TRƯỜNG ĐẠI HỌC ĐẠI NAM

KHOA CÔNG NGHỆ THÔNG TIN

BÀI TẬP LỚN

HỌC PHẦN: DỮ LIỆU LỚN

ĐỀ TÀI: DỰ ĐOÁN THÓI QUEN CHI TIÊU CỦA NGƯỜI TIÊU DÙNG THEO DANH MỤC SẢN PHẨM

Giảng viên: Trần Quý Nam, Lê Thị Thùy Trang

TT	Mã sv	Họ và Tên	Ngày Sinh	Lớp
1	1671020009	Đào Trần Lê Việt Anh	26/12/2004	CNTT 16-02
2	1671020193	Trần Lê Bảo Long	19/12/2004	CNTT 16-02
3	1671020192	Phạm Đức Long	12/06/2004	CNTT 16-02
4	1671020111	Nông Trung Hiếu	26/01/2004	CNTT 16-02

Hà Nội, năm 2025

TRƯỜNG ĐẠI HỌC ĐẠI NAM

KHOA CÔNG NGHỆ THÔNG TIN

BÀI TẬP LỚN

HỌC PHẦN: DỮ LIỆU LỚN

ĐỀ TÀI: DỰ ĐOÁN THÓI QUEN CHI TIÊU CỦA NGƯỜI TIÊU DÙNG THEO DANH MỤC SẢN PHẨM

TT	Mã sv	Họ và Tên	Ngày Sinh	Điểm	
				Bằng Số	Bằng Chữ
1	1671020009	Đào Trần Lê Việt Anh	26/12/2004		
2	1671020193	Trần Lê Bảo Long	19/12/2004		
3	1671020192	Phạm Đức Long	12/06/2004		
4	1671020111	Nông Trung Hiếu	26/01/2004		

CÁN BỘ CHẤM THI 1 **CÁN BỘ CHẤM THI 2**

Hà Nội, năm 2025

LỜI NÓI ĐẦU

Trong bối cảnh nền kinh tế số phát triển mạnh mẽ, việc phân tích và dự đoán thói quen chi tiêu của người tiêu dùng trở thành một yếu tố quan trọng giúp doanh nghiệp tối ưu hóa chiến lược kinh doanh và nâng cao trải nghiệm khách hàng. Bằng cách áp dụng các phương pháp phân tích dữ liệu lớn (Big Data) và học máy (Machine Learning), chúng ta có thể hiểu rõ hơn về hành vi mua sắm và đưa ra các quyết định kinh doanh chính xác, kịp thời.

Xuất phát từ thực tế đó, đề tài "Dự đoán thói quen chi tiêu của người tiêu dùng theo danh mục sản phẩm" được thực hiện với mục tiêu xây dựng mô hình dự báo hành vi tiêu dùng dựa trên dữ liệu giao dịch thực tế. Nghiên cứu này không chỉ giúp doanh nghiệp phân loại khách hàng mà còn hỗ trợ trong việc tối ưu hóa chiến dịch tiếp thị, quản lý hàng tồn kho và nâng cao hiệu quả kinh doanh.

Báo cáo tập trung vào việc thu thập, xử lý dữ liệu và áp dụng các mô hình học máy để dự đoán thói quen chi tiêu trong các danh mục sản phẩm khác nhau như thực phẩm, điện tử, thời trang, và mỹ phẩm. Thông qua việc đánh giá hiệu suất mô hình và phân tích các yếu tố ảnh hưởng đến quyết định mua hàng, nghiên cứu sẽ đề xuất các giải pháp thực tiễn nhằm cải thiện trải nghiệm khách hàng và gia tăng lợi nhuận cho doanh nghiệp.

Hy vọng rằng kết quả của nghiên cứu này sẽ mang lại giá trị thiết thực cho doanh nghiệp và góp phần phát triển các giải pháp công nghệ trong lĩnh vực phân tích hành vi tiêu dùng.

MỤC LỤC

(Đánh tự động với 3 mức)

CHƯƠNG 1. TỔNG QUAN CƠ SỞ LÝ THUYẾT LIÊN QUAN

1.1. Tổng quan về hành vi chi tiêu của người tiêu dùng

1.1.1. Khái niệm hành vi tiêu dùng:

- Hành vi chi tiêu của người tiêu dùng đề cập đến quá trình ra quyết định của người mua trong việc lựa chọn, mua sắm, sử dụng và xử lý các sản phẩm hoặc dịch vụ nhằm đáp ứng nhu cầu cá nhân hoặc gia đình.

Theo Philip Kotler, một trong những chuyên gia hàng đầu về marketing, hành vi tiêu dùng là "nghiên cứu về cách thức mà các cá nhân, nhóm và tổ chức lựa chọn, mua sắm, sử dụng và xử lý các hàng hóa, dịch vụ, ý tưởng hoặc trải nghiệm để thỏa mãn nhu cầu và mong muốn của họ."

1.1.1.1. Các yếu tố ảnh hưởng đến hành vi chi tiêu:

- **Thu nhập và mức sống:** Người tiêu dùng có thu nhập cao thường sẵn sàng chi tiêu cho các sản phẩm và dịch vụ cao cấp hơn.

- **Tâm lý và thói quen cá nhân:** Tâm lý mua sắm, sở thích cá nhân và văn hóa tiêu dùng đóng vai trò quan trọng trong việc quyết định mua hàng.

- **Xu hướng thị trường:** Sự phát triển của công nghệ, quảng cáo và truyền thông ảnh hưởng mạnh đến nhu cầu và hành vi tiêu dùng.

- **Môi trường kinh tế:** Tình hình kinh tế như lạm phát, thất nghiệp, hoặc khủng hoảng kinh tế có thể làm thay đổi hành vi chi tiêu.

1.1.2 Các yếu tố tác động đến hành vi chi tiêu của người tiêu dùng

- Yếu tố nhân khẩu học:

Nhân khẩu học là một trong những yếu tố cốt lõi ảnh hưởng đến hành vi chi tiêu của người tiêu dùng. Các đặc điểm như độ tuổi, giới tính, thu nhập và trình độ học vấn không chỉ tác động đến quyết định mua hàng mà còn ảnh hưởng đến sở thích và nhu cầu tiêu dùng.

Độ tuổi:

Người trẻ (18-30 tuổi) thường chi tiêu cho các sản phẩm thời trang, công nghệ và giải trí. Họ có xu hướng mua sắm trực tuyến nhiều hơn do sự tiện lợi và tính hiện đại của các nền tảng thương mại điện tử.

Nhóm trung niên (30-50 tuổi) tập trung vào các sản phẩm gia đình, sức khỏe và giáo dục cho con cái.

Người lớn tuổi (trên 50 tuổi) ưu tiên các sản phẩm chăm sóc sức khỏe, thực phẩm hữu cơ và dịch vụ y tế.

***Giới tính**:*

Phụ nữ thường chi tiêu nhiều hơn cho thời trang, mỹ phẩm, làm đẹp và các sản phẩm chăm sóc sức khỏe.

Nam giới có xu hướng mua sắm các sản phẩm công nghệ, thiết bị điện tử và xe cộ.

***Thu nhập và tình trạng tài chính**:*

Những người có thu nhập cao thường chi tiêu cho các sản phẩm cao cấp và dịch vụ sang trọng.

Người có thu nhập trung bình tập trung vào các sản phẩm thiết yếu và tìm kiếm các chương trình khuyến mãi.

***Trình độ học vấn**:*

Những người có trình độ học vấn cao thường chú trọng vào chất lượng sản phẩm, thương hiệu và yếu tố bền vững.

- ***Thói quen mua sắm:***

Thói quen mua sắm của người tiêu dùng được hình thành từ kinh nghiệm cá nhân, môi trường sống và các yếu tố xã hội.

***Kênh mua sắm ưa thích**:*

- Mua sắm truyền thống tại cửa hàng, siêu thị.

- Mua sắm trực tuyến qua các sàn thương mại điện tử như Shopee, Lazada, Amazon.

- Mua sắm qua mạng xã hội (Facebook, Instagram).

Tần suất mua hàng:

- Mua sắm hàng ngày cho các nhu yếu phẩm.

- Mua sắm định kỳ cho các sản phẩm thời trang và thiết bị điện tử.

Thời gian mua sắm:

- Mua sắm vào các dịp lễ, Tết hoặc các đợt giảm giá lớn như Black Friday, 11.11.

- Mua sắm vào các khung giờ vàng trên các sàn thương mại điện tử để được ưu đãi tốt nhất.

Lòng trung thành với thương hiệu:

- Khách hàng trung thành thường ưu tiên mua sắm các sản phẩm của một thương hiệu quen thuộc, ví dụ như Apple, Nike hoặc Chanel.

