DỰ ÁN
PHÂN TÍCH HÀNH VI KHÁCH HÀNG & PHÂN KHÚC (CUSTOMER SEGMENTATION)

1. Tổng quan dự án

Dự án này tập trung vào việc phân tích và phân khúc khách hàng dựa trên hành vi mua sắm của họ trong lĩnh vực thương mại điện tử. Bằng cách sử dụng phương pháp RFM (Recency, Frequency, Monetary) và thuật toán K-Means Clustering, mục tiêu là xác định các nhóm khách hàng khác nhau để hỗ trợ các chiến dịch marketing mục tiêu, tối ưu hóa mức độ tương tác của khách hàng và thúc đẩy tăng trưởng doanh thu.

2. Quy trình phân tích dữ liệu từng bước

Dự án được thực hiện theo một quy trình phân tích dữ liệu chuẩn mực, bao gồm các giai đoạn chính sau:

2.1. Nhận định vấn đề (Problem Definition & Objectives)

- **Vấn đề:** Trong một thị trường cạnh tranh cao như thương mại điện tử, việc hiểu rõ khách hàng là tối quan trọng. Tuy nhiên, việc áp dụng các chiến lược marketing "một kích cỡ cho tất cả" (one-size-fits-all) thường không hiệu quả. Cần một cách tiếp cận cá nhân hóa hơn.
- **Mục tiêu:**
    - Phân loại khách hàng thành các nhóm riêng biệt dựa trên hành vi mua hàng của họ.
    - Xác định các phân khúc khách hàng có giá trị cao, khách hàng tiềm năng, và khách hàng có nguy cơ rời bỏ.
    - Cung cấp thông tin chi tiết để định hướng các chiến lược marketing mục tiêu và giữ chân khách hàng hiệu quả.

2.2. Thu thập dữ liệu (Data Collection)

- **Nguồn dữ liệu:** Bộ dữ liệu "Online Retail Dataset" được tải từ Kaggle (nguồn gốc từ UCI Machine Learning Repository).
- **Mô tả:** Dữ liệu chứa thông tin về các giao dịch bán lẻ trực tuyến của một công ty có trụ sở tại Vương quốc Anh trong giai đoạn từ 01/12/2010 đến 09/12/2011.

- **Các trường dữ liệu quan trọng:** InvoiceNo, InvoiceDate, Quantity, UnitPrice, và CustomerID.

2.3. Làm sạch & Tiền xử lý dữ liệu (Data Cleaning & Preprocessing)

Mục tiêu: đảm bảo chất lượng dữ liệu trước khi phân tích:

- **Xử lý giá trị thiếu:** Loại bỏ các bản ghi không có CustomerID (khoảng 25% dữ liệu gốc) vì không thể phân khúc khách hàng mà không có định danh duy nhất.
- **Xử lý giá trị không hợp lệ:** Lọc bỏ các giao dịch có Quantity hoặc UnitPrice nhỏ hơn hoặc bằng 0, vì chúng không đại diện cho hành vi mua hàng thực tế (ví dụ: trả hàng, ghi nhận lỗi).
- **Chuyển đổi kiểu dữ liệu:** Đảm bảo cột InvoiceDate được chuyển đổi sang định dạng datetime để phục vụ cho các phép tính liên quan đến thời gian.
- **Tạo cột TotalPrice:** Tính toán tổng giá trị mỗi mặt hàng trong một giao dịch (Quantity * UnitPrice) để phục vụ cho chỉ số Monetary.

2.4. Xây dựng đặc trưng (Feature Engineering)

Để nắm bắt hành vi khách hàng, mô hình RFM đã được xây dựng:

- **Recency (R):** Số ngày kể từ lần mua hàng cuối cùng của khách hàng. Khách hàng càng mua gần đây, Recency càng thấp và giá trị càng cao.
- **Frequency (F):** Tổng số giao dịch duy nhất của khách hàng. Khách hàng mua sắm càng thường xuyên, Frequency càng cao.
- **Monetary (M):** Tổng số tiền mà khách hàng đã chi tiêu. Khách hàng chi tiêu càng nhiều, Monetary càng cao.
- **Phương pháp:** Dữ liệu giao dịch được nhóm theo CustomerID và các hàm tổng hợp (max cho Recency, nunique cho Frequency, sum cho Monetary) được áp dụng để tạo ra tập dữ liệu RFM cho mỗi khách hàng.

