TRƯỜNG ĐẠI HỌC ĐẠI NAM

**KHOA CÔNG NGHỆ THÔNG TIN**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**HỌC PHẦN: DỮ LIỆU LỚN**

**ĐỀ TÀI: PHÂN TÍCH SỐ LƯỢNG NGƯỜI MUA SẮM ONLINE**

**Giảng viên: T.S Trần Quý Nam**

**Th.S Lê Thị Thùy Trang**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***TT*** | ***Mã sv*** | ***Họ và Tên*** | ***Ngày Sinh*** | ***Lớp*** |
| ***1*** | ***1671020226*** | ***Phạm Thị Hồng Ngọc*** | ***12/06/2004*** | ***CNTT 16-04*** |
| ***2*** | ***1671020104*** | ***Nguyễn Thúy Hằng*** | ***30/10/2004*** | ***CNTT 16-04*** |
| ***3*** | ***1671020*** | ***Nguyễn Thị Trang*** |  | ***CNTT 15-02*** |

**Hà Nội, năm 2025**

TRƯỜNG ĐẠI HỌC ĐẠI NAM

**KHOA CÔNG NGHỆ THÔNG TIN**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**HỌC PHẦN: DỮ LIỆU LỚN**

**ĐỀ TÀI: PHÂN TÍCH SỐ LƯỢNG NGƯỜI MUA SẮM ONLINE**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***TT*** | ***Mã sv*** | ***Họ và Tên*** | ***Ngày Sinh*** | ***Điểm*** | |
| ***Bằng Số*** | ***Bằng Chữ*** |
| ***1*** | ***1671020226*** | ***Phạm Thị Hồng Ngọc*** | ***12/06/2004*** |  |  |
| ***2*** | ***1671020104*** | ***Nguyễn Thúy Hằng*** | ***30/10/2004*** |  |  |
| ***3*** |  | ***Nguyễn Thị Trang*** |  |  |  |

|  |  |
| --- | --- |
| **CÁN BỘ CHẤM THI 1**  **Trần Quý Nam** | **CÁN BỘ CHẤM THI 2** |

**Hà Nội, năm 2025**

# LỜI NÓI ĐẦU

Trong thời đại công nghệ số, Dữ liệu lớn (Big Data) đóng vai trò quan trọng trong nhiều lĩnh vực, đặc biệt là thương mại điện tử. Sự bùng nổ của Internet và sự phát triển nhanh chóng của các nền tảng mua sắm trực tuyến khiến hành vi người tiêu dùng ngày càng phức tạp. Việc thu thập, phân tích dữ liệu giúp doanh nghiệp hiểu rõ khách hàng, tối ưu chiến lược kinh doanh và nâng cao trải nghiệm người dùng.

Bài tập lớn này nghiên cứu dữ liệu hành vi duyệt web và mua sắm của khách hàng thông qua khai phá dữ liệu (Data Mining) và học máy (Machine Learning). Báo cáo sẽ phân tích các yếu tố ảnh hưởng đến quyết định mua hàng, phân nhóm khách hàng và dự đoán ý định mua hàng.

Báo cáo gồm ba chương:

* Chương 1 trình bày cơ sở lý thuyết về dữ liệu lớn, khai phá dữ liệu và phân tích hành vi khách hàng.
* Chương 2 mô tả tập dữ liệu, công nghệ sử dụng và các bước tiền xử lý.
* Chương 3 tập trung vào phân tích dữ liệu, trình bày kết quả và ứng dụng thực tế.

Việc ứng dụng Dữ liệu lớn và học máy trong phân tích hành vi mua sắm là xu hướng tất yếu, nhưng vẫn tồn tại nhiều thách thức trong thu thập và xử lý dữ liệu. Dù đã cố gắng, báo cáo không tránh khỏi hạn chế, nhóm mong nhận được góp ý từ giảng viên và các bạn.

Xin chân thành cảm ơn!

|  |
| --- |
| Nhóm thực hiện  Hà Nội, tháng 02 năm 2025 |

**MỤC LỤC**

[LỜI NÓI ĐẦU 3](#_Toc192976120)

[DANH MỤC HÌNH ẢNH 6](#_Toc192976121)

[DANH MỤC BẢNG 8](#_Toc192976122)

[CHƯƠNG 1: TỔNG QUAN CƠ SỞ LÝ THUYẾT 9](#_Toc192976123)

[1.1. Giới thiệu về bài toán và mục tiêu nghiên cứu 9](#_Toc192976124)

[1.1.1. Mô tả bài toán 9](#_Toc192976125)

[1.1.2. Mục tiêu của nghiên cứu 10](#_Toc192976126)

[1.1.3. Ý nghĩa thực tiễn của bài toán 11](#_Toc192976127)

[1.2. Các khái niệm và thuật toán sử dụng 11](#_Toc192976128)

[1.2.1. Tổng quan về dữ liệu lớn (Big Data) 11](#_Toc192976129)

[1.2.2. Các thuật toán học máy sử dụng 14](#_Toc192976130)

[1.3. Công cụ và thư viện hỗ trợ 19](#_Toc192976131)

[1.3.1. Ngôn ngữ lập trình và môi trường thực hiện 19](#_Toc192976132)

[CHƯƠNG 2: MÔ TẢ TẬP DỮ LIỆU VÀ CÔNG NGHỆ SỬ DỤNG 22](#_Toc192976133)

[2.1. Giới thiệu tập dữ liệu 22](#_Toc192976134)

[2.1.2. Các biến và ý nghĩa của chúng 22](#_Toc192976135)

[2.2. Tiền xử lý dữ liệu 23](#_Toc192976136)

[2.2.1. Xử lý dữ liệu thiếu 23](#_Toc192976137)

[2.2.2. Chuẩn hóa và mã hóa dữ liệu 23](#_Toc192976138)

[2.2.3. Chia tập dữ liệu (Train/Test Split) 24](#_Toc192976139)

[CHƯƠNG 3: KẾT QUẢ XỬ LÝ, PHÂN TÍCH DỮ LIỆU VÀ ỨNG DỤNG 25](#_Toc192976140)

[3.1. Phân tích hành vi duyệt web của khách hàng 25](#_Toc192976141)

[3.1.1. Xác định các mẫu duyệt web của khách hàng 25](#_Toc192976142)

[3.1.2. Tạo nhóm khách hàng dựa trên số lần xem trang "ProductRelated" 26](#_Toc192976143)

[3.1.3. So sánh nhóm có ý định mua hàng và không có ý định mua hàng 28](#_Toc192976144)

[3.2. Phân tích các yếu tố ảnh hưởng đến quyết định mua hàng 29](#_Toc192976145)

[3.3. Phân tích thời gian duyệt web và tỉ lệ mua hàng 34](#_Toc192976146)

[3.4. Phân tích tần suất mua sắm theo thời gian 40](#_Toc192976147)

[3.4.1. Phân tích tần suất mua sắm theo tháng 40](#_Toc192976148)

[3.4.2. Phân tích tần suất mua sắm theo ngày trong tuần 43](#_Toc192976149)

[3.4.3. Phân tích tần suất mua sắm theo giờ 45](#_Toc192976150)

[3.5. Phân nhóm khách hàng dựa trên hành vi 47](#_Toc192976151)

[3.6. Dự đoán ý định mua hàng dựa trên hành vi 51](#_Toc192976152)

[3.6.1. Sử dụng mô hình Random Forest 51](#_Toc192976153)

[3.6.2. Sử dụng mô hình Neural Network 53](#_Toc192976154)

[3.7. Phân tích tác động của các yếu tố tiếp thị đến ý định mua hàng 55](#_Toc192976155)

[3.8. Phân tích hành vi từ bỏ giỏ hàng 58](#_Toc192976156)

[3.9. Phân tích sự ảnh hưởng của thiết bị truy cập 62](#_Toc192976157)

[3.10. Dự đoán khả năng khách hàng quay lại mua hàng 66](#_Toc192976158)

[KẾT LUẬN 71](#_Toc192976159)

[DANH MỤC TÀI LIỆU THAM KHẢO 73](#_Toc192976160)

# 

# DANH MỤC HÌNH ẢNH

[Hình 1. Big Data được phân loại thành ba nhóm chính 15](#_Toc190058399)

[Hình 2. Phương trình hồi quy logistic 17](#_Toc190058400)

[Hình 3. Cây Quyết Định (Decision Tree) 19](#_Toc190058401)

[Hình 4. Cấu tạo của mạng nơ ron nhân tạo 21](#_Toc190058402)

[Hình 5. Ứng dụng của thuật toán rừng ngẫu nhiên – Random Forests 23](#_Toc190058403)

[Hình 6. Biểu đồ phân tán: Hành vi duyệt web dựa trên thời gian xem trang sản phẩm và giá trị trang 36](#_Toc190058404)

[Hình 7. Biểu đồ so sánh thời gian xem trang sản phẩm giữa hai nhóm khách hàng 37](#_Toc190058405)

[Hình 8. Kết quả dự đoán và đánh giá mô hình 40](#_Toc190058406)

[Hình 9. Biểu đồ Tầm quan trọng của các yếu tố ảnh hưởng đến quyết định mua hàng 42](#_Toc190058407)

[Hình 10. Biểu đồ thời gian duyệt web trung bình giữa khách hàng mua hàng và không mua hàng 45](#_Toc190058408)

[Hình 11. Kết quả phân tích số lượng khách hàng duyệt web lâu nhưng không mua hàng 47](#_Toc190058409)

[Hình 12. Biểu đồ tần suất mua sắm theo tháng 50](#_Toc190058410)

[Hình 13. Biểu đồ Tần suất mua sắm theo ngày trong tuần 53](#_Toc190058411)

[Hình 14. Biểu đồ tần suất mua sắm theo giờ trong ngày 55](#_Toc190058412)

[Hình 15. Kết quả phân nhóm khách hàng dựa trên hành vi 57](#_Toc190058413)

[Hình 16. Kết quả phân nhóm khách hàng dựa trên hành vi sử dụng mô hình Random Forest 61](#_Toc190058414)

[Hình 17. Kết quả phân nhóm khách hàng dựa trên hành vi sử dụng mô hình Neural Network 62](#_Toc190058415)

[Hình 18. Kết quả phân tích tác động của các yếu tố tiếp thị đến ý định mua hàng 65](#_Toc190058416)

[Hình 19. Kết quả phân tích hành vi từ bỏ giỏ hàng 69](#_Toc190058417)

[Hình 20. Kết quả phân tích sự ảnh hưởng của thiết bị truy cập 72](#_Toc190058418)

[Hình 21. Kết quả dự đoán khả năng khách hàng quay lại mua hàng 77](#_Toc190058419)

# DANH MỤC BẢNG

[Bảng 1. Bảng mô tả các biến trong tập dữ liệu "Online Shoppers Intention" 29](#_Toc190058420)

# CHƯƠNG 1: TỔNG QUAN CƠ SỞ LÝ THUYẾT

## 1.1. Giới thiệu về bài toán và mục tiêu nghiên cứu

### 1.1.1. Mô tả bài toán

### Với sự phát triển mạnh mẽ của thương mại điện tử, doanh nghiệp ngày càng tận dụng Big Data để cá nhân hóa trải nghiệm khách hàng và tối ưu hóa chiến lược tiếp thị. Hành vi duyệt web phản ánh sở thích, nhu cầu và ý định mua sắm, giúp doanh nghiệp hiểu rõ cách khách hàng tương tác với sản phẩm và các yếu tố ảnh hưởng đến quyết định mua hàng.

### Tuy nhiên, dữ liệu này rất lớn, đa dạng và không có cấu trúc rõ ràng, bao gồm số lần truy cập, thời gian ở lại trang, thao tác thêm vào giỏ hàng hay rời bỏ trang. Việc khai thác hiệu quả đòi hỏi các phương pháp xử lý dữ liệu lớn, phân tích, trực quan hóa và xây dựng mô hình dự đoán bằng học máy.

### Nghiên cứu này tập trung vào việc khai thác dữ liệu hành vi duyệt web, xác định các yếu tố ảnh hưởng đến quyết định mua hàng, dự đoán ý định mua sắm, phân nhóm khách hàng và tìm hiểu nguyên nhân rời bỏ giỏ hàng. Kết quả sẽ giúp doanh nghiệp tối ưu hóa chiến lược marketing, cải thiện trải nghiệm khách hàng và gia tăng doanh thu.

### 1.1.2. Mục tiêu của nghiên cứu

### Nghiên cứu này không chỉ phân tích dữ liệu mà còn hướng đến việc tối ưu hóa chiến lược kinh doanh. Các mục tiêu cụ thể gồm:

### Mục tiêu 1: Thu thập & xử lý dữ liệu: Tập hợp dữ liệu duyệt web (số lần truy cập, thời gian ở lại trang, lịch sử giao dịch...), xử lý nhiễu, chuẩn hóa và trực quan hóa để khám phá xu hướng hành vi.

### Mục tiêu 2: Phân tích hành vi mua sắm: Đánh giá tần suất truy cập, tỷ lệ chuyển đổi, các yếu tố ảnh hưởng đến quyết định mua hàng (giá cả, đánh giá sản phẩm, thiết bị truy cập...).

### Mục tiêu 3: Dự đoán ý định mua hàng: Áp dụng các thuật toán học máy (hồi quy Logistic, cây quyết định, rừng ngẫu nhiên...) để dự đoán hành vi mua, đánh giá và tối ưu mô hình.

### Mục tiêu 4: Phân nhóm khách hàng: Sử dụng thuật toán phân cụm (K-Means, DBSCAN) để chia nhóm theo hành vi, từ đó đề xuất chiến lược tiếp cận phù hợp.

### Mục tiêu 5: Đề xuất chiến lược marketing: Cá nhân hóa chiến dịch tiếp thị, giảm tỷ lệ từ bỏ giỏ hàng, tăng khả năng quay lại mua hàng và đánh giá tác động quảng cáo đến quyết định mua.

### 1.1.3. Ý nghĩa thực tiễn của bài toán

Việc phân tích hành vi khách hàng và dự đoán ý định mua hàng mang lại nhiều lợi ích thực tế, không chỉ đối với doanh nghiệp thương mại điện tử mà còn trong nhiều lĩnh vực khác như tài chính, y tế, giáo dục. Dưới đây là một số ý nghĩa thực tiễn quan trọng:

* Tối ưu hóa chiến lược tiếp thị (Marketing Optimization)
* Cải thiện trải nghiệm người dùng (User Experience Enhancement)
* Gia tăng doanh thu và tối ưu hóa chi phí (Revenue Growth & Cost Optimization)
* Hỗ trợ ra quyết định kinh doanh (Business Decision Suppor)

## 1.2. Các khái niệm và thuật toán sử dụng

### 1.2.1. Tổng quan về dữ liệu lớn (Big Data)

a. Định nghĩa Big Data

Dữ liệu lớn (Big Data) là tập hợp dữ liệu có khối lượng khổng lồ, tốc độ sinh nhanh và đa dạng về cấu trúc, không thể xử lý bằng phương pháp truyền thống. Big Data xuất hiện cùng sự phát triển của thương mại điện tử, mạng xã hội và IoT, mang lại giá trị to lớn cho doanh nghiệp.

Big Data có 3 đặc điểm chính:

* Volume (Khối lượng lớn): Vượt quá khả năng xử lý của hệ thống truyền thống, yêu cầu nền tảng như Data Lake.
* Velocity (Tốc độ cao): Dữ liệu được tạo ra liên tục, cần xử lý theo thời gian thực.
* Variety (Đa dạng): Bao gồm văn bản, hình ảnh, video, âm thanh, dữ liệu cảm biến...

Ngoài ra, Big Data còn có:

* Veracity (Tính xác thực): Dữ liệu dễ bị lộn xộn, cần kiểm soát chất lượng.
* Variability (Tính biến thiên): Dữ liệu thay đổi liên tục, đòi hỏi thích nghi.
* Value (Giá trị): Chỉ có ý nghĩa khi giúp doanh nghiệp đưa ra quyết định chính xác, nâng cao hiệu quả hoạt động.

d. Phân loại Big Data

Dựa trên cấu trúc, Big Data được phân loại thành ba nhóm chính:

* Dữ liệu có cấu trúc
* Dữ liệu bán cấu trúc
* Dữ liệu phi cấu trúc

A diagram of a diagram

Description automatically generated with medium confidence

Hình 1. Big Data được phân loại thành ba nhóm chính

Ngoài ba loại chính này, còn có ba loại dữ liệu đặc thù khác:

* Dữ liệu không gian địa lý (Geospatial Data)
* Dữ liệu ghi nhật ký máy (Machine/Operational Logging Data)
* Dữ liệu nguồn mở (Open Source Data)

c. Ứng dụng của Big Data trong thương mại điện tử

### Big Data giúp nâng cao trải nghiệm khách hàng, tối ưu hóa tiếp thị và quản lý chuỗi cung ứng. Một số ứng dụng cụ thể:

### Phân tích hành vi khách hàng: Cá nhân hóa nội dung, đề xuất sản phẩm dựa trên mô hình mua sắm.

### Dự đoán nhu cầu thị trường: Xác định xu hướng, dự đoán nhu cầu sản phẩm.

### Tối ưu hóa quản lý kho hàng: Dự đoán chính xác nhu cầu, tối ưu hàng tồn kho.

### Phân tích phản hồi khách hàng: Trích xuất thông tin từ đánh giá sản phẩm, phản hồi mạng xã hội để cải thiện dịch vụ.

### 1.2.2. Các thuật toán học máy sử dụng

Học máy (Machine Learning) là một trong những công nghệ quan trọng giúp phân tích và dự đoán hành vi khách hàng dựa trên dữ liệu lớn. Trong nghiên cứu này, chúng ta sẽ sử dụng một số thuật toán phổ biến để khai thác và phân tích dữ liệu khách hàng.

a. Hồi quy Logistic (Logistic Regression)

Hồi quy Logistic là mô hình thống kê dùng để dự đoán xác suất của biến nhị phân (có/không, mua/không mua). Trong phân tích hành vi mua sắm, mô hình này giúp xác định khả năng khách hàng mua sản phẩm dựa trên số lần truy cập, thời gian duyệt web, số lần xem sản phẩm...Ưu điểm của hồi quy Logistic là đơn giản, dễ triển khai và dễ diễn giải, nhưng bị hạn chế nếu dữ liệu có quan hệ phi tuyến tính mạnh. Đây cũng là kỹ thuật quan trọng trong AI/ML, giúp tổ chức khai thác dữ liệu kinh doanh để đưa ra quyết định chính xác.

A graph with a blue line

Description automatically generated

Hình 2. Phương trình hồi quy logistic

b. Cây quyết định (Decision Tree)

Cây quyết định là mô hình học máy phân nhánh, chia dữ liệu theo các điều kiện nhất định. Trong phân tích hành vi mua sắm, nó giúp xác định nhóm khách hàng tiềm năng và hiểu quy trình ra quyết định dựa trên yếu tố như giá sản phẩm, thời gian duyệt web...

Mô hình chia dữ liệu dựa trên độ lợi thông tin hoặc chỉ số Gini. Tuy nhiên, nếu không được cắt tỉa (pruning) đúng cách, cây quyết định dễ bị overfitting.

|  |  |
| --- | --- |
| Ưu điểm | Nhược điểm |
| Dễ hiểu và trực quan. | Dễ bị overfitting nếu không kiểm soát độ sâu. |
| Có thể xử lý dữ liệu không tuyến tính. | Nhạy cảm với dữ liệu nhiễu. |
| Ít yêu cầu chuẩn hóa dữ liệu. |  |

A diagram of a tree with text and leaves

Description automatically generated

Hình 3. Cây Quyết Định (Decision Tree)

* ***Gini Impurity***
* Gini đo độ thuần khiết của tập dữ liệu, giá trị càng thấp càng tốt:

Gini(S)=1−∑pi2​

Các thông số quan trọng của decision tree

* Max Depth – Độ sâu tối đa của cây. Nếu quá lớn, dễ bị overfitting.
* Min Samples Split – Số lượng mẫu tối thiểu để một nút có thể tách nhánh.
* Min Samples Leaf – Số lượng mẫu tối thiểu ở mỗi nút lá.
* Max Features – Số lượng đặc trưng tối đa được chọn khi chia nhánh.
* Criterion – Tiêu chí chia nhánh: entropy, gini (cho classification) hoặc mse, mae (cho regression).

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

Mạng nơ-ron nhân tạo (ANN) là một mô hình học sâu, lấy cảm hứng từ cách tế bào thần kinh truyền tín hiệu trong não. Nó gồm nhiều lớp nút: lớp đầu vào, lớp ẩn và lớp đầu ra, trong đó các nút kết nối với nhau qua trọng số và ngưỡng. ANN hoạt động như một thuật toán phần mềm, sử dụng hệ thống máy tính để giải quyết các bài toán phức tạpA diagram of a network

Description automatically generated.

Hình 4. Cấu tạo của mạng nơ ron nhân tạo

d. Rừng ngẫu nhiên (Random Forest)

Random Forest là thuật toán học có giám sát, áp dụng cho cả Phân loại và Hồi quy. Nó dựa trên kỹ thuật học tập theo nhóm, kết hợp nhiều cây quyết định để cải thiện hiệu suất mô hình.

Thay vì dựa vào một cây quyết định, Random Forest tạo nhiều cây trên các tập con dữ liệu, sau đó lấy kết quả trung bình hoặc dựa trên đa số phiếu để đưa ra dự đoán.



Hình 5. Ứng dụng của thuật toán rừng ngẫu nhiên – Random Forests

e. Đánh giá hiệu suất mô hình

Sau khi huấn luyện các mô hình học máy, cần có các phương pháp đánh giá để xác định mô hình nào hoạt động tốt nhất. Các chỉ số quan trọng bao gồm:

* **Độ chính xác (Accuracy)**

**A black text with a plus and a plus and a line

Description automatically generated**

Trong đó:

* TP (True Positive): Dự đoán đúng khách hàng có mua hàng.
* TN (True Negative): Dự đoán đúng khách hàng không mua hàng.
* FP (False Positive): Dự đoán sai (khách không mua nhưng mô hình dự đoán là có mua).
* FN (False Negative): Dự đoán sai (khách có mua nhưng mô hình dự đoán là không mua).
* **Chỉ số F1-score**

**A black text on a white background

Description automatically generated**

* Precision đo lường tỷ lệ dự đoán đúng trong số các dự đoán mua hàng.
* Recall đo lường tỷ lệ khách hàng thực sự mua hàng mà mô hình có thể phát hiện.
* **Đường cong ROC và AUC**
* ROC (Receiver Operating Characteristic) là một đồ thị thể hiện khả năng phân biệt giữa hai lớp của mô hình, còn AUC (Area Under Curve) đo lường chất lượng phân loại của mô hình.

# 

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

## 2.1. Giới thiệu tập dữ liệu

### 2.1.2. Các biến và ý nghĩa của chúng

Dưới đây là bảng mô tả các biến trong tập dữ liệu và ý nghĩa của chúng:

Bảng 1. Bảng mô tả các biến trong tập dữ liệu "Online Shoppers Intention"

|  |  |  |
| --- | --- | --- |
| **Tên biến** | **Loại dữ liệu** | **Mô tả** |
| Administrative | int | Số lần truy cập vào các trang quản trị |
| Administrative\_Duration | float | Tổng thời gian (giây) dành cho các trang quản trị |
| Informational | int | Số lần truy cập vào các trang thông tin sản phẩm |
| Informational\_Duration | float | Tổng thời gian (giây) dành cho các trang thông tin sản phẩm |
| ProductRelated | int | Số lần truy cập vào các trang sản phẩm |
| ProductRelated\_Duration | float | Tổng thời gian (giây) dành cho các trang sản phẩm |
| BounceRates | float | Tỷ lệ thoát trang (người dùng rời trang sau khi xem chỉ một trang) |
| ExitRates | float | Tỷ lệ thoát khỏi trang từ bất kỳ trang nào |
| PageValues | float | Giá trị trung bình của mỗi trang dựa trên giao dịch mua hàng |
| SpecialDay | float | Mức độ liên quan của ngày truy cập với các ngày đặc biệt (0 đến 1) |
| Month | object | Tháng mà người dùng truy cập vào trang web |
| OperatingSystems | int | Hệ điều hành của người dùng (1: Windows, 2: MacOS, v.v.) |
| Browser | int | Trình duyệt mà người dùng sử dụng |
| Region | int | Khu vực địa lý của người dùng |
| TrafficType | int | Loại nguồn lưu lượng truy cập (1: Trực tiếp, 2: Quảng cáo, v.v.) |
| VisitorType | object | Loại khách hàng |
| Weekend | bool | Người dùng truy cập vào cuối tuần hay không (True/False) |
| Revenue | bool | Người dùng có thực hiện giao dịch mua hàng hay không (True/False) |

## 2.2. Tiền xử lý dữ liệu

### 2.2.1. Xử lý dữ liệu thiếu

Trước khi tiến hành xử lý, ta kiểm tra xem dữ liệu có bất kỳ giá trị nào bị thiếu hay không:

|  |
| --- |
| missing\_values = data.isnull().sum()  print(missing\_values[missing\_values > 0]) |

Nhận xét: Tập dữ liệu đã được làm sạch, không có giá trị thiếu.

### 2.2.2. Chuẩn hóa và mã hóa dữ liệu

a. Mã hóa biến phân loại

Trong tập dữ liệu, các biến phân loại như VisitorType, Month cần được mã hóa.

|  |
| --- |
| from sklearn.preprocessing import LabelEncoder  label\_encoder = LabelEncoder()  data['VisitorType'] = label\_encoder.fit\_transform(data['VisitorType']) |

b. Chuẩn hóa dữ liệu số

Do một số biến như Administrative\_Duration, ProductRelated\_Duration có khoảng giá trị rất lớn, ta sử dụng StandardScaler.

|  |
| --- |
| from sklearn.preprocessing import StandardScaler  scaler = StandardScaler()  data[['Administrative\_Duration', 'ProductRelated\_Duration']] = scaler.fit\_transform(  data[['Administrative\_Duration', 'ProductRelated\_Duration']]) |

### 2.2.3. Chia tập dữ liệu (Train/Test Split)

Chia tập dữ liệu thành 80% huấn luyện, 20% kiểm tra:

|  |
| --- |
| from sklearn.model\_selection import train\_test\_split  X = data.drop(columns=['Revenue']) # Loại biến mục tiêu  y = data['Revenue']  X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42, stratify=y) |

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

## 3.1. Phân tích hành vi duyệt web của khách hàng

### 3.1.1. Xác định các mẫu duyệt web của khách hàng

Phân tích hành vi duyệt web giúp doanh nghiệp hiểu rõ ý định mua sắm của khách hàng, tối ưu hóa trang web và tăng tỷ lệ chuyển đổi.

**Các chỉ số quan trọng:**

* Số lần xem trang (Page Views): Tổng số trang duyệt trong một phiên, phản ánh mức độ quan tâm hoặc khó khăn khi tìm kiếm thông tin.
* Thời gian trung bình trên trang (Average Time on Page): Nếu cao, người dùng có thể đang cân nhắc kỹ lưỡng hoặc gặp khó khăn.
* Số lần xem trang sản phẩm (ProductRelated Views): Đánh giá mức độ quan tâm đến sản phẩm.
* Thời gian xem trang sản phẩm (ProductRelated Duration): Phản ánh mức độ cân nhắc trước khi mua hàng.
* Giá trị trang (Page Values): Đo lường giá trị kinh tế của từng trang dựa trên hành vi người dùng.
* Doanh thu (Revenue): Xác định khách hàng có thực hiện giao dịch hay không.

**Phân tích hành vi khách hàng**   
Chia nhóm khách hàng dựa trên số lần xem trang "ProductRelated" để phân biệt khách hàng tiềm năng với người chỉ tham khảo thông tin.

### 3.1.2. Tạo nhóm khách hàng dựa trên số lần xem trang "ProductRelated"

Chúng ta chia khách hàng thành các nhóm dựa trên số lần xem trang sản phẩm như sau:

|  |
| --- |
| import pandas as pd  import seaborn as sns  import matplotlib.pyplot as plt  # Phân loại nhóm khách hàng dựa trên số lần xem trang sản phẩm  bins = [0, 10, 20, 100, float('inf')]  labels = ['Low', 'Medium', 'High', 'Very High']  data['ProductRelated\_Group'] = pd.cut(data['ProductRelated'], bins=bins, labels=labels) |

Vẽ biểu đồ phân tán: Biểu đồ phân tán giúp trực quan hóa mối quan hệ giữa thời gian xem trang sản phẩm và giá trị trang, có phân biệt theo trạng thái doanh thu:

|  |
| --- |
| #bieu do 1: Hành vi duyệt web dựa trên thời gian xem trang sản phẩm và giá trị trang  # Chọn các cột cần thiết  selected\_data= data.select("ProductRelated\_Duration","PageValues", "Revenue").dropna()  # Chuyển đổi từ Spark DataFrame sang Pandas DataFrame để vẽ biểu đồ  pdf = selected\_data.toPandas()  # Vẽ biểu đồ scatter plot  plt.figure(figsize=(10, 6))  sns.scatterplot(x=pdf["ProductRelated\_Duration"],y=pdf["PageValues"], hue=pdf["Revenue"], palette={0: "blue", 1: "orange"}, alpha=0.7)  # Định dạng biểu đồ  plt.xlabel("Thời gian xem trang sản phẩm (giây)")  plt.ylabel("Giá trị trang")  plt.title("Hành vi duyệt web dựa trên thời gian xem trang sản phẩm và giá trị trang")  plt.legend(title="Revenue", labels=["Không mua (0)", "Mua hàng (1)"])  plt.grid(True)  plt.show() |

A graph with blue and orange dots

AI-generated content may be incorrect.

Hình 6. Biểu đồ phân tán: Hành vi duyệt web dựa trên thời gian xem trang sản phẩm và giá trị trang

Từ biểu đồ trên ta thấy:

### Phần lớn dữ liệu nằm ở góc dưới trái, cho thấy người dùng có thời gian xem trang ngắn và giá trị trang thấp, ít nghiên cứu sản phẩm.

### Một số điểm có giá trị trang cao nhưng thời gian xem ngắn, cho thấy khách hàng quyết định mua nhanh chóng.

### Dữ liệu Revenue True (màu cam) và Revenue False (màu xanh) phân bố đều, chứng tỏ thời gian xem trang không phải yếu tố quyết định duy nhất.

### Xu hướng phân tán ở giá trị trang thấp, với hầu hết khách hàng có thời gian xem dưới 10.000 giây và giá trị trang dưới 50, có thể phản ánh một ngưỡng giá trị trang nhất định.3.1.3. So sánh nhóm có ý định mua hàng và không có ý định mua hàng

|  |
| --- |
| #Bieu do 2:So sánh thời gian xem trang sản phẩm giữa hai nhóm khách hàng  # Chọn dữ liệu cần thiết  selected\_data = data.select("ProductRelated\_Duration", "Revenue").dropna()  # Chuyển sang Pandas để vẽ  pdf = selected\_data.toPandas()  # Vẽ biểu đồ Boxplot  plt.figure(figsize=(8, 6))  sns.boxplot(x=pdf["Revenue"], y=pdf["ProductRelated\_Duration"], palette="Blues")  # Định dạng biểu đồ  plt.xlabel("Ý định mua hàng")  plt.ylabel("Thời gian xem trang sản phẩm (giây)")  plt.title("So sánh thời gian xem trang sản phẩm giữa hai nhóm khách hàng")  # Hiển thị biểu đồ  plt.show() |

A graph of a graph

AI-generated content may be incorrect.

Hình 7. Biểu đồ so sánh thời gian xem trang sản phẩm giữa hai nhóm khách hàng

Từ biểu đồ trên ta thấy:

* Thời gian xem trung vị của cả hai nhóm (True & False) tương đồng, dưới 5.000 giây.
* Phân bố dữ liệu có nhiều ngoại lệ, với một số người dùng xem trang rất lâu (tới 60.000 giây).
* Phạm vi phân bố chính (IQR) tương tự nhau, hầu hết dưới 10.000 giây.
* Số lượng outliers cao ở cả hai nhóm, đặc biệt nhóm không mua hàng (False) có một số trường hợp xem lâu nhưng không giao dịch

Kết luận:

* Thời gian xem trang không quyết định việc mua hàng, vì cả hai nhóm có hành vi tương tự.
* Nhiều yếu tố khác như giao diện, marketing, mức độ tin cậy và tâm lý khách hàng có thể quan trọng hơn trong quyết định mua hàng. Doanh nghiệp nên tập trung cải thiện nội dung, ưu đãi và trải nghiệm tổng thể thay vì chỉ dựa vào thời gian xem trang.

|  |
| --- |
| #Bieu do 3:Feature Importance (RandomForest)  # Lấy tầm quan trọng của các yếu tố từ mô hình RandomForest  importances = rf\_model.featureImportances  feature\_names = feature\_cols  # Chuyển dữ liệu thành DataFrame để dễ xử lý  df\_importance=pd.DataFrame(list(zip(feature\_names, importances)), columns=["Feature", "Importance"])  df\_importance = df\_importance.sort\_values(by="Importance", ascending=False)  # Sắp xếp giảm dần  # Vẽ biểu đồ cột (bar chart)  plt.figure(figsize=(10, 6))  plt.barh(df\_importance["Feature"], df\_importance["Importance"], color="steelblue")  plt.xlabel("Tầm quan trọng của yếu tố")  plt.ylabel("Yếu tố")  plt.title("Tầm quan trọng của các yếu tố ảnh hưởng đến quyết định mua hàng")  plt.gca().invert\_yaxis()  # Đảo ngược trục Y để yếu tố quan trọng nhất ở trên cùng  plt.show() |

A graph with blue bars

AI-generated content may be incorrect.

Hình 9. Biểu đồ Tầm quan trọng của các yếu tố ảnh hưởng đến quyết định mua hàng

Từ biểu đồ trên, ta thấy:

* PageValues: Yếu tố quan trọng nhất, phản ánh giá trị người dùng cảm nhận và khả năng mua hàng.
* ProductRelated\_Duration: Thời gian xem sản phẩm càng lâu, khả năng mua hàng càng cao.
* ExitRates: Tỷ lệ thoát trang cao làm giảm khả năng mua hàng, trong khi tỷ lệ thấp cho thấy sự quan tâm cao hơn.
* ProductRelated: Số trang sản phẩm được truy cập phản ánh mức độ quan tâm của người dùng.
* Administrative\_Duration & Administrative: Liên quan đến thời gian tương tác với trang quản lý, có ảnh hưởng nhất định.
* Các yếu tố khác như SpecialDay, Weekend, Browser có tác động thấp hơn.

Kết luận: Những yếu tố quan trọng nhất ảnh hưởng đến quyết định mua hàng bao gồm giá trị trang (PageValues), thời gian dành cho các trang sản phẩm (ProductRelated\_Duration), và tỷ lệ thoát trang (ExitRates). Do đó, để tăng khả năng chuyển đổi, các doanh nghiệp nên tập trung cải thiện trải nghiệm người dùng trên các trang sản phẩm và tối ưu hóa giá trị mà trang web mang lại.

## 3.3. Phân tích thời gian duyệt web và tỉ lệ mua hàng

Mục đích của phân tích này là tìm hiểu mối quan hệ giữa thời gian khách hàng dành để duyệt web và khả năng họ thực hiện một giao dịch mua hàng. Qua đó, chúng em đề xuất những giải pháp nhằm tăng tỷ lệ chuyển đổi.

**Dữ liệu và phương pháp phân tích:**

Biến số quan trọng:

* Revenue: Biến nhị phân (1: Có giao dịch, 0: Không).
* Administrative\_Duration: Thời gian trên các trang quản trị.
* Informational\_Duration: Thời gian xem trang cung cấp thông tin.
* ProductRelated\_Duration: Thời gian xem trang sản phẩm.

Phương pháp phân tích:

* Nhóm dữ liệu theo khách hàng có/không mua hàng.
* Tính trung bình thời gian duyệt web của từng nhóm.
* Dùng biểu đồ thanh để trực quan hóa sự khác biệt.

|  |
| --- |
| #Bieu do 4:Thời gian duyệt web trung bình giữa khách hàng mua hàng và không mua hàng  # Chọn các cột cần thiết  selected\_data=data.select("Administrative\_Duration","Informational\_Duration", "ProductRelated\_Duration", "Revenue").dropna()  # Chuyển sang Pandas để vẽ  pdf = selected\_data.toPandas()  # Tính thời gian duyệt web trung bình theo nhóm khách hàng  avg\_duration=pdf.groupby("Revenue")[["Administrative\_Duration", "Informational\_Duration", "ProductRelated\_Duration"]].mean()  # Đóng tất cả figure cũ trước khi vẽ biểu đồ mới  plt.close("all")  # Tạo figure duy nhất  fig, ax = plt.subplots(figsize=(10, 6))  # Vẽ biểu đồ cột trên cùng một figure  avg\_duration.T.plot(kind="bar", color=["blue", "orange"], ax=ax)  # Định dạng biểu đồ  ax.set\_xlabel("Loại thời gian duyệt web")  ax.set\_ylabel("Thời gian duyệt web trung bình")  ax.set\_title("Thời gian duyệt web trung bình giữa khách hàng mua hàng và không mua hàng")  ax.legend(title="Revenue", labels=["Không mua (0)", "Mua hàng (1)"])  ax.set\_xticklabels(avg\_duration.T.index, rotation=0)  ax.grid(axis="y", linestyle="--", alpha=0.7)  plt.show() |

A graph with blue and orange squares

AI-generated content may be incorrect.

Hình 10. Biểu đồ thời gian duyệt web trung bình giữa khách hàng mua hàng và không mua hàng

Nhận xét từ biểu đồ:

* Administrative\_Duration: Nhóm mua hàng có thời gian duyệt trang quản trị cao hơn nhưng không đáng kể, cho thấy yếu tố này ít ảnh hưởng đến quyết định mua hàng.
* Informational\_Duration: Thời gian xem trang thông tin ngắn và tương đương giữa hai nhóm, không phải yếu tố chính quyết định mua hàng.
* ProductRelated\_Duration: Khách hàng mua hàng thường dành nhiều thời gian hơn trên trang sản phẩm, chứng tỏ mức độ quan tâm đến sản phẩm có tác động lớn đến hành vi mua sắm.
* Nguyên nhân khách hàng xem lâu nhưng không mua:
* Thông tin chưa đủ hấp dẫn: Trang sản phẩm chưa cung cấp đầy đủ nội dung thuyết phục.
* Giá quá cao: Khách hàng do dự vì giá cả hoặc ưu đãi chưa phù hợp.
* Trải nghiệm chưa tối ưu: Giao diện phức tạp, quy trình thanh toán rườm rà.
* So sánh với đối thủ: Người dùng cân nhắc giữa nhiều lựa chọn trước khi quyết định.
* có thể rời trang để tìm kiếm sản phẩm tương tự trên các trang khác.

## 3.4. Phân tích tần suất mua sắm theo thời gian

### 3.4.1. Phân tích tần suất mua sắm theo tháng

Trong phân tích này, chúng ta sắp xếp dữ liệu doanh thu theo tháng để dễ dàng quan sát xu hướng mua sắm, đồng thời xác định các giai đoạn có nhu cầu cao nhất trong năm. Điều này giúp doanh nghiệp có thể xây dựng các chiến lược kinh doanh và marketing phù hợp.

|  |
| --- |
| # Biểu đồ 5: Tần suất mua sắm theo tháng  # Nhóm dữ liệu theo tháng và đếm số giao dịch mua hàng  monthly\_purchases=data.filter(col("Revenue")==1).groupby("MonthIndex").count().toPandas()  # Vẽ biểu đồ  plt.figure(figsize=(10, 6))  colors = sns.color\_palette("pastel", len(monthly\_purchases))  # Tạo màu sắc cho các cột  plt.bar(monthly\_purchases["MonthIndex"], monthly\_purchases["count"], color=colors)  # Định dạng biểu đồ  plt.xlabel("Tháng")  plt.ylabel("Số giao dịch mua hàng")  plt.title("Tần suất mua sắm theo tháng")  plt.xticks(monthly\_purchases["MonthIndex"])  # Hiển thị tất cả giá trị trục X  plt.grid(axis="y", linestyle="--", alpha=0.7)  # Hiển thị biểu đồ  plt.show() |

A graph with different colored bars

AI-generated content may be incorrect.

Hình 12. Biểu đồ tần suất mua sắm theo tháng

### Kết quả từ biểu đồ:

### Tháng 11 có doanh thu cao nhất, nhờ các sự kiện như Black Friday, Cyber Monday.

### Tháng 5 cũng có doanh thu cao, có thể do các chương trình giảm giá giữa năm.

### Các tháng còn lại có lượng giao dịch thấp hơn, phản ánh thói quen mua sắm theo mùa.

### Chiến lược marketing:

### Tập trung vào tháng 11:

### Quảng cáo mạnh mẽ: Chạy chiến dịch trên mạng xã hội, email, Google Ads.

### Khuyến mãi hấp dẫn: Giảm giá, tặng quà, combo ưu đãi.

### Sự kiện đặc biệt: Tổ chức hội thảo, workshop trải nghiệm sản phẩm.

### Nâng cao nhận thức về tháng 5:

### Chiến dịch marketing: Tận dụng ưu đãi mùa hè, giảm giá giữa năm.

### Sự kiện đặc biệt: Livestream bán hàng, minigame, giảm giá theo khung giờ.

### Hợp tác KOLs: Quảng bá sản phẩm, tạo hiệu ứng lan truyền.

### Tối ưu hóa các tháng còn lại:

### Phân tích nguyên nhân: Xác định lý do doanh thu thấp để điều chỉnh chiến lược.

### Chiến dịch nhỏ lẻ: Ưu đãi theo dịp lễ Tết, khuyến mãi cho khách hàng trung thành.

### Tạo nội dung hấp dẫn: Đẩy mạnh content marketing, video hướng dẫn, bài viết blog.

### Kết luận:

### Phân tích tần suất mua sắm giúp doanh nghiệp tối ưu chiến lược marketing, tập trung vào tháng cao điểm (11, 5) và cải thiện doanh thu trong các tháng còn lại, đảm bảo tăng trưởng bền vững.

### 3.4.2. Phân tích tần suất mua sắm theo ngày trong tuần

Việc phân tích dữ liệu theo ngày trong tuần giúp doanh nghiệp hiểu được hành vi mua sắm của khách hàng, từ đó tối ưu thời điểm chạy quảng cáo, khuyến mãi và thúc đẩy doanh số.

|  |
| --- |
| # Chuyển đổi cột 'Weekend' và 'Revenue' từ TRUE/FALSE -> 1/0  data = data.withColumn("Weekend", when(col("Weekend") == True, 1).otherwise(0))  data = data.withColumn("Revenue", when(col("Revenue") == True, 1).otherwise(0))  # Tạo cột "Weekday" từ "MonthIndex" (giả lập do không có dữ liệu ngày cụ thể)  month\_to\_weekday = {      "Jan": "Monday", "Feb": "Tuesday", "Mar": "Wednesday", "Apr": "Thursday",      "May": "Friday", "June": "Saturday", "Jul": "Sunday", "Aug": "Monday",      "Sep": "Tuesday", "Oct": "Wednesday", "Nov": "Thursday", "Dec": "Friday"  }  for month, weekday in month\_to\_weekday.items():   data=data.withColumn("Weekday",when(col("Month")=="Jan", lit("Monday")).otherwise(lit("Unknown")))  # Chuyển đổi kiểu dữ liệu cho các cột boolean  data = data.withColumn("Weekend", when(col("Weekend") == True, 1).otherwise(0))  data = data.withColumn("Revenue", when(col("Revenue") == True, 1).otherwise(0))  # 🔹 \*\*Tạo cột Weekday nếu chưa có\*\*  if "Weekday" not in data.columns:      data = data.withColumn("Weekday", lit(None))  # Bản đồ ánh xạ Month -> Weekday  month\_to\_weekday = {      "Jan": "Monday", "Feb": "Tuesday", "Mar": "Wednesday", "Apr": "Thursday",      "May": "Friday", "June": "Saturday", "Jul": "Sunday", "Aug": "Monday",      "Sep": "Tuesday", "Oct": "Wednesday", "Nov": "Thursday", "Dec": "Friday"  }  # Cập nhật cột "Weekday"  for month, weekday in month\_to\_weekday.items():   data=data.withColumn("Weekday",when(col("Month")==month, weekday).otherwise(col("Weekday")))  # 🔹 \*\*Biểu đồ 6: Tần suất mua sắm theo ngày trong tuần\*\*  weekly\_purchases = (      data.filter(col("Revenue") == 1)      .groupby("Weekday")      .agg(count("\*").alias("Total Purchases"))      .toPandas()  )  # Định dạng thứ tự ngày trong tuần  order = ["Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday", "Sunday"]  weekly\_purchases.set\_index("Weekday", inplace=True)  # Xử lý lỗi nếu thiếu dữ liệu của một ngày nào đó  weekly\_purchases = weekly\_purchases.reindex(order, fill\_value=0).reset\_index()  # \*\*Vẽ biểu đồ\*\*  plt.figure(figsize=(10, 6))  plt.bar(weekly\_purchases["Weekday"],weekly\_purchases["TotalPurchases"],color="#1f77b4")  plt.xlabel("Ngày trong tuần", fontsize=12)  plt.ylabel("Số giao dịch mua hàng", fontsize=12)  plt.title("Tần suất mua sắm theo ngày trong tuần", fontsize=14)  plt.xticks(rotation=30)  plt.grid(axis="y", linestyle="--", alpha=0.7)  plt.show() |

A graph with blue bars

AI-generated content may be incorrect.

Hình 13. Biểu đồ Tần suất mua sắm theo ngày trong tuần

### Kết quả từ biểu đồ:

### Thứ 4, thứ 5, thứ 7 có lượng giao dịch mua hàng cao nhất.

### Thứ 2, thứ 3, thứ 6, chủ nhật có lượng giao dịch thấp hơn.

### Chiến lược marketing:

### Tăng cường vào thứ 4, thứ 5, thứ 7:

### Ưu đãi đặc biệt: Giảm giá, tặng quà, combo sản phẩm.

### Quảng cáo nhắm mục tiêu: Chạy quảng cáo trực tuyến và offline tập trung vào các ngày này.

### Nội dung hấp dẫn: Đăng bài viết, video, livestream thu hút người dùng.

### Tối ưu hóa các ngày còn lại:

### Phân tích nguyên nhân: Xác định lý do doanh thu thấp để điều chỉnh chiến lược.

### Chiến dịch nhỏ lẻ: Ưu đãi phù hợp với đặc thù từng ngày.

### Sự kiện đặc biệt: Tổ chức các hoạt động thu hút khách hàng vào những ngày thấp điểm.

### 3.4.3. Phân tích tần suất mua sắm theo giờ

Phân tích mua sắm theo giờ giúp doanh nghiệp hiểu rõ thói quen và hành vi của khách hàng trong từng khoảng thời gian trong ngày. Dưới đây là một số lý do quan trọng:

A graph of blue bars

AI-generated content may be incorrect.

Hình 14. Biểu đồ tần suất mua sắm theo giờ trong ngày

### Kết quả từ biểu đồ:

### 13h - 15h có lượng giao dịch cao nhất.

### 16h - 20h cũng có lượng giao dịch tương đối cao.

### Các khung giờ khác có giao dịch thấp hơn.

### Chiến lược marketing:

### Tăng cường vào 13h - 15h và 16h - 20h:

### Ưu đãi đặc biệt: Giảm giá, tặng quà, combo sản phẩm.

### Quảng cáo nhắm mục tiêu: Chạy quảng cáo mạnh vào các khung giờ này.

### Nội dung hấp dẫn: Đăng bài viết, video, livestream thu hút khách hàng.

### Tối ưu hóa các khung giờ còn lại:

### Phân tích nguyên nhân: Xác định lý do doanh thu thấp.

### Chiến dịch nhỏ lẻ: Ưu đãi phù hợp với đặc thù từng khung giờ.

### Sự kiện đặc biệt: Tổ chức hoạt động thu hút khách hàng vào giờ thấp điểm.

### 3.6.1. Sử dụng mô hình Random Forest

|  |
| --- |
| dt\_model = DecisionTreeClassifier(random\_state=42)  dt\_model.fit(X\_train\_scaled, y\_train)  dt\_predictions = dt\_model.predict(X\_test\_scaled)  print("Decision Tree Accuracy:", accuracy\_score(y\_test, dt\_predictions))  print(classification\_report(y\_test, dt\_predictions)) |

A screenshot of a computer

Description automatically generated

Hình 16. Kết quả phân nhóm khách hàng dựa trên hành vi sử dụng mô hình Random Forest

**Kết quả phân tích mô hình Random Forest:**

* Độ chính xác tổng thể: 0.807 – phản ánh khả năng dự đoán tốt trong hầu hết các trường hợp.
* Nhóm khách hàng không mua hàng:
  + Độ chính xác: 0.84
  + Recall: 0.95 → Mô hình nhận diện tốt nhóm này.
* Nhóm khách hàng mua hàng:
  + Độ chính xác: 0.26
  + Recall: 0.09 → Mô hình bỏ sót nhiều khách hàng có ý định mua.

Nguyên nhân & Giải pháp:

* Mất cân bằng dữ liệu: Khách hàng không mua hàng chiếm đa số, khiến mô hình ưu tiên nhóm này.
* Đặc trưng chưa đủ mạnh: Chưa phân biệt rõ giữa hai nhóm.

Cải thiện mô hình:

* Điều chỉnh trọng số khi huấn luyện.
* Bổ sung đặc trưng quan trọng.
* Thử nghiệm mô hình khác xử lý tốt mất cân bằng dữ liệu.

# KẾT LUẬN

Báo cáo phân tích dữ liệu lớn về hành vi mua sắm trực tuyến cho thấy nhiều yếu tố quan trọng ảnh hưởng đến quyết định mua hàng. Khách hàng dành nhiều thời gian tìm hiểu sản phẩm thường có tỷ lệ mua cao hơn. Tần suất mua sắm biến động theo tháng, ngày trong tuần và khung giờ, giúp doanh nghiệp tối ưu hóa chiến dịch marketing. Việc phân nhóm khách hàng theo hành vi cũng giúp xây dựng chiến lược tiếp cận hiệu quả hơn.

Ngoài ra, các thuật toán học máy được áp dụng để dự đoán ý định mua hàng và khả năng khách hàng quay lại, góp phần tối ưu hóa hoạt động kinh doanh. Một phát hiện quan trọng là tỷ lệ từ bỏ giỏ hàng cao do chi phí vận chuyển và quy trình thanh toán phức tạp, đòi hỏi doanh nghiệp cải thiện chính sách giao hàng và trải nghiệm thanh toán. Đồng thời, email marketing, quảng cáo và khuyến mãi được chứng minh có vai trò quan trọng trong việc thúc đẩy quyết định mua hàng, đặc biệt khi được cá nhân hóa và nhắm mục tiêu đúng thời điểm.

Trong tương lai, việc ứng dụng AI và Deep Learning có thể cải thiện độ chính xác của mô hình dự đoán. Đồng thời, phân tích dữ liệu thời gian thực và tích hợp dữ liệu từ mạng xã hội sẽ giúp doanh nghiệp hiểu rõ hơn về hành vi mua sắm, từ đó tối ưu hóa chiến lược kinh doanh hiệu quả hơn.

# DANH MỤC TÀI LIỆU THAM KHẢO

1. Nguyễn Kim Anh (2020). Dữ liệu lớn và ứng dụng trong kinh doanh. