng.
- Ứng dụng thực tiễn:
  - Tối ưu hóa hoạt động kinh doanh.
  - Cải thiện dịch vụ khách hàng.
  - Nâng cao hiệu quả marketing.
- Ý nghĩa phát triển chiến lược:
  - Hỗ trợ doanh nghiệp nắm bắt xu hướng.
  - Phát triển chiến lược kinh doanh bền vững.

Mục tiêu đề tài

- Thu thập và xử lý dữ liệu.
- Phân cụm dữ liệu.
- Tối ưu hóa hiệu quả phân cụm.
- Phân tích kết quả phân cụm.
- Đề xuất chiến lược phù hợp.

Chuẩn bị dữ liệu

Thu thập và tiền xử lý dữ liệu

Thu thập dữ liệu:

- Dữ liệu thu thập từ sàn thương mại điện tử Tiki bằng phương pháp API Scraping.
- Dữ liệu ban đầu gồm: 73.000 mẫu chứa thông tin về các cửa hàng, thông tin về sản phẩm của cửa hàng và các đánh giá của khách hàng trên từng sản phẩm.
- Dữ liệu sau khi tiền xử lý dữ liệu.

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: Bảng dữ liệu hoàn chỉnh sau khi áp dụng định dạng.

Thu thập và tiền xử lý dữ liệu

Tiền xử lý dữ liệu cột "Revenue":

$$\text{Revenue}_{\text{store}} = \sum \left((\text{QuantitySold}_{1\ \text{month}} - \text{QuantitySold}_{\text{initial}}) \cdot \text{price}_{\text{product}} \right)$$

- Trong đó:

- **Revenue**$_{store}$: Doanh thu ước tính của từng cửa hàng trong 1 tháng.
- **QuantitySold**$_{initial}$: Số lượng bán sau thu thập lần đầu.
- **QuantitySold**$_{1month}$: Số lượng bán sau thu thập lần đầu sau 1 tháng.
- **price**$_{product}$: Giá sản phẩm cửa hàng bán.

Thu thập và tiền xử lý dữ liệu

Tiền xử lý dữ liệu cột "RatingQuality":

- Là sự kết hợp với 2 biến **ShopRating** và **CounterRating**.
- Sử dụng phương pháp **Wilson score Interval** để đánh giá độ tin cậy của các cửa hàng để tạo ra thứ hạng khách quan hơn.
- Ưu tiên các cửa hàng có số lượng đánh giá lớn hơn và đáng tin cậy hơn.
- Ví dụ:
  - Một cửa hàng có **ShopRating** là 5.0 nhưng chỉ có 1 **CounterRating** thì độ tin cậy thấp.
  - Một cửa hàng khác có **ShopRating** là 4.5 nhưng có 500 **CounterRating** thì đáng tin cậy hơn.

Thu thập và tiền xử lý dữ liệu

Tiền xử lý dữ liệu cột "PositiveQuality":

- Ban đầu dữ liệu gồm các bình luận của khách hàng về sản phẩm.
- Sử dụng thư viện **underthesea** để:
  - Chuẩn hóa dữ liệu tiếng Việt (Normalize).
  - Phân đoạn câu (Tokenize).
  - Phân tích cảm xúc đánh giá của khách hàng (Sentiment).

==> **Positive, Negative, TotalFeedback:** Số lượng đánh giá tích cực, tiêu cực và tổng số lượng nhận xét của từng cửa hàng.