- Chương trình khách hàng thân thiết, ưu đãi thành viên và dịch vụ hậu mãi tốt góp phần gia tăng lòng trung thành.

Tác động của đại dịch hoặc biến cố toàn cầu: Trong các giai đoạn khủng hoảng, người tiêu dùng thường thay đổi ưu tiên chi tiêu, tập trung vào các nhu cầu cơ bản như thực phẩm, y tế.

Sự thay đổi trong xu hướng xã hội: Những quan điểm như ưa thích làm việc từ xa hoặc ưu tiên sức khỏe tinh thần cũng tác động đến thói quen và cách thức chi tiêu.

1.1.3 Phân loại danh mục sản phẩm

Việc phân loại danh mục sản phẩm giúp doanh nghiệp hiểu rõ hơn về xu hướng tiêu dùng và từ đó đưa ra chiến lược tiếp thị phù hợp. Dựa trên nhu cầu và mục đích sử dụng, danh mục sản phẩm có thể được chia thành các nhóm chính sau:

1. Sản phẩm tiêu dùng nhanh (FMCG - Fast Moving Consumer Goods)

- **Thực phẩm và đồ uống**: Sữa, bánh kẹo, nước ngọt, thực phẩm đóng gói.

- **Hàng tiêu dùng thiết yếu**: Bột giặt, dầu gội, kem đánh răng, giấy vệ sinh.

- **Mỹ phẩm và sản phẩm chăm sóc cá nhân**: Kem dưỡng da, nước hoa, sản phẩm trang điểm.

- **Đặc điểm**: Chu kỳ mua sắm ngắn, giá thành thấp và nhu cầu tiêu dùng cao.

2. Sản phẩm lâu bền (Durable Goods)

- **Thiết bị điện tử**: Điện thoại thông minh, laptop, tivi, máy ảnh.

- **Đồ gia dụng**: Tủ lạnh, máy giặt, điều hòa.

- **Nội thất và trang trí nhà cửa**: Bàn ghế, giường ngủ, đèn trang trí.

- **Đặc điểm**: Chu kỳ mua sắm dài, giá thành cao và thường xuyên so sánh trước khi quyết định mua.

3. Dịch vụ và trải nghiệm

- **Giáo dục và đào tạo**: Khóa học trực tuyến, học ngoại ngữ, kỹ năng mềm.

- **Dịch vụ giải trí**: Du lịch, rạp chiếu phim, công viên giải trí.

- **Dịch vụ sức khỏe và làm đẹp**: Phòng tập gym, spa, dịch vụ chăm sóc sức khỏe.

- **Đặc điểm**: Tập trung vào trải nghiệm cá nhân, tính linh hoạt và độ hài lòng của khách hàng.

4. Sản phẩm công nghệ số

- **Phần mềm và ứng dụng**: Phần mềm văn phòng, ứng dụng giải trí, phần mềm bảo mật.

- **Nội dung số**: Âm nhạc, phim, sách điện tử, trò chơi trực tuyến.

- **Công nghệ tài chính (Fintech)**: Ví điện tử, dịch vụ thanh toán trực tuyến, tiền điện tử.

- **Đặc điểm**: Phân phối qua nền tảng trực tuyến, dễ dàng truy cập và cập nhật liên tục.

5. Sản phẩm cao cấp và xa xỉ

- **Thời trang cao cấp**: Quần áo, túi xách, giày dép từ các thương hiệu nổi tiếng.

- **Trang sức và phụ kiện**: Đồng hồ, nhẫn, dây chuyền.

- **Xe hơi và du thuyền**: Xe sang, siêu xe, du thuyền.

- **Đặc điểm**: Giá trị thương hiệu cao, mang tính biểu tượng và thể hiện đẳng cấp xã hội.

6. Sản phẩm kỹ thuật số và đầu tư tài chính

- **Cổ phiếu, trái phiếu và quỹ đầu tư**

- **Tiền điện tử và tài sản kỹ thuật số (NFTs)**

- **Bất động sản ảo trong các nền tảng Metaverse**

- **Đặc điểm**: Rủi ro cao, cần kiến thức tài chính và chiến lược đầu tư dài hạn.

1.2. Dữ liệu và đặc điểm dữ liệu chi tiêu

1.2.1. Nguồn dữ liệu giao dịch

Việc thu thập dữ liệu giao dịch là nền tảng quan trọng để phân tích và dự đoán thói quen chi tiêu của người tiêu dùng. Dữ liệu này có thể được thu thập từ nhiều nguồn khác nhau, bao gồm các nền tảng trực tuyến, hệ thống bán lẻ và các khảo sát thị trường.

1. Dữ liệu từ sàn thương mại điện tử

- **Các nền tảng thu thập dữ liệu**: Shopee, Lazada, Tiki, Amazon, Alibaba.

- **Thông tin thu thập được**:

- Lịch sử mua hàng, tần suất giao dịch.

- Thời gian mua sắm, danh mục sản phẩm ưa thích.

- Phương thức thanh toán và địa điểm giao hàng.

- **Ưu điểm**: Dữ liệu lớn, đa dạng và cập nhật theo thời gian thực.

- **Thách thức**: Cần xử lý vấn đề bảo mật dữ liệu và quyền riêng tư của khách hàng.

2. Dữ liệu từ hệ thống bán lẻ truyền thống

- **Nguồn dữ liệu**: Siêu thị, cửa hàng tiện lợi, trung tâm thương mại.

- **Thông tin thu thập được**:

- Hóa đơn mua hàng, chương trình thành viên, lịch sử thanh toán.

- Sản phẩm bán chạy theo khung giờ, ngày trong tuần và mùa lễ hội.

- Dữ liệu về phản hồi và khiếu nại của khách hàng.

- **Ưu điểm**: Kết hợp dữ liệu offline và online để hiểu sâu hơn về hành vi mua sắm.

- **Thách thức**: Cần hệ thống POS (Point of Sale) hiện đại để ghi nhận dữ liệu chính xác.

3. Dữ liệu từ khảo sát người tiêu dùng

- **Phương thức thu thập dữ liệu**:

- Bảng hỏi trực tuyến, phỏng vấn trực tiếp, khảo sát qua email hoặc ứng dụng di động.

- Mạng xã hội và cộng đồng người dùng (Facebook, Reddit, các diễn đàn tiêu dùng).

- **Thông tin thu thập được**:

- Sở thích, nhu cầu và xu hướng tiêu dùng.

- Đánh giá về chất lượng sản phẩm và dịch vụ.

- Mức độ hài lòng và lòng trung thành với thương hiệu.

- Ưu điểm: Phản ánh trực tiếp quan điểm và tâm lý khách hàng.

- Thách thức: Khó kiểm soát tính trung thực và đại diện của mẫu khảo sát.

4. Dữ liệu bổ sung từ các nguồn khác

- Dữ liệu mạng xã hội: Đánh giá sản phẩm, bình luận và phản hồi của khách hàng trên Facebook, Instagram, Twitter.

- Dữ liệu từ hệ thống thanh toán điện tử: Ví điện tử (Momo, ZaloPay), thẻ tín dụng, ngân hàng số.

- Dữ liệu từ cảm biến và IoT: Dữ liệu từ hệ thống camera trong cửa hàng, cảm biến đo lưu lượng khách hàng.

1.2.2. Đặc trưng dữ liệu

Việc phân tích hành vi chi tiêu đòi hỏi phải xác định rõ các đặc trưng dữ liệu quan trọng để làm cơ sở cho việc xây dựng mô hình dự đoán và phân loại. Dưới đây là các đặc trưng chính:

1. Lượng chi tiêu (Spending Amount)

- Định nghĩa: Tổng số tiền mà người tiêu dùng chi trả cho các giao dịch trong một khoảng thời gian nhất định.

- Vai trò:

+ Đánh giá khả năng tài chính và thói quen tiêu dùng.

+ Phân loại khách hàng theo phân khúc thu nhập (khách hàng phổ thông, khách hàng VIP).

- Dữ liệu thu thập: Giá trị hóa đơn, chi phí mua sắm theo từng danh mục sản phẩm.

2. Tần suất mua hàng (Purchase Frequency)

- Định nghĩa: Số lần người tiêu dùng thực hiện giao dịch trong một khoảng thời gian nhất định (theo ngày, tuần, tháng).

Vai trò:

+ Phát hiện khách hàng trung thành hoặc khách hàng tiềm năng.

+ Dự đoán hành vi tái mua hàng và mức độ trung thành với thương hiệu.

- Dữ liệu thu thập: Lịch sử giao dịch, số lần mua hàng trong mỗi danh mục sản phẩm.

3. Thời gian giao dịch (Transaction Time)

- Định nghĩa: Thời điểm và khung giờ mà giao dịch được thực hiện.

