BỘ CÔNG THƯƠNG

TRƯỜNG ĐẠI HỌC KINH TẾ KỸ THUẬT CÔNG NGHIỆP

KHOA KHOA HỌC ỨNG DỤNG

BÁO CÁO TỔNG KẾT

HỌC PHẦN: ĐỒ ÁN 2

PHÂN TÍCH HÀNH VI MUA SẮM, PHÂN NHÓM KHÁCH HÀNG VÀ DỰ ĐOÁN KHẢ NĂNG QUAY LẠI CỦA KHÁCH HÀNG TỪ DỮ LIỆU BÁN LẺ

Sinh viên thực hiện:

NGUYỄN KHANG	DHKL16A1HN	22174600062
NGUYỄN VĂN HOÀNG	DHKL16A1HN	22174600023
LÊ THỊ LAN	DHKL16A1HN	22174600093
PHÙNG THỊ LINH	DHKL16A1HN	22174600001
NGUYỄN THỊ THANH HOA	DHKL16A1HN	22174600052

Giáo viên giảng dạy: Lê Hằng Anh

Hà Nội, 05/2025

BỘ CÔNG THƯƠNG
TRƯỜNG ĐẠI HỌC KINH TẾ KỸ THUẬT CÔNG NGHIỆP

KHOA KHOA HỌC ỨNG DỤNG

BÁO CÁO TỔNG KẾT

HỌC PHẦN: ĐỒ ÁN 2

PHÂN TÍCH HÀNH VI MUA SẮM, PHÂN NHÓM KHÁCH HÀNG VÀ DỰ ĐOÁN KHẢ NĂNG QUAY LẠI CỦA KHÁCH HÀNG TỪ DỮ LIỆU BÁN LẺ

Sinh viên thực hiện:

NGUYỄN KHANG	DHKL16A1HN	22174600062
NGUYỄN VĂN HOÀNG	DHKL16A1HN	22174600023
LÊ THỊ LAN	DHKL16A1HN	22174600093
PHÙNG THỊ LINH	DHKL16A1HN	22174600001
NGUYỄN THỊ THANH HOA	DHKL16A1HN	22174600052

Giáo viên giảng dạy: Lê Hằng Anh

Hà Nội, 05/2025

PHIẾU ĐĂNG KÝ ĐỀ TÀI

1. Tên đề tài: Phân tích hành vi mua sắm, phân nhóm khách hàng và dự đoán khả năng quay lại của khách hàng từ dữ liệu bán lẻ

2. Thông tin nhóm sinh viên:

Sinh viên 1 (Nhóm trưởng):

- **Họ và tên:** Nguyễn Khang
- **Mã sinh viên:** 22174600062
- **Điện thoại:** 0862648906
- **Email:** nkhang.dhkl16a1hn@sv.uneti.edu.vn

Sinh viên 2:

- **Họ và tên:** Nguyễn Văn Hoàng
- **Mã sinh viên:** 22174600023
- **Điện thoại:** 0365586740
- **Email:** nvhoang.dhkl16a1hn@sn.uneti.edu.vn

Sinh viên 3 :

- **Họ và tên:** Lê Thị Lan
- **Mã sinh viên:** 22174600093
- **Điện thoại:** 0583467602
- **Email:** ltlan.dhkl16a1hn@sv.uneti.edu.vn

Sinh viên 4:

- **Họ và tên:** Phùng Thị Linh
- **Mã sinh viên:** 22174600001
- **Điện thoại:** 0329869246
- **Email:** ptlinh.dhkl16a1hn@sv.uneti.edu.vn

Sinh viên 5:

- **Họ và tên:** Nguyễn Thị Thanh Hoa
- **Mã sinh viên:** 22174600052
- **Điện thoại:** 0352307901

- **Email:** ntthoa.dhkl16a1hn@sv.uneti.edu.vn

3. Tóm tắt nội dung đề tài:

Đề tài "Phân tích hành vi mua sắm, phân nhóm khách hàng và dự đoán khả năng quay lại của khách hàng từ dữ liệu bán lẻ" tập trung vào việc ứng dụng các kỹ thuật học máy để phân tích dữ liệu bán lẻ từ bộ dữ liệu Retail Sales Dataset (Kaggle) lấy từ link: https://www.kaggle.com/datasets/mohammadtalib786/retail-sales-dataset?, bao gồm 1.000 giao dịch với thông tin về khách hàng (độ tuổi, giới tính), sản phẩm (danh mục, giá), và giao dịch (số lượng, tổng chi tiêu, ngày mua). Mục tiêu chính là để hiểu rõ hành vi mua sắm của khách hàng, phân nhóm họ thành các phân khúc tiềm năng và dự đoán khả năng quay lại trong tương lai.

Phương pháp:

- Phân tích hành vi: Khám phá các yếu tố ảnh hưởng đến hành vi mua sắm như nhân khẩu học, sản phẩm, thời gian và giá cả, sử dụng các biểu đồ và bảng thống kê.
- Phân nhóm khách hàng: Áp dụng mô hình RFM (Recency, Frequency, Monetary) và thuật toán phân cụm (K-Means) để phân loại khách hàng thành các nhóm như khách hàng thân thiết, tiềm năng, cần chú ý và rời bỏ.
- Dự đoán khả năng quay lại: Xây dựng mô hình dự đoán bằng Random Forest, Logistic Regression và Decision Tree sau đó đánh giá bằng các chỉ số Accuracy, Precision, Recall, F1-score và ROC AUC.

Kết quả nghiên cứu có thể được ứng dụng để cá nhân hóa trải nghiệm khách hàng, triển khai chiến dịch marketing, xây dựng chương trình khách hàng thân thiết và tối ưu hóa hoạt động kinh doanh.

Ngày 11 tháng 04 năm 2025

Nhóm trưởng

Nguyễn Khang

ĐỀ CƯƠNG CHI TIẾT ĐỀ TÀI

1. Tên đề tài: Phân tích hành vi mua sắm, phân nhóm khách hàng và dự đoán khả năng quay lại của khách hàng từ dữ liệu bán lẻ

2. Mục tiêu đề tài:

Mục tiêu chính của đề tài là để hiểu rõ hành vi mua sắm của khách hàng, phân nhóm họ thành các phân khúc tiềm năng và dự đoán khả năng quay lại trong tương lai.

Cụ thể, đề tài hướng tới:

- Thứ nhất, phân tích và mô tả hành vi mua sắm của khách hàng dựa trên các yếu tố như nhân khẩu học (độ tuổi, giới tính), sản phẩm ưa chuộng, thời gian mua sắm, và giá cả. Thông qua việc khám phá các xu hướng và mô hình ẩn chứa trong dữ liệu, đề tài giúp doanh nghiệp hiểu rõ hơn về đặc điểm và nhu cầu của từng phân khúc khách hàng.
- Thứ hai, phân nhóm khách hàng thành các phân khúc tiềm năng dựa trên mô hình RFM (Recency, Frequency, Monetary). Việc này cho phép doanh nghiệp xác định các nhóm khách hàng quan trọng như khách hàng thân thiết, khách hàng tiềm năng, khách hàng cần chú ý và khách hàng có nguy cơ rời bỏ. Từ đó, doanh nghiệp có thể điều chỉnh chiến lược tiếp thị và chăm sóc khách hàng một cách hiệu quả hơn.
- Thứ ba, xây dựng mô hình dự đoán khả năng quay lại của khách hàng trong tương lai. Bằng cách áp dụng Random Forest, Logistic Regression và Decision Tree, đề tài giúp doanh nghiệp dự đoán được khách hàng nào có khả năng cao sẽ quay lại mua hàng và khách hàng nào có nguy cơ rời bỏ. Thông tin này là vô cùng quý giá để doanh nghiệp triển khai các chiến dịch marketing targeted và các chương trình chăm sóc khách hàng phù hợp, nhằm tăng tỷ lệ giữ chân khách hàng và tối ưu hóa doanh thu.

Tóm lại, mục tiêu của đề tài là cung cấp cho doanh nghiệp một cái nhìn toàn diện và sâu sắc về khách hàng, từ hành vi mua sắm, phân khúc tiềm năng đến khả năng quay lại trong tương lai.

3. Tổng quan tình hình nghiên cứu thuộc lĩnh vực đề tài:

Lĩnh vực phân tích hành vi mua sắm, phân nhóm khách hàng và dự đoán khả năng quay lại đã thu hút sự quan tâm đáng kể từ giới nghiên cứu và các doanh nghiệp trong ngành bán lẻ. Với sự phát triển của công nghệ thu thập và xử lý dữ liệu, cùng với sự tiến bộ của các kỹ thuật học máy, lĩnh vực này đang chứng kiến những bước tiến vượt bậc.

Phân tích hành vi mua sắm là một chủ đề nghiên cứu truyền thống, tập trung vào việc hiểu rõ các yếu tố ảnh hưởng đến quyết định mua hàng của khách hàng. Các nghiên cứu thường sử dụng các phương pháp thống kê và phân tích dữ liệu để khám phá các mô hình và xu hướng trong hành vi mua sắm. Gần đây, với sự phổ biến của dữ liệu lớn và học máy, các phương pháp phân tích hành vi đã được nâng cao, cho phép phân tích dữ liệu phức tạp và đa dạng hơn.

Phân nhóm khách hàng là một kỹ thuật quan trọng giúp doanh nghiệp phân chia khách hàng thành các nhóm nhỏ hơn dựa trên các đặc điểm và hành vi tương đồng. Mô hình RFM (Recency, Frequency, Monetary) là một trong những mô hình phổ biến nhất được sử dụng trong phân nhóm khách hàng. Các nghiên cứu đã chỉ ra rằng việc phân nhóm khách hàng hiệu quả có thể giúp doanh nghiệp cá nhân hóa trải nghiệm khách hàng, tối ưu hóa chiến dịch tiếp thị và tăng doanh thu.

Dự đoán khả năng quay lại là một lĩnh vực nghiên cứu tương đối mới, nhưng đang phát triển nhanh chóng. Các thuật toán học máy như Random Forest, Logistic Regression, Support Vector Machine và Neural Networks đã được sử dụng để dự đoán khả năng khách hàng sẽ quay lại mua hàng trong tương lai. Các nghiên cứu đã chứng minh rằng việc dự đoán chính xác khả năng quay lại có thể giúp doanh nghiệp giảm thiểu chi phí marketing, tăng tỷ lệ giữ chân khách hàng và nâng cao lợi nhuận.

Xu hướng nghiên cứu hiện nay tập trung vào việc kết hợp các phương pháp phân tích hành vi, phân nhóm khách hàng và dự đoán khả năng quay lại để xây dựng một hệ thống quản lý khách hàng toàn diện. Các nhà nghiên cứu cũng đang khám phá tiềm năng của các công nghệ mới như trí tuệ nhân tạo (AI) và học sâu (Deep Learning) để nâng cao hiệu quả của các mô hình dự đoán và phân tích.

Tổng quan tình hình nghiên cứu cho thấy lĩnh vực phân tích hành vi mua sắm, phân nhóm khách hàng và dự đoán khả năng quay lại đang phát triển mạnh mẽ và có nhiều tiềm năng ứng dụng trong thực tế. Đề tài này góp phần vào việc mở rộng hiểu biết về

lĩnh vực này và cung cấp các giải pháp thực tiễn cho doanh nghiệp trong ngành bán lẻ.

4. Nội dung đề tài:

Đề tài này được triển khai theo một quy trình logic, bao gồm các bước sau:

- Thu thập và tiền xử lý dữ liệu:
  - Nguồn dữ liệu: Sử dụng bộ dữ liệu Retail Sales Dataset từ Kaggle, chứa thông tin về 1.000 giao dịch bán lẻ.
  - Tiền xử lý: Làm sạch dữ liệu, xử lý các giá trị thiếu, chuyển đổi dữ liệu sang định dạng phù hợp cho phân tích.
  - Tạo đặc trưng: Tạo các biến mới từ dữ liệu gốc, ví dụ như biến RFM (Recency, Frequency, Monetary) cho mỗi khách hàng.
- Phân tích thăm dò dữ liệu (EDA):
  - Thống kê mô tả: Phân tích các đặc trưng cơ bản của dữ liệu như độ tuổi, giới tính, danh mục sản phẩm, giá cả, số lượng mua,...
  - Trực quan hóa dữ liệu: Sử dụng các biểu đồ (histogram, bar chart, line chart, scatter plot), heatmap để khám phá các xu hướng, mô hình và mối quan hệ giữa các biến.
- Phân tích hành vi mua sắm: Nghiên cứu các yếu tố ảnh hưởng đến hành vi mua sắm của khách hàng, ví dụ như mối quan hệ giữa độ tuổi và danh mục sản phẩm ưa chuộng, xu hướng mua sắm theo thời gian,...
- Phân nhóm khách hàng:
  - Áp dụng mô hình RFM: Tính toán các chỉ số RFM cho mỗi khách hàng và sử dụng chúng để phân nhóm khách hàng.
  - Thuật toán phân cụm: Sử dụng thuật toán K-Means, Spectral Clustering để phân chia khách hàng thành các nhóm dựa trên các đặc trưng RFM.
  - Đánh giá phân cụm: Sử dụng các chỉ số như Silhouette Score để đánh giá chất lượng phân cụm và xác định số lượng cụm tối ưu.
  - Phân tích đặc điểm từng nhóm: Nghiên cứu đặc điểm và hành vi mua sắm của từng nhóm khách hàng để hiểu rõ hơn về các phân khúc khách hàng.
- Dự đoán khả năng quay lại:

- Lựa chọn đặc trưng: Xác định các biến đầu vào quan trọng để dự đoán khả năng quay lại, bao gồm các đặc trưng RFM, nhân khẩu học, và hành vi mua sắm.
- Xây dựng mô hình: Huấn luyện mô hình dự đoán sử dụng các thuật toán học máy như Random Forest, Logistic Regression và Decision Tree.
- Đánh giá mô hình: Đánh giá hiệu quả của mô hình dựa trên các chỉ số như Accuracy, Precision, Recall, F1-score và ROC AUC.
- So sánh mô hình: So sánh hiệu suất của các mô hình khác nhau để lựa chọn mô hình tốt nhất.

- Kết luận và đề xuất:
  - Tổng kết kết quả: Tóm tắt các kết quả chính của phân tích hành vi, phân nhóm khách hàng và dự đoán khả năng quay lại.
  - Rút ra kết luận: Đưa ra các kết luận quan trọng về hành vi mua sắm của khách hàng và các yếu tố ảnh hưởng đến khả năng quay lại.
  - Đề xuất ứng dụng: Đề xuất các chiến lược kinh doanh và marketing dựa trên kết quả nghiên cứu để nâng cao hiệu quả kinh doanh, tăng tỷ lệ giữ chân khách hàng và tối ưu hóa doanh thu.

5. Phương pháp thực hiện:

Trong quá trình thực hiện đề tài, dựa trên kiến thức đã học trong môn Đồ án 2 nhóm dự kiến sử dụng các mô hình và phương pháp trong lĩnh vực khoa học dữ liệu và học máy, phù hợp với từng mục tiêu cụ thể của đề tài.

Để thực hiện phân nhóm khách hàng dựa trên hành vi mua sắm, nhóm sử dụng các thuật toán phân cụm như K-Means và Spectral Clustering. Thuật toán K-Means sẽ được áp dụng trên các đặc trưng hành vi khách hàng theo mô hình RFM (Recency - tần suất gần nhất mua hàng, Frequency - số lần mua, Monetary - tổng số tiền chi tiêu).

Trong phần dự đoán khả năng quay lại của khách hàng, ba thuật toán được áp dụng bao gồm Random Forest, Logistic Regression và Decision Tree. Hiệu quả của các mô hình được đánh giá thông qua các chỉ số như Accuracy, Precision, Recall, F1-score và ROC AUC. Cuối cùng, dựa trên kết quả đánh giá, các mô hình sẽ được so sánh để lựa chọn phương pháp tối ưu nhất.

6. Phân công công việc:

STT	Họ và tên	Mã sinh viên	Nội dung công việc được phân công
1	Nguyễn Khang	22174600062	Điều phối và giám sát toàn bộ quá trình thực hiện đề tài đảm bảo tiến độ thực hiện theo đúng kế hoạch. Góp ý và hỗ trợ các thành viên giải quyết các vấn đề kỹ thuật phát sinh và đảm bảo sự phối hợp nhịp nhàng giữa các phần việc. Tìm cơ sở lý thuyết đưa vào báo cáo Tổng hợp và kết luận chung
2	Lê Thị Lan	22174600093	Xây dựng tập đặc trưng RFM (Recency, Frequency, Monetary) cho từng khách hàng. Áp dụng thuật toán phân cụm K-Means, Spectral Clustering. Đánh giá chất lượng phân cụm bằng Silhouette Score. Trực quan hóa kết quả phân nhóm khách hàng và mô tả đặc điểm từng nhóm. Viết báo cáo
3	Nguyễn Văn Hoàng	22174600023	Thực hiện phân tích thăm dò dữ liệu (EDA): Trực quan hóa xu hướng bán hàng theo thời gian, loại sản phẩm, khu vực.

			Phân tích mối quan hệ giữa số lượng bán và doanh thu. Phân tích hành vi mua sắm của khách hàng theo nhóm (giới tính, độ tuổi nếu có). Phân tích và đề xuất chiến lược kinh doanh dựa trên kết quả phân tích.
4	Phùng Thị Linh	22174600001	Chuẩn bị dữ liệu cho dự đoán Xây dựng & tinh chỉnh mô hình dự đoán (Random Forest, Logistic Regression, Decision Tree) Đánh giá mô hình (Accuracy, Precision, Recall, F1, ROC AUC) Hỗ trợ viết báo cáo
5	Nguyễn Thị Thanh Hoa	22174600052	Tìm hiểu bộ dữ liệu "Retail Sales Dataset" từ Kaggle: Mô tả cấu trúc dữ liệu và ý nghĩa các biến: mã đơn hàng, mã khách hàng, loại sản phẩm, doanh thu,... Phân tích đặc điểm định tính và định lượng của dữ liệu. Thu thập & tiền xử lý dữ liệu (cleaning, feature engineering) Làm slide báo cáo

7. Dự kiến kết quả đạt được:

Dự kiến, nhóm sẽ xây dựng được một hệ thống phân tích dữ liệu bán lẻ hoàn chỉnh, bao gồm các bước tiền xử lý, phân tích mô tả, mô hình hóa và dự đoán dữ liệu. Hệ

thống này cho phép khám phá và trực quan hóa các đặc điểm mua sắm của khách hàng, doanh số theo thời gian, loại sản phẩm và khu vực kinh doanh.

Ngoài ra, nhóm dự kiến sẽ phân chia khách hàng thành các nhóm mục tiêu dựa trên hành vi mua sắm sử dụng các kỹ thuật phân cụm như K-Means hoặc Spectral Clustering. Đồng thời, nhóm sẽ xây dựng và đánh giá các mô hình dự đoán khả năng quay lại của khách hàng, bao gồm các thuật toán như Random Forest, Logistic Regression và Decision Tree, nhằm giúp doanh nghiệp chủ động nhận diện khách hàng tiềm năng quay lại.

Kết quả này giúp doanh nghiệp xác định các phân khúc khách hàng tiềm năng để xây dựng chiến lược marketing và chăm sóc phù hợp, tăng hiệu quả giữ chân khách hàng.

Ngày 11 tháng 04 năm 2025

Nhóm trưởng

Nguyễn Khang

MỞ ĐẦU

Trong những năm gần đây, ngành bán lẻ đã có nhiều bước phát triển vượt bậc nhờ vào sự bùng nổ của công nghệ số và xu hướng mua sắm trực tuyến. Tuy nhiên, cùng với sự phát triển nhanh chóng đó, các doanh nghiệp bán lẻ cũng phải đối mặt với nhiều thách thức như: cạnh tranh gay gắt, biến động nhu cầu khách hàng, quản lý hàng tồn kho, và tối ưu hóa chiến lược tiếp thị. Trong bối cảnh đó, việc khai thác và phân tích dữ liệu bán lẻ trở thành một nhu cầu tất yếu, giúp doanh nghiệp nắm bắt xu hướng, dự báo doanh số và đưa ra quyết định kinh doanh chính xác hơn.

Vấn đề đặt ra hiện nay là làm sao để các doanh nghiệp bán lẻ có thể tận dụng hiệu quả nguồn dữ liệu khổng lồ đang có để hiểu rõ hành vi khách hàng, tăng doanh thu, giảm thiểu chi phí vận hành và đặc biệt là dự đoán được khả năng quay lại của khách hàng? Làm thế nào để xây dựng một hệ thống phân tích thông minh, có khả năng phân nhóm khách hàng, dự báo hành vi và đề xuất chiến lược kinh doanh tối ưu?

Xuất phát từ thực tiễn đó, nhóm em đã thực hiện đề tài "Phân tích hành vi mua sắm, phân nhóm khách hàng và dự đoán khả năng quay lại của khách hàng từ dữ liệu bán lẻ". Mục tiêu của đề tài là vận dụng các kỹ thuật trong lĩnh vực học máy (machine learning) để phân tích dữ liệu bán lẻ thực tế, từ đó giúp doanh nghiệp hiểu rõ hành vi tiêu dùng, phân khúc khách hàng hiệu quả và dự đoán khả năng khách hàng sẽ quay lại mua sắm. Nhóm em tin tưởng rằng đây là một đề tài có tính thực tiễn cao và hoàn toàn có thể áp dụng vào hoạt động kinh doanh thực tế của nhiều công ty bán lẻ hiện nay.

Trong quá trình thực hiện đề tài, em đã nhận được sự hướng dẫn tận tình, sâu sắc và đầy tâm huyết từ Cô Lê Hằng Anh. Nhóm em xin chân thành cảm ơn Cô vì đã truyền đạt cho chúng em những kiến thức quý báu và giúp chúng em từng bước hoàn thiện đề tài. Tuy nhiên, do hạn chế về kiến thức, kinh nghiệm cũng như thời gian thực hiện, bài làm của nhóm em chắc chắn không tránh khỏi những thiếu sót. Chúng em rất mong nhận được sự góp ý chân thành của Cô để có thể hoàn thiện đề tài một cách tốt hơn.