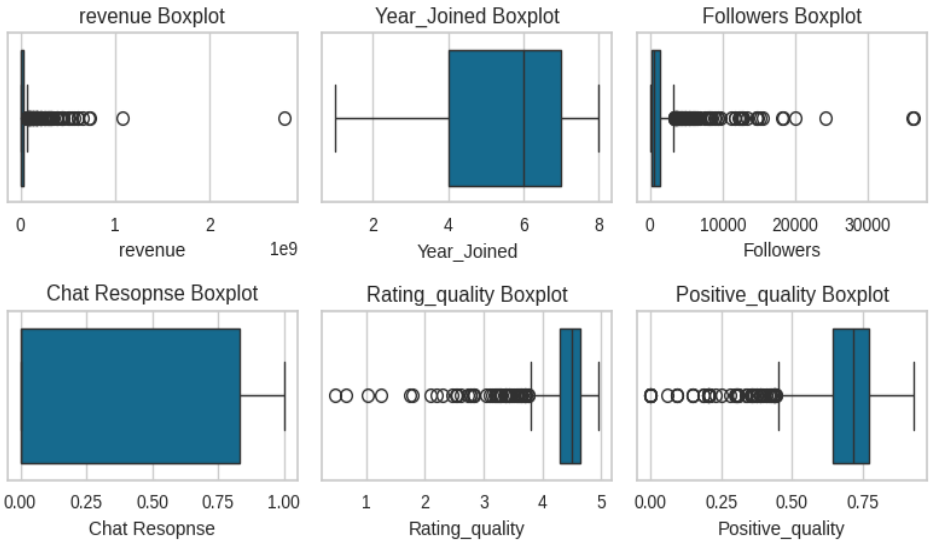
Thu thập và tiền xử lý dữ liệu

Tiền xử lý dữ liệu cột "PositiveQuality":

- Là sự kết hợp với 2 biến **Positive** và **TotalFeedback**.
- Sử dụng phương pháp **Wilson score Interval** để đánh giá độ tin cậy của các cửa hàng để tạo ra thứ hạng khách quan hơn.
- Ưu tiên các cửa hàng có số lượng nhận xét lớn hơn và đáng tin cậy hơn.
- Ví dụ:
  - Một cửa hàng có **Positive** là 20 trên 29 **TotalFeedback** thì độ tin cậy thấp.
  - Một cửa hàng khác có **Positive** là 20 nhưng có 24 **TotalFeedback** thì đáng tin cậy hơn.

Khám phá dữ liệu

Trực quan hóa dữ liệu



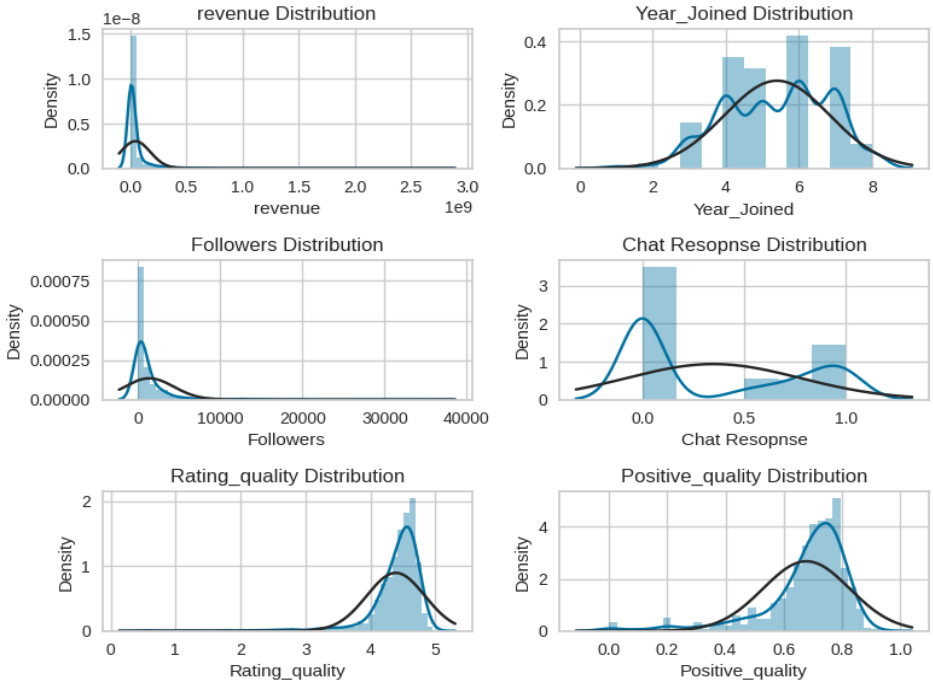
Ảnh 1: Biểu đồ BoxPlot.