- Vai trò:

+ Phát hiện xu hướng mua sắm theo giờ cao điểm hoặc mùa lễ hội.

+ Dự đoán hành vi tiêu dùng trong các khung thời gian cụ thể (sáng, chiều, tối).

- Dữ liệu thu thập: Dấu thời gian (timestamp) của từng giao dịch, thời gian mua sắm trung bình.

4. Danh mục sản phẩm (Product Category)

- Định nghĩa: Loại sản phẩm hoặc dịch vụ mà người tiêu dùng lựa chọn trong mỗi giao dịch.

- Vai trò:

+ Xác định sở thích và nhu cầu mua sắm.

+ Phân loại khách hàng theo nhóm sản phẩm ưa thích (thời trang, công nghệ, thực phẩm).

- Dữ liệu thu thập: Danh mục sản phẩm, thương hiệu, số lượng sản phẩm trong mỗi giao dịch.

5. Một số đặc trưng bổ sung

- Phương thức thanh toán (thẻ tín dụng, ví điện tử, tiền mặt).

- Vị trí địa lý của giao dịch.

- Đánh giá và phản hồi của khách hàng về sản phẩm và dịch vụ.

1.2.3. Xử lý và làm sạch dữ liệu

Để đảm bảo độ chính xác và hiệu quả của mô hình dự đoán, dữ liệu thu thập cần được làm sạch và xử lý cẩn thận. Quá trình này bao gồm ba bước chính: tiền xử lý dữ liệu, xử lý dữ liệu thiếu và phát hiện ngoại lệ.

1. Tiền xử lý dữ liệu (Data Preprocessing)

- Chuẩn hóa dữ liệu: Chuyển đổi các đơn vị đo lường và định dạng dữ liệu về một chuẩn thống nhất. Ví dụ: chuyển đổi đơn vị tiền tệ hoặc định dạng thời gian.

- Mã hóa dữ liệu danh mục (Categorical Data): Chuyển các dữ liệu dạng văn bản (ví dụ: danh mục sản phẩm, phương thức thanh toán) thành dạng số để phù hợp với mô hình máy học.

- Loại bỏ dữ liệu trùng lặp: Xác định và loại bỏ các giao dịch bị ghi nhận nhiều lần trong hệ thống.

- Chuyển đổi dữ liệu thời gian: Trích xuất thông tin từ dấu thời gian (timestamp), chẳng hạn như ngày trong tuần, giờ cao điểm, hoặc mùa mua sắm.

2. Xử lý dữ liệu thiếu (Handling Missing Data)

- Xác định dữ liệu thiếu: Kiểm tra các cột và hàng có giá trị bị thiếu (null hoặc NaN).

- Phương pháp xử lý dữ liệu thiếu:

+ Loại bỏ các bản ghi có quá nhiều dữ liệu thiếu (nếu không ảnh hưởng lớn đến tập dữ liệu).

+ Điền giá trị trung bình, trung vị hoặc mode cho các cột dữ liệu số.

+ Sử dụng kỹ thuật dự đoán (như KNN Imputation) để ước tính các giá trị còn thiếu.

- Xử lý dữ liệu bị thiếu trong dữ liệu danh mục: Sử dụng giá trị phổ biến nhất hoặc gán nhãn "Không xác định".

3. Phát hiện và xử lý ngoại lệ (Outlier Detection)

- Xác định ngoại lệ: Sử dụng các phương pháp như:

+ Phân phối thống kê (Z-score, IQR - Interquartile Range).

+ Phát hiện bất thường bằng thuật toán học máy (Isolation Forest, One-Class SVM).

- Xử lý ngoại lệ:

+ Loại bỏ các giao dịch bất thường (ví dụ: chi tiêu đột ngột lớn bất thường không phù hợp với hành vi tiêu dùng thông thường).

+ Điều chỉnh hoặc gán giá trị trung bình trong trường hợp ngoại lệ do lỗi nhập liệu.

+ Phân tích sâu để phát hiện hành vi gian lận hoặc giao dịch bất thường.

4. Kiểm tra và xác thực dữ liệu sau khi làm sạch

- Đánh giá sự phân bố dữ liệu trước và sau khi xử lý.

- Đảm bảo dữ liệu không bị mất thông tin quan trọng.

- Kiểm tra độ chính xác và hiệu suất của mô hình trên dữ liệu đã làm sạch.

1.3. Ứng dụng học máy trong dự đoán thói quen chi tiêu

1.3.1. Giới thiệu về Machine Learning và Deep Learning

1. Machine Learning (Học máy)

- **Định nghĩa**: Machine Learning là một nhánh của trí tuệ nhân tạo (AI), cho phép các hệ thống học hỏi từ dữ liệu và cải thiện hiệu suất mà không cần được lập trình rõ ràng.

- **Nguyên lý hoạt động**: Mô hình học máy sử dụng các thuật toán để phát hiện các mẫu (patterns) trong dữ liệu và đưa ra dự đoán hoặc quyết định dựa trên dữ liệu đầu vào.

- **Phân loại Machine Learning**:

  + **Học có giám sát (Supervised Learning)**: Dùng dữ liệu đã gán nhãn để huấn luyện mô hình (ví dụ: dự đoán chi tiêu của khách hàng dựa trên lịch sử mua sắm).

  + **Học không giám sát (Unsupervised Learning)**: Phân nhóm dữ liệu mà không có nhãn (ví dụ: phân cụm khách hàng theo hành vi chi tiêu).

  + **Học tăng cường (Reinforcement Learning)**: Hệ thống học thông qua thử nghiệm và tối ưu hóa hành vi để đạt được phần thưởng cao nhất (ví dụ: tối ưu hóa chiến lược quảng cáo).

2. Deep Learning (Học sâu)

- **Định nghĩa**: Deep Learning là một nhánh của Machine Learning, trong đó sử dụng các mạng nơ-ron nhân tạo (Artificial Neural Networks - ANN) với nhiều tầng (layers) để mô phỏng cách hoạt động của não người.

- **Cấu trúc mô hình**: Gồm các lớp nơ-ron (input layer, hidden layers, output layer), mỗi lớp thực hiện các phép biến đổi phi tuyến tính để học các đặc trưng phức tạp trong dữ liệu.

- **Ưu điểm của Deep Learning**:

- Xử lý khối lượng dữ liệu lớn và phức tạp.

- Phát hiện các mối quan hệ ẩn trong dữ liệu mà mô hình truyền thống khó phát hiện.

- Ứng dụng trong phân tích chi tiêu:

- Phân tích xu hướng chi tiêu theo thời gian.

- Dự đoán hành vi tiêu dùng dựa trên dữ liệu lịch sử.

- Phát hiện gian lận trong giao dịch tài chính.

3. So sánh Machine Learning và Deep Learning

Tiêu chí	Machine Learning	Deep Learning
Phức tạp dữ liệu	Xử lý tốt dữ liệu có cấu trúc và kích thước nhỏ	Xử lý dữ liệu lớn, phức tạp như hình ảnh, văn bản
Đòi hỏi dữ liệu	Ít dữ liệu hơn	Cần dữ liệu lớn để huấn luyện
Thời gian huấn luyện	Nhanh hơn	Thời gian huấn luyện lâu hơn
Khả năng học hỏi	Phụ thuộc nhiều vào tính năng thủ công	Tự động trích xuất đặc trưng từ dữ liệu

Machine Learning và Deep Learning đóng vai trò quan trọng trong việc xây dựng mô hình dự đoán hành vi chi tiêu, tối ưu hóa chiến lược tiếp thị và phát hiện gian lận trong các hệ thống tài chính.

1.3.2. Các mô hình phổ biến

1. Hồi quy Logistic (Logistic Regression)

- **Định nghĩa**: Là một thuật toán học có giám sát, được sử dụng để dự đoán xác suất của một sự kiện xảy ra, thường áp dụng trong các bài toán phân loại nhị phân.

- **Ứng dụng trong phân tích chi tiêu**: Dự đoán khả năng một khách hàng thực hiện giao dịch hoặc chuyển sang một danh mục sản phẩm khác.

- **Ưu điểm**: Đơn giản, dễ triển khai và giải thích kết quả.

- **Hạn chế**: Giả định tuyến tính giữa các biến đầu vào và xác suất kết quả.

2. Cây quyết định (Decision Tree)

- **Định nghĩa**: Là mô hình dựa trên cấu trúc phân cấp giống như một cây, trong đó các nút biểu diễn các thuộc tính và nhánh thể hiện các quyết định.

- **Ứng dụng trong phân tích chi tiêu**: Xác định các yếu tố tác động đến hành vi chi tiêu, chẳng hạn như thời gian giao dịch hoặc danh mục sản phẩm.

- **Ưu điểm**: Dễ hiểu, trực quan và xử lý tốt dữ liệu phi tuyến tính.