Đồ án bao gồm các phần được phân chương như sau:

Chương 1: Đặt vấn đề

Chương 2: Cơ sở lý thuyết

Chương 3: Thực nghiệm

Chương 4: Kết quả đạt được

MỤC LỤC

MỤC LỤC HÌNH VẼ

MỤC LỤC BẢNG BIỂU

DANH MỤC CÁC CHỮ VIẾT TẮT

AUC	Area Under the Curve – Diên tích dưới đường cong ROC
DT	Decision Tree – Cây quyết định
EDA	Exploratory Data Analysis – Phân tích thăm dò dữ liệu
F1-score	Trung bình điều hòa giữa Precision và Recall
FN	False Negative – Dự đoán sai lớp tiêu cực
FP	False Positive – Dự đoán sai lớp tích cực
F_Score	Điểm Frequency – Tần suất mua hàng
IQR	Interquartile Range – Khoảng tứ phân vị
LR	Logistic Regression – Hồi quy logistic
M_Score	Điểm Monetary – Giá trị chi tiêu
ML	Machine Learning – Học máy
Precision	Độ chính xác dương – Tỉ lệ dự đoán đúng trong số dự đoán dương tính
Recall	Độ nhạy – Tỉ lệ phát hiện đúng trường hợp dương tính
RF	Random Forest – Rừng ngẫu nhiên
RFM	Recency, Frequency, Monetary – Bộ ba chỉ số hành vi mua hàng
R_Score	Điểm Recency – Đánh giá độ gần đây của lần mua cuối
ROC	Receiver Operating Characteristic – Đường cong đặc trưng hoạt động
SC	Spectral Clustering – Phân cụm phổ
TP	True Positive – Dự đoán đúng lớp tích cực
TN	True Negative – Dự đoán đúng lớp tiêu cực

CHƯƠNG 1: ĐẶT VẤN ĐỀ

1.1. BỐI CẢNH VÀ LÝ DO CHỌN ĐỀ TÀI

Trong bối cảnh kinh tế toàn cầu hóa và cách mạng công nghiệp 4.0 đang diễn ra mạnh mẽ, ngành bán lẻ trên thế giới bao gồm cả tại Việt Nam đã có những bước chuyển mình toàn diện. Sự phát triển vượt bậc của công nghệ số, hệ sinh thái thương mại điện tử và các nền tảng số hóa đã tạo ra một môi trường kinh doanh đầy cạnh tranh nhưng cũng đồng thời mở ra vô vàn cơ hội mới cho các doanh nghiệp. Song song với đó, thói quen tiêu dùng của khách hàng cũng thay đổi rõ nét khi họ ngày càng quan tâm đến trải nghiệm mua sắm thuận tiện, nhanh chóng và cá nhân hóa. Do đó, việc thấu hiểu hành vi, nhu cầu, sở thích cũng như dự đoán xu hướng hành vi của khách hàng ngày càng trở nên quan trọng, trở thành nhân tố quyết định giúp doanh nghiệp giữ vững và phát triển thị phần trong ngành bán lẻ.

Tuy nhiên, nhiều doanh nghiệp bán lẻ quy mô vừa và nhỏ tại Việt Nam hiện vẫn chưa khai thác hiệu quả nguồn dữ liệu khổng lồ mà họ đang nắm giữ. Dữ liệu bán hàng và thông tin khách hàng nếu không được xử lý, phân tích một cách khoa học sẽ trở nên vô dụng hoặc thậm chí là gây tốn kém chi phí lưu trữ, thời gian và nhân lực. Việc thiếu các công cụ phân tích dữ liệu chuyên sâu và ít áp dụng các phương pháp học máy khiến các quyết định kinh doanh thường lệ thuộc vào kinh nghiệm chủ quan, dự báo không chính xác, dẫn đến sai sót trong chiến lược marketing, quản lý tồn kho và hoạch định nguồn lực.

Ngoài ra, chi phí vận hành cũng đang ngày càng gia tăng, đặc biệt là chi phí dành cho hoạt động marketing, quản lý và tối ưu tồn kho. Nếu không dự báo chính xác nhu cầu hoặc không biết phân phối nguồn lực đúng cách, doanh nghiệp dễ rơi vào tình trạng hàng tồn kho cao, lãng phí vốn và giảm khả năng xoay vòng vốn nhanh. Hơn thế nữa, những chương trình chăm sóc khách hàng chưa hiệu quả sẽ làm tăng nguy cơ khách hàng rời bỏ, giảm tỷ lệ giữ chân khách hàng trung thành. Điều này chính là thách thức lớn đòi hỏi các doanh nghiệp bán lẻ phải xây dựng được hệ thống phân tích và dự báo hành vi khách hàng tối ưu, đáp ứng tốt hơn nhu cầu thị trường và nâng cao hiệu quả vận hành.

Việc áp dụng các kỹ thuật học máy và phân tích dữ liệu lớn không chỉ giúp doanh nghiệp nhận diện được các nhóm khách hàng có đặc điểm hành vi tương đồng mà còn dự báo

chính xác khả năng quay lại của khách hàng trong tương lai. Từ đó, doanh nghiệp có thể phát triển các chiến lược tiếp thị cá nhân hóa, gia tăng sự hài lòng và lòng trung thành của khách hàng, đồng thời tối ưu hóa chi phí marketing và nâng cao doanh thu bền vững. Chính vì vậy, lựa chọn và triển khai một hệ thống phân tích dữ liệu bán lẻ toàn diện, thông minh, vừa mang tính dự báo vừa đề xuất chiến lược phù hợp là yêu cầu cấp thiết trong bối cảnh hiện nay.

1.2. MỤC TIÊU BÀI TOÁN

Dựa trên bối cảnh thực tiễn của ngành bán lẻ, đề tài "Phân tích hành vi mua sắm, phân nhóm khách hàng và dự đoán khả năng quay lại của khách hàng từ dữ liệu bán lẻ" được triển khai nhằm đạt được những mục tiêu cụ thể và chi tiết sau:

- Phân tích sâu sắc hành vi mua sắm của khách hàng: Trên cơ sở dữ liệu giao dịch bán hàng thực tế, đề tài sẽ thực hiện khai thác các thông tin về thời gian mua hàng, danh mục sản phẩm khách hàng yêu thích, tần suất mua hàng, giá trị chi tiêu và các đặc điểm nhân khẩu học như giới tính, độ tuổi để xây dựng bức tranh toàn cảnh về thói quen tiêu dùng khác nhau trong từng nhóm khách hàng.
- Phân nhóm khách hàng dựa trên đặc trưng hành vi: Áp dụng mô hình RFM để xây dựng các đặc trưng quan trọng phản ánh mức độ tương tác của khách hàng với sản phẩm dịch vụ. Sử dụng các thuật toán phân cụm tiến tiến như K-Means và Spectral Clustering để phân tách khách hàng thành các phân khúc có đặc điểm và giá trị kinh tế tương tự, hỗ trợ cho việc cá nhân hóa các chiến dịch marketing.
- Dự đoán khả năng quay lại của khách hàng: Mục tiêu quan trọng là xây dựng mô hình học máy có khả năng dự báo chính xác xác suất khách hàng sẽ tiếp tục mua sắm hoặc thoái lui, từ đó giúp doanh nghiệp triển khai các biện pháp giữ chân kịp thời, giảm thiểu tỷ lệ khách hàng rời bỏ.
- Hỗ trợ ra quyết định quản trị và xây dựng chiến lược: Từ kết quả phân tích và mô hình dự báo, đề tài sẽ đề xuất các giải pháp và chiến lược kinh doanh cụ thể, nhằm tối ưu hóa hoạt động tiếp thị, quản lý tồn kho, hoạch định nguồn lực, từ đó gia tăng doanh thu, giảm chi phí và nâng cao khả năng cạnh tranh trên thị trường bán lẻ đầy biến động ngày nay.
- Xây dựng quy trình và hệ thống phân tích dữ liệu tự động: Đề tài hướng tới xây dựng một hệ thống phân tích toàn diện, từ tiền xử lý, trực quan hóa, phân nhóm, dự báo

cho đến đề xuất chiến lược nhằm hỗ trợ doanh nghiệp triển khai ứng dụng thực tế một cách bài bản, hiệu quả.

1.3. ĐỐI TƯỢNG VÀ PHẠM VI NGHIÊN CỨU

Đề tài sử dụng bộ dữ liệu "Retail Sales Dataset" từ nền tảng Kaggle làm nguồn dữ liệu chính, bao gồm khoảng 1.000 giao dịch bán lẻ với các trường thông tin: mã đơn hàng, ngày giao dịch, danh mục sản phẩm, số lượng bán, tổng giá trị đơn hàng, cùng với các đặc điểm nhân khẩu học của khách hàng như độ tuổi và giới tính.

Phạm vi nghiên cứu tập trung vào phân tích hành vi mua sắm, phân nhóm khách hàng theo mô hình RFM kết hợp hai thuật toán phân cụm K-Means và Spectral Clustering, cũng như xây dựng mô hình dự đoán khả năng quay lại của khách hàng bằng ba thuật toán học máy là Random Forest, Logistic Regression và Decision Tree dựa trên dữ liệu lịch sử mua hàng. Đề tài không mở rộng phân tích sang các yếu tố bên ngoài như vùng địa lý hay yếu tố thời tiết nhằm tập trung sâu vào khai thác đặc điểm hành vi và dự báo khả năng duy trì khách hàng.

Trong quá trình phân tích, nhóm cũng thực hiện tiền xử lý dữ liệu, trực quan hóa dữ liệu và đánh giá hiệu quả các mô hình qua các chỉ số đánh giá phù hợp.

1.4. Ý NGHĨA THỰC TIỄN VÀ ĐÓNG GÓP CỦA ĐỀ TÀI

Đề tài không chỉ mang lại ý nghĩa học thuật trong việc vận dụng kiến thức khoa học dữ liệu và các thuật toán học máy vào phân tích hành vi mua sắm, phân nhóm khách hàng và dự đoán khả năng quay lại của khách hàng từ dữ liệu bán lẻ, mà còn có giá trị ứng dụng cao trong môi trường kinh doanh hiện đại. Việc xây dựng hệ thống phân tích giúp doanh nghiệp hiểu rõ hơn về đặc điểm tiêu dùng và hành vi của từng nhóm khách hàng, từ đó tạo nền tảng xây dựng các chiến lược marketing cá nhân hóa, nâng cao hiệu quả tiếp thị và tăng khả năng giữ chân khách hàng.

Đồng thời, việc dự báo khả năng quay lại của khách hàng giúp doanh nghiệp có thể xác định sớm những khách hàng tiềm năng và những khách hàng có nguy cơ rời bỏ, từ đó triển khai các chương trình chăm sóc, ưu đãi phù hợp nhằm tối ưu hóa chi phí marketing và tăng lợi nhuận. Ngoài ra, dự báo doanh số theo thời gian hỗ trợ nhà quản lý lập kế hoạch điều phối hàng hóa, nhân lực và tài chính một cách hiệu quả, phù hợp với nhu cầu thực tế của thị trường.

Những đóng góp của đề tài không chỉ nâng cao hiệu quả hoạt động kinh doanh mà còn góp phần xây dựng nền tảng quản trị dữ liệu bài bản và bền vững, cung cấp cơ sở tin cậy cho các hoạt động ra quyết định chiến lược của doanh nghiệp trong dài hạn.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1. KHÁM PHÁ VÀ XỬ LÝ DỮ LIỆU

2.1.1. Giới thiệu chung

Khám phá và xử lý dữ liệu là bước đầu tiên và quan trọng trong quy trình phân tích dữ liệu. Qua việc khám phá (data exploration), người phân tích có thể hiểu rõ cấu trúc, đặc điểm và mối quan hệ giữa các biến trong tập dữ liệu, đồng thời phát hiện các vấn đề bất thường như giá trị thiếu hay ngoại lệ. Phần xử lý dữ liệu tập trung vào làm sạch, chuẩn hóa để đảm bảo dữ liệu đầu vào có chất lượng, từ đó nâng cao hiệu quả phân tích và mô hình hóa.

2.1.2. Các bước cơ bản trong khám phá dữ liệu

- Kiểm tra cấu trúc dữ liệu: Sử dụng các hàm trong pandas như shape để xem số hàng, số cột; dtypes để xác định kiểu dữ liệu từng biến. Điều này giúp xác định bước xử lý thích hợp cho từng loại dữ liệu.
- Quan sát mẫu dữ liệu: Dùng head() và tail() để quan sát vài dòng đầu và cuối nhằm phát hiện các dữ liệu bất thường hoặc thiếu hợp lý về định dạng.
- Phân tích thống kê mô tả: Tính các chỉ số như trung bình (mean), trung vị (median), độ lệch chuẩn (standard deviation), các phân vị để đánh giá phân bố và biến thiên của các biến trong dữ liệu.
- Trực quan hóa dữ liệu:
    - Histogram: Biểu đồ phân phối tần suất của biến liên tục, giúp nhận diện phân bố dữ liệu, các giá trị outlier.
    - Boxplot: Hiển thị phân bố và phát hiện ngoại lệ (outliers) một cách trực quan.
    - Heatmap ma trận tương quan: Biểu diễn hệ số tương quan giữa các biến, hỗ trợ xác định các biến có mối quan hệ mạnh hoặc tiềm năng đa cộng tuyến.

2.1.3. Xử lý dữ liệu

- Xác định và xử lý dữ liệu thiếu:
    - Kiểm tra dữ liệu thiếu dụng isnull().sum() hoặc isna().sum() trong pandas.

    - Các phương pháp xử lý tùy theo tỷ lệ và ngữ cảnh: loại bỏ hàng (drop), thay thế bằng trung bình/trung vị/mode hoặc sử dụng nội suy (interpolation) để ước tính.
- Chuẩn hóa và kiểm tra tính nhất quán:
    - Đảm bảo các biến có định dạng thống nhất, ví dụ chuyển đổi dữ liệu dạng chuỗi ngày tháng sang kiểu datetime.
    - Loại bỏ bản ghi trùng lặp bằng hàm drop_duplicates() để giữ tính duy nhất của dữ liệu.
- Xử lý giá trị ngoại lệ (outliers):
    - Phát hiện thông qua biểu đồ boxplot hoặc các chỉ số thống kê như IQR (interquartile range).
    - Có thể loại bỏ, chuyển đổi hoặc giữ nguyên các ngoại lệ tùy thuộc vào đặc trưng dữ liệu và mục đích phân tích.
- Chuyển đổi dữ liệu:
    - Mã hóa biến phân loại thành dạng số (one-hot encoding, label encoding) để thuận tiện trong xử lý mô hình học máy.
    - Chuẩn hóa hoặc tỷ lệ hóa biến số liên tục để đảm bảo tính đồng bộ trong phân tích.

2.1.4. Ý nghĩa của khám phá và xử lý dữ liệu

Việc khám phá giúp nhà phân tích nắm bắt đặc điểm và các vấn đề tiềm ẩn trong dữ liệu trước khi thực hiện phân tích sâu hoặc xây dựng mô hình, từ đó thiết kế phương pháp xử lý hiệu quả. Làm sạch và chuẩn hóa dữ liệu đảm bảo dữ liệu đầu vào có độ chính xác cao, giảm rủi ro sai lệch kết quả phân tích, đồng thời giúp tiết kiệm thời gian và công sức trong các bước tiếp theo của quy trình phân tích dữ liệu. Đây là nền tảng thiết yếu cho mọi dự án dữ liệu thành công.

2.2. TRỰC QUAN HÓA DỮ LIỆU

2.2.1. Tầm quan trọng của trực quan hóa dữ liệu

Trực quan hóa dữ liệu là một bước không thể thiếu trong quy trình phân tích dữ liệu. Đây không chỉ là công cụ hỗ trợ mà còn là phương pháp quan trọng giúp khai thác và

truyền đạt giá trị từ dữ liệu một cách hiệu quả. Các vai trò chính của trực quan hóa dữ liệu bao gồm:

- Khám phá và hiểu dữ liệu: Bằng các biểu đồ, đồ thị, trực quan giúp phát hiện các mẫu, xu hướng, và mối quan hệ phức tạp trong dữ liệu mà khi chỉ quan sát số liệu thô sẽ rất khó nhận biết. Ví dụ, biểu đồ phân tán giúp người dùng dễ dàng nhận diện mối tương quan giữa hai biến.
- Truyền đạt thông tin hiệu quả: Dữ liệu khi chuyển hóa thành hình ảnh trực quan giúp thông tin trở nên dễ hiểu và nhanh chóng tiếp cận. Các biểu đồ, đồ thị hoặc bản đồ nhiệt cung cấp cách trình bày rõ ràng, làm nổi bật điểm chính mà không yêu cầu người xem phải phân tích số liệu chi tiết.
- Hỗ trợ ra quyết định: Trực quan hóa cung cấp góc nhìn tổng quan về các chỉ số quan trọng và xu hướng dữ liệu, giúp nhà quản lý hoặc người ra quyết định nhanh chóng nắm bắt tình hình và lựa chọn giải pháp phù hợp.
- Kể câu chuyện dữ liệu (Data storytelling): Trực quan hóa dữ liệu còn kết hợp các yếu tố hình ảnh với bối cảnh, tạo nên câu chuyện có ý nghĩa, giúp người xem dễ dàng tiếp nhận và liên kết thông tin một cách sâu sắc hơn.

2.2.2. Các loại đồ thị cơ bản trong trực quan hóa dữ liệu

Biểu đồ cột (Bar Chart)

- Đặc điểm: Sử dụng các cột đứng hoặc nằm ngang với chiều cao hoặc độ dài khác nhau để thể hiện giá trị của các nhóm hoặc danh mục khác nhau.
- Phù hợp dùng để:
    - So sánh giá trị giữa các nhóm hoặc danh mục rời rạc.
    - Hiển thị phân phối tần suất của các danh mục (ví dụ: số lượng khách hàng theo nhóm tuổi).
    - Biểu diễn dữ liệu theo thời gian khi số điểm thời gian ít (như doanh thu hàng tháng).

Biểu đồ đường (Line Chart)

- Đặc điểm: Kết nối các điểm dữ liệu theo thứ tự, thường dùng để thể hiện sự thay đổi của dữ liệu theo thời gian.
- Phù hợp dùng để:

- Biểu diễn dữ liệu dạng chuỗi thời gian (time series) như giá cổ phiếu, nhiệt độ hàng ngày.
- Thể hiện xu hướng tăng, giảm hoặc ổn định theo thời gian.
- So sánh nhiều chuỗi dữ liệu (vẽ nhiều đường trên cùng biểu đồ).

Biểu đồ tròn (Pie Chart)

- Đặc điểm: Chia hình tròn thành các phần tương ứng tỷ lệ phần trăm của từng nhóm trong tổng thể.
- Phù hợp dùng để:
  - Thể hiện tỷ lệ phần trăm của các phần trong tổng thể.
  - So sánh các phần tương đối trong tổng thể.

Lưu ý: Nên giới hạn số phần không vượt quá 6-7 để đảm bảo dễ hiểu.

Biểu đồ phân tán (Scatter Plot)

- Đặc điểm: Hiển thị vị trí các điểm dữ liệu trên mặt phẳng theo hai chiều biến số, thể hiện mối quan hệ giữa hai biến.
- Phù hợp dùng để:
  - Phân tích mối tương quan, xác định xu hướng liên hệ giữa hai biến.
  - Phát hiện giá trị bất thường (outliers).
  - Nhận diện các nhóm hoặc phân khúc trong dữ liệu (clusters)

2.2.2.1. Nguyên tắc chọn loại đồ thị phù hợp

Việc lựa chọn loại biểu đồ phù hợp giúp truyền đạt thông tin một cách hiệu quả và chính xác. Một số nguyên tắc cơ bản được dựa trên loại dữ liệu và mục tiêu phân tích:

- Dữ liệu phân loại (Categorical data):
  - Sử dụng biểu đồ cột để so sánh giá trị giữa các nhóm.
  - Dùng biểu đồ tròn để thể hiện tỷ lệ phần trăm trong tổng thể với số phần không nhiều (≤7).
- Dữ liệu số (Numerical data):
  - Chọn biểu đồ đường để hiển thị xu hướng theo thời gian.
  - Sử dụng biểu đồ phân tán để phân tích mối quan hệ giữa hai biến liên tục.
- Dữ liệu phân phối (Distribution data):

- Sử dụng biểu đồ hộp (Box Plot) để mô tả phân phối dữ liệu và phát hiện ngoại lệ.
- Dùng histogram thể hiện tần suất phân bố của biến số liên tục.

- Mục đích truyền đạt:
  - So sánh giá trị: Biểu đồ cột là lựa chọn tối ưu.
  - Biểu diễn xu hướng theo thời gian: Biểu đồ đường trực quan và dễ hiểu.
  - Phân tích tương quan: Biểu đồ phân tán giúp biểu diễn mối liên hệ.
  - Thể hiện cấu trúc phần-tổng: Biểu đồ tròn hoặc biểu đồ cột chồng giúp làm rõ tỉ trọng các phần trong tổng thể.

Trực quan hóa dữ liệu không chỉ giúp nhà phân tích phát hiện insight một cách trực quan mà còn là công cụ đắc lực để truyền tải thông tin phức tạp một cách dễ hiểu, hỗ trợ cho việc ra quyết định dựa trên dữ liệu chính xác. Việc lựa chọn đúng loại biểu đồ phù hợp với loại dữ liệu và mục tiêu phân tích sẽ tối ưu hóa hiệu quả của quá trình trực quan hóa, giúp câu chuyện dữ liệu trở nên thuyết phục và có sức ảnh hưởng lớn hơn.

2.3. MÔ HÌNH PHÂN KHÚC KHÁCH HÀNG RFM

2.3.1. Khái niệm RFM

RFM là một mô hình phân tích hành vi khách hàng phổ biến trong lĩnh vực kinh doanh và marketing, dùng để phân khúc khách hàng dựa trên ba yếu tố quan trọng thể hiện giá trị và mức độ tương tác của khách hàng với doanh nghiệp:

Recency (R): Thời gian kể từ lần mua hàng gần nhất của khách hàng.

Frequency (F): Tần suất khách hàng mua hàng trong một khoảng thời gian nhất định.

Monetary (M): Tổng giá trị tiền mà khách hàng đã chi tiêu trong khoảng thời gian đó.

Mục tiêu của mô hình RFM là xác định mức độ quan trọng, giá trị và hành vi của từng nhóm khách hàng để từ đó xây dựng các chiến lược marketing và chăm sóc phù hợp, nâng cao hiệu quả giữ chân và phát triển khách hàng.

2.3.2. Ý nghĩa và vai trò của các thành phần trong RFM

Recency (R): Khách hàng mua hàng gần đây có xu hướng tiếp tục mua nhiều hơn so với khách hàng không tương tác trong thời gian dài. Do đó, giá trị recency càng nhỏ (gần đây) thì khách hàng càng có giá trị, cần được ưu tiên chăm sóc.

Frequency (F): Khách hàng mua thường xuyên thể hiện sự trung thành và mức độ tương tác cao. Tần suất mua lớn cho thấy khách hàng tin tưởng và hài lòng với sản phẩm/dịch vụ.

Monetary (M): Tổng tiền chi tiêu thể hiện giá trị kinh tế mà khách hàng mang lại. Khách hàng có giá trị monetary cao là nguồn doanh thu quan trọng.

2.3.3. Cách xây dựng và sử dụng mô hình RFM

Bước 1: Thu thập dữ liệu: Thu thập dữ liệu giao dịch của khách hàng bao gồm ngày mua gần nhất, số lần mua và tổng tiền chi tiêu.

Bước 2: Tính toán chỉ số R, F, M

- Recency: Tính khoảng cách ngày giữa ngày khảo sát (hoặc hiện tại) và ngày mua gần nhất.
- Frequency: Đếm số lần mua hàng trong khoảng thời gian phân tích.
- Monetary: Tính tổng giá trị giao dịch trong khoảng thời gian đó.

Bước 3: Phân loại điểm R, F, M: Thông thường, mỗi chỉ số được chia thành 3-5 nhóm điểm (score) dựa trên phân vị (percentile) hoặc ngưỡng quy định sao cho các khách hàng được xếp hạng theo thứ tự từ thấp đến cao.

Bước 4: Phân nhóm khách hàng (Segmentation): Kết hợp 3 điểm R, F và M để tạo thành các nhóm khách hàng khác nhau

2.3.4. Ưu điểm và hạn chế của RFM

Ưu điểm:

- Đơn giản, dễ hiểu và triển khai trên dữ liệu giao dịch cơ bản.
- Giúp phân loại khách hàng dựa trên hành vi thực tế mua hàng.
- Có thể áp dụng rộng rãi trong nhiều ngành và mô hình kinh doanh.

Hạn chế:

- Chỉ dựa trên dữ liệu giao dịch, không phản ánh được các yếu tố ngoại cảnh, mức độ hài lòng hay xu hướng tiêu dùng.
- Cần kết hợp thêm các biến động thời gian hoặc dữ liệu phức tạp khác để nâng cao hiệu quả phân tích sâu hơn.