Ảnh 2:

- Quan sát xu hướng phân phối của dữ liệu.
- Xác định ngoại lai.

Ảnh 1:

- Tóm tắt dữ liệu(interquantile, min, max).
- Phát hiện các điểm ngoại lai.



Ảnh 2: Biểu đồ phân phối.

Chuẩn hóa dữ liệu

Chuẩn hóa dữ liệu bằng ScalerStandard (Z-score): Đưa dữ liệu về trung bình (=0) và độ lệch chuẩn (=1).

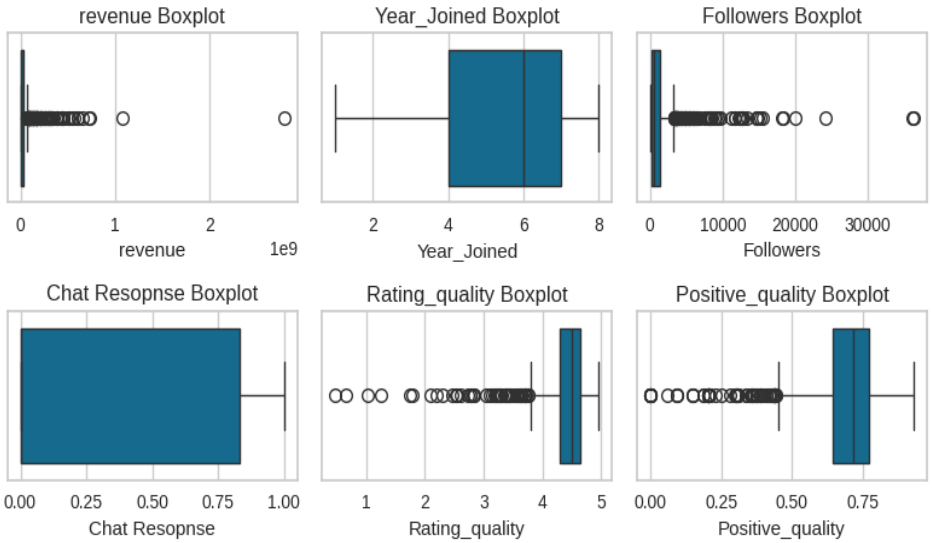
- **Đưa dữ liệu về thang đo chung:** Loại bỏ sự khác biệt về đơn vị đo lường.
- **Cải thiện hiệu suất mô hình:** Đặc biệt quan trọng với các thuật toán nhạy cảm với thang đo.

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 phương pháp Winsorization:



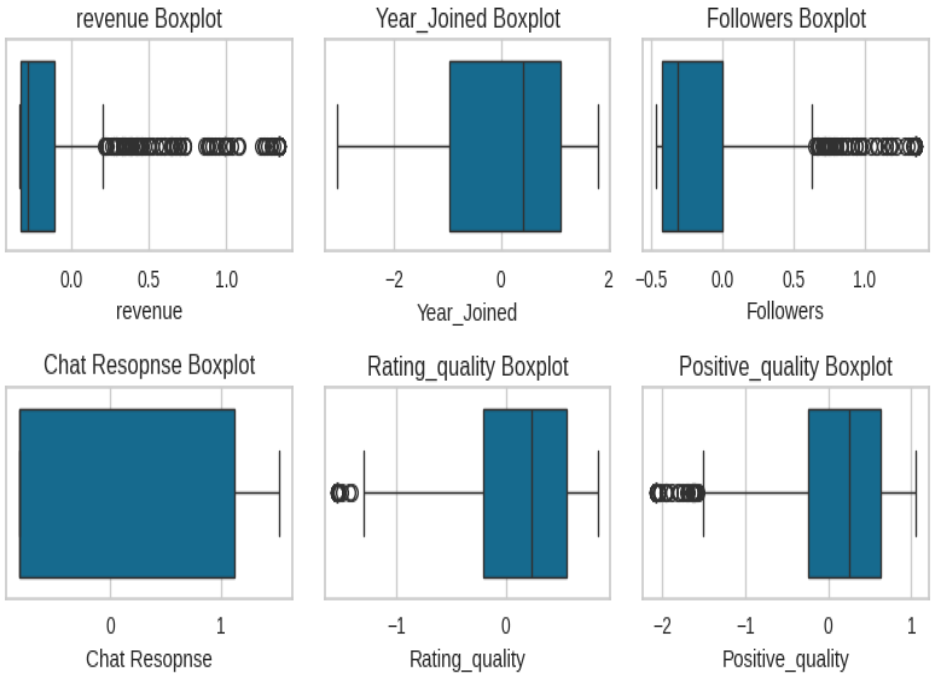
Ảnh 1: Trước khi áp dụng Winsorization.

Ưu điểm:

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

Nhược điểm:

- Không giải quyết được mọi vấn đề ngoại lệ.
- Không phù hợp với dữ liệu ngoại lệ có ý nghĩa.



Ảnh 2: Sau khi áp dụng Winsorization.

Xử lý Outliers

Xử lý Outliers bằng phương pháp Isolation Forest:

Isolation Forest là gì?

- Thuật toán phát hiện bất thường dựa trên nguyên tắc ”cô lập” (isolation).
- Các điểm bất thường dễ bị cô lập hơn vì chúng nằm cách xa các cụm chính.
- Sử dụng cây ngẫu nhiên để chia không gian dữ liệu.

Lợi ích:

- Loại bỏ triệt để nhiễu từ bước Winsorization.
- Làm việc tốt với dữ liệu đa chiều.
- Tăng độ tin cậy của phân tích.

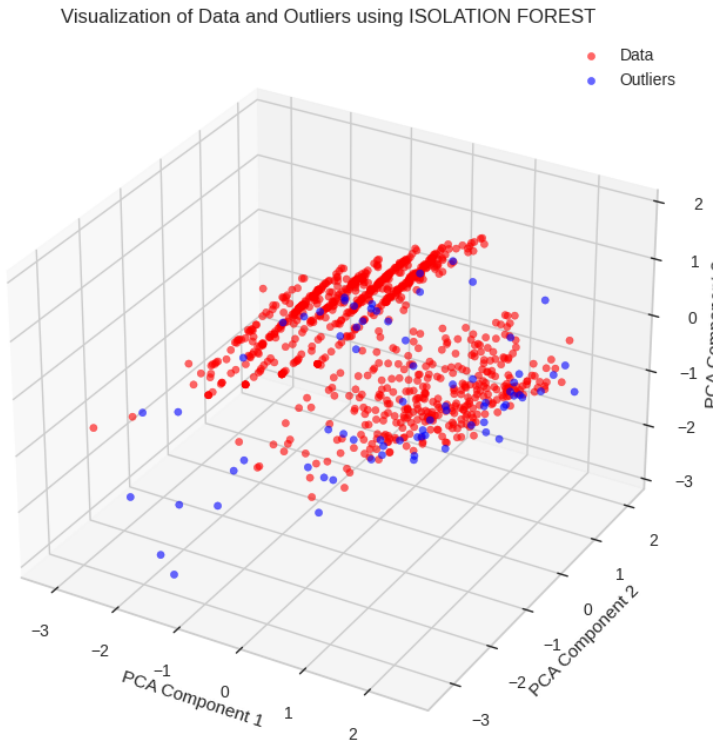


Figure: Isolation Forest.

Phân tích độ tương quan

Phân cụm trước khi sử dụng khai thác mẫu thường xuyên

Loại bỏ nhiễu từ DBSCAN

Tìm số cụm tối ưu bằng Silhouette, AIC, BIC

Phân cụm K-Means

Phân cụm GMM (Gaussian Mixture Model)

Phân cụm Agglomerative

Phân cụm sau khi sử dụng khai thác mẫu thường xuyên

Rời rạc hóa dữ liệu K-Bins Discretizer

Áp dụng khai thác mẫu thường xuyên FP-Max

Phân cụm K-Means

Phân cụm GMM (Gaussian Mixture Model)

Phân cụm Agglomerative

Kết luận