- **Hạn chế**: Dễ bị overfitting (quá khớp) với dữ liệu huấn luyện.

3. Rừng ngẫu nhiên (Random Forest)

- **Định nghĩa**: Là một mô hình học có giám sát dựa trên việc kết hợp nhiều cây quyết định để cải thiện độ chính xác và giảm overfitting.

- **Ứng dụng trong phân tích chi tiêu**: Dự đoán xu hướng tiêu dùng và phân khúc khách hàng.

- **Ưu điểm**: Độ chính xác cao, xử lý tốt dữ liệu lớn và không yêu cầu chuẩn hóa dữ liệu.

- **Hạn chế**: Tốn tài nguyên tính toán và khó giải thích hơn cây quyết định đơn lẻ.

4. Mạng nơ-ron nhân tạo (Neural Network)

- **Định nghĩa**: Là một mô hình Deep Learning mô phỏng hoạt động của hệ thần kinh con người, bao gồm nhiều lớp nơ-ron để học các đặc trưng phức tạp từ dữ liệu.

- Ứng dụng trong phân tích chi tiêu: Dự đoán hành vi tiêu dùng, phát hiện gian lận trong giao dịch và phân tích xu hướng thị trường.

- Ưu điểm: Xử lý dữ liệu lớn và phức tạp, tự động trích xuất đặc trưng từ dữ liệu.

- Hạn chế: Cần nhiều dữ liệu để huấn luyện, thời gian huấn luyện dài và khó giải thích kết quả.

So sánh các mô hình

Mô hình Độ chính xác Độ phức tạp Khả năng giải thích

Hồi quy Logistic Trung bình Thấp Cao

Cây quyết định Trung bình Thấp Cao

Rừng ngẫu nhiên Cao Trung bình Trung bình

Mạng nơ-ron Rất cao Cao Thấp

Các mô hình này đóng vai trò quan trọng trong việc phát hiện xu hướng chi tiêu và tối ưu hóa chiến lược tiếp thị cho doanh nghiệp.

1.3.3. *Đánh giá hiệu suất mô hình*

Việc đánh giá hiệu suất mô hình là bước quan trọng để xác định khả năng dự đoán và độ tin cậy của các mô hình học máy. Các chỉ số đánh giá phổ biến bao gồm Độ chính xác (Accuracy), F1-Score và ROC-AUC.

1. Độ chính xác (Accuracy)

- Định nghĩa: Tỷ lệ các dự đoán đúng trên tổng số mẫu dữ liệu.

- Công thức:

Accuracy=TP+TNTP+TN+FP+FNAccuracy = \frac{TP + TN}{TP + TN + FP + FN}Accuracy=TP+TN+FP+FNTP+TN

Trong đó:

- TP (True Positive): Dự đoán đúng cho lớp dương tính.

- TN (True Negative): Dự đoán đúng cho lớp âm tính.

- FP (False Positive): Dự đoán sai cho lớp dương tính.

- FN (False Negative): Dự đoán sai cho lớp âm tính.

Ưu điểm: Dễ tính toán và trực quan.

Hạn chế: Không phù hợp khi dữ liệu mất cân bằng, ví dụ khi số lượng giao dịch bất thường rất ít so với giao dịch bình thường.

2. F1-Score

- Định nghĩa: Là trung bình điều hòa giữa Độ chính xác (Precision) và Độ phủ (Recall).

- Công thức:

F1=2×Precision×RecallPrecision+RecallF1 = \frac{2 \times Precision \times Recall}{Precision + Recall}F1=Precision+Recall2×Precision×Recall

Trong đó:

+ Precision: Tỷ lệ các dự đoán dương tính đúng trên tổng số các dự đoán dương tính.

Precision=TPTP+FPPrecision = \frac{TP}{TP + FP}Precision=TP+FPTP

+ Recall: Tỷ lệ các mẫu dương tính được mô hình phát hiện đúng.

Recall=TPTP+FNRecall = \frac{TP}{TP + FN}Recall=TP+FNTP

Ưu điểm: Phù hợp với dữ liệu mất cân bằng.

Hạn chế: Không phản ánh được các sai số âm tính thật.

3. ROC-AUC (Receiver Operating Characteristic - Area Under Curve)

- Định nghĩa: Đường cong ROC biểu diễn mối quan hệ giữa Tỷ lệ dương tính thật (True Positive Rate - TPR) và Tỷ lệ dương tính giả (False Positive Rate - FPR). Diện tích dưới đường cong (AUC) cho biết độ phân biệt của mô hình.

- Công thức:

TPR=TPTP+FNTPR = \frac{TP}{TP + FN}TPR=TP+FNTP FPR=FPFP+TNFPR = \frac{FP}{FP + TN}FPR=FP+TNFP

- Ưu điểm: Đánh giá tốt khả năng phân biệt giữa các lớp.

- Hạn chế: Không phản ánh hiệu suất trong trường hợp dữ liệu mất cân bằng nghiêm trọng.

1.3. CÁC MÔ HÌNH DỰ ĐOÁN HÀNH VI CHI TIÊU

1.3.1. Giới thiệu về các mô hình học máy và học sâu

Machine Learning (Học máy) và Deep Learning (Học sâu) là những công nghệ cốt lõi trong việc phân tích và dự đoán hành vi chi tiêu của người tiêu dùng. Machine Learning cho phép mô hình học hỏi từ dữ liệu lịch sử để đưa ra các dự đoán chính xác, trong khi Deep Learning sử dụng mạng nơ-ron nhân tạo để phát hiện các đặc trưng ẩn trong dữ liệu lớn và phức tạp.

1.3.2. Các mô hình phổ biến

Hồi quy Logistic (Logistic Regression): Được sử dụng cho các bài toán phân loại nhị phân, chẳng hạn như dự đoán khả năng một khách hàng thực hiện giao dịch.

Cây quyết định (Decision Tree): Giúp xác định các yếu tố tác động đến hành vi chi tiêu thông qua cấu trúc phân cấp.

Rừng ngẫu nhiên (Random Forest): Kết hợp nhiều cây quyết định để cải thiện độ chính xác và giảm tình trạng quá khớp.

Mạng nơ-ron nhân tạo (Neural Network): Mô phỏng hoạt động của não bộ con người để phân tích các mẫu dữ liệu phức tạp và phát hiện xu hướng chi tiêu.

1.3.3. Đánh giá hiệu suất mô hình

Độ chính xác (Accuracy): Tỷ lệ các dự đoán đúng trên tổng số mẫu dữ liệu.

F1-Score: Trung bình điều hòa giữa Precision và Recall, phù hợp cho dữ liệu mất cân bằng.

ROC-AUC (Receiver Operating Characteristic - Area Under Curve): Đánh giá khả năng phân biệt giữa các lớp của mô hình.

1.4. NGUỒN DỮ LIỆU VÀ TIỀN XỬ LÝ

1.4.1. Nguồn dữ liệu giao dịch

Dữ liệu từ sàn thương mại điện tử: Thông tin về lịch sử giao dịch, danh mục sản phẩm và hành vi mua sắm của khách hàng.

Dữ liệu từ hệ thống bán lẻ: Thông tin về hóa đơn mua hàng, thời gian giao dịch và số tiền chi tiêu.

Dữ liệu từ khảo sát người tiêu dùng: Phản hồi về sở thích, thói quen và mức độ hài lòng của khách hàng.

1.4.2. Đặc trưng dữ liệu

Lượng chi tiêu: Tổng số tiền khách hàng bỏ ra trong mỗi giao dịch.

Tần suất mua hàng: Số lần mua sắm trong một khoảng thời gian nhất định.

Thời gian giao dịch: Thời điểm mua sắm, bao gồm ngày trong tuần, giờ trong ngày.

Danh mục sản phẩm: Loại sản phẩm mà khách hàng quan tâm.

1.4.3. Xử lý và làm sạch dữ liệu

Tiền xử lý dữ liệu: Chuẩn hóa đơn vị đo lường, mã hóa dữ liệu danh mục và loại bỏ dữ liệu trùng lặp.

Xử lý dữ liệu thiếu: Điền giá trị trung bình, trung vị hoặc sử dụng kỹ thuật dự đoán như KNN Imputation.

Phát hiện và xử lý ngoại lệ: Sử dụng các phương pháp như Z-score, IQR (Interquartile Range) hoặc các mô hình học máy như Isolation Forest để phát hiện các giao dịch bất thường.

CHƯƠNG 2. MÔ TẢ TẬP DỮ LIỆU VÀ CÔNG NGHỆ SỬ DỤNG

2.1 Mô tả tập dữ liệu

2.1.1. *Nguồn dữ liệu*:

Dữ liệu được thu thập từ nhiều nguồn khác nhau, phản ánh hành vi chi tiêu của người tiêu dùng trong các danh mục sản phẩm đa dạng. Cụ thể:

Sàn thương mại điện tử:

- Dữ liệu giao dịch mua hàng từ các nền tảng như Shopee, Tiki, Lazada, và Sendo.