2.4. THUẬT TOÁN PHÂN CỤM

2.4.1. Khái niệm phân cụm

Phân cụm là kỹ thuật học máy không giám sát (unsupervised learning) dùng để nhóm các đối tượng (điểm dữ liệu) thành các cụm sao cho các đối tượng trong cùng một cụm có sự tương đồng cao về đặc điểm, trong khi các đối tượng thuộc các cụm khác nhau có sự khác biệt lớn.

Mục tiêu chính của phân cụm là khám phá cấu trúc tiềm ẩn trong dữ liệu, nhằm mục đích phân đoạn, rút trích thông tin, hoặc hỗ trợ các bước phân tích và mô hình hóa sau này.

2.4.2. Nguyên tắc hoạt động chung

Định nghĩa tiêu chí nhóm: Xác định cách đo khoảng cách hay độ tương đồng giữa các điểm dữ liệu (điển hình là khoảng cách Euclid, khoảng cách Manhattan, cosine similarity, v.v).

Khởi tạo cụm: Bắt đầu bằng cách xác định số cụm hoặc các điểm trung tâm ban đầu.

Phân bổ điểm dữ liệu: Phân loại mỗi điểm dữ liệu vào cụm phù hợp dựa trên tiêu chí khoảng cách hoặc độ tương đồng.

Cập nhật cụm: Điều chỉnh vị trí trung tâm cụm hoặc cấu trúc cụm dựa trên các điểm đã phân bổ.

Lặp lại: Thực hiện phân bổ và cập nhật cho đến khi mô hình hội tụ hoặc đạt chuẩn dừng.

2.4.3. Một số thuật toán phân cụm phổ biến

2.4.3.1. K-Means

Mô tả: Thuật toán phân cụm dựa trên khoảng cách phổ biến nhất, chia dữ liệu thành K cụm.

Cách hoạt động:

- Khởi tạo ngẫu nhiên K điểm trung tâm (centroids).
- Gán mỗi điểm dữ liệu vào cụm có centroid gần nhất.
- Cập nhật vị trí centroid bằng cách tính trung bình các điểm trong cụm đó.
- Lặp lại cho đến khi centroid không đổi hoặc đạt số lần lặp tối đa.

Ưu điểm: Dễ triển khai, hiệu quả tính toán nhanh.

Hạn chế:

- Cần xác định trước số cụm K.
- Nhạy cảm với điểm ngoại lai và vị trí khởi tạo centroid.
- Không phù hợp với dữ liệu không có cấu trúc hình cầu hoặc có mật độ khác biệt.

2.4.3.2. Phân cụm phổ (Spectral Clustering)

Mô tả: Thuật toán dựa trên lý thuyết đồ thị, phù hợp với dữ liệu có cấu trúc phức tạp không tuyến tính.

Cách hoạt động:

- Tạo ma trận tương đồng giữa các điểm dữ liệu (ví dụ ma trận kề hoặc Gaussian similarity matrix).
- Xây dựng ma trận Laplacian từ ma trận tương đồng.
- Tính các vector riêng đặc trưng (eigenvectors) của Laplacian.
- Chuyển dữ liệu sang không gian vector riêng và áp dụng thuật toán K-Means hoặc các thuật toán phân cụm khác để phân cụm.

Ưu điểm: Có thể phát hiện các cụm có hình dạng phức tạp, không giới hạn tuyến tính.

Hạn chế:

- Tính toán phức tạp với tập dữ liệu lớn.
- Cần chọn tham số phù hợp (ví dụ số cụm, tham số Gaussian kernel).

2.4.3.3. Đánh giá chất lượng phân cụm

Không giống như phân loại, phân cụm thiếu nhãn nên việc đánh giá không dựa vào độ chính xác tuyệt đối mà dùng các chỉ số thống kê hoặc nội bộ.

Silhouette Score: Đo lường mức độ phù hợp của các điểm với cụm được gán. Giá trị từ -1 đến $+1$, càng cao càng tốt.

2.5. DỰ ĐOÁN KHẢ NĂNG QUAY LẠI CỦA KHÁCH HÀNG

2.5.1. Thuật toán

Để xây dựng mô hình dự đoán rời bỏ khách hàng, các thuật toán áp dụng gồm:

Random Forest (Rừng ngẫu nhiên): Thuật toán dựa trên tập hợp nhiều cây quyết định (decision trees). Mỗi cây được xây dựng trên tập dữ liệu con ngẫu nhiên khác nhau

và kết quả cuối cùng lấy theo nguyên tắc bỏ phiếu đa số. Random Forest có khả năng xử lý tốt dữ liệu phức tạp, giảm overfitting và cho kết quả ổn định.

Logistic Regression (Hồi quy Logistic): Là mô hình hồi quy tuyến tính dùng để dự đoán xác suất của một biến nhị phân (ví dụ: rời bỏ – không rời bỏ). Logistic Regression chuyển giá trị tuyến tính qua hàm sigmoid để ra xác suất thuộc nhóm tích cực (ví dụ: khách hàng rời bỏ).

Decision Tree (Cây quyết định): Sử dụng cấu trúc cây để phân loại dữ liệu dựa trên chuỗi các quy tắc điều kiện (nút phân chia), từ đó xác định lớp của từng mẫu dữ liệu. Cây quyết định dễ giải thích, trực quan và triển khai.

2.5.2. Đánh giá hiệu quả mô hình

Accuracy (Độ chính xác): Xác định tỷ lệ dự đoán đúng trên tổng số dự đoán.

Công thức:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad \textit{(2.1)}$$

Precision (Độ chính xác dương trong dự đoán tích cực): Tỷ lệ dự đoán đúng trên tổng số dự đoán dương tính (rời bỏ) mà mô hình đưa ra.

Công thức:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad \textit{(2.2)}$$

Recall (Độ nhạy hay tỷ lệ phát hiện đúng):

Công thức:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad \textit{(2.3)}$$

F1-score: Là trung bình điều hòa giữa precision và recall, giúp cân bằng giữa hai chỉ số này.

Công thức:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}} \quad \textit{(2.4)}$$

ROC AUC: Đo lường khả năng phân biệt giữa hai lớp mô hình. Giá trị ROC AUC nằm trong khoảng [0,1], với 1 là mô hình hoàn hảo, 0.5 là mô hình phân loại ngẫu nhiên.

2.5.3. Ý nghĩa các ký hiệu trong các công thức đánh giá

True Positive (TP) – Dự đoán đúng lớp tích cực: Là số lượng mẫu được mô hình dự đoán thuộc lớp tích cực và thực sự thuộc lớp tích cực.

True Negative (TN) – Dự đoán đúng lớp tiêu cực: Là số lượng mẫu được mô hình dự đoán thuộc lớp tiêu cực và thực sự thuộc lớp tiêu cực.

False Positive (FP) – Dự đoán sai lớp tích cực (Dương tính giả): Là số lượng mẫu được mô hình dự đoán thuộc lớp tích cực nhưng thực tế thuộc lớp tiêu cực.

False Negative (FN) – Dự đoán sai lớp tiêu cực (Âm tính giả): Là số lượng mẫu được mô hình dự đoán thuộc lớp tiêu cực nhưng thực tế thuộc lớp tích cực.

2.6. CÁC THƯ VIỆN CẦN THIẾT

Trong các bài toán phân tích dữ liệu và học máy, Python cung cấp nhiều thư viện đa dạng và mạnh mẽ hỗ trợ từ xử lý dữ liệu đến xây dựng mô hình và trực quan hóa kết quả. Dưới đây là các thư viện cơ bản thường được sử dụng:

- Pandas: Thư viện mạnh mẽ hỗ trợ xử lý và phân tích dữ liệu dạng bảng (DataFrame). pandas cung cấp các tính năng đọc/ghi nhiều định dạng dữ liệu (CSV, Excel, SQL...), thao tác, lọc, gộp nhóm, xử lý dữ liệu thiếu và tổng hợp số liệu hiệu quả.
- Numpy: Thư viện cung cấp các hàm toán học, thao tác trên mảng (arrays), các phép toán đại số tuyến tính nhanh chóng và hiệu quả. numpy là nền tảng cho các tính toán số học trên dữ liệu lớn trong Python.
- Scikit-learn: Thư viện học máy phổ biến với bộ công cụ đa dạng cho các thuật toán phân loại, hồi quy, phân cụm, cũng như các hàm tiền xử lý dữ liệu, chọn lọc tính năng và đánh giá mô hình. scikit-learn có API dễ sử dụng, phù hợp cho nhiều bài toán thực tế.
- Matplotlib: Thư viện cơ bản để vẽ biểu đồ tĩnh và tùy biến phong phú. Là nền tảng cho nhiều thư viện trực quan hóa cao cấp hơn.
- Seaborn: Xây dựng trên matplotlib, seaborn cung cấp các hàm đơn giản hơn để tạo các biểu đồ thống kê đẹp mắt và dễ hiểu như heatmap, boxplot, violin plot,... Phù hợp để trực quan hóa các đặc tính phân phối và mối quan hệ giữa các biến.

Tóm lại, các thư viện trên là công cụ nền tảng trong toàn bộ quy trình phân tích dữ liệu, từ bước thu thập, làm sạch, mô hình hóa đến trình bày và diễn giải kết quả một cách trực quan, hiệu quả.

2.7. PIPELINE TRONG MACHINE LEARNING

Pipeline là một công cụ trong thư viện scikit-learn được sử dụng để tự động hóa quy trình xử lý và mô hình hóa dữ liệu trong một chuỗi các bước tuần tự. Mỗi bước trong pipeline có thể là một bước tiền xử lý dữ liệu (như chuẩn hóa, mã hóa, biến đổi đặc trưng) hoặc một bước học máy (ví dụ, huấn luyện mô hình).

Mục đích và lợi ích chính của Pipeline:

- Tự động hóa quy trình: Giúp gộp nhiều bước xử lý dữ liệu và mô hình vào một đối tượng duy nhất, tránh việc phải thực hiện thủ công từng bước riêng lẻ.
- Tính tái lập cao: Đảm bảo các bước xử lý và mô hình hóa được thực hiện chính xác theo đúng quy trình, dễ dàng tái sử dụng và chia sẻ.
- Giảm lỗi: Hạn chế sai sót trong việc áp dụng đồng bộ các bước xử lý cho cả tập huấn luyện và tập kiểm tra.
- Đơn giản hóa công việc chuẩn bị mô hình: Đặc biệt hữu ích khi kết hợp với các kỹ thuật như chọn tham số (Grid Search, Random Search) giúp tối ưu hóa mô hình hiệu quả.

Cách hoạt động:

- Các bước trong pipeline được định nghĩa theo thứ tự: ví dụ, bước đầu làm sạch dữ liệu, bước tiếp theo chuẩn hóa, bước kế tiếp áp dụng mô hình.
- Khi gọi .fit() trên pipeline, dữ liệu lần lượt được đưa qua từng bước xử lý và cuối cùng là huấn luyện mô hình.
- Khi gọi .predict() hoặc .transform(), pipeline tự động thực hiện tuần tự các bước xử lý tương ứng trước khi ra dự đoán.

CHƯƠNG 3: THỰC NGHIỆM

Tất cả các đoạn code và dữ liệu sử dụng được tổng hợp trong đường link sau:

https://github.com/NguyenKhang0062/DO_AN_2/blob/main/README.md

3.1. THU THẬP VÀ LÀM SẠCH DỮ LIỆU

3.1.1. Tải và kiểm tra dữ liệu ban đầu

Trong giai đoạn đầu của quy trình phân tích, bộ dữ liệu "Retail Dataset" được truy xuất từ nền tảng Kaggle thông qua thư viện hỗ trợ chuyên dụng, cho phép tải về bộ dữ liệu một cách thuận tiện và có kiểm soát. Sau khi hoàn tất quá trình tải, tập tin dữ liệu chính được nạp vào môi trường phân tích dưới dạng một bảng dữ liệu (DataFrame), giúp dễ dàng thao tác và xử lý.

Ngay sau đó, hệ thống tiến hành kiểm tra tổng quan về kích thước của dữ liệu, bao gồm số dòng (tương ứng với số giao dịch) và số cột (đại diện cho các thuộc tính của từng giao dịch). Việc xác minh này không chỉ đảm bảo rằng dữ liệu đã được nạp thành công mà còn đóng vai trò như một bước kiểm tra ban đầu về tính toàn vẹn và cấu trúc dữ liệu, tạo tiền đề cho các bước tiền xử lý và phân tích chuyên sâu trong các phần tiếp theo.

Dưới đây là kết quả hiển thị 5 dòng đầu của bộ dữ liệu:

Bảng 3-1: *Bảng kết quả hiển thị 5 dòng đầu của bộ dữ liệu*

Transaction ID	Date	Customer ID	Gender	Age	Product Category	Quantity	Price per Unit	Total Amount
1	2023-04-12	CUST482	Male	47	Electronics	4	30	120
2	2023-11-24	CUST419	Male	34	Beauty	3	50	150
3	2023-02-27	CUST245	Female	26	Clothing	2	500	1000
4	2023-01-13	CUST178	Male	50	Electronics	1	30	30

5	2023-05-21	CUST206	Male	37	Clothing	1	500	500

Bộ dữ liệu Retail Dataset từ Kaggle chứa thông tin về:

- 1000 khách hàng
- 9 cột đặc trưng bao gồm những thông tin về khách hàng , thông tin giao dịch , thông tin sản phẩm , thông tin mua hàng .
- Thông tin khách hàng:
    - Mã khách hàng (Customer ID): Mã định danh cho mỗi khách hàng
    - Giới tính (Gender): Giới tính của khách hàng (Nam/Nữ), cung cấp thông tin chi tiết về mô hình mua hàng theo giới tính.
    - Độ tuổi (Age): Độ tuổi của khách hàng, giúp phân khúc và khám phá những ảnh hưởng liên quan đến độ tuổi.
- Thông tin giao dịch:
    - Mã giao dịch (Transaction ID): Mã định danh duy nhất cho mỗi giao dịch, cho phép theo dõi và tham chiếu.
    - Ngày thực hiện (Date): Ngày giao dịch diễn ra, cung cấp thông tin chi tiết về xu hướng bán hàng theo thời gian.
- Thông tin sản phẩm:
    - Danh mục sản phẩm (Product Category): Danh mục sản phẩm đã mua (ví dụ: Đồ điện tử, Quần áo, Làm đẹp), giúp hiểu được sở thích về sản phẩm.
- Thông tin mua hàng:
    - Số lượng mua (Quantity): Số lượng sản phẩm đã mua, góp phần cung cấp thông tin chi tiết về khối lượng mua hàng.
    - Đơn giá (Price per Unit): Giá của một đơn vị sản phẩm, hỗ trợ tính toán liên quan đến tổng chi tiêu.
    - Tổng số tiền thanh toán cho giao dịch đó (Total Amount):Tổng giá trị tiền tệ của giao dịch, thể hiện tác động tài chính của mỗi lần mua hàng.

3.1.2. Kiểm tra chất lượng dữ liệu

Sau khi dữ liệu được nạp thành công, bước tiếp theo trong quy trình xử lý là tiến hành kiểm tra sơ bộ nhằm đánh giá chất lượng và cấu trúc của tập dữ liệu. Cụ thể, ba yếu tố chính được xem xét gồm:

- Giá trị thiếu (Missing values): Việc thống kê số lượng giá trị bị thiếu trong từng cột giúp xác định mức độ hoàn chỉnh của dữ liệu, từ đó đưa ra các chiến lược xử lý phù hợp như loại bỏ, thay thế hoặc ước lượng lại giá trị.
- Dữ liệu trùng lặp (Duplicated records): Việc phát hiện các bản ghi bị lặp lại là cần thiết để đảm bảo tính duy nhất và độ tin cậy của dữ liệu, tránh gây nhiễu trong quá trình phân tích và huấn luyện mô hình.
- Kiểu dữ liệu (Data types): Việc xác định chính xác kiểu dữ liệu của từng cột là cơ sở để lựa chọn các phương pháp xử lý và phân tích thích hợp, đồng thời giúp phát hiện sớm các cột có định dạng chưa đúng đặc biệt là các trường ngày tháng hoặc định lượng.
- Ngoại lai: Việc xử lý các giá trị ngoại lai là vô cùng quan trọng, vì chúng có thể làm sai lệch các chỉ số thống kê và ảnh hưởng đến kết quả phân tích hoặc huấn luyện mô hình. Loại bỏ hoặc điều chỉnh ngoại lai một cách phù hợp giúp mô hình phản ánh chính xác xu hướng chủ đạo của dữ liệu, đồng thời ngăn ngừa những sai số hoặc kết luận sai lầm.

Thông qua bước kiểm tra này, nhóm chúng em thực hiện đảm bảo rằng dữ liệu đầu vào sạch, đầy đủ và có cấu trúc rõ ràng điều kiện tiên quyết để triển khai hiệu quả các bước phân tích, mô hình hóa và trực quan hóa trong các giai đoạn tiếp theo.

Kết quả hiển thị:

Kiểm tra giá trị thiếu

Bảng 3-2: *Bảng hiển thị số lượng giá trị thiếu của bộ dữ liệu*

Transaction ID	0
Date	0
Customer ID	0
Gender	0
Age	0
Product category	0
Quantity	0
Price per Unit	0
Total Amount	0

Từ kết quả trên, ta thấy được mỗi cột trong bộ dữ liệu đều hoàn chỉnh và không có bất kỳ giá trị nào bị thiếu. Điều này cho thấy dữ liệu được thu thập và xử lý đầy đủ, sẵn sàng để tiến hành các bước phân tích và mô hình hóa tiếp theo mà không cần phải xử lý giá trị thiếu.

Kiểm tra dữ liệu trùng lặp

Bên cạnh đó kết quả trên cho thấy không có bản ghi trùng lặp trong tập dữ liệu. Điều này chứng tỏ rằng mỗi giao dịch là duy nhất, giúp tránh được sự nhiễu và sai lệch trong kết quả phân tích và huấn luyện mô hình. Việc không có bản sao cũng giúp giảm thiểu rủi ro khi áp dụng các mô hình dự báo hay phân tích hành vi khách hàng.

Kiểu dữ liệu

Bảng 3-3: *Bảng hiển thị kiểu dữ liệu của các đặc trưng*

Transaction ID	int64
Date	object
Customer ID	object
Gender	object
Age	int64
Product category	object
Quantity	int64
Price per Unit	int64
Total Amount	int64

Thống kê mô tả

Bảng 3-4: *Bảng hiển thị thống kê mô tả của bộ dữ liệu*

Statistic	Age	Quantity	Price per Unit	Total Amount
count	1000.00000	1000.000000	1000.000000	1000.000000
mean	41.39200	2.514000	179.890000	456.000000
std	13.68143	1.132734	189.681400	559.997600
min	18.00000	1.000000	25.000000	25.000000
25%	29.00000	1.000000	30.000000	60.000000
50% (median)	42.00000	3.000000	50.000000	135.000000
75%	53.00000	4.000000	300.00000	900.00000
max	64.00000	4.000000	500.00000	2000.00000

Phân tích thống kê mô tả đối với các biến định lượng trong tập dữ liệu giao dịch bán lẻ mang lại những thông tin giá trị, làm cơ sở cho việc hiểu sâu về hành vi tiêu dùng cũng như xây dựng các mô hình phân tích hiệu quả hơn. Dưới đây là những nhận xét chuyên sâu cho từng biến:

- Tuổi khách hàng (Age): Độ tuổi trung bình của khách hàng là khoảng 41 tuổi, với độ lệch chuẩn là 13,68, cho thấy sự phân bố tuổi khá đa dạng trong tập dữ liệu. Nhóm khách hàng trẻ nhất ở độ tuổi 18 và lớn nhất là 64 tuổi. Phân vị thứ nhất (Q1) là 29 tuổi, trung vị (Q2) là 42 tuổi và phân vị thứ ba (Q3) là 53 tuổi, hàm ý rằng phần lớn khách hàng nằm trong độ tuổi trưởng thành và trung niên – một phân khúc có tiềm năng tiêu dùng ổn định và đáng lưu ý trong các chiến lược tiếp thị.
- Số lượng mua (Quantity): Trung bình mỗi giao dịch bao gồm khoảng 2,5 đơn vị sản phẩm, với số lượng tối đa là 4. Mức phân bố khá hẹp (std = 1,13) và số lượng tối đa không lớn, cho thấy hành vi mua sắm chủ yếu là nhỏ lẻ, phù hợp với mô hình bán lẻ trực tiếp đến người tiêu dùng. Phân vị 25% là 1 và 75% là 4 phản ánh sự chênh lệch nhất định về số lượng mua giữa các nhóm khách hàng.
- Giá mỗi đơn vị sản phẩm (Price per Unit): Giá sản phẩm dao động từ 25 đến 500 đơn vị tiền tệ, với mức trung bình là khoảng 180. Tuy nhiên, độ lệch chuẩn lớn (xấp xỉ 190) cho thấy sự biến thiên đáng kể giữa các mặt hàng, phù hợp với tính chất đa dạng của danh mục sản phẩm bao gồm *Clothing*, *Beauty* và *Electronics*. Mức giá phân vị thứ ba đạt 300, cho thấy một phần đáng kể giao dịch tập trung vào các sản phẩm giá trị trung bình đến cao.
- Tổng số tiền giao dịch (Total Amount): Tổng giá trị mỗi giao dịch có mức trung bình là 456, nhưng cũng thể hiện sự phân tán khá cao (std = 560). Giá trị tối đa lên đến 2000 cho thấy một số giao dịch có quy mô tài chính lớn, có thể đến từ việc mua sản phẩm điện tử cao cấp với số lượng nhiều. Trong khi đó, 50% các giao dịch có giá trị dưới 135, phản ánh rằng đa số khách hàng chi tiêu ở mức trung bình hoặc thấp. Sự chênh lệch này là cơ sở để tiến hành phân khúc khách hàng và đề xuất chiến lược giá phù hợp.

Kiểm tra ngoại lại

Để đảm bảo tính toàn vẹn của dữ liệu bán lẻ, nhóm đã thực hiện kiểm tra ngoại lai trên các đặc trưng số quan trọng như tuổi, số lượng, đơn giá và tổng tiền. Phương pháp kiểm tra này dựa trên việc tính toán khoảng tứ phân vị (IQR) và xác định các điểm dữ liệu nằm ngoài giới hạn cho phép. Cụ thể, chúng ta đã tính toán Q1 (tứ phân vị thứ nhất) và Q3 (tứ phân vị thứ ba), từ đó xác định IQR bằng hiệu của Q3 và Q1. Giới hạn dưới và

giới hạn trên được tính bằng cách cộng/trừ 1.5 lần IQR vào Q1/Q3 tương ứng. Kết quả kiểm tra cho thấy không có ngoại lai nào trong bất kỳ đặc trưng số đã phân tích. Điều này khẳng định rằng các giá trị của các đặc trưng này đều nằm trong phạm vi phân bố thông thường, góp phần đảm bảo tính tin cậy cho các phân tích và mô hình hóa tiếp theo. Tuy nhiên, cần lưu ý rằng việc không tìm thấy ngoại lai không đồng nghĩa với việc dữ liệu đã hoàn toàn "sạch".

Kết luận: Những thống kê mô tả trên không chỉ giúp xác định đặc điểm hành vi mua sắm của khách hàng mà còn cung cấp nền tảng vững chắc cho việc chuẩn hóa dữ liệu, phát hiện ngoại lệ và thiết kế các mô hình phân tích phù hợp. Các đặc điểm như độ lệch chuẩn cao, phân phối không đồng đều giữa các biến cũng là tín hiệu cho thấy cần áp dụng các kỹ thuật xử lý như chuẩn hóa hoặc biến đổi dữ liệu để đảm bảo hiệu quả mô hình hóa trong các bước tiếp theo.

3.1.3. Xử lý dữ liệu

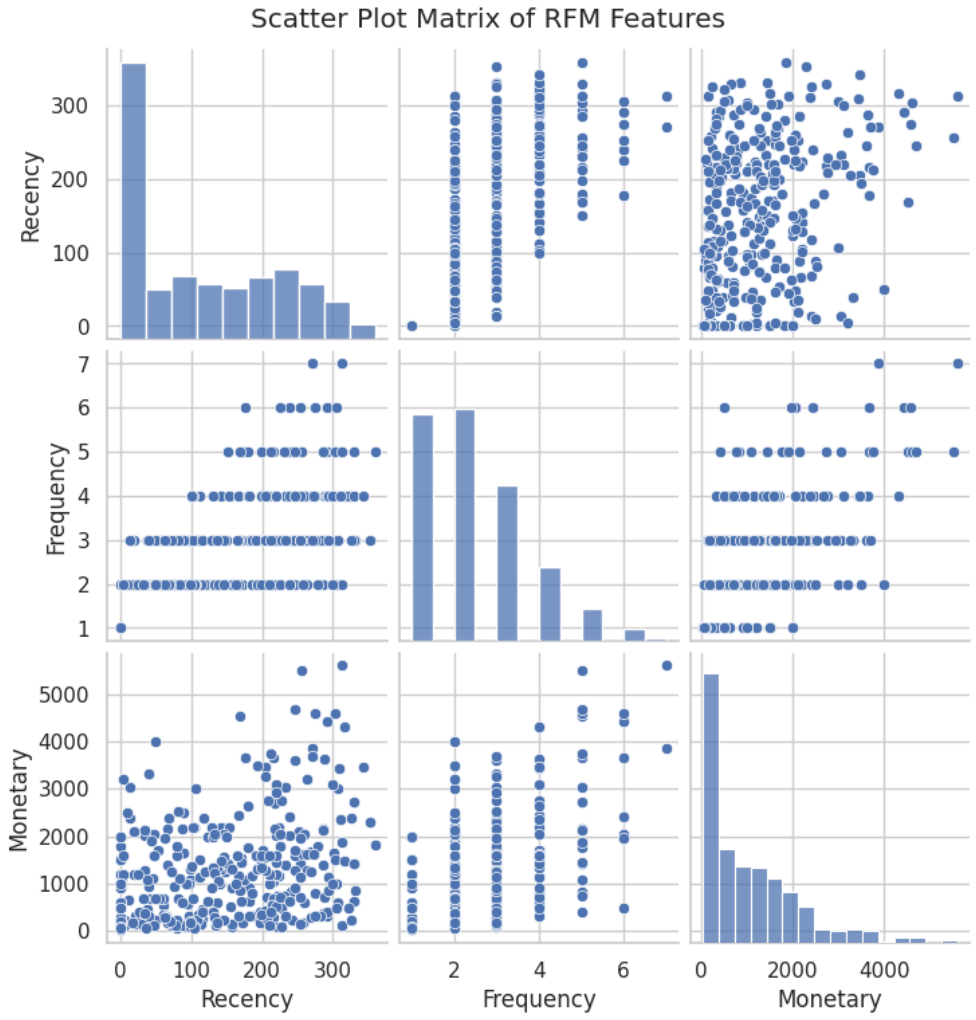
Trong quá trình xử lý dữ liệu, cột ngày giao dịch ban đầu được chuyển đổi từ định dạng chuỗi văn bản sang định dạng ngày-tháng-năm chuẩn, nhằm phục vụ cho các thao tác phân tích chuỗi thời gian như trích xuất tháng, năm hoặc phân tích xu hướng theo thời điểm. Sau đó, các giao dịch phát sinh trong năm 2024 được loại bỏ khỏi tập dữ liệu, do chưa đủ dữ liệu đại diện và không nằm trong phạm vi phân tích của nghiên cứu. Để đảm bảo tính an toàn và linh hoạt trong quá trình phân tích tiếp theo, một bản sao của bộ dữ liệu sau xử lý được tạo ra, giúp lưu giữ phiên bản dữ liệu sạch đã qua lọc và chuyển đổi, đồng thời hỗ trợ kiểm nghiệm và so sánh khi cần thiết. Cách tiếp cận này phản ánh quy trình xử lý dữ liệu có hệ thống, khoa học và hướng đến độ tin cậy cao trong các bước phân tích tiếp theo.

Tiếp theo nhóm tạo thêm đặc trưng RFM cho mỗi khách hàng cụ thể dữ liệu được nhóm theo từng mã khách hàng (Customer ID) và tính toán ba chỉ số chính:

- Recency được đo bằng khoảng cách số ngày giữa lần mua hàng gần nhất và xa nhất, thể hiện mức độ gần đây của các giao dịch.
- Frequency được tính dựa trên tổng số lượng giao dịch mà mỗi khách hàng đã thực hiện, phản ánh tần suất mua sắm.

- Monetary là tổng số tiền mà khách hàng đã chi tiêu trong toàn bộ lịch sử giao dịch, biểu thị giá trị tài chính mà khách hàng mang lại cho doanh nghiệp.

Sau đó vẽ biểu đồ hiển thị trực quan dữ liệu RFM (Recency, Frequency, Monetary) đã được tính toán trước đó. Mục đích là để hiểu rõ hơn về phân bố của các đặc trưng RFM này và mối quan hệ giữa chúng.



Hình 3-1: *Biểu đồ trực quan dữ liệu RFM*

Nhận xét:

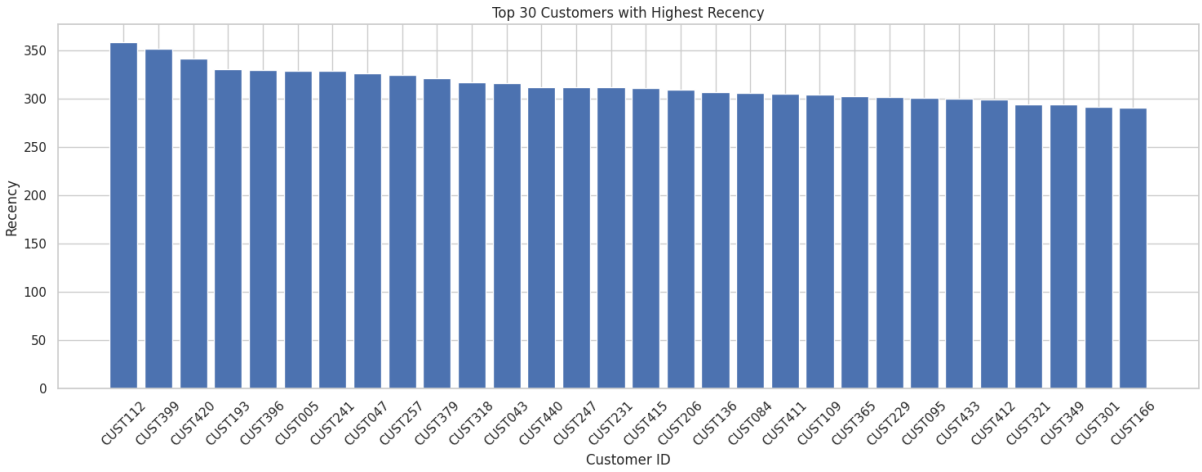
Phân phối từng biến:

- Recency: Phân phối lệch phải, phần lớn khách hàng có giá trị Recency nhỏ (gần đây mua hàng).
- Frequency: Phần lớn khách hàng mua hàng ở tần suất thấp (1-3 lần).
- Monetary: Giá trị chi tiêu lệch phải, nhiều khách hàng tương đối chi tiêu thấp, nhưng có một số người chi rất nhiều.

Mối quan hệ giữa các biến:

- Recency và Frequency: Mối quan hệ âm yếu, khách hàng mua nhiều lần có xu hướng mua gần hơn (giá trị recency nhỏ hơn).
- Recency và Monetary: Một sự tương quan dương yếu, khách hàng mua gần đây có xu hướng chi tiêu nhiều hơn.

- Frequency và Monetary: Tương quan dương rõ ràng hơn, khách hàng mua nhiều lần thường có giá trị chi tiêu cao hơn.



Hình 3-2: *Biểu đồ trực quan 30 khách hàng lâu không mua hàng nhất*

Nhận xét:

- Những khách hàng trong biểu đồ là nhóm khách hàng lâu không mua hàng nhất.
- Đây có thể là nhóm khách hàng cần được chú ý để tiến hành các chiến dịch tái kích hoạt hoặc chăm sóc đặc biệt.
- Giá trị Recency tương đối đồng đều chứng tỏ nhóm khách này lâu ngày không giao dịch và mức thời gian cách biệt không nhiều.

3.2. PHÂN TÍCH VÀ THĂM DÒ DỮ LIỆU

3.2.1. Phân phối số lượng, đơn giá, doanh thu

Trong phần này, chúng ta sẽ phân tích trực quan phân phối của ba biến chính liên quan đến giao dịch bao gồm số lượng sản phẩm (Quantity), đơn giá sản phẩm (Price per Unit) và tổng doanh thu (Total Amount). Các biểu đồ histogram giúp quan sát được mức độ tập trung, sự phân bố cũng như sự xuất hiện của các giá trị bất thường (nếu có) trong

từng biến. Qua đó, ta có thể hiểu rõ hơn về đặc điểm cơ bản của dữ liệu trước khi tiến hành các bước phân tích sâu hơn hoặc xây dựng các mô hình dự báo phù hợp.

Hình 3-3: *Biểu đồ phân phối biến số lượng, đơn giá và tổng doanh thu*

Nhận xét:

Histogram của Quantity (Số lượng):

- Số lượng hàng hóa đa phần nằm trong các giá trị nguyên từ 1 đến 4.
- Các giá trị này phân bố khá đều đặn, không thấy sự lệch rõ ràng.
- Điều này cho thấy khách hàng chủ yếu mua số lượng hàng hóa từ 1 đến 4 đơn vị, tập trung phổ biến các mức mua nhỏ.

Histogram của Price per Unit (Giá mỗi đơn vị):

- Có một số mức giá tập trung rõ ràng ở quanh khoảng dưới 50, 300 và 500.
- Giá cả phân bố không đều, cho thấy có vài nhóm sản phẩm rõ rệt với các mức giá khác nhau.
- Phần lớn sản phẩm có giá thấp, còn một nhóm sản phẩm có giá cao hơn (300 hoặc 500), có thể là các mặt hàng cao cấp hoặc đặc biệt.

Histogram của Total Amount (Tổng doanh thu):

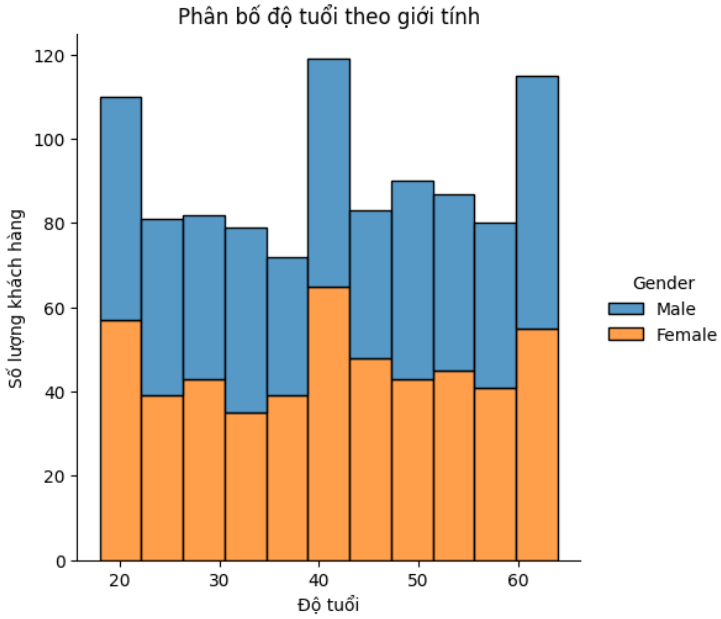
- Phân phối lệch phải rõ rệt, với số lượng lớn giao dịch ở mức tiền nhỏ .
- Các giá trị lớn hơn xuất hiện ít hơn, thể hiện các đơn hàng giá trị cao hiếm hơn.

3.2.2. Phân bố độ tuổi theo giới tính

Trong giai đoạn phân tích thăm dò dữ liệu (Exploratory Data Analysis - EDA), việc trực quan hóa phân bố độ tuổi theo giới tính được thực hiện nhằm khám phá đặc điểm nhân khẩu học của tập khách hàng hiện tại. Chúng em tiến hành vẽ biểu đồ histogram xếp chồng (stacked histogram) để thể hiện cách thức phân bổ độ tuổi giữa hai nhóm giới

tính, qua đó giúp nhận diện những khoảng tuổi chiếm ưu thế trong từng nhóm. Phân tích này đóng vai trò quan trọng trong việc hiểu rõ hơn về cấu trúc khách hàng, từ đó hỗ trợ việc phân khúc thị trường, xây dựng hồ sơ khách hàng điển hình và đề xuất các chiến lược marketing cá nhân hóa phù hợp với từng độ tuổi và giới tính. Việc trực quan hóa dưới dạng phân bố cũng giúp phát hiện các điểm bất thường hoặc xu hướng đặc biệt có thể ảnh hưởng đến hành vi tiêu dùng.

Dưới đây là kết quả hiển thị:



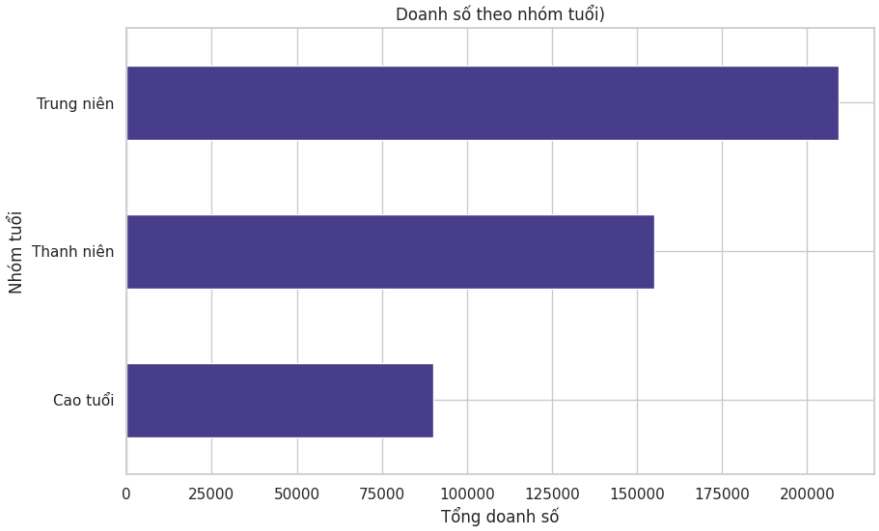
Hình 3-4: *Phân bố độ tuổi theo giới tính*

Nhận xét:

Biểu đồ thể hiện phân bố độ tuổi của khách hàng theo giới tính cho thấy dữ liệu được chia thành các nhóm tuổi từ khoảng 18 đến 64. Trong hầu hết các nhóm tuổi, số lượng khách hàng nam (màu xanh) chiếm tỷ lệ cao hơn nữ (màu cam). Đặc biệt, các nhóm tuổi khoảng 20, 40 và 60 ghi nhận tổng số khách hàng cao nhất, dao động từ 110 đến 120 người. Ngược lại, các nhóm tuổi khoảng 35 và 45 có số lượng khách hàng thấp hơn, chỉ khoảng 70–80 người. Sự chênh lệch giữa nam và nữ rõ rệt nhất ở nhóm tuổi khoảng 20 và 40, trong đó nam giới vượt trội. Tuy nhiên, ở một số nhóm tuổi như 50 và 55, tỷ lệ nam – nữ tương đối cân bằng. Biểu đồ cho thấy sự phân bố khách hàng khá đều theo độ tuổi, với xu hướng nam giới chiếm ưu thế trong phần lớn các nhóm.

3.2.3. Doanh số theo nhóm tuổi

Đầu tiên, dữ liệu độ tuổi của khách hàng được phân thành ba nhóm chính là Thanh niên (17-34 tuổi), Trung niên (34-55 tuổi) và Cao tuổi (trên 55 tuổi) bằng cách sử dụng hàm pd.cut. Tiếp theo, tổng doanh số (Total Amount) của từng nhóm tuổi được tính toán bằng phương pháp nhóm (groupby) và tổng hợp (sum). Kết quả được sắp xếp theo thứ tự doanh số tăng dần để dễ dàng quan sát. Cuối cùng, biểu đồ thanh ngang (horizontal bar plot) được vẽ thể hiện tổng doanh số tương ứng với từng nhóm tuổi, giúp trực quan hóa đóng góp doanh thu của mỗi phân khúc khách hàng theo độ tuổi.



Hình 3-5: *Biểu đồ doanh số theo nhóm tuổi*

Nhận xét:

Nhóm khách hàng "Trung niên" có tổng doanh số cao nhất, khoảng hơn 200,000.

Nhóm "Thanh niên" đứng thứ hai với doanh số khoảng 150,000, thấp hơn nhóm trung niên nhưng vẫn chiếm tỷ trọng lớn.

Nhóm "Cao tuổi" có doanh số thấp nhất, dưới 100,000.

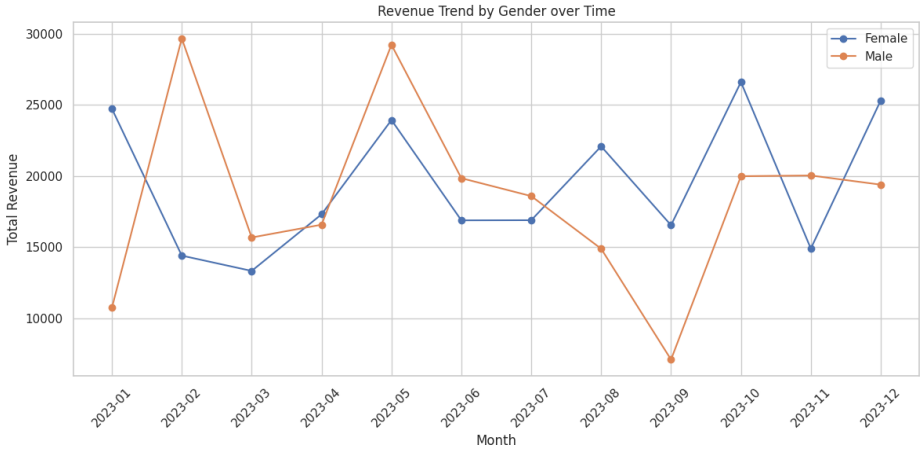
Điều này cho thấy nhóm khách hàng trung niên đóng góp lớn nhất vào doanh số bán hàng, có thể do khả năng chi tiêu hoặc nhu cầu mua sắm cao hơn.

Kết luận:

Doanh nghiệp nên tập trung khai thác và xây dựng các chương trình tiếp thị, chăm sóc khách hàng dành riêng cho nhóm trung niên và thanh niên để tối ưu hóa doanh thu. Đồng thời có thể nghiên cứu nguyên nhân doanh số thấp ở nhóm cao tuổi để cải thiện hoặc điều chỉnh phù hợp.

3.2.4. Doanh thu theo giới tính và thời gian

Dữ liệu được nhóm theo tháng và giới tính bằng cách chuyển đổi cột ngày tháng sang định dạng tháng (Period('M')), sau đó tính tổng doanh thu (Total Amount) cho từng nhóm. Kết quả được đưa về dạng bảng (DataFrame) để thuận tiện cho việc vẽ biểu đồ. Tiếp theo, biểu đồ đường (line plot) thể hiện xu hướng doanh thu theo thời gian được xây dựng, với mỗi đường biểu diễn doanh thu của một nhóm giới tính riêng biệt. Trục hoành biểu diễn tháng, trục tung biểu diễn tổng doanh thu, đồng thời nhãn và chú thích giúp người xem dễ dàng theo dõi sự biến động doanh thu theo giới tính trong từng tháng. Biểu đồ cũng được tùy chỉnh để hiển thị rõ ràng hơn với việc xoay nhãn trục x.



Hình 3-6: *Biểu đồ thể hiện doanh thu theo giới tính và thời gian*

Nhận xét:

Doanh thu của cả hai nhóm giới tính biến động không ổn định qua các tháng, không có xu hướng tăng hoặc giảm dài hạn rõ ràng.

Doanh thu của nam giới có những tháng rất cao như tháng 2, tháng 5, tuy nhiên cũng có tháng doanh thu thấp đáng kể như tháng 9.

Doanh thu của nữ giới tương đối ổn định hơn so với nam, nhưng vẫn có sự dao động đáng kể qua các tháng.

Trong nhiều tháng (ví dụ tháng 1, tháng 8, tháng 10, tháng 12), doanh thu của nữ giới cao hơn nam giới.

Ngược lại, ở các tháng 2, 5, 6, nam giới tạo doanh thu cao hơn nữ giới rõ rệt.

Sự đảo chiều doanh thu giữa hai giới tính qua các tháng cho thấy khả năng tiêu dùng hoặc hành vi mua sắm của nam và nữ có sự khác biệt theo thời điểm trong năm.

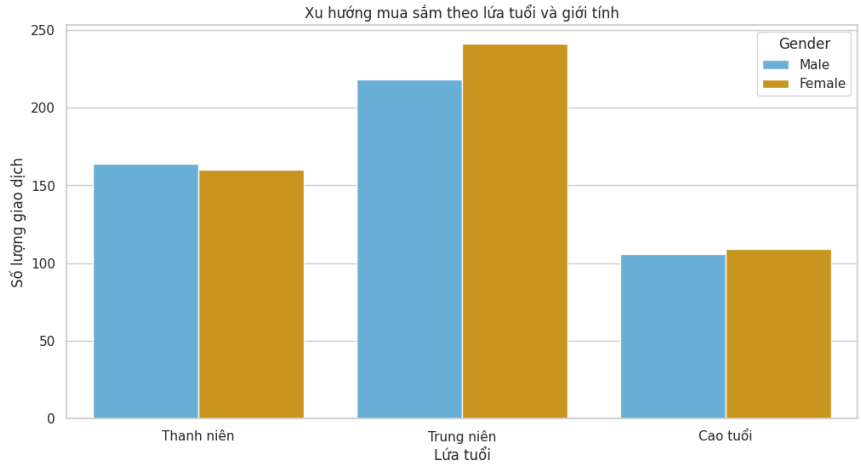
Tóm lại, biểu đồ cho thấy sự cạnh tranh và thay đổi doanh thu rõ nét giữa hai nhóm khách hàng nam và nữ theo tháng, điều này có thể giúp doanh nghiệp điều chỉnh chiến

lược marketing theo mùa hoặc theo giới tính để tận dụng tốt hơn từng phân khúc khách hàng.

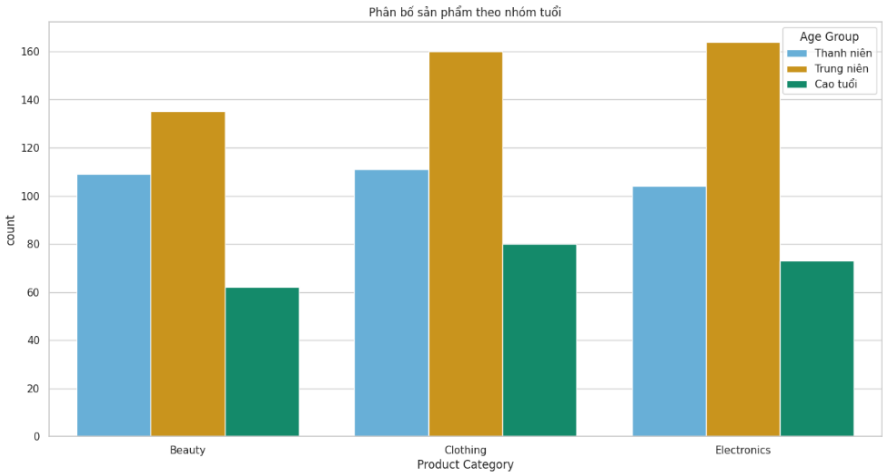
3.2.5. Xu hướng mua sắm theo lứa tuổi và giới tính

Đầu tiên sử dụng biểu đồ cột (countplot) để hiển thị xu hướng mua sắm của khách hàng theo các nhóm tuổi khác nhau, phân biệt theo giới tính. Biểu đồ thể hiện số lượng giao dịch của từng nhóm tuổi, giúp quan sát sự khác biệt về hành vi mua sắm giữa nam và nữ trong từng phân khúc tuổi.

Tiếp theo sử dụng biểu đồ cột tương tự để minh họa phân bố các sản phẩm theo nhóm tuổi khác nhau. Qua đó, ta có thể nhận diện những nhóm tuổi ưu tiên lựa chọn các danh mục sản phẩm cụ thể, giúp hỗ trợ trong việc xây dựng chiến lược tiếp thị và phát triển sản phẩm phù hợp với từng đối tượng khách hàng.



Hình 3-7: *Biểu đồ thể hiện xu hướng theo lứa tuổi và giới tính*



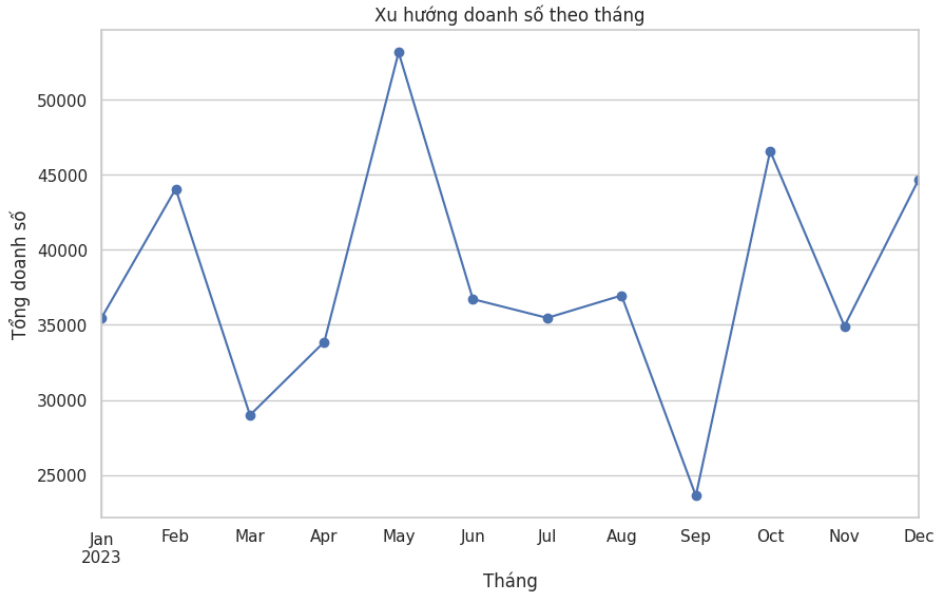
Hình 3-8: *Biểu đồ phân bố sản phẩm theo nhóm tuổi*

Nhận xét:

Nhóm trung niên là nhóm mua sắm nhiều nhất ở cả ba danh mục sản phẩm (Beauty, Electronics, Clothing) và ở cả hai giới tính (nam và nữ). Điều này cho thấy trung niên là đối tượng khách hàng tiềm năng nhất về tần suất mua hàng. cũng có tần suất mua sắm khá cao, đặc biệt là với sản phẩm Clothing ở nam giới – phản ánh sự quan tâm của nhóm tuổi này đến thời trang.

Nhóm cao tuổi có xu hướng mua sắm ít hơn rõ rệt so với hai nhóm còn lại, nhưng vẫn giữ sự hiện diện ở cả ba danh mục – cho thấy nhóm này không hoàn toàn bị động trong tiêu dùng.

3.2.6. Doanh số và số lượng bán được theo tháng



Hình 3-9: *Biểu đồ thể hiện doanh số theo tháng*

Nhận xét:

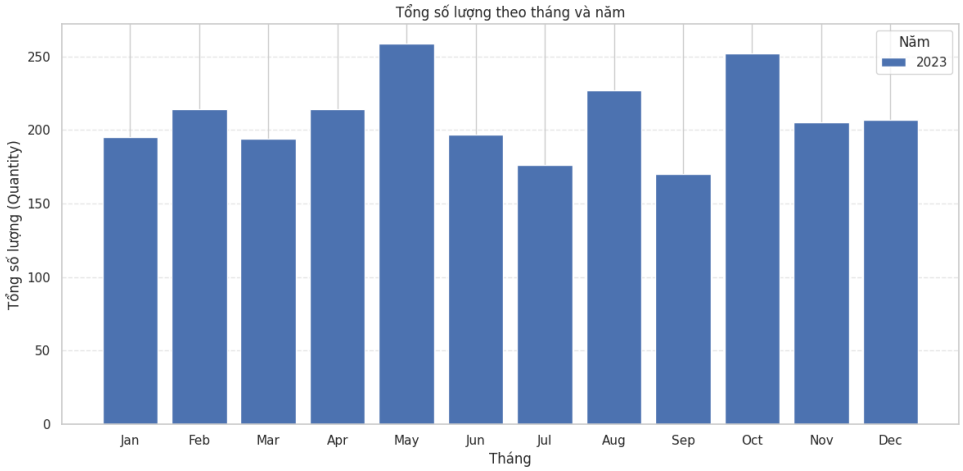
Biểu đồ thể hiện biến động tổng doanh số bán hàng theo từng tháng trong năm 2023, giúp nhận diện các mùa cao điểm và thấp điểm trong hoạt động kinh doanh.

Mùa cao điểm

- Tháng 5 ghi nhận mức doanh số cao nhất trong năm, vượt mốc 53,000, cho thấy đây có thể là thời điểm diễn ra các chiến dịch khuyến mãi lớn, ngày lễ hoặc nhu cầu thị trường tăng mạnh.
- Tháng 2, 10 và 12 cũng đạt mức doanh số cao đáng kể, dao động quanh 44,000 – 47,000, thường trùng với các thời điểm mua sắm như Tết Nguyên Đán, ngày 20/10, Giáng sinh, hoặc các đợt sale như 10/10 và 12/12.

Mùa thấp điểm

- Tháng 9 là tháng có doanh số thấp nhất trong năm (trừ tháng 1/2024 gần như không có số liệu), chỉ khoảng 24,000, cho thấy nhu cầu mua sắm yếu hơn vào giai đoạn này.
- Doanh số cũng có xu hướng giảm nhẹ trong tháng 6–7, nhiều khả năng do hết mùa lễ và rơi vào giai đoạn giữa năm ít hoạt động khuyến mãi.



Hình 3-10: *Biểu đồ số lượng theo tháng*

Nhận xét:

Biểu đồ thể hiện tổng số lượng sản phẩm theo từng tháng trong hai năm 2023. Dữ liệu cho thấy trong năm 2023, các tháng có mức tiêu thụ cao nhất là tháng 5 và tháng 10 với tổng lượng sản phẩm lần lượt đạt đỉnh khoảng 260 và 250 đơn vị. Ngược lại, tháng 9 và tháng 7 ghi nhận mức tiêu thụ thấp nhất, dao động quanh mức 170–180 đơn vị.

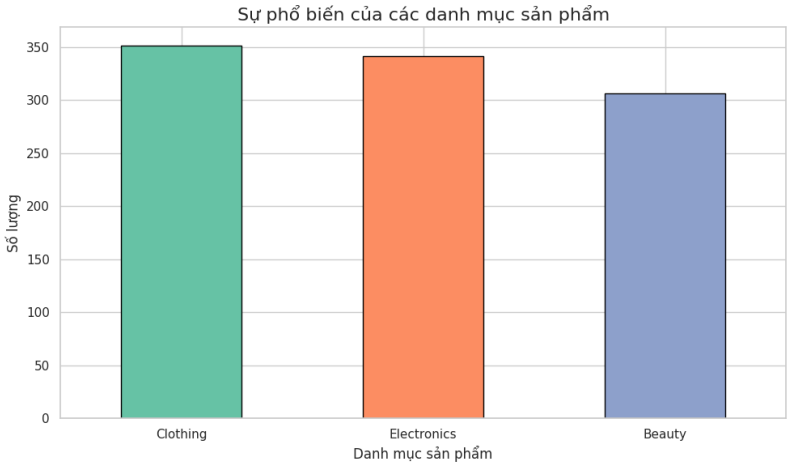
Nhìn chung, xu hướng tiêu thụ trong năm 2023 khá ổn định với các đợt tăng trưởng rõ rệt vào giữa năm (tháng 5) và cuối năm (tháng 10), có thể trùng khớp với các sự kiện khuyến mãi hoặc mùa cao điểm tiêu dùng.

3.2.7. Phân tích hành vi mua sắm qua danh mục sản phẩm

Trong quá trình phân tích xu hướng mua sắm, nhóm nghiên cứu đã thực hiện việc xác định sự phổ biến của các danh mục sản phẩm thông qua việc trực quan hóa số lượng giao dịch theo từng danh mục. Cụ thể, dữ liệu được xử lý để tính toán số lượng các giao dịch bán hàng trong mỗi danh mục sản phẩm, từ đó xác định mức độ phổ biến của từng loại sản phẩm trong cơ sở dữ liệu.

Thông qua phân tích này, doanh nghiệp có thể nhanh chóng nhận diện được các danh mục sản phẩm phổ biến và có lượng giao dịch cao nhất, từ đó đưa ra các chiến lược nhập hàng và tiếp thị phù hợp. Điều này giúp tối ưu hóa việc quản lý kho hàng, đồng thời cải thiện chiến lược kinh doanh hướng đến các nhóm sản phẩm có tiềm năng lớn.

Dưới đây là kết quả hiển thị:



Hình 3-11: *Biểu đồ thể hiện sự phổ biến của các danh mục sản phẩm*

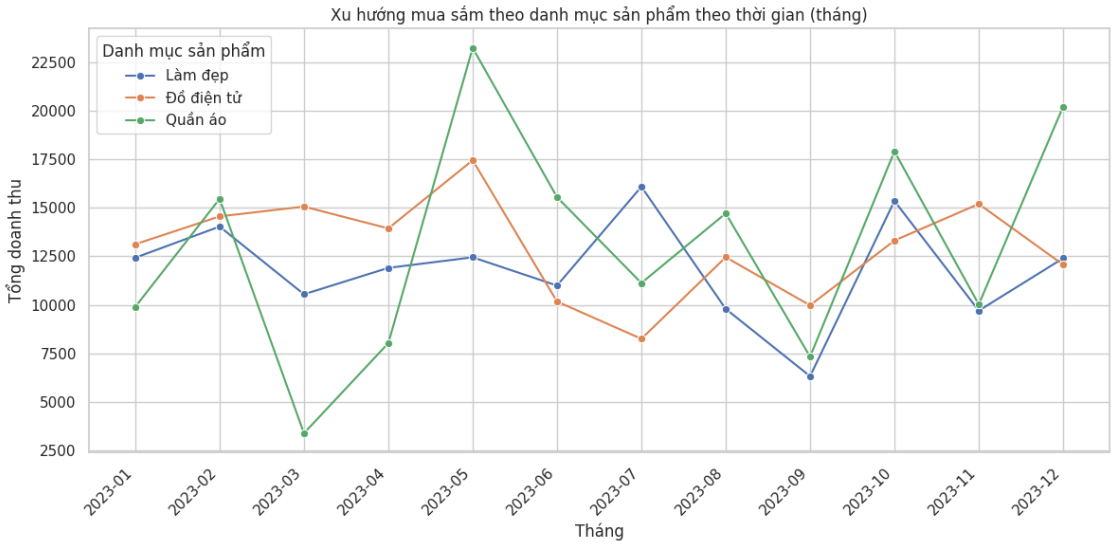
Nhận xét:

Biểu đồ thể hiện số lượng sản phẩm theo từng danh mục cho thấy sự phân bố tương đối đồng đều giữa các nhóm. Cụ thể:

Danh mục 1 có số lượng sản phẩm cao nhất với khoảng 350 mục.

Danh mục 2 đứng thứ hai, chỉ thấp hơn một chút.

Danh mục 0 có số lượng thấp nhất, nhưng vẫn duy trì ở mức trên 300.



Hình 3-12: *Biểu đồ xu hướng mua sắm theo danh mục sản phẩm theo thời gian*

Nhân xét:

- Quần áo có biến động doanh thu lớn nhất qua các tháng, với các đỉnh cao rõ rệt như tháng 5, tháng 10 và tháng 12. Đây cũng là danh mục có doanh thu cao nhất vào những tháng này.

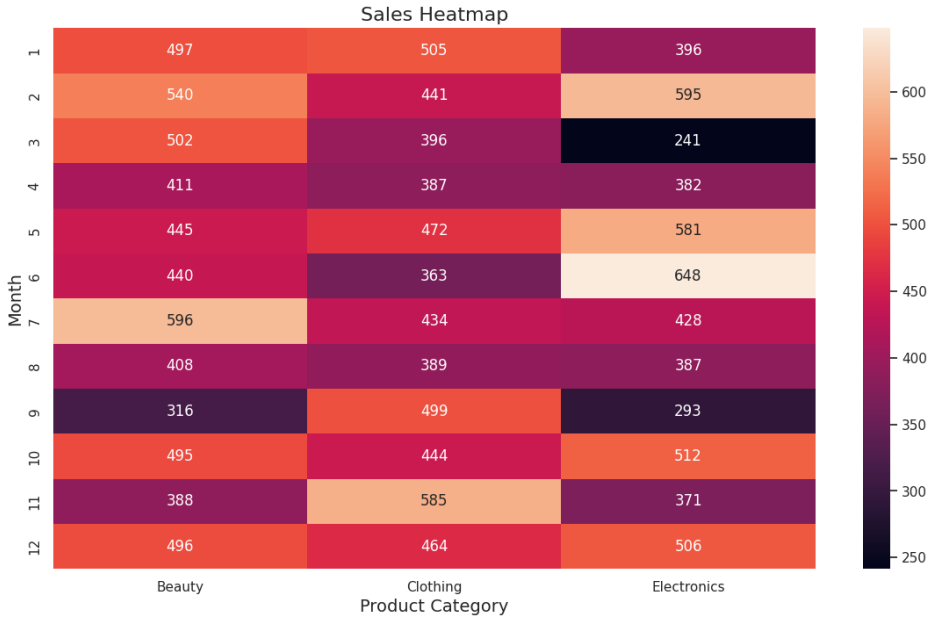
- Đồ điện tử có xu hướng doanh thu ổn định hơn, dao động nhẹ quanh mức từ 8,000 đến 17,000, với đỉnh cao vào tháng 5 và tháng 11.
- Làm đẹp có doanh thu ổn định trung bình so với hai nhóm còn lại, nhưng dao động không quá mạnh, với mức thấp nhất vào tháng 9 và cao điểm vào tháng 7 và tháng 10.
- Tất cả các danh mục đều trải qua sự giảm mạnh doanh số vào tháng 9, có thể do mùa vụ hoặc yếu tố bên ngoài ảnh hưởng chung.
- Danh mục quần áo có sự phục hồi mạnh mẽ vào cuối năm, đặc biệt tháng 12, phù hợp với mùa mua sắm cuối năm và các dịp lễ hội.

Tổng thể, danh mục quần áo và đồ điện tử có mức doanh thu cao và biến động rõ nét hơn so với làm đẹp.

Kết luận:

- Doanh nghiệp cần chú ý các tháng có doanh thu thấp (như tháng 9) để thực hiện các chiến dịch kích cầu.
- Tập trung phát triển mạnh danh mục quần áo trong những tháng cao điểm cuối năm để tận dụng tối đa tiềm năng thị trường.
- Chiến lược marketing nên điều chỉnh linh hoạt theo từng danh mục phù hợp với xu hướng mua sắm theo mùa.

Tiếp theo dữ liệu doanh số được tổng hợp theo từng tháng và từng danh mục sản phẩm. Sau đó, biểu đồ heatmap được sử dụng để trực quan hoá mức độ doanh số của mỗi danh mục sản phẩm theo từng tháng. Biểu đồ cho phép dễ dàng nhận diện các xu hướng biến động doanh thu theo thời gian và mức độ phổ biến của từng nhóm sản phẩm trong các tháng khác nhau, hỗ trợ việc ra quyết định trong hoạch định chiến lược kinh doanh và quản lý sản phẩm.



Hình 3-13: *Biểu đồ nhiệt trực quan doanh số mỗi danh mục theo từng tháng*

Nhận xét:

Danh mục Electronics (Đồ điện tử):

- Có doanh số dao động khá lớn, với những tháng doanh số rất cao như tháng 6 (648), tháng 2 (595), tháng 5 (581).
- Một số tháng doanh số thấp hơn như tháng 3 (241), tháng 9 (293), cho thấy sự biến động lớn theo thời gian.

Danh mục Clothing (Quần áo):

- Doanh số tương đối ổn định trong các tháng, mức doanh số nằm trong khoảng 360-585.
- Có tháng doanh số cao như tháng 11 (585), tháng 1 (505), tháng 2 (441).

Tháng 9 doanh số cao (499) tương đối nổi bật so với các tháng lân cận.

Danh mục Beauty (Làm đẹp):

- Doanh số khá ổn định, thường ở mức từ 388 đến 596 (cao điểm tháng 7).
- Có tháng doanh số thấp nhất là tháng 9 với 316.

Xu hướng tổng thể:

- Tháng 9 có doanh số thấp khá phổ biến ở các danh mục, đặc biệt ở danh mục Beauty và Electronics.
- Tháng 6 và tháng 7 được xem là thời điểm doanh số cao nhất, đặc biệt ở Electronics và Beauty.

- Các danh mục sản phẩm có sự biến động doanh số khác nhau, Electronics biến động nhiều nhất trong khi Beauty tương đối ổn định hơn.

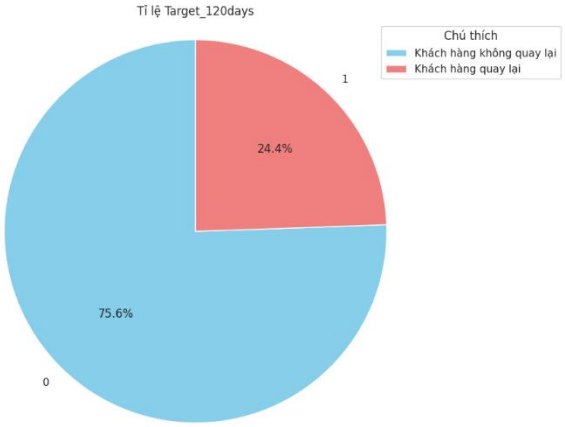
3.3. DỰ ĐOÁN KHẢ NĂNG QUAY LẠI CỦA KHÁCH HÀNG

3.3.1. Chuẩn bị dữ liệu

3.3.1.1. Tạo biến mục tiêu

Trong bước chuẩn bị dữ liệu, chúng ta tiến hành xác định ngày mua hàng đầu tiên và ngày mua hàng cuối cùng của mỗi khách hàng thông qua việc nhóm dữ liệu theo khách hàng và sử dụng các hàm lấy giá trị nhỏ nhất và lớn nhất của ngày giao dịch. Khách hàng chỉ có một giao dịch duy nhất sẽ được loại bỏ khỏi phân tích, vì không đủ dữ liệu để đánh giá hành vi mua sắm liên tục. Tiếp theo, khoảng thời gian giữa lần mua cuối cùng và lần mua đầu tiên được tính toán để làm cơ sở tạo biến mục tiêu. Biến mục tiêu Target_120days được gán giá trị 1 nếu khoảng cách giữa hai lần mua này nhỏ hơn hoặc bằng 120 ngày, thể hiện khách hàng có mức độ tương tác liên tục trong khoảng thời gian này; ngược lại, giá trị 0 biểu thị khách hàng không duy trì hoạt động mua trong 120 ngày hoặc chỉ có một giao dịch.

Kết quả sau khi tạo biến mục tiêu cho thấy trong tổng số khách hàng, có 754 khách hàng thuộc nhóm Target_120days = 0, tức là những khách hàng không duy trì hoạt động mua sắm liên tục trong vòng 120 ngày hoặc chỉ thực hiện một giao dịch duy nhất. Ngược lại, có 244 khách hàng được gán Target_120days = 1, phản ánh nhóm khách hàng có hành vi mua sắm liên tục và ổn định trong khoảng thời gian này.



Hình 3-14: *Biểu đồ trực quan tỉ lệ biến mục tiêu*

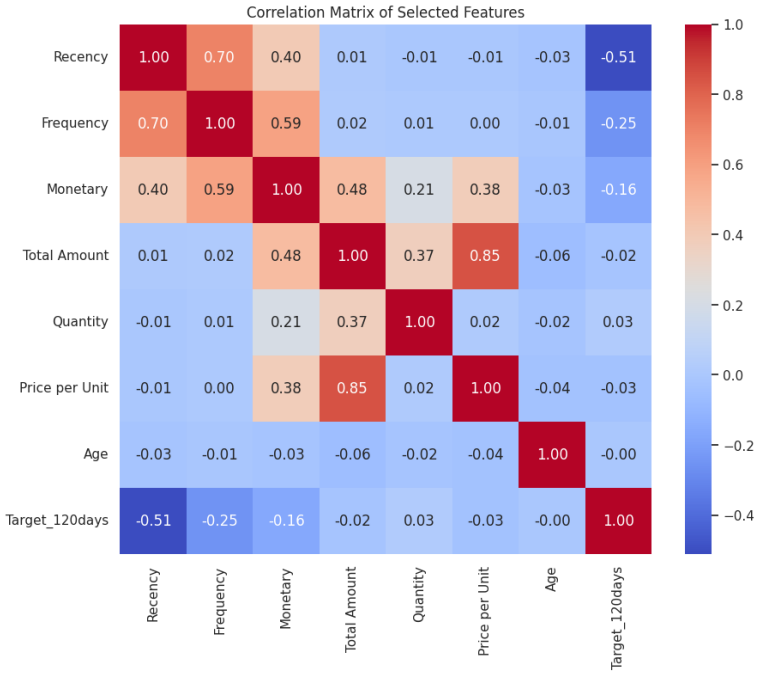
Nhận xét:

Có thể thấy, nhóm khách hàng không quay lại chiếm phần lớn với 75.6%, trong khi chỉ có 24.4% khách hàng duy trì hoạt động mua sắm liên tục trong khoảng thời gian này.

3.3.1.2. Kiểm tra tương quan

Để hiểu rõ hơn mối quan hệ giữa các biến số quan trọng trong dữ liệu, nhóm thực hiện phân tích tương quan giữa các đặc trưng chính bao gồm Recency, Frequency, Monetary, Total Amount, Quantity, Price per Unit, Age và biến mục tiêu Target_120days. Bằng cách kết hợp dữ liệu RFM với các thông tin giao dịch và đặc điểm khách hàng, ma trận tương quan giúp khám phá các liên hệ tiềm năng giữa các biến số, đồng thời hỗ trợ trong việc lựa chọn các đặc trưng ảnh hưởng đáng kể cho các mô hình dự báo hoặc phân tích sâu hơn.

Dữ liệu chính được kết hợp với bảng RFM theo cột ‘Customer ID’ để tổng hợp đầy đủ các đặc trưng cần thiết. Sau đó, ma trận tương quan giữa các biến số được tính toán để đánh giá mức độ liên hệ tuyến tính giữa chúng.



Hình 3-15: *Biểu đồ ma trận tương quan giữa các đặc trưng số*

Nhận xét:

Ma trận tương quan cho thấy mối liên hệ đáng chú ý giữa các biến đặc trưng và biến mục tiêu Target_120days. Cụ thể, biến Recency có tương quan âm khá mạnh (-0.51) với biến mục tiêu, điều này cho thấy khách hàng có khoảng thời gian kể từ lần mua hàng

gần nhất càng lớn thì khả năng quay lại trong 120 ngày càng thấp. Biến Frequency và Monetary cũng có tương quan âm (-0.25 và -0.16) tuy yếu hơn nhưng rõ ràng, cho thấy khách hàng ít mua hàng hoặc chi tiêu thấp cũng có nguy cơ không quay lại cao hơn.

Ngoài ra, các biến như Total Amount và Price per Unit có mối tương quan rất thấp và gần như không đáng kể với biến mục tiêu, điều này gợi ý chúng có ít liên quan tới khả năng khách hàng quay lại trong 120 ngày. Biến Age cũng không có sự liên hệ rõ ràng với hành vi quay lại.

Về mặt tương quan giữa các biến độc lập, Total Amount và Price per Unit cho thấy mức tương quan rất cao (0.85), điều này có thể gây đa cộng tuyến nếu cùng đưa vào mô hình. Biến Recency và Frequency cũng có tương quan tích cực cao (0.70), phản ánh mối liên hệ chặt chẽ trong hành vi mua hàng.

Dựa trên kết quả ma trận tương quan, chúng ta có thể xác định các đặc trưng đóng vai trò quan trọng và có ảnh hưởng rõ nét đối với biến mục tiêu Target_120days. Những biến có hệ số tương quan cao (gần ±1) với mục tiêu nên được ưu tiên giữ lại trong mô hình để tăng khả năng dự báo và giảm thiểu sự dư thừa thông tin.

Ngược lại, những biến có tương quan rất thấp hoặc gần như không có mối quan hệ với biến mục tiêu có thể bị loại trừ nhằm đơn giản hóa mô hình và giảm thiểu nguy cơ gây nhiễu. Ngoài ra, cần chú ý đến các biến có tương quan rất cao với nhau (đa cộng tuyến), bởi việc giữ lại nhiều biến có tính tương quan cao có thể làm giảm hiệu quả và độ ổn định của mô hình. Trong trường hợp này, có thể lựa chọn biến đại diện hoặc áp dụng các phương pháp giảm chiều như PCA.

Việc lựa chọn đặc trưng kỹ càng không chỉ giúp tăng hiệu suất của mô hình học máy mà còn hỗ trợ xác định được những yếu tố quan trọng ảnh hưởng đến hành vi khách hàng, từ đó nâng cao hiệu quả trong các chiến lược kinh doanh và tiếp thị.

3.3.1.3. Chia dữ liệu

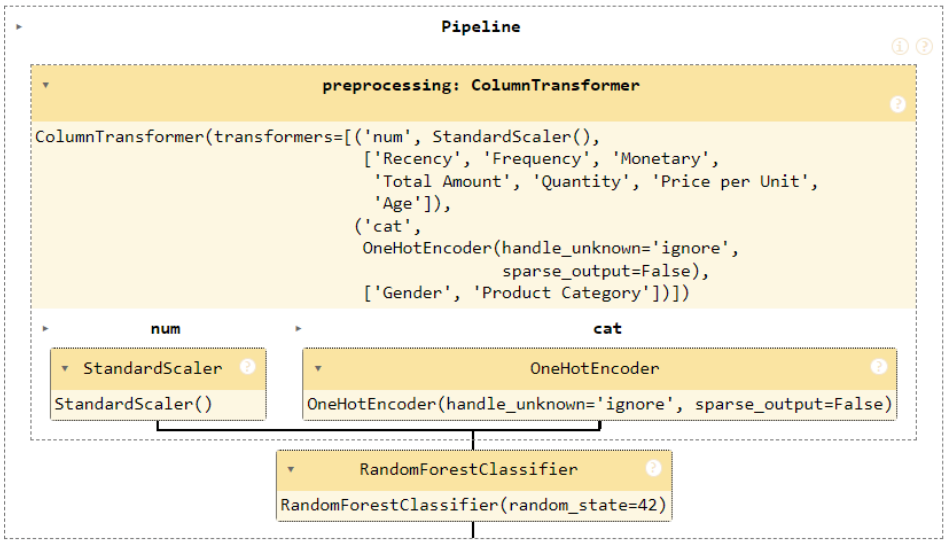
Dữ liệu được tách thành tập huấn luyện (80%) và tập kiểm tra (20%) với tham số random_state=42 để đảm bảo khả năng tái lập kết quả.

Để tăng tính đa dạng và tránh quá khớp (overfitting), mỗi biến số trong tập huấn luyện được thêm thêm nhiễu Gaussian ở mức 30% độ lệch chuẩn của biến đó.

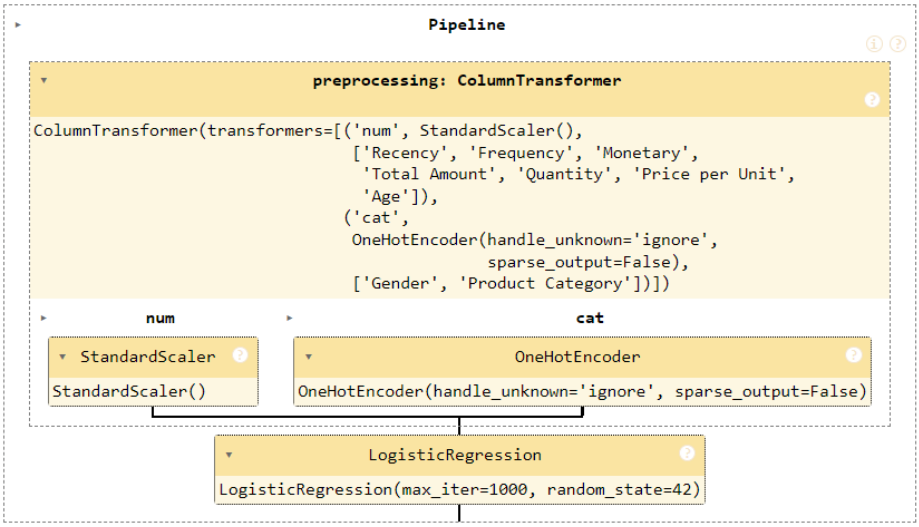
3.3.1.4. Xây dựng pipeline tiền xử lý dữ liệu cho các mô hình

Ba mô hình phân loại được lựa chọn để xây dựng và đánh giá gồm có: Random Forest với 100 cây quyết định, Logistic Regression được cấu hình với số vòng lặp tối đa là 1000 nhằm đảm bảo hội tụ, và Decision Tree đơn giản. Mỗi mô hình này đều được tích hợp vào pipeline bao gồm bước tiền xử lý dữ liệu tự động, giúp chuẩn hóa các biến số liên tục và mã hóa các biến phân loại trước khi dữ liệu được đưa vào huấn luyện. Việc sử dụng pipeline không chỉ đảm bảo tính tự động và thống nhất trong quy trình xử lý dữ liệu mà còn giúp giảm thiểu sai sót và thuận tiện trong việc thử nghiệm và so sánh hiệu suất của các mô hình khác nhau.

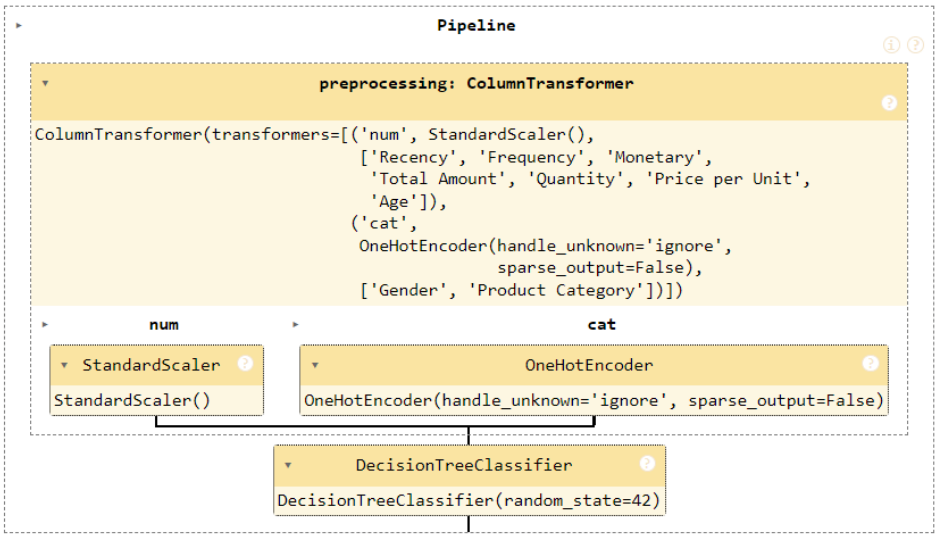
Kết quả như sau:



Hình 3-16: *Pipeline cho RandomForestClassifier*



Hình 3-17: *Pipeline cho LogisticRegression*



Hình 3-18: *Pipeline cho DecisionTreeClassifier*

Pipeline này thể hiện một quy trình xử lý dữ liệu và xây dựng mô hình rõ ràng, mạch lạc và chuẩn mực trong học máy hiện đại. Ở bước tiền xử lý, ColumnTransformer được sử dụng để tách riêng hai nhóm đặc trưng: nhóm biến số liên tục như 'Recency', 'Frequency', 'Monetary', 'Total Amount', 'Quantity', 'Price per Unit' và 'Age' được chuẩn hóa bằng StandardScaler nhằm đảm bảo dữ liệu thống nhất về thang đo; trong khi nhóm biến phân loại như 'Gender' và 'Product Category' được mã hóa dưới dạng one-hot thông qua OneHotEncoder với tham số handle_unknown='ignore', giúp mô hình linh hoạt khi gặp các giá trị chưa xuất hiện trong dữ liệu huấn luyện. Sau bước tiền xử lý, dữ liệu được đưa trực tiếp vào các mô hình học máy cụ thể như Logistic Regression, Decision Tree hoặc Random Forest với random_state=42 để đảm bảo kết quả huấn luyện có thể tái lập. Việc gộp toàn bộ các bước từ tiền xử lý đến huấn luyện thành một chuỗi tuần tự duy nhất trong pipeline không chỉ tạo điều kiện thuận lợi cho việc huấn luyện, đánh giá và triển khai mô hình mà còn giảm thiểu sai sót trong quy trình. Đây là cách tiếp cận chuẩn mực, đảm bảo tính tự động hóa và hiệu quả cao trong phân tích dữ liệu và xây dựng các hệ thống học máy hiện đại.

3.3.2. Xây dựng và huấn luyện mô hình

3.3.2.1. RandomForestClassifier

Xác định tên đặc trưng sau tiền xử lý

Đầu tiên, sau khi pipeline xử lý dữ liệu, nhóm lấy tên các biến số liên tục giữ nguyên, còn các biến phân loại được one-hot encode sẽ có nhiều cột mới tương ứng. Do đó,

nhóm dùng phương thức get_feature_names_out trên đối tượng onehotencoder để lấy tên đầy đủ các cột đặc trưng phân loại đã mã hóa, sau đó kết hợp với tên biến số liên tục thành một danh sách tên đặc trưng hoàn chỉnh.

Trích xuất và trực quan hóa feature importance

Sau khi huấn luyện pipeline, truy cập thuộc tính feature_importances_ của mô hình random forest để lấy mức độ quan trọng từng đặc trưng, giá trị này cho biết đặc trưng nào góp phần lớn nhất trong quyết định dự đoán của mô hình. Tiếp đến tạo một bảng dữ liệu (dataframe) gồm tên đặc trưng và mức độ quan trọng, sau đó sắp xếp giảm dần và chọn ra 10 đặc trưng quan trọng nhất để trực quan bằng biểu đồ cột (barplot) bằng thư viện seaborn.

Đánh giá hiệu suất mô hình trên tập kiểm tra

Sử dụng pipeline đã huấn luyện để dự đoán nhãn (y_pred) và xác suất dự đoán (y_prob) trên tập test, từ đó tính toán các chỉ số đánh giá như accuracy (độ chính xác tổng thể), precision (độ chính xác dự đoán dương), recall (khả năng tìm đúng các mẫu dương), f1-score (trung bình điều hòa giữa precision và recall) và roc auc (đánh giá khả năng phân biệt của mô hình trên nhiều ngưỡng). các giá trị sau khi tính được in ra để thuận tiện cho việc phân tích.

Ma trận nhầm lẫn (confusion matrix)

Tạo ma trận nhầm lẫn từ nhãn thật và nhãn dự đoán trên tập test, sau đó sử dụng seaborn heatmap để trực quan hóa với số lượng dự đoán đúng và sai ở mỗi ô, giúp quan sát rõ hơn điểm mạnh và điểm yếu của mô hình trong phân loại từng nhãn.

Đường cong ROC (receiver operating characteristic curve)

Hàm roc_curve được sử dụng để tính toán tỷ lệ dương tính giả (false positive rate) và tỷ lệ dương tính thật (true positive rate) ở các ngưỡng phân loại khác nhau. Dựa trên các giá trị này, ta vẽ biểu đồ đường cong ROC thể hiện hiệu suất phân loại của mô hình. Trên biểu đồ này còn có thêm đường chéo đại diện cho mô hình phân loại ngẫu nhiên để làm chuẩn so sánh. Ngoài ra, giá trị ROC AUC cũng được hiển thị trên biểu đồ như một thước đo tổng quát cho khả năng phân biệt giữa các lớp của mô hình.

3.3.2.2. LogisticRegression

Trong bước tiền xử lý và lựa chọn đặc trưng, từ kết quả mô hình Random Forest trước đó, nhóm đã trích xuất 8 đặc trưng có mức độ quan trọng cao nhất. Việc chọn lựa này nhằm đơn giản hóa mô hình Logistic Regression đồng thời giữ lại những thông tin quan trọng nhất ảnh hưởng đến dự đoán. Dữ liệu đầu vào bao gồm 8 đặc trưng này được chuẩn hóa với biến số liên tục qua StandardScaler để đưa về cùng thang đo, còn các biến phân loại được mã hóa one-hot sử dụng OneHotEncoder với tham số handle_unknown='ignore' nhằm đảm bảo xử lý tốt các giá trị mới chưa xuất hiện trong tập huấn luyện.

Sau khi tiền xử lý, mô hình Logistic Regression được huấn luyện trên 80% dữ liệu và đánh giá trên 20% dữ liệu còn lại. Kết quả đánh giá thể hiện thông qua các chỉ số chính như sau:

- Accuracy (Độ chính xác): Mô hình đạt độ chính xác cao, thể hiện tỷ lệ dự đoán đúng trên tổng số mẫu.
- Precision (Độ chính xác dự đoán dương): Mô hình có khả năng dự đoán chính xác các trường hợp dương tính.
- Recall (Độ nhạy): Mô hình có tỷ lệ phát hiện đúng các mẫu dương tính cao.
- F1-score: Đây là chỉ số trung hòa thể hiện sự cân bằng giữa precision và recall, đảm bảo mô hình không bị thiên lệch quá mức về một trong hai.
- ROC AUC: Giá trị này thể hiện khả năng phân biệt giữa các lớp của mô hình trên toàn bộ các ngưỡng phân loại; giá trị càng gần 1, mô hình hoạt động càng tốt.

Ngoài các chỉ số đánh giá, ma trận nhầm lẫn được hiển thị dưới dạng biểu đồ heatmap giúp trực quan hóa số lượng dự đoán đúng và nhầm lẫn giữa các lớp, cung cấp cái nhìn chi tiết về các dạng lỗi mô hình có thể mắc phải. Đồng thời, đường cong ROC được vẽ để thể hiện rõ khả năng phân biệt của mô hình qua nhiều ngưỡng phân loại khác nhau, với diện tích dưới đường cong (AUC) được chú thích trực tiếp trên biểu đồ.

3.3.2.3. DicisionTreeClassifier

Quy trình xây dựng mô hình Decision Tree với thêm nhiễu Gaussian:

Chọn lựa đặc trưng: Từ bảng importance_df chứa mức độ quan trọng của các đặc trưng được trích xuất từ mô hình Random Forest, ta chọn ra 8 đặc trưng hàng đầu có ảnh

hưởng nhiều nhất đến mô hình. Việc lựa chọn đặc trưng giúp giảm độ phức tạp mô hình và tập trung vào những biến có giá trị dự báo cao.

Chuẩn bị dữ liệu:

- Dữ liệu đầu vào được lọc theo 8 đặc trưng đã chọn, đồng thời nhãn mục tiêu y được giữ nguyên. Dữ liệu sau đó được chia thành tập huấn luyện và tập kiểm tra theo tỷ lệ 80%-20% với tham số random_state=42 để đảm bảo kết quả có thể tái lập.
- Thêm nhiễu Gaussian vào tập huấn luyện: Để tăng độ đa dạng và giúp mô hình không bị quá phù hợp với tập huấn luyện (overfitting), ta chèn thêm nhiễu Gaussian (được lấy mẫu từ phân phối chuẩn với trung bình 0 và độ lệch chuẩn bằng 10% độ lệch chuẩn của từng biến số) vào các biến số liên tục trong tập huấn luyện. Việc này giúp mô hình có khả năng tổng quát hóa tốt hơn trên các dữ liệu mới.

Xây dựng và huấn luyện mô hình Decision Tree: Mô hình cây quyết định (DecisionTreeClassifier) được khởi tạo với random_state=42 để có kết quả nhất quán khi chạy lại. Mô hình được huấn luyện trên dữ liệu đã được thêm nhiễu.

Đánh giá mô hình:

- Sau khi huấn luyện, mô hình được sử dụng để dự đoán nhãn và xác suất dự đoán trên tập kiểm tra. Các chỉ số đánh giá chất lượng gồm:
    - Accuracy: Độ chính xác tổng thể.
    - Precision: Độ chính xác dự đoán dương tính.
    - Recall: Khả năng phát hiện đúng các mẫu dương tính.
    - F1-score: Sự cân bằng giữa precision và recall.
    - ROC AUC: Đánh giá tổng quát khả năng phân biệt của mô hình.

Trực quan hóa kết quả:

- Ma trận nhầm lẫn (Confusion matrix): Vẽ dưới dạng heatmap thể hiện số lượng dự đoán đúng và sai theo từng lớp.
- Đường cong ROC (Receiver Operating Characteristic Curve): Vẽ biểu đồ ROC để thể hiện khả năng phân biệt của mô hình qua nhiều ngưỡng khác nhau, với đường chéo biểu thị mức hiệu quả của mô hình ngẫu nhiên làm chuẩn so sánh. Trên biểu đồ cũng hiển thị giá trị ROC AUC giúp đánh giá tổng quan.

3.4. PHÂN CỤM KHÁCH HÀNG

3.4.1. Chuẩn bị dữ liệu

Trong các bước trước, chúng ta đã xây dựng và chuẩn bị bộ dữ liệu RFM, tổng hợp các chỉ số Recency, Frequency và Monetary thể hiện hành vi mua hàng của khách hàng. Bộ dữ liệu RFM này sẽ là nền tảng chính để tiến hành phân cụm khách hàng nhằm phân loại và hiểu rõ hơn các nhóm khách hàng dựa trên hành vi tiêu dùng của họ. Việc sử dụng dữ liệu RFM đã được chuẩn hóa và xử lý đầy đủ sẽ giúp các thuật toán phân cụm hoạt động hiệu quả, từ đó đưa ra những phân nhóm khách hàng có ý nghĩa thực tiễn trong việc xây dựng chiến lược tiếp thị và chăm sóc khách hàng.

Dưới đây là 5 dòng đầu của kết quả hiển thị sau khi nhóm từng mã khách hàng và tính toán các chỉ số:

Bảng 3-5: *Bảng hiện thị 5 dòng đầu tiên dữ liệu RFM*

Customer ID	Recency	Frequency	Monetary
CUST000	295	5	4400
CUST001	315	3	2225
CUST002	121	3	2500
CUST003	148	4	360
CUST005	172	2	950

Bảng RFM này đóng vai trò cốt lõi trong việc phân loại và đánh giá giá trị của từng nhóm khách hàng. Thông qua các chỉ số này, doanh nghiệp có thể nhận diện nhóm khách hàng trung thành, khách hàng có giá trị cao, hoặc những khách hàng đang có dấu hiệu rời bỏ, từ đó xây dựng các chiến lược marketing và chăm sóc phù hợp để tối ưu hóa mối quan hệ và doanh thu lâu dài.

Trong bước tiếp theo của quy trình phân tích RFM, nhóm tiến hành chuẩn hóa các chỉ số Recency, Frequency và Monetary bằng cách gán điểm số từ 1 đến 5 cho từng thành phần, dựa trên các phân vị (quantile) thống kê của toàn bộ tập dữ liệu.

Cụ thể, các khách hàng có Recency càng nhỏ (nghĩa là mua hàng gần đây) sẽ được chấm điểm càng cao, ngược lại, những người đã lâu không mua hàng sẽ nhận điểm thấp hơn.

Ngược lại, với Frequency và Monetary, khách hàng có tần suất mua sắm cao và tổng giá trị chi tiêu lớn sẽ được chấm điểm cao hơn, phản ánh mức độ tương tác và giá trị kinh tế mà họ mang lại. Cách chấm điểm này giúp chuẩn hóa và làm nổi bật sự khác biệt giữa các khách hàng trên cả ba khía cạnh.

Sau khi điểm số cho từng thành phần được xác định, nhóm tiếp tục tạo ra một mã số tổng hợp có tên RFM_Score bằng cách kết hợp các điểm R, F và M của mỗi khách hàng. Thang điểm ba chữ số này đóng vai trò là chỉ số định danh, hỗ trợ việc phân loại khách hàng vào các nhóm hành vi cụ thể như khách hàng trung thành, khách hàng mới, khách hàng tiềm năng hay nhóm cần được kích hoạt lại. Việc gán điểm RFM một cách có hệ thống như vậy giúp doanh nghiệp dễ dàng thực hiện các chiến lược tiếp thị mục tiêu, tối ưu hóa hiệu quả chăm sóc và khai thác từng phân khúc khách hàng một cách khoa học và chính xác.

Dưới đây là kết quả hiển thị:

Bảng 3-6: *Bảng hiển thị kết quả thang điểm RFM*

Customer ID	Recency	Frequency	Monetary	R_Score	F_Score	M_Score	RFM_Score
CUST000	295	5	4400	3	4	1	341
CUST001	315	3	2225	3	3	1	331
CUST002	121	3	2500	3	3	1	331
CUST003	148	4	360	3	4	4	344
CUST005	172	2	950	3	2	3	323

Chuẩn hóa dữ liệu RFM

Sau khi đã gán điểm cho từng thành phần trong mô hình RFM, nhóm tiến hành bước chuẩn hóa dữ liệu nhằm đảm bảo sự đồng nhất về thang đo giữa các biến số trước khi đưa vào thuật toán phân cụm. Cụ thể, các điểm R_Score, F_Score và M_Score được chuyển đổi về cùng một phân phối chuẩn (mean = 0, standard deviation = 1) thông qua công cụ StandardScaler.

Việc chuẩn hóa này đóng vai trò then chốt trong quá trình phân tích, vì nếu các biến đầu vào có đơn vị đo lường khác nhau hoặc chênh lệch về độ lớn, kết quả phân cụm sẽ bị sai lệch – thuật toán sẽ thiên lệch về những biến có giá trị tuyệt đối lớn hơn. Do đó, việc

chuẩn hóa không chỉ giúp mô hình phân cụm hoạt động hiệu quả và công bằng hơn mà còn đảm bảo rằng mỗi thành phần của RFM đều được cân nhắc với mức độ ảnh hưởng tương đương trong quá trình xác định nhóm khách hàng.

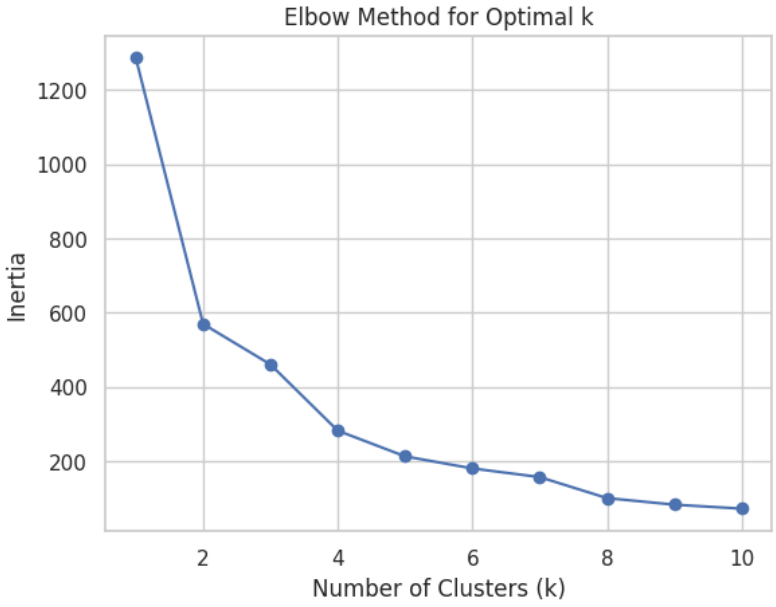
3.4.2. Áp dụng thuật toán Kmeans

Trong bước tiếp theo của quá trình phân tích và phân nhóm khách hàng theo mô hình RFM, nhóm tiến hành áp dụng thuật toán K-Means để xác định các phân khúc khách hàng tiềm năng. Tuy nhiên, trước khi lựa chọn số cụm (số lượng nhóm khách hàng) tối ưu, nhóm sử dụng phương pháp Elbow — một kỹ thuật trực quan thường được sử dụng để xác định giá trị k phù hợp nhất trong thuật toán K-Means.

Cụ thể, nhóm cho thuật toán K-Means chạy lặp lại nhiều lần với số cụm k thay đổi từ 1 đến 10. Ở mỗi giá trị k, thuật toán sẽ tính toán chỉ số inertia – đại diện cho tổng bình phương khoảng cách từ mỗi điểm dữ liệu đến tâm cụm gần nhất. Chỉ số này giúp đánh giá mức độ đồng nhất bên trong các cụm được tạo ra.

Sau khi thu thập được các giá trị inertia tương ứng với từng k, nhóm trực quan hóa kết quả bằng biểu đồ đường.

Dưới đây là kết quả hiển thị:



Hình 3-19: *Biểu đồ Elbow để xác định số cụm k tối ưu*

Nhận xét:

Quan sát biểu đồ Elbow, độ lệch (inertia) giảm mạnh từ k = 1 đến k = 4, sau đó giảm chậm dần. Điểm khuỷu (elbow) rõ ràng tại k = 4, cho thấy đây là số lượng cụm hợp lý

nhất. Việc tăng thêm số cụm sau k = 4 không cải thiện đáng kể độ chặt của cụm, nên không cần thiết.

Sau khi xác định được số lượng cụm tối ưu từ biểu đồ Elbow, nhóm tiến hành áp dụng thuật toán K-Means với số cụm đã chọn là 4 để phân loại khách hàng dựa trên các chỉ số RFM đã được tính toán và chuẩn hóa. Mỗi khách hàng trong tập dữ liệu được gán vào một cụm tương ứng, phản ánh sự tương đồng trong hành vi mua sắm giữa các cá nhân trong cùng một nhóm.

Việc huấn luyện mô hình K-Means với số cụm cụ thể giúp chia toàn bộ tập khách hàng thành bốn phân khúc rõ ràng dựa trên ba chỉ số: mức độ gần nhất với lần mua cuối (Recency), tần suất giao dịch (Frequency), và tổng giá trị chi tiêu (Monetary). Sau khi phân cụm, nhóm tiến hành tổng hợp và tính toán giá trị trung bình của từng thành phần R, F, M trong mỗi cụm.

Dưới đây là kết quả hiển thị:

Bảng 3-7: *Bảng tổng hợp các giá trị trung bình của R, F, M trong mỗi cụm*

	Cluster_Kmeans	R_Score	F_Score	M_Score
0	0	3.557823	2.408163	2.823129
1	1	4.248804	1.665072	4.535885
2	2	3.421875	2.781250	1.000000
3	3	3.086957	3.521739	1.000000

Kết quả số lượng mỗi cụm:

Bảng 3-8: *Bảng thống kê số lượng khách hàng thuộc từng cụm (Kmeans)*

Cluster_Kmeans	count
1	209
0	147
2	64
3	23

3.4.3. Áp dụng thuật toán Spectral Clustering

Nhằm đánh giá tính hiệu quả và độ nhất quán của việc phân cụm khách hàng, nhóm tiếp tục áp dụng thuật toán Spectral Clustering với cùng số lượng cụm là 4 như đã thực hiện trong mô hình K-Means. Thuật toán này sử dụng phương pháp ánh xạ dữ liệu sang không gian phổ của ma trận liên kết (affinity matrix), sau đó thực hiện phân cụm trên không gian mới, giúp giải quyết tốt các trường hợp dữ liệu không tách biệt rõ ràng tuyến tính hoặc có hình dạng phân cụm phức tạp.

Trong quá trình triển khai, nhóm sử dụng lựa chọn affinity='nearest_neighbors', tức là dựa trên mối quan hệ lân cận gần nhất giữa các điểm dữ liệu để xây dựng ma trận liên kết. Mỗi khách hàng sẽ được gán vào một cụm nhất định theo kết quả của thuật toán, và thông tin này được lưu lại dưới dạng cột mới trong bảng dữ liệu.

Sau khi hoàn tất quá trình phân cụm khách hàng bằng thuật toán Spectral Clustering, nhóm tiến hành thống kê số lượng khách hàng thuộc về từng cụm bằng cách đếm tần suất xuất hiện của từng nhãn cụm trong tập dữ liệu. Mục tiêu của bước này là kiểm tra mức độ phân bố của các cụm nhằm đánh giá tính hợp lý và hiệu quả của kết quả phân cụm.

Dưới đây là kết quả hiển thị:

Bảng 3-9: *Bảng thống kê số lượng khách hàng từng cụm (Spectral Clustering)*

Cluster_Spectral	count
0	186
2	160
1	83
3	14

3.4.4. Đánh giá bằng Sihouette Score

Trong giai đoạn đánh giá chất lượng phân cụm, nhóm nghiên cứu đã sử dụng chỉ số silhouette để so sánh hiệu quả của hai thuật toán phân cụm: K-Means và Spectral Clustering. Chỉ số silhouette đo lường mức độ phù hợp của các điểm dữ liệu trong mỗi cụm, với giá trị càng cao cho thấy các điểm trong cùng một cụm càng có sự tương đồng mạnh mẽ, trong khi đó sự khác biệt giữa các cụm càng rõ rệt.

Cụ thể, nhóm đã tính toán chỉ số silhouette cho kết quả phân cụm của thuật toán K-Means và Spectral Clustering. Bằng cách so sánh các giá trị silhouette score từ cả hai phương pháp, nhóm có thể đánh giá được phương pháp phân cụm nào tạo ra các nhóm có sự phân biệt rõ ràng và độ đồng nhất cao hơn. Kết quả từ chỉ số này giúp xác định sự

phù hợp của các mô hình phân cụm đối với dữ liệu khách hàng, từ đó đưa ra những nhận định và quyết định quan trọng trong việc tối ưu hóa các chiến lược phân nhóm trong phân tích hành vi người tiêu dùng.

Dưới đây là kết quả hiển thị:

Bảng 3-10: *Bảng hiển thị chỉ số silhouette đánh kết quả phân cụm*

	Silhouette Score
K-Means	0.6390936259161152
Spectral Clustering	0.24020923506019068

Kết quả phân cụm cho thấy:

- **K-Means:** đạt giá trị Silhouette Score khoảng 0.64, cho thấy các cụm được phân chia khá rõ ràng, mô hình phân cụm này tạo ra các nhóm khách hàng có tính đồng nhất cao bên trong và phân biệt rõ với các nhóm khác. Đây được xem là kết quả tốt, phù hợp để áp dụng trong các phân tích và chiến lược tiếp theo.
- **Spectral Clustering:** đạt giá trị Silhouette Score khoảng 0.24, thấp hơn rất nhiều so với K-Means. Điều này cho thấy các cụm được tạo ra bởi phương pháp này có thể không thực sự đồng nhất hoặc có sự giao thoa nhiều giữa các cụm, làm giảm độ rõ ràng và tính giải thích của từng nhóm khách hàng.

Từ các giá trị này, có thể kết luận rằng phương pháp K-Means phù hợp và hiệu quả hơn trong việc phân cụm dữ liệu khách hàng hiện tại so với Spectral Clustering. Tuy nhiên, tùy thuộc vào mục tiêu phân tích và đặc điểm dữ liệu, việc thử nghiệm thêm các phương pháp khác hoặc điều chỉnh tham số vẫn có thể mang lại kết quả tốt hơn.

CHƯƠNG 4: KẾT QUẢ ĐẠT ĐƯỢC

4.1. PHÂN TÍCH HÀNH VI MUA SẮM CỦA KHÁCH HÀNG

Qua quá trình phân tích chi tiết hành vi mua sắm của khách hàng dựa trên các chỉ số giao dịch, đặc điểm nhân khẩu học và xu hướng theo thời gian cùng danh mục sản phẩm, chúng ta đã thu được những kết quả quan trọng như sau:

- Khách hàng chủ yếu mua với số lượng nhỏ (1-4 sản phẩm), tập trung vào các mức giá thấp đến trung bình, tuy nhiên vẫn tồn tại nhóm sản phẩm cao cấp với mức giá đáng kể.
- Phân bố khách hàng theo độ tuổi và giới tính cho thấy nhóm trung niên và thanh niên là nhóm khách hàng lớn và có tầm ảnh hưởng quan trọng trong việc chi tiêu, trong khi nhóm cao tuổi đóng góp tương đối ít hơn.
- Doanh số và số lượng bán hàng có sự biến động theo mùa, đặc biệt cao điểm vào các tháng 2, 5, 10 và 12, đồng thời thấp điểm rơi vào tháng 9.
- Danh mục sản phẩm Quần áo và Điện tử chiếm phần lớn doanh số và có biến động theo mùa rõ nét hơn so với danh mục Làm đẹp, vốn có doanh thu tương đối ổn định nhưng vẫn chịu ảnh hưởng bởi yếu tố mùa vụ.

Từ những kết quả này, doanh nghiệp có thể đưa ra các đề xuất chiến lược nhằm tối ưu hóa hoạt động kinh doanh và gia tăng hiệu suất tiếp thị, cụ thể:

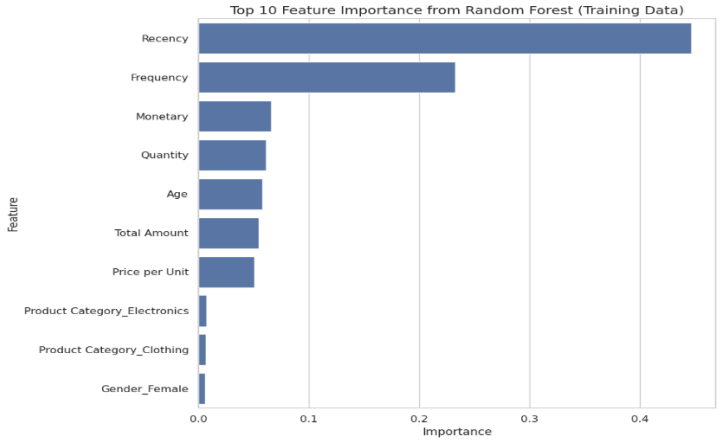
- Tập trung phát triển và thúc đẩy các chương trình tiếp thị cho nhóm khách hàng trung niên và thanh niên, tận dụng khả năng chi tiêu và tần suất mua sắm cao của các nhóm này.
- Xây dựng các chương trình khuyến mãi, ưu đãi theo mùa vụ, đặc biệt trong các tháng cao điểm (2, 5, 10, 12) để kích thích tiêu dùng, đồng thời triển khai các chiến dịch kích cầu vào tháng thấp điểm (tháng 9) nhằm giảm thiểu tác động mùa vụ.
- Ưu tiên phát triển danh mục sản phẩm Quần áo và Điện tử, đặc biệt vào các dịp mua sắm lớn và cuối năm, đồng thời duy trì và gia tăng sự đa dạng, chất lượng cho danh mục Làm đẹp để giữ chân khách hàng và tăng trưởng bền vững.
- Phân tích sâu hơn về hành vi và nhu cầu của nhóm khách hàng cao tuổi để hiểu rõ nguyên nhân doanh số thấp, từ đó thiết kế các chính sách phù hợp nhằm mở rộng thị trường tiềm năng này.

- Tối ưu hóa quản lý tồn kho dựa trên các phân tích xu hướng doanh số theo tháng và danh mục, tránh tồn kho ứ đọng, đồng thời đảm bảo nguồn hàng đáp ứng kịp thời nhu cầu thị trường.

4.2. DỰ ĐOÁN KHẢ NĂNG QUAY LẠI CỦA KHÁCH HÀNG

4.2.1. RandomForestClassifier

Sau khi hoàn tất việc huấn luyện từ mô hình Random Forest ta xác định được mức độ đóng góp của từng đặc trưng trong việc đưa ra dự đoán . Giá trị này phản ánh tầm quan trọng tương đối của mỗi đặc trưng đối với hiệu suất mô hình. Tiếp theo, các đặc trưng cùng với mức độ quan trọng của chúng được tổng hợp thành một bảng dữ liệu (dataframe), sau đó sắp xếp theo thứ tự giảm dần nhằm phân loại và chọn ra 10 đặc trưng xuất sắc nhất. Các đặc trưng này được trực quan hóa dưới dạng biểu đồ cột (barplot) sử dụng thư viện seaborn, giúp dễ dàng nhận diện những yếu tố quan trọng nhất ảnh hưởng đến mô hình.



Hình 4-1: *Biểu đồ 10 đặc trưng quan trong nhất từ Random Forest*

Nhận xét:

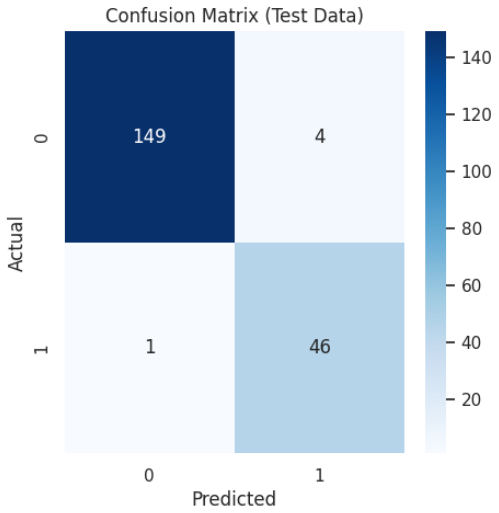
Biểu đồ cho thấy rõ ràng sự khác biệt về mức độ ảnh hưởng của các đặc trưng đến kết quả dự đoán của mô hình. Trong đó, đặc trưng Recency có mức độ quan trọng vượt trội so với các biến còn lại, cho thấy thời gian kể từ lần mua hàng gần nhất là yếu tố hàng đầu quyết định hành vi khách hàng. Tiếp theo, Frequency cũng giữ vai trò quan trọng đáng kể, phản ánh tần suất mua hàng là yếu tố cần được chú ý. Các biến liên quan đến giá trị và số lượng giao dịch như Monetary, Quantity, Total Amount và Price per Unit có mức độ ảnh hưởng vừa phải, cho thấy yếu tố chi tiêu và khối lượng mua hàng cũng góp phần không nhỏ. Các đặc trưng cá nhân như Age cùng với một số biến phân

loại như danh mục sản phẩm và giới tính mặc dù có mức độ quan trọng thấp hơn nhưng vẫn giữ vai trò hỗ trợ trong việc tăng cường hiệu quả dự đoán của mô hình. Tổng thể, biểu đồ giúp làm sáng tỏ các yếu tố quyết định trong dữ liệu, từ đó hướng đến việc lựa chọn đặc trưng hiệu quả cho các phân tích và mô hình hóa tiếp theo.

Các chỉ số đánh giá mô hình:

- Accuracy: 0.975
- Precision: 0.920
- Recall: 0.979
- F1-score: 0.948
- ROC AUC: 0.997

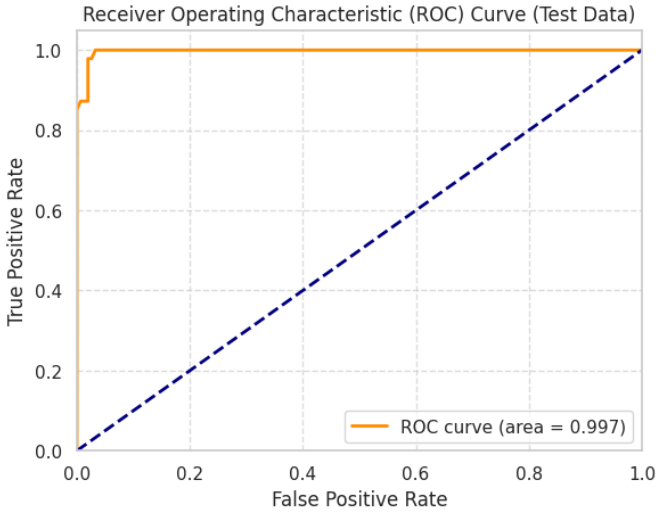
Mô hình đạt được hiệu suất rất ấn tượng trên tập kiểm tra với các chỉ số đánh giá đều ở mức cao. Cụ thể, độ chính xác (Accuracy) lên đến 97.5% cho thấy phần lớn các dự đoán của mô hình đều chính xác. Giá trị Precision đạt 92.0% phản ánh khả năng mô hình dự đoán đúng các trường hợp tích cực, hạn chế đáng kể các dự đoán sai lệch. Recall ở mức 97.9% chứng tỏ mô hình có khả năng phát hiện hầu hết các mẫu dương tính thực sự, giảm thiểu số lượng bỏ sót quan trọng. F1-score đạt 94.8% là minh chứng cho sự cân bằng tốt giữa Precision và Recall, đảm bảo mô hình không bị thiên lệch về một trong hai chỉ số này. Đặc biệt, chỉ số ROC AUC gần như tuyệt đối (0.997) cho thấy mô hình có khả năng phân biệt mạnh mẽ giữa các lớp, hoạt động hiệu quả trên mọi ngưỡng phân loại.



Hình 4-2: *Biểu đồ ma trận nhầm lẫn (RandomForest)*

Nhận xét:

Ma trận nhầm lẫn cho thấy mô hình có khả năng phân loại chính xác cao trên tập kiểm tra. Cụ thể, với lớp 0 (nhãn âm), mô hình dự đoán đúng 149 trường hợp trong tổng số 153 mẫu, chỉ có 4 trường hợp bị sai lệch thành nhãn dương tính (false positives). Đối với lớp 1 (nhãn dương), mô hình dự đoán chính xác 46 trường hợp trên tổng số 47 mẫu, chỉ có 1 trường hợp bị bỏ sót (false negative). Tỷ lệ lỗi thấp ở cả hai lớp thể hiện sự cân bằng tốt trong việc phát hiện cả hai loại mẫu, giúp mô hình duy trì độ chính xác và khả năng nhạy bén cao với các mẫu quan trọng. Đây là minh chứng cho hiệu quả của mô hình trong việc phân biệt chính xác các lớp, giảm thiểu sai sót dự đoán không mong muốn.



Hình 4-3: *Biểu đồ đường cong ROC (Random Forest)*

Nhận xét:

Đường cong ROC thể hiện khả năng phân biệt giữa hai lớp của mô hình trên tập dữ liệu kiểm tra. Đường cong gần như chạm sát góc trên bên trái, cho thấy mô hình đạt được tỷ lệ True Positive Rate rất cao trong khi giữ False Positive Rate ở mức rất thấp. Diện tích dưới đường cong (AUC) đạt giá trị 0.997, gần với giá trị tối đa 1, chứng tỏ mô hình có hiệu suất phân loại xuất sắc, có khả năng phân biệt rõ rệt giữa các mẫu thuộc các lớp khác nhau trên tất cả các ngưỡng phân loại.

4.2.2. LogisticRegression

Kết quả các chỉ số đánh giá mô hình:

- Accuracy: 0.770
- Precision: 0.514

- Recall: 0.404
- F1-score: 0.452
- ROC AUC: 0.823

Accuracy (Độ chính xác) = 0.770: Mô hình dự đoán đúng 77% tổng số trường hợp.

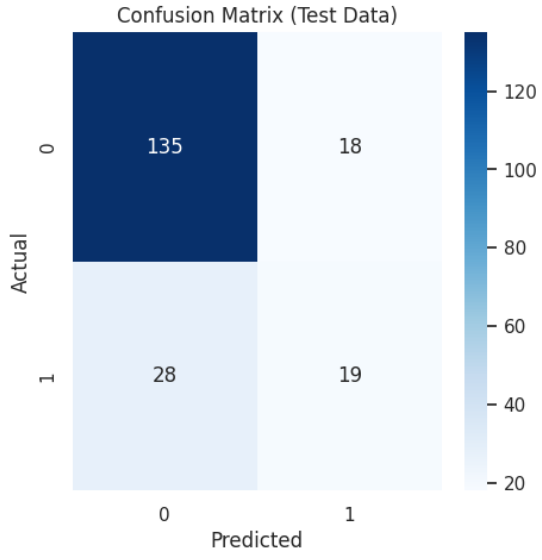
Precision (Độ chính xác khi dự đoán dương) = 0.514: Trong số các dự đoán khách hàng mua hàng, chỉ có 51.4% là chính xác.

Recall (Độ nhạy) = 0.404: Mô hình chỉ phát hiện được 40.4% khách hàng thực sự mua hàng.

F1-score = 0.452: Đánh giá tổng hợp giữa Precision và Recall ở mức trung bình thấp.

ROC AUC = 0.823: Khả năng phân biệt khách hàng mua và không mua của mô hình khá tốt.

Tóm lại, mặc dù mô hình có độ chính xác tổng thể khá cao, nhưng khả năng phát hiện chính xác nhóm khách hàng mua hàng (Recall) và độ chính xác trong dự đoán nhóm này (Precision) còn khiêm tốn, cho thấy mô hình có xu hướng bỏ sót nhiều khách hàng tiềm năng hoặc dự đoán sai nhóm khách hàng mua. Đây là điểm cần cải thiện nếu mục tiêu là nhắm đúng những khách hàng có khả năng mua cao.



Hình 4-4: *Biểu đồ ma trận nhầm lẫn (LogisticRegression)*

Nhận xét :

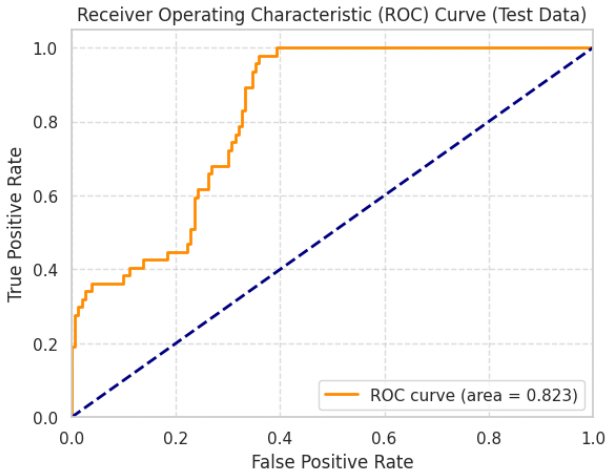
- True Negative (TN) = 135: Mô hình dự đoán đúng 135 trường hợp thực tế là lớp 0 (không mua).

- False Positive (FP) = 18: Mô hình dự đoán nhầm 18 trường hợp không mua thành mua.
- False Negative (FN) = 28: Mô hình bỏ sót 28 trường hợp khách hàng thực sự mua nhưng dự đoán không mua.
- True Positive (TP) = 19: Mô hình dự đoán đúng 19 trường hợp khách hàng thực sự mua.

Nhận xét:

- Mô hình có khả năng dự đoán chính xác nhóm khách hàng không mua khá tốt (135 trường hợp đúng so với chỉ 18 sai).
- Tuy nhiên, số lượng khách hàng thực sự mua bị dự đoán sai không nhỏ (28 trường hợp FN), cho thấy mô hình bị bỏ sót nhiều khách hàng mua tiềm năng.
- Số trường hợp dự đoán đúng khách hàng mua thấp (19 TP), làm ảnh hưởng đến Precision và Recall của mô hình.
- Mô hình có xu hướng thiên về dự đoán nhóm khách hàng không mua hơn, có thể do sự mất cân bằng dữ liệu giữa hai lớp hoặc do ngưỡng phân loại chưa tối ưu.

Kết luận: Nếu mục tiêu là gửi các chiến dịch marketing nhắm đúng khách hàng có khả năng mua, mô hình hiện tại chưa thực sự hiệu quả do tỷ lệ bỏ sót khách mua cao. Cần xem xét các kỹ thuật cải thiện như điều chỉnh ngưỡng phân loại, sử dụng phương pháp xử lý mất cân bằng lớp (oversampling, undersampling), hoặc thử các thuật toán khác nhằm nâng cao khả năng nhận diện khách hàng mua.



Hình 4-5: *Biểu đồ đường cong ROC (Logistic Regression)*

Nhận xét:

- Đường ROC (màu cam) nằm phía trên đường chéo chấm (đường tham chiếu), điều này cho thấy mô hình có khả năng phân biệt giữa hai lớp (khách hàng mua và không mua) tốt hơn so với việc dự đoán ngẫu nhiên.
- Diện tích dưới đường cong (AUC) là 0.823, đây là giá trị khá tốt, thể hiện mô hình có khả năng phân loại chính xác ở mức tương đối cao. Cụ thể, mô hình có khoảng 82.3% khả năng đúng khi phân biệt một khách hàng mua với một khách hàng không mua được chọn ngẫu nhiên.
- ROC curve cũng cho thấy khả năng cân bằng giữa True Positive Rate (Recall) và False Positive Rate. Ở mức FPR thấp, tỉ lệ TPR cũng được giữ ổn định, điều này là tích cực cho việc phát hiện khách hàng mua mà không làm tăng quá nhiều dự đoán sai là khách không mua.

4.2.3. DecisionTreeClassifier

Kết quả các chỉ số đánh giá mô hình:

- Accuracy: 0.98
- Precision: 0.9574468085106383
- Recall: 0.9574468085106383
- F1-score: 0.9574468085106383
- ROC AUC: 0.972187456542901

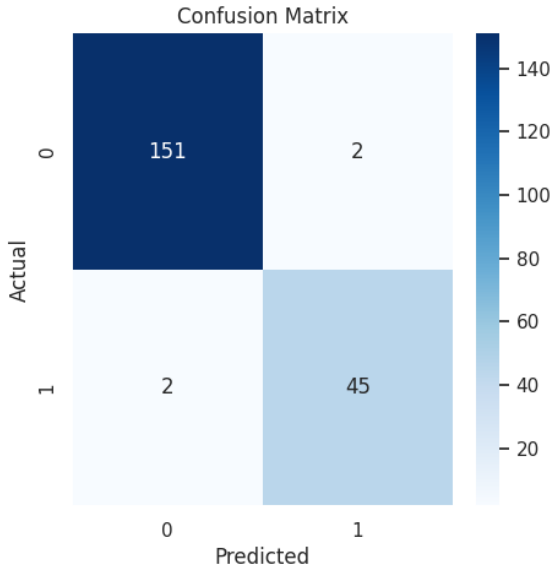
Accuracy (Độ chính xác) = 0.98: Mô hình dự đoán đúng 98% tổng số trường hợp, rất cao.

Precision = 0.957: Trong số các dự đoán khách hàng mua, có đến 95.7% là chính xác, cho thấy mô hình rất ít dự đoán sai khách không mua thành mua.

Recall = 0.957: Mô hình phát hiện được 95.7% khách hàng thực sự mua, tức tỷ lệ bỏ sót rất thấp.

F1-score = 0.957: Điểm cân bằng giữa Precision và Recall ở mức rất cao, thể hiện hiệu quả tổng thể của mô hình rất tốt.

ROC AUC = 0.972: Khả năng phân biệt khách hàng mua và không mua xuất sắc, gần đạt mức hoàn hảo.



Hình 4-6: *Biểu đồ ma trận nhầm lẫn (Decision Tree)*

True Negative (TN) = 151: Mô hình dự đoán chính xác 151 trường hợp khách hàng không mua.

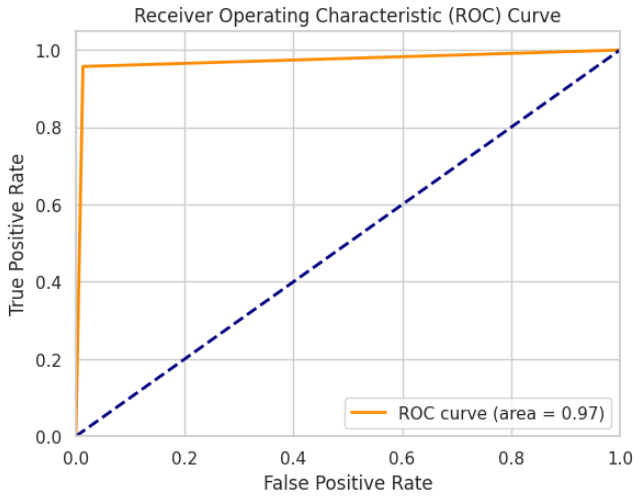
False Positive (FP) = 2: Mô hình chỉ sai dự đoán 2 trường hợp khách hàng không mua thành mua.

False Negative (FN) = 2: Mô hình bỏ sót rất ít, chỉ 2 trường hợp khách hàng thực sự mua nhưng dự đoán không mua.

True Positive (TP) = 45: Mô hình dự đoán đúng 45 trường hợp khách hàng thực sự mua.

Nhận xét:

- Mô hình có khả năng phân loại chính xác rất cao cả hai lớp, với tỷ lệ nhầm lẫn rất nhỏ (chỉ 2 trường hợp sai ở mỗi phía).
- Tỷ lệ bỏ sót khách hàng mua (FN) rất thấp, điều này cực kỳ quan trọng trong bài toán dự báo hành vi mua sắm vì giảm thiểu nguy cơ bỏ qua khách hàng tiềm năng.
- Số lượng dự đoán sai nhóm không mua thành mua (FP) cũng rất ít, giảm thiểu lãng phí khi tiếp cận nhầm đối tượng không có khả năng mua hàng.
- Ma trận nhầm lẫn phản ánh rõ hiệu suất vượt trội của Decision Tree so với Logistic Regression trước đó, đồng thời rất phù hợp cho ứng dụng thực tế.



Hình 4-7: *Biểu đồ đương cong ROC (Decision Tree)*

Nhận xét:

Mô hình Decision Tree có hiệu suất xuất sắc với AUC gần 0.97, thể hiện khả năng phân biệt hai lớp rất tốt. ROC curve nằm sát góc trên trái cho thấy mô hình vừa phát hiện chính xác khách mua, vừa giảm thiểu sai sót. Đây là mô hình rất tin cậy và phù hợp cho dự báo hành vi khách hàng.

4.2.4. So sánh các mô hình

Bảng 4-1: *Bảng tổng hợp kết quả đánh giá các mô hình*

Chỉ số	**Logistic Regression**	**Decision Tree**	**Random Forest**
Accuracy	0.770	0.980	0.975
Precision	0.514	0.957	0.920
Recall	0.404	0.957	0.979
F1-score	0.452	0.957	0.948
ROC AUC	0.823	0.972	0.997

Nhận xét:

- Logistic Regression có hiệu suất thấp nhất, đặc biệt về Precision và Recall, dễ bỏ sót khách hàng mua.
- Decision Tree cân bằng tốt giữa các chỉ số với hiệu suất rất cao.
- Random Forest có ROC AUC cao nhất và Recall tốt nhất, thể hiện khả năng phát hiện khách hàng mua xuất sắc, dù Precision hơi thấp hơn Decision Tree chút ít.

Dựa vào kết quả so sánh giữa ba mô hình sử dụng mô hình tốt nhất Random Forest thực hiện dự báo hành vi mua hàng cho một khách hàng được huấn luyện với 8 đặc trưng quan trọng như Recency, Frequency, Monetary, Age, giới tính (one-hot encoding) và danh mục sản phẩm. Dữ liệu khách hàng ví dụ gồm: độ trễ gần nhất 100 ngày, 5 lần mua, chi tiêu 5000, tổng số lượng 20, giá trung bình 50, tuổi 30, giới tính nam và thuộc nhóm danh mục sản phẩm quần áo. Mô hình được sử dụng để dự đoán khả năng mua hàng (0 hoặc 1) kèm theo xác suất dự đoán, giúp doanh nghiệp nắm bắt chính xác khả năng khách hàng này sẽ mua trong tương lai gần.

Kết quả:

```
Prediction: 0
```

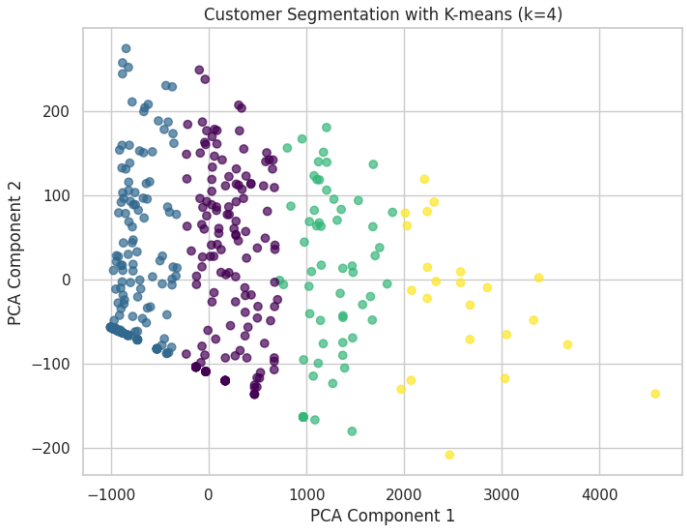
Prediction: 0 cho thấy mô hình dự báo khách hàng mẫu này không có khả năng mua hàng trong khoảng thời gian dự báo (120 ngày).

4.3. PHÂN CỤM KHÁCH HÀNG

4.3.1. Trực quan hóa kết quả phân cụm

Để trực quan hóa kết quả phân cụm và so sánh giữa hai thuật toán phân cụm K-Means và Spectral Clustering, nhóm nghiên cứu đã sử dụng phương pháp phân tích thành phần chính (PCA) nhằm giảm số chiều của dữ liệu xuống còn hai thành phần chính. Điều này giúp tái hiện dữ liệu phức tạp trong không gian hai chiều, từ đó dễ dàng quan sát và so sánh các nhóm khách hàng.

Dưới đây là kết quả hiển thị:



Hình 4-8: *Biểu đồ trực quan phân khúc khách hàng với thuật toán K-Means*

Nhận xét:

Dữ liệu khách hàng được chia thành 4 nhóm phân biệt khá rõ ràng trên không gian hai thành phần chính, cho thấy K-means đã nắm bắt được các nhóm đặc trưng khách hàng khác biệt.

Hình 4-9: *Biểu đồ trực quan phân khúc khách hàng với Spectral Clustering*

Nhận xét:

Một số cụm khá tập trung, trong khi các cụm khác có phần trải rộng và không đồng đều, điều này cho thấy Spectral Clustering có thể phát hiện các cấu trúc dữ liệu phức tạp hơn, không nhất thiết là các hình dạng cầu đều như K-means.

4.3.2. Gán tên cụm cho từng nhóm

Tiếp đến nhóm thực hiện việc gán tên phân khúc khách hàng bằng tiếng Việt dựa trên kết quả phân cụm K-means. Trước tiên, một từ điển segment_names được định nghĩa để ánh xạ các nhãn cụm số thành các tên phân khúc có ý nghĩa như "Nguy cơ rời bỏ", "Khách hàng mới", "Chi tiêu lớn" và "Chi tiêu thấp nhưng thường xuyên". Sau đó, phương thức .map() được sử dụng để chuyển đổi các giá trị trong cột Cluster_Kmeans của bảng dữ liệu rfm_data thành các tên phân khúc tương ứng, kết quả được lưu vào cột mới Segment.

- Nguy cơ rời bỏ: Các khách hàng có xu hướng ít tương tác và mua sắm ít.
- Khách hàng mới: Những khách hàng mới tham gia và có tần suất mua sắm thấp.

- Chi tiêu lớn: Các khách hàng có mức chi tiêu cao, thể hiện sự trung thành và giá trị lớn đối với doanh nghiệp.
- Chi tiêu thấp nhưng thường xuyên: Các khách hàng mua sắm thường xuyên nhưng với mức chi tiêu thấp.

Việc gán tên các phân khúc khách hàng như "Nguy cơ rời bỏ", "Khách hàng mới", "Chi tiêu lớn" và "Chi tiêu thấp nhưng thường xuyên" được thực hiện dựa trên đặc điểm và hành vi tiêu biểu của từng nhóm khách hàng sau khi phân cụm bằng thuật toán K-means. Cụ thể, chúng tôi đã phân tích các chỉ số Recency (thời gian kể từ lần mua gần nhất), Frequency (tần suất mua) và Monetary (giá trị chi tiêu) của từng cụm để nhận diện rõ đặc trưng hành vi khách hàng trong mỗi nhóm. Ví dụ, nhóm có giá trị Recency cao, nghĩa là khách hàng lâu không tương tác, được đặt tên là “Nguy cơ rời bỏ”; nhóm thường xuyên mua nhưng chi tiêu thấp được gọi là “Chi tiêu thấp nhưng thường xuyên”; còn nhóm có giá trị chi tiêu lớn được gọi là “Chi tiêu lớn”.

Bảng 4-2: Bảng dữ liệu sau khi gán tên cụm

Customer ID	R_Score	F_Score	M_Score	RFM_Score	Cluster_Kmeans	Segment
CUST000	3	4	1	341	3	Chi tiêu thấp nhưng thường xuyên
CUST001	3	3	1	331	2	Chi tiêu lớn
CUST002	3	3	1	331	2	Chi tiêu lớn
CUST003	3	4	4	344	1	Khách hàng mới
CUST005	3	2	3	323	0	Nguy cơ rời bỏ

4.3.3. Phân phối phân khúc khách hàng

Nhằm đánh giá quy mô tương đối của từng phân khúc khách hàng sau quá trình phân cụm bằng K-Means, nhóm đã tiến hành thống kê số lượng khách hàng thuộc mỗi phân khúc. Chúng em thực hiện việc đếm số lượng khách hàng trong từng nhóm, sau đó tổ chức lại dữ liệu thành một bảng hiển thị rõ ràng hai cột: tên phân khúc và số lượng khách

hàng tương ứng. Việc xác định quy mô từng phân khúc đóng vai trò quan trọng trong việc ưu tiên nguồn lực và hoạch định chiến lược tiếp cận.

Dưới đây là kết quả hiển thị:

Segment	Count
Khách hàng mới	209
Nguy cơ rời bỏ	147
Chi tiêu lớn	64
Chi tiêu thấp nhưng thường xuyên	23

Hình 4-10: *Biểu đồ phân phối số lượng từng nhóm khách hàng*

Nhận xét:

Biểu đồ phân phối cho thấy số lượng khách hàng trong từng phân khúc có sự khác biệt rõ rệt. Nhóm "Khách hàng mới" chiếm số lượng lớn nhất với hơn 200 khách, cho thấy doanh nghiệp đang thu hút được nhiều khách hàng mới nhưng cũng cần lưu ý giữ chân nhóm này. Nhóm "Nguy cơ rời bỏ" đứng thứ hai với khoảng 150 khách, báo hiệu doanh nghiệp cần triển khai các biện pháp chăm sóc, khuyến mãi để gia tăng sự gắn bó của nhóm này. Nhóm "Chi tiêu lớn" có số lượng khách ít hơn đáng kể, chỉ khoảng 60 khách, tuy nhiên đây là nhóm mang lại giá trị cao cần được tập trung phát triển. Cuối cùng, nhóm "Chi tiêu thấp nhưng thường xuyên" là nhóm nhỏ nhất với khoảng 20 khách, doanh nghiệp có thể cân nhắc các chiến lược duy trì hoặc thúc đẩy tăng chi tiêu nhóm này. Tổng thể, biểu đồ giúp doanh nghiệp phân bổ nguồn lực hợp lý để tối ưu hóa hiệu quả marketing và chăm sóc khách hàng theo từng phân khúc.

KẾT LUẬN VÀ KIẾN NGHỊ

Kết luận

Đề tài "Phân tích hành vi mua sắm, phân nhóm khách hàng và dự đoán khả năng quay lại của khách hàng từ dữ liệu bán lẻ" đã khai thác hiệu quả bộ dữ liệu giao dịch thực tế để làm sáng tỏ đặc điểm hành vi tiêu dùng của khách hàng trong môi trường bán lẻ. Thông qua việc áp dụng mô hình RFM kết hợp các thuật toán phân cụm như K-Means và Spectral Clustering, nhóm nghiên cứu đã phân loại thành công khách hàng thành các phân khúc có hành vi mua sắm khác biệt rõ rệt. Đồng thời, việc xây dựng và đánh giá các mô hình dự báo như Logistic Regression, Decision Tree và Random Forest đã giúp dự đoán chính xác khả năng quay lại mua hàng của khách hàng với hiệu suất cao, trong đó mô hình Random Forest thể hiện độ chính xác và khả năng phân biệt xuất sắc.

Việc chuẩn hóa dữ liệu, kết hợp phân tích thống kê mô tả và trực quan hóa dữ liệu đã cung cấp cái nhìn toàn diện về giá trị, tần suất cũng như xu hướng tiêu dùng theo độ tuổi, giới tính và danh mục sản phẩm của từng nhóm khách hàng. Quá trình so sánh giữa các phương pháp phân cụm cùng đánh giá qua chỉ số silhouette đã củng cố tính tin cậy của kết quả phân tích phân nhóm. Tổng thể, đề tài đã đạt được mục tiêu đề ra, cung cấp cơ sở dữ liệu định lượng vững chắc giúp doanh nghiệp nhận diện khách hàng tiềm năng, xây dựng chiến lược chăm sóc phù hợp và tối ưu hóa hiệu quả kinh doanh trong bối cảnh cạnh tranh gay gắt hiện nay.

Tuy vậy, đề tài vẫn có một số hạn chế như phạm vi dữ liệu đầu vào tương đối hạn chế, chưa khai thác sâu các nguồn dữ liệu đa chiều như hành vi trên nền tảng số hoặc phản hồi từ khách hàng. Bên cạnh đó, một số mô hình dự báo vẫn còn tiềm năng cải tiến về khả năng cân bằng giữa độ nhạy và độ chính xác trong một số trường hợp cụ thể.

Kiến nghị

Trên cơ sở kết quả đã đạt được, nhóm nghiên cứu kiến nghị một số hướng phát triển và hoàn thiện sau:

- Mở rộng phạm vi dữ liệu đầu vào bằng cách tích hợp thêm các nguồn dữ liệu bổ sung như phản hồi khách hàng, hành vi tương tác trên nền tảng số và mạng xã hội. Việc này giúp xây dựng bức tranh toàn diện hơn về hành trình mua sắm và tâm lý khách hàng.

- Nâng cấp mô hình dự báo bằng cách áp dụng các thuật toán học máy tiên tiến như Random Forest, XGBoost hoặc các mô hình học sâu (Deep Learning) để khai thác hiệu quả các mối quan hệ phức tạp trong dữ liệu, từ đó tăng cường độ chính xác và tính ổn định của dự báo.
- Xây dựng hệ thống dashboard phân tích dữ liệu theo thời gian thực, giúp nhà quản trị theo dõi kịp thời các chỉ số khách hàng và hiệu suất kinh doanh, từ đó linh hoạt điều chỉnh chiến lược phù hợp với biến động thị trường và hành vi người tiêu dùng.
- Khai thác hiệu quả kết quả phân nhóm khách hàng cho các chiến dịch marketing cá nhân hóa. Tối ưu hóa nội dung, kênh truyền thông và chính sách chăm sóc dựa trên đặc điểm hành vi cụ thể của từng phân khúc sẽ nâng cao trải nghiệm, gia tăng giá trị vòng đời khách hàng và tỷ lệ giữ chân một cách bền vững.

Với các kiến nghị trên, nhóm nghiên cứu kỳ vọng mô hình sẽ tiếp tục được cải tiến và áp dụng rộng rãi, trở thành công cụ đắc lực giúp các doanh nghiệp đưa ra quyết định chiến lược dựa trên dữ liệu một cách chính xác, nhanh chóng và hiệu quả trong thời đại số hóa.

TÀI LIỆU THAM KHẢO

[1]. TS. Chu Bình Minh, TS Lê Xuân Huy (2025), *Học máy dành cho ngành Khoa học dữ liệu,* Trường Đại học Kinh tế - Kỹ thuật Công nghiệp.

[2]. Trần Thị Hoàng Yến, Bùi Văn Tân, Chu Bình Minh (2024), *Tài liệu học tập Nhập môn Trí tuệ Nhân tạo*, Trường Đại học Kinh tế - Kỹ thuật Công nghiệp.

[3]. Vũ Hữu Tiệp (2018), *Machine Learning Cơ Bản,* NXB Khoa học Kỹ thuật

[4]. Đinh Mạnh Tường (2016), *Học máy các kỹ thuật cơ bản và hiện đại*, NXB Đại học Quốc gia Hà Nội.

[5]. Foster Provost, Tom Fawcett (2013), *Data Science for Business*, O'Reilly Media, Sebastopol, California.

[6]. Michael J. A. Berry, Gordon S. Linoff (2004), *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*, Wiley.

[7]. S. S. Iyengar, M. R. Gupta (2020*), Customer Segmentation using Machine Learning*, International Journal of Advanced Research in Computer Science.

[8]. J. Liu (2019), *Predicting Customer Churn using Machine Learning*, Journal of Business Research.