2.5. Khám phá dữ liệu & Tiền xử lý cho mô hình (Exploratory Data Analysis & Preprocessing for Modeling)

Trước khi áp dụng thuật toán K-Means, dữ liệu RFM được chuẩn bị kỹ lưỡng để tối ưu hóa hiệu suất mô hình:

- **Phân tích phân phối:** Trực quan hóa (sử dụng biểu đồ Histograms) cho thấy các chỉ số RFM có phân phối lệch (skewed), đặc biệt là Frequency và Monetary, và chứa các giá trị ngoại lai.
- **Biến đổi Logarit:** Áp dụng phép biến đổi logarit (sử dụng np.log1p) cho các cột RFM để giảm thiểu ảnh hưởng của các giá trị ngoại lai và làm cho phân phối của dữ liệu gần với phân phối chuẩn hơn, giúp K-Means hoạt động hiệu quả.
- **Chuẩn hóa dữ liệu (Scaling):** Sử dụng StandardScaler từ scikit-learn để chuẩn hóa dữ liệu RFM đã biến đổi logarit. Điều này đảm bảo rằng tất cả các đặc trưng có cùng thang đo (mean=0, standard deviation=1), ngăn chặn các thuộc tính có giá trị lớn hơn chi phối quá trình phân cụm.

2.6. Xây dựng mô hình (Model Building)

Thuật toán K-Means Clustering được sử dụng để phân nhóm khách hàng:

- **Xác định số lượng cụm tối ưu (K):** Phương pháp Elbow Method được sử dụng để tính toán WCSS (Within-Cluster Sum of Squares) cho một dải các giá trị K (từ 1 đến 10). Dựa trên biểu đồ WCSS, điểm "khuỷu tay" được xác định là K=4, cho thấy đây là số lượng cụm tối ưu để phân tách dữ liệu.
- **Huấn luyện mô hình K-Means:** Mô hình K-Means được huấn luyện với n_clusters=4 trên dữ liệu RFM đã chuẩn hóa.
- **Gán nhãn cụm:** Nhãn cụm (Cluster ID) được gán cho từng khách hàng, thêm vào DataFrame RFM gốc.

2.7. Đánh giá & Diễn giải kết quả (Model Evaluation & Interpretation)

Chuyển đổi kết quả thống kê thành thông tin kinh doanh có giá trị:

- **Phân tích đặc điểm cụm:** Tính toán giá trị trung bình của Recency, Frequency, và Monetary cho mỗi cụm. Đồng thời, xác định số lượng và tỷ lệ phần trăm khách hàng trong mỗi cụm.
- **Trực quan hóa:** Sử dụng biểu đồ thanh (Bar charts) để so sánh các giá trị RFM trung bình giữa các cụm, giúp nhận diện rõ ràng sự khác biệt về hành vi.
- **Diễn giải & Đặt tên cụm (Customer Personas):** Dựa trên đặc điểm RFM của từng nhóm, các cụm được đặt tên có ý nghĩa kinh doanh, ví dụ:
    - **Cụm 1: "Khách hàng ưu tú / Trung thành & Giá trị cao nhất"** (Recency rất thấp, Frequency rất cao, Monetary rất cao).

- **Cụm 0: "Khách hàng tiềm năng / Mới nổi"** (Recency thấp, Frequency thấp, Monetary thấp).
- **Cụm 2: "Khách hàng cần sự chú ý / Ngủ đông"** (Recency trung bình, Frequency trung bình, Monetary trung bình).
- **Cụm 3: "Khách hàng có nguy cơ rời bỏ / Đã rời bỏ"** (Recency rất cao, Frequency rất thấp, Monetary rất thấp).

2.8. Triển khai & Đề xuất (Deployment & Recommendations)

- **Xuất kết quả:** DataFrame RFM với nhãn cụm (customer_segmentation_rfm_results.csv) và bảng tóm tắt đặc điểm cụm (cluster_summary.csv) được xuất ra file CSV.
- **Trực quan hóa bằng Power BI:** Dữ liệu CSV được nhập vào Power BI để xây dựng một dashboard tương tác. Dashboard bao gồm:
  - Biểu đồ Donut chart về phân phối khách hàng theo cụm.
  - Bảng chi tiết đặc điểm RFM trung bình của từng cụm.
  - Biểu đồ phân tán (Scatter plot) trực quan hóa sự phân tách cụm trên không gian Recency vs. Monetary.
  - Thẻ (Card) tổng số khách hàng.
- **Đề xuất chiến lược:** Dựa trên các đặc điểm của từng cụm, đưa ra các đề xuất cụ thể về chiến lược marketing mục tiêu và cá nhân hóa trải nghiệm khách hàng cho từng nhóm.

---

3. Công cụ và Công nghệ

- Ngôn ngữ lập trình: Python
- Thư viện Python:
  - pandas: Xử lý và thao tác dữ liệu.
  - numpy: Hỗ trợ các phép toán số học.
  - scikit-learn: Thực hiện K-Means Clustering và chuẩn hóa dữ liệu (StandardScaler).
  - matplotlib & seaborn: Trực quan hóa dữ liệu và kết quả phân cụm.

    - openpyxl: Đọc file Excel.
- Công cụ Business Intelligence (BI): Power BI Desktop
- Hệ thống kiểm soát phiên bản: Github.