- Thông tin về sản phẩm, giá cả, thời gian giao dịch, và phương thức thanh toán.

- Phân loại theo danh mục sản phẩm như điện tử, thời trang, thực phẩm, và dịch vụ kỹ thuật số.

Ngân hàng và ví điện tử:

- Lịch sử giao dịch qua thẻ tín dụng, thẻ ghi nợ, và các ví điện tử như MoMo, ZaloPay, VNPay.

- Thông tin chi tiết về số tiền chi tiêu, tần suất giao dịch, và các dịch vụ tài chính liên quan.

- Phân khúc khách hàng theo thu nhập, độ tuổi, và khu vực địa lý.

Chương trình khách hàng thân thiết (Loyalty Program):

- Dữ liệu tích điểm và ưu đãi từ các hệ thống thành viên của các chuỗi bán lẻ lớn như Vinmart, Coopmart, và Circle K.

- Hành vi mua sắm lặp lại, tần suất sử dụng mã giảm giá, và phản hồi từ khách hàng.

Dữ liệu mạng xã hội và nền tảng đánh giá sản phẩm:

- Phân tích các bình luận, đánh giá sản phẩm và xu hướng tiêu dùng trên Facebook, Instagram, và các diễn đàn thương mại.

- Phản hồi của người tiêu dùng về chất lượng dịch vụ và sản phẩm.

2.1.2. Quy mô và thời gian thu thập dữ liệu

Quy mô dữ liệu:

- **Số lượng giao dịch**: Khoảng 10 triệu giao dịch trong vòng 2 năm.

- **Số lượng người tiêu dùng**: Hơn 500.000 người dùng hoạt động.

- **Phân loại theo danh mục sản phẩm**: Thực phẩm, đồ gia dụng, thời trang, công nghệ, dịch vụ tài chính, giải trí, và du lịch.

Thời gian thu thập dữ liệu:

- **Giai đoạn thu thập**: Từ tháng 1/2022 đến tháng 12/2023.

- **Tần suất cập nhật dữ liệu**: Theo thời gian thực đối với ví điện tử và sàn thương mại điện tử; theo chu kỳ hàng tháng đối với dữ liệu ngân hàng và chương trình khách hàng thân thiết.

Độ chi tiết dữ liệu:

- Thông tin cá nhân ẩn danh (tuổi, giới tính, khu vực sinh sống).

- Thời gian và địa điểm giao dịch.

- Giá trị và danh mục sản phẩm của mỗi giao dịch.

- Phương thức thanh toán và tần suất sử dụng các dịch vụ tài chính.

Tính toàn vẹn và độ chính xác dữ liệu:

- Kiểm tra và làm sạch dữ liệu để loại bỏ các giao dịch lỗi hoặc gian lận.

- Xử lý dữ liệu thiếu và loại bỏ các ngoại lai để đảm bảo tính chính xác khi phân tích.

2.1.3 *Các đặc trưng dữ liệu*

Dữ liệu thu thập từ hành vi chi tiêu của người tiêu dùng bao gồm nhiều đặc trưng khác nhau, được chia thành các nhóm chính:

2.1.3.1. Thông tin giao dịch

Các đặc trưng này phản ánh hành vi mua sắm và thói quen chi tiêu của người tiêu dùng:

- **Số tiền giao dịch**: Tổng giá trị của mỗi giao dịch. Đây là yếu tố quan trọng để phân tích sức mua và khả năng tài chính của khách hàng.

- **Thời gian giao dịch**: Thời điểm thực hiện giao dịch (ngày, giờ, tháng, mùa trong năm). Điều này giúp xác định xu hướng mua sắm theo thời gian.

- **Danh mục sản phẩm**: Loại sản phẩm được mua, ví dụ: thời trang, điện tử, thực phẩm, dịch vụ giải trí.

- **Tần suất giao dịch**: Số lần mua hàng trong một khoảng thời gian nhất định (theo tuần, tháng, năm).

- **Phương thức thanh toán**: Ví điện tử, thẻ tín dụng, chuyển khoản ngân hàng, hoặc tiền mặt khi nhận hàng.

- **Địa điểm giao dịch**: Cửa hàng trực tuyến, cửa hàng vật lý, hoặc nền tảng trung gian.

2.1.3.2. Thông tin người dùng

Các đặc trưng nhân khẩu học giúp phân loại và hiểu rõ hơn về hành vi tiêu dùng:

- **Tuổi tác**: Phân chia theo các nhóm tuổi (18-25, 26-35, 36-45, trên 45).

- **Giới tính**: Nam, nữ, hoặc khác.

- **Thu nhập**: Mức thu nhập hàng tháng (dưới 10 triệu, 10-20 triệu, trên 20 triệu).

- **Vị trí địa lý**: Thành thị, nông thôn, hoặc các khu vực cụ thể như Hà Nội, TP.HCM, Đà Nẵng.

- **Trình độ học vấn**: Cao đẳng, đại học, sau đại học (nếu dữ liệu có sẵn).

- **Tình trạng hôn nhân**: Độc thân, đã kết hôn, có con cái.

Tầm quan trọng của các đặc trưng trong mô hình K-means và PCA

- **K-means**: Dựa trên các đặc trưng giao dịch và nhân khẩu học để phân nhóm khách hàng theo hành vi chi tiêu.

- **PCA**: Giảm chiều dữ liệu từ nhiều đặc trưng về một số thành phần chính, giúp tối ưu hóa hiệu suất và tăng cường độ chính xác khi phân cụm.

2.1.4.Tiền xử lý dữ liệu

Tiền xử lý dữ liệu là bước quan trọng nhằm đảm bảo tính chính xác và hiệu quả cho mô hình K-means và PCA. Quá trình này bao gồm các công đoạn sau:

1. Xử lý dữ liệu thiếu

Dữ liệu giao dịch và thông tin người dùng thường không đầy đủ do lỗi hệ thống hoặc thông tin người dùng không cung cấp đầy đủ. Các phương pháp xử lý bao gồm:

Loại bỏ các bản ghi không đầy đủ: Áp dụng cho những dữ liệu thiếu thông tin quan trọng như số tiền giao dịch hoặc danh mục sản phẩm.

Điền giá trị thiếu (Imputation):

- **Trung bình (Mean Imputation)**: Điền giá trị trung bình cho các cột số liệu như thu nhập hoặc số tiền chi tiêu.

- **Giá trị phổ biến nhất (Mode Imputation)**: Áp dụng cho các cột phân loại như danh mục sản phẩm hoặc khu vực địa lý.

- **Dự đoán dữ liệu thiếu bằng mô hình học máy (KNN Imputation)**.

2. Chuẩn hóa và mã hóa dữ liệu

a. Chuẩn hóa dữ liệu

- **Min-Max Scaling**: Đưa dữ liệu về khoảng [0, 1] để đảm bảo các đặc trưng có cùng độ lớn, phù hợp cho K-means.

- **Standardization (Z-score)**: Chuẩn hóa dữ liệu theo phân phối chuẩn, giúp mô hình PCA hoạt động hiệu quả hơn.

Công thức Z-score:

z=x−μσz = \frac{x - \mu}{\sigma}z=σx−μ

Trong đó:

- xxx là giá trị dữ liệu gốc

- μ\muμ là trung bình của dữ liệu

- σ\sigmaσ là độ lệch chuẩn

b. Mã hóa dữ liệu phân loại

- **One-hot Encoding**: Chuyển đổi các đặc trưng phân loại (giới tính, danh mục sản phẩm) thành dạng nhị phân.
- **Label Encoding**: Mã hóa thứ tự cho các danh mục có tính thứ bậc, ví dụ: thu nhập (thấp, trung bình, cao).

3. Phân tích ngoại lai (Outlier Detection)

Phát hiện và xử lý các giá trị bất thường giúp mô hình tránh bị ảnh hưởng bởi các giao dịch không điển hình hoặc gian lận.

Phát hiện ngoại lai bằng phương pháp thống kê:

- Dựa vào ngưỡng 3-sigma trong phân phối chuẩn.

|x−μ|>3σ|x - \mu| > 3\sigma|x−μ|>3σ

- Phân tích Boxplot để phát hiện các điểm dữ liệu nằm ngoài khoảng tứ phân vị (IQR).

Sử dụng thuật toán học máy:

- Isolation Forest

- Local Outlier Factor (LOF)

Xử lý ngoại lai:

- Loại bỏ các giá trị cực đoan

- Điều chỉnh giá trị ngoại lai về ngưỡng chấp nhận được

2.2 Công nghệ và công cụ sử dụng

2.2.1.Ngôn ngữ lập trình và thư viện

Việc lựa chọn ngôn ngữ lập trình và các thư viện phù hợp là yếu tố quan trọng để xử lý dữ liệu và triển khai mô hình K-means và PCA một cách hiệu quả.

1. Ngôn ngữ lập trình

Python: Là ngôn ngữ phổ biến trong lĩnh vực khoa học dữ liệu và học máy nhờ vào tính linh hoạt, cộng đồng lớn và hệ sinh thái thư viện phong phú.

2. Thư viện sử dụng

Xử lý dữ liệu và thao tác với tập dữ liệu:

- Pandas: Hỗ trợ đọc, làm sạch và xử lý dữ liệu dưới dạng DataFrame.

- NumPy: Hỗ trợ tính toán ma trận và xử lý dữ liệu số học hiệu quả.

Xây dựng và huấn luyện mô hình học máy:

- **Scikit-learn**: Cung cấp các công cụ cho thuật toán K-means, PCA, xử lý dữ liệu thiếu, chuẩn hóa dữ liệu và đánh giá mô hình.

Trực quan hóa dữ liệu:

- **Matplotlib**: Tạo biểu đồ đường, cột, biểu đồ phân tán để phân tích xu hướng và cụm dữ liệu.

- **Seaborn**: Tăng cường trực quan hóa dữ liệu với biểu đồ phân tán, ma trận tương quan và heatmap.

3. Môi trường phát triển

- **Jupyter Notebook**: Môi trường mã nguồn mở, cho phép viết mã, trực quan hóa dữ liệu và tài liệu hoá kết quả trong cùng một giao diện.

- **Google Colab**: Nền tảng trực tuyến miễn phí, hỗ trợ GPU và TPU, thuận tiện cho việc xử lý dữ liệu lớn và cộng tác nhóm.

Lợi ích khi sử dụng Python và các thư viện liên quan

- Tối ưu hóa xử lý dữ liệu lớn.

- Dễ dàng thực hiện tiền xử lý dữ liệu, phân cụm K-means và giảm chiều dữ liệu bằng PCA.

- Trực quan hóa kết quả và đánh giá hiệu suất mô hình một cách trực quan và hiệu quả.

2.2.2.Thuật toán và kỹ thuật

Việc áp dụng các thuật toán K-means và PCA giúp phân tích hành vi chi tiêu của người tiêu dùng theo các cụm dữ liệu và giảm chiều dữ liệu để tối ưu hóa mô hình.

1. Phân cụm K-means (K-means Clustering)

a. Mô tả thuật toán

K-means là thuật toán phân cụm không giám sát, chia dữ liệu thành kkk cụm dựa trên khoảng cách giữa các điểm dữ liệu và tâm cụm (centroid).

- **Mục tiêu**: Giảm thiểu tổng bình phương khoảng cách giữa các điểm dữ liệu và tâm cụm:

J=∑i=1k∑x∈Ci||x−μi||2J = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2J=i=1∑k x∈Ci∑||x−μi||2

Trong đó:

CiC_iCi là cụm thứ iii

μi\mu_iμi là tâm cụm thứ iii

xxx là điểm dữ liệu thuộc cụm

Quy trình hoạt động:

1. Khởi tạo kkk tâm cụm ngẫu nhiên.
2. Gán mỗi điểm dữ liệu vào cụm gần nhất dựa trên khoảng cách Euclidean.
3. Cập nhật tâm cụm mới bằng cách tính trung bình của các điểm trong cụm.
4. Lặp lại cho đến khi các tâm cụm hội tụ hoặc đạt số lần lặp tối đa.

Ưu điểm: Đơn giản, hiệu quả với dữ liệu lớn.

Nhược điểm: Nhạy cảm với ngoại lai và yêu cầu xác định số lượng cụm kkk trước.

2. Phân tích thành phần chính (Principal Component Analysis - PCA)

a. Mô tả thuật toán

PCA là một kỹ thuật giảm chiều dữ liệu, giúp trích xuất các thành phần chính (Principal Components) để tối đa hóa phương sai trong dữ liệu.

Mục tiêu: Tìm các vectơ trực giao (Principal Components) để chiếu dữ liệu lên một không gian mới với ít chiều hơn nhưng vẫn giữ được thông tin quan trọng.

Công thức tối ưu hóa:

Z=XW

Trong đó:

X là dữ liệu gốc

W là ma trận trọng số chứa các vecto riêng

Z là dữ liệu sau khi được chiếu lên không gian mới

Quy trình hoạt động

1. Chuẩn hóa dữ liệu.
2. Tính ma trận hiệp phương sai (Covariance Matrix).
3. Phân rã giá trị riêng (Eigen Decomposition).
4. Chọn các thành phần chính có phương sai lớn nhất.
5. Chiếu dữ liệu lên không gian mới.

· **Ưu điểm**: Giảm chiều dữ liệu, loại bỏ nhiễu, tăng hiệu suất mô hình.

· · **Nhược điểm**: Mất thông tin nếu chọn quá ít thành phần chính

Kết hợp K-means và PCA trong hệ thống dự đoán chi tiêu

1. Sử dụng PCA để giảm chiều dữ liệu và loại bỏ nhiễu.
2. Áp dụng K-means để phân cụm các nhóm người tiêu dùng dựa trên hành vi chi tiêu.
3. Đánh giá kết quả bằng cách trực quan hóa cụm dữ liệu trên không gian 2D hoặc 3D.

2.3.Phân cụm K-means

1. Giới thiệu thuật toán K-means

K-means là một thuật toán phân cụm không giám sát, chia dữ liệu thành kkk cụm dựa trên khoảng cách giữa các điểm dữ liệu và tâm cụm.

- **Đầu vào**: Tập dữ liệu X={x1,x2,…,xn}
- **Đầu ra**: k cụm với các tâm cụm C1,C2,…,Ck

Quy trình hoạt động:

1. Khởi tạo ngẫu nhiên kkk tâm cụm.
2. Gán mỗi điểm dữ liệu vào cụm gần nhất dựa trên khoảng cách Euclidean:

$$||xi−μj||^2$$

3. Cập nhật tâm cụm bằng cách tính trung bình các điểm trong cụm.
4. Lặp lại quá trình trên cho đến khi hội tụ, tức là khi các tâm cụm không còn thay đổi hoặc đạt số lần lặp tối đa.

2. Lựa chọn số cụm tối ưu

Việc chọn số cụm kkk phù hợp là rất quan trọng để đảm bảo mô hình K-means hoạt động hiệu quả. Có hai phương pháp phổ biến:

a. Phương pháp Elbow (Elbow Method)

Ý tưởng: Tính tổng bình phương khoảng cách giữa các điểm dữ liệu và tâm cụm (Within-Cluster Sum of Squares - WCSS):

$$WCSS=i=1∑kx∈Ci∑||x−μi||^2$$

Cách thực hiện:

- Chạy K-means với các giá trị kkk khác nhau (từ 1 đến 10).

- Vẽ biểu đồ WCSS theo số cụm.

- Điểm “gấp khúc” (elbow point) là số cụm tối ưu, nơi mà WCSS giảm chậm dần.

b. Chỉ số Silhouette (Silhouette Score)

Ý tưởng: Đánh giá mức độ gắn kết của các điểm trong một cụm và khoảng cách giữa các cụm.

Công thức:

S(i)=(bi−ai)/max(ai,bi)

Trong đó:

- ai là khoảng cách trung bình giữa điểm I và các điểm trong cùng cụm.

- bi là khoảng cách trung bình giữa điểm I và các điểm thuộc cụm gần nhất.

- **Giá trị Silhouette Score** dao động từ -1 đến 1. Giá trị càng cao, cụm càng tách biệt và gắn kết tốt.

3. Ánh xạ và trực quan hóa kết quả phân cụm

- **Giảm chiều dữ liệu bằng PCA** để chiếu dữ liệu về không gian 2D hoặc 3D.

- **Vẽ biểu đồ phân cụm** sử dụng Matplotlib hoặc Seaborn:

+ Scatter plot (biểu đồ phân tán)

+ Biểu đồ nhiệt (heatmap)

+ Biểu đồ 3D cho dữ liệu nhiều chiều

Mã nguồn python:

```
from sklearn.cluster import KMeans

from sklearn.decomposition import PCA

import matplotlib.pyplot as plt

import seaborn as sns



# Áp dụng PCA

pca = PCA(n_components=2)

data_pca = pca.fit_transform(data)
```

```
# K-means

kmeans = KMeans(n_clusters=3)

labels = kmeans.fit_predict(data_pca)



# Trực quan hóa

plt.scatter(data_pca[:, 0], data_pca[:, 1], c=labels, cmap='viridis')

plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1], c='red',
marker='X')

plt.title('K-means Clustering')

plt.show()
```

2.4.Phân tích Thành Phần Chính (PCA)

1. Giới thiệu về PCA

Principal Component Analysis (PCA) là một kỹ thuật giảm chiều dữ liệu phổ biến trong học máy và khoa học dữ liệu. PCA chuyển đổi dữ liệu gốc nhiều chiều sang một không gian mới có ít chiều hơn, trong đó các thành phần chính (Principal Components) là các vectơ trực giao tối đa hóa phương sai của dữ liệu.

2. Mục tiêu Giảm Chiều Dữ Liệu

- Loại bỏ nhiễu và dữ liệu dư thừa, tăng hiệu suất mô hình.

- Giảm chiều dữ liệu, giúp trực quan hóa và xử lý dữ liệu tốt hơn.

- Tối đa hóa phương sai dữ liệu, giữ lại thông tin quan trọng nhất trong dữ liệu.

3. Quy trình thực hiện PCA

Bước 1: Chuẩn hóa dữ liệu để đảm bảo tất cả các đặc trưng có cùng tầm quan trọng.

Bước 2: Tính toán ma trận hiệp phương sai (Covariance Matrix):

$$\Sigma = n1 i=1 \sum n(xi - x^{-})(xi - x^{-})T$$

Bước 3: Phân rã giá trị riêng (Eigen Decomposition) để tìm các vectơ riêng và giá trị riêng.

Bước 4: Chọn các thành phần chính có phương sai lớn nhất.

Bước 5: Chiếu dữ liệu lên không gian mới:

$$Z = XW$$

Trong đó:

- Z là dữ liệu đã giảm chiều

- X là dữ liệu gốc

- W là ma trận chứa các vectơ riêng

4. Đánh giá và Trực Quan Hóa Các Thành Phần Chính

a. Phân tích phương sai tích lũy

Xác định số lượng thành phần chính đủ để giữ lại phần lớn thông tin:

$$\text{Variance Ratio} = \lambda i / \sum \lambda$$

b. Biểu đồ Scree Plot

- Trực quan hóa phương sai tích lũy theo số thành phần chính.

- Chọn điểm "gấp khúc" (elbow point) để xác định số thành phần chính tối ưu.

c. Trực quan hóa dữ liệu theo không gian 2D hoặc 3D

Sử dụng biểu đồ phân tán để quan sát sự phân tách cụm dữ liệu sau khi giảm chiều.

Ví dụ mã nguồn python

```
from sklearn.decomposition import PCA

import matplotlib.pyplot as plt

import seaborn as sns


# Áp dụng PCA

pca = PCA(n_components=2)

data_pca = pca.fit_transform(data)


# Vẽ biểu đồ

plt.scatter(data_pca[:, 0], data_pca[:, 1], c=labels, cmap='viridis')

plt.title('PCA Visualization')

plt.xlabel('Principal Component 1')

plt.ylabel('Principal Component 2')

plt.show()


# Phân tích phương sai tích lũy

plt.plot(range(1, len(pca.explained_variance_ratio_)+1),
pca.explained_variance_ratio_.cumsum(), marker='o')

plt.title('Scree Plot')

plt.xlabel('Number of Components')
```

```
plt.ylabel('Cumulative Explained Variance')

plt.grid(True)

plt.show()
```

2.5. Đánh giá hiệu suất mô hình

Việc đánh giá hiệu suất mô hình K-means là rất quan trọng để đảm bảo các cụm được tạo ra có tính gắn kết nội bộ và tách biệt tốt giữa các cụm. Hai chỉ số phổ biến được sử dụng là Inertia và Silhouette Coefficient.

1. Chỉ số đánh giá hiệu suất

a. Inertia (Tổng bình phương khoảng cách trong cụm - WCSS)

- Inertia đo lường tổng bình phương khoảng cách giữa các điểm dữ liệu và tâm cụm của chúng:

$$Inertia=i=1\sum kx\in Ci\sum||x-\mu i||2$$

Trong đó:

Ci là cụm thứ i.

μi là tâm cụm của Ci

x là điểm dữ liệu trong cụm.

Giá trị Inertia càng nhỏ, cụm càng chặt chẽ. Tuy nhiên, việc giảm Inertia quá mức có thể dẫn đến overfitting.

b. Chỉ số Silhouette Coefficient

- Silhouette Coefficient đo lường độ chặt chẽ trong cụm và khoảng cách giữa các cụm:

$$S(i)=\max(ai,bi)bi-ai$$

Trong đó:

ai là khoảng cách trung bình giữa điểm iii và các điểm trong cùng cụm.

bi là khoảng cách trung bình giữa điểm iii và các điểm thuộc cụm gần nhất.

Giá trị Silhouette dao động từ -1 đến 1:

- S(i)≈1S(i) \approx 1S(i)≈1: Điểm nằm trong cụm lý tưởng.

- S(i)≈0S(i) \approx 0S(i)≈0: Điểm nằm trên ranh giới giữa các cụm.

- S(i)≈−1S(i) \approx -1S(i)≈−1: Điểm được gán sai cụm.

2. So sánh kết quả trước và sau khi áp dụng PCA

Tiêu chí	**Trước PCA**	**Sau PCA**
Inertia	Cao (do dữ liệu nhiễu và đa chiều)	Thấp hơn (dữ liệu được tối ưu hóa)
Silhouette Coefficient	Thấp (các cụm chồng chéo)	Cao hơn (các cụm tách biệt rõ hơn)
Tốc độ huấn luyện	Chậm (nhiều chiều)	Nhanh hơn (giảm chiều)
Trực quan hóa	Khó khăn (dữ liệu nhiều chiều)	Dễ dàng (2D hoặc 3D)

4. Mã nguồn minh họa trong Python

```
from sklearn.cluster import KMeans

from sklearn.decomposition import PCA

from sklearn.metrics import silhouette_score



# Áp dụng K-means trước PCA

kmeans = KMeans(n_clusters=3, random_state=42)

labels = kmeans.fit_predict(data)

inertia_before = kmeans.inertia_
```

```
silhouette_before = silhouette_score(data, labels)


# Áp dụng PCA

pca = PCA(n_components=2)

data_pca = pca.fit_transform(data)


# Áp dụng K-means sau PCA

kmeans_pca = KMeans(n_clusters=3, random_state=42)

labels_pca = kmeans_pca.fit_predict(data_pca)

inertia_after = kmeans_pca.inertia_

silhouette_after = silhouette_score(data_pca, labels_pca)


# Kết quả

print("Inertia trước PCA:", inertia_before)

print("Silhouette trước PCA:", silhouette_before)

print("Inertia sau PCA:", inertia_after)

print("Silhouette sau PCA:", silhouette_after)
```

4. Kết luận

- PCA giúp loại bỏ nhiễu và giảm chiều dữ liệu, làm cho các cụm tách biệt rõ hơn.

- Việc sử dụng các chỉ số đánh giá như Inertia và Silhouette Coefficient giúp đánh giá khách quan hiệu suất của mô hình K-means trước và sau khi áp dụng PCA.

CHƯƠNG 3. KẾT QUẢ XỬ LÝ, PHÂN TÍCH DỮ LIỆU VÀ ỨNG DỤNG

3.1. Tiền xử lý dữ liệu

Tiền xử lý dữ liệu là bước quan trọng nhằm cải thiện chất lượng dữ liệu trước khi đưa vào mô hình phân tích. Quy trình bao gồm các giai đoạn chính như làm sạch dữ liệu, xử lý giá trị ngoại lai và chuẩn hóa dữ liệu.

3.1.1. Thu thập dữ liệu và làm sạch dữ liệu

- **Nguồn dữ liệu**: Sàn thương mại điện tử, ngân hàng, ví điện tử, v.v.

· **Loại bỏ dữ liệu thiếu**:

- Kiểm tra các giá trị null hoặc NaN.

- Áp dụng các phương pháp như loại bỏ dòng/cột có dữ liệu thiếu hoặc điền giá trị trung bình (mean), trung vị (median) hoặc mô hình dự đoán.

· **Loại bỏ dữ liệu trùng lặp**:

- Phát hiện các dòng dữ liệu trùng lặp.

- Loại bỏ các bản ghi dư thừa để tránh làm sai lệch kết quả phân cụm.

3.1.2. Xử lý giá trị ngoại lai

- Phát hiện ngoại lai:

- Phân phối dữ liệu qua biểu đồ Boxplot, Histogram.

- Sử dụng các phương pháp thống kê như IQR (Interquartile Range) hoặc Z-score.

· **Xử lý ngoại lai**:

- Loại bỏ các điểm dữ liệu bất thường.

- Giới hạn giá trị trong một ngưỡng cụ thể.

3.1.3.Chuẩn hóa và mã hóa dữ liệu

a. Chuẩn hóa dữ liệu

- **Min-Max Scaling**: Áp dụng khi dữ liệu không theo phân phối chuẩn:

- $x' = \max(x) - \min(x) x - \min(x)$

- **Standardization (Z-score normalization)**: Áp dụng khi dữ liệu theo phân phối chuẩn:

$x' = \sigma x - \mu$

b. Mã hóa dữ liệu danh mục (Categorical Encoding)

- **One-hot Encoding**: Biến đổi các thuộc tính dạng phân loại (ví dụ: danh mục sản phẩm) thành các vector nhị phân.
- **Label Encoding**: Áp dụng cho các trường hợp dữ liệu có thứ tự (Ordinal Data).

Mã nguồn python

```
import pandas as pd

from sklearn.preprocessing import MinMaxScaler, StandardScaler, OneHotEncoder



# Dữ liệu mẫu

data = pd.DataFrame({

    'Age': [25, 30, 45, 50],

    'Income': [4000, 5000, 10000, 12000],

    'Category': ['A', 'B', 'A', 'C']
```

```
})


# Chuẩn hóa dữ liệu

scaler = MinMaxScaler()

data[['Age', 'Income']] = scaler.fit_transform(data[['Age', 'Income']])


# Mã hóa One-hot Encoding

encoder = OneHotEncoder(sparse=False)

encoded = encoder.fit_transform(data[['Category']])

encoded_df = pd.DataFrame(encoded, columns=encoder.get_feature_names_out(['Category']))


# Ghép dữ liệu

data = pd.concat([data, encoded_df], axis=1).drop('Category', axis=1)


print(data)
```

3.2. Ứng dụng kết quả phân tích vào thực tiễn

Việc áp dụng thuật toán K-means trong phân tích thói quen chi tiêu của người tiêu dùng mang lại nhiều lợi ích cho doanh nghiệp, đặc biệt trong lĩnh vực thương mại điện tử và dịch vụ tài chính.

1. Dự đoán xu hướng chi tiêu của người tiêu dùng

- Nhận diện mô hình chi tiêu theo các danh mục sản phẩm như thời trang, thực phẩm, hoặc điện tử.

- Dự đoán hành vi mua sắm trong tương lai dựa trên tần suất và giá trị giao dịch trong từng cụm khách hàng.

- Phát hiện các nhóm khách hàng có xu hướng tăng trưởng cao để tối ưu chiến lược đầu tư và mở rộng thị trường.

2. Cá nhân hóa chiến lược marketing cho từng phân khúc khách hàng

- Thiết kế chiến dịch quảng cáo phù hợp với từng cụm khách hàng dựa trên sở thích và hành vi chi tiêu.

- Gửi ưu đãi và khuyến mãi cá nhân hóa dựa trên thói quen mua hàng của từng nhóm. Ví dụ: nhóm khách hàng chi tiêu cao sẽ nhận được các chương trình ưu đãi VIP.

- Tối ưu hóa nội dung và trải nghiệm người dùng trên các nền tảng số, tăng tỷ lệ chuyển đổi và giữ chân khách hàng.

3.Tối ưu hóa chính sách khuyến mãi và dịch vụ chăm sóc khách hàng

- Xây dựng các chương trình giảm giá phù hợp với từng nhóm khách hàng như nhóm săn khuyến mãi, nhóm trung thành, hoặc nhóm mua sắm cao cấp.

- Nâng cao trải nghiệm khách hàng thông qua việc cá nhân hóa dịch vụ chăm sóc, hỗ trợ tư vấn theo sở thích và nhu cầu cụ thể.

- Tối ưu hóa chiến lược giữ chân khách hàng tiềm năng thông qua phân tích hành vi rời bỏ (churn rate) và đưa ra các biện pháp ngăn chặn kịp thời.

3.3.Hạn Chế và Hướng Phát Triển Trong Tương Lai

Hạn chế của mô hình K-means và PCA

- Độ nhạy với dữ liệu ngoại lai: K-means dễ bị ảnh hưởng bởi các điểm dữ liệu ngoại lai, làm sai lệch tâm cụm.

- Phụ thuộc vào việc chọn số cụm (k): Việc xác định số cụm tối ưu đòi hỏi phải sử dụng các phương pháp bổ trợ như Elbow Method hay Silhouette Score, dẫn đến tính chủ quan.

- Giảm chiều dữ liệu bằng PCA có thể làm mất thông tin quan trọng: PCA tập trung vào tối đa hóa phương sai nhưng không đảm bảo giữ lại tất cả thông tin hữu ích cho phân cụm.

- Thiếu khả năng xử lý dữ liệu phi tuyến tính: K-means hoạt động tốt trên dữ liệu tuyến tính và khó khăn khi dữ liệu có cấu trúc phức tạp hoặc dạng cụm không tròn.

2. Hướng phát triển trong tương lai

Kết hợp các thuật toán phân cụm khác:

- Áp dụng các mô hình như DBSCAN hoặc Hierarchical Clustering để xử lý dữ liệu có cấu trúc phi tuyến tính hoặc dữ liệu không đồng nhất.

- Sử dụng thuật toán K-means++ để cải thiện việc khởi tạo tâm cụm.

Tăng cường kỹ thuật tiền xử lý dữ liệu:

- Sử dụng các phương pháp phát hiện và xử lý ngoại lai tiên tiến.

- Áp dụng các kỹ thuật mở rộng dữ liệu (Data Augmentation) để cải thiện tính đa dạng của dữ liệu.

Nâng cao khả năng giải thích kết quả phân cụm:

- Sử dụng các phương pháp trực quan hóa cụm dữ liệu (t-SNE, UMAP) để hiểu rõ hơn về đặc điểm hành vi chi tiêu của từng nhóm khách hàng.

Tích hợp với hệ thống dự báo và đề xuất sản phẩm:

- Áp dụng các mô hình học sâu (Deep Learning) để phát hiện xu hướng chi tiêu và dự đoán nhu cầu sản phẩm theo thời gian thực.

- Phát triển hệ thống gợi ý sản phẩm cá nhân hóa dựa trên kết quả phân cụm.

Mở rộng quy mô và xử lý dữ liệu lớn (Big Data):

- Sử dụng các nền tảng xử lý dữ liệu phân tán như Apache Spark hoặc Hadoop để xử lý dữ liệu lớn từ các sàn thương mại điện tử và hệ thống thanh toán trực tuyến.

3. Kết luận

Việc khắc phục các hạn chế và hướng đến các giải pháp phân tích dữ liệu tiên tiến sẽ giúp mô hình K-means và PCA phát huy tối đa hiệu quả trong việc dự đoán thói quen chi tiêu và tối ưu hóa chiến lược kinh doanh.

KẾT LUẬN

Trong bối cảnh thị trường cạnh tranh và nhu cầu cá nhân hóa trải nghiệm khách hàng ngày càng cao, việc ứng dụng mô hình phân cụm K-means và phân tích thành phần chính (PCA) đã mang lại những kết quả đáng kể trong việc dự đoán thói quen chi tiêu của người tiêu dùng.

Thông qua quá trình thu thập và xử lý dữ liệu, hệ thống đã xác định được các đặc trưng quan trọng trong hành vi mua sắm, từ đó phân nhóm khách hàng theo các đặc điểm chi tiêu khác nhau. Việc áp dụng PCA giúp giảm chiều dữ liệu và tối ưu hóa hiệu suất mô hình K-means, cho phép doanh nghiệp hiểu rõ hơn về các phân khúc khách hàng tiềm năng.

Kết quả phân tích không chỉ hỗ trợ trong việc cá nhân hóa chiến lược marketing mà còn giúp tối ưu hóa các chương trình khuyến mãi và dịch vụ chăm sóc khách hàng. Tuy nhiên, mô hình vẫn gặp một số thách thức như xử lý dữ liệu ngoại lai và phụ thuộc vào việc chọn số cụm tối ưu.

Trong tương lai, việc kết hợp các mô hình phân cụm tiên tiến khác như DBSCAN, áp dụng công nghệ Big Data và hệ thống học sâu (Deep Learning) sẽ là hướng phát triển tiềm năng, giúp doanh nghiệp nâng cao độ chính xác trong dự báo và gia tăng lợi thế cạnh tranh trên thị trường.

Với những kết quả đạt được, nghiên cứu này không chỉ cung cấp cơ sở lý thuyết và thực tiễn vững chắc mà còn mở ra nhiều cơ hội để tiếp tục cải tiến và mở rộng trong các lĩnh vực liên quan.

DANH MỤC TÀI LIỆU THAM KHẢO

[1]. *https://vitaichinh.vn/cong-nghe-ai-trong-tai-chinh-ca-nhan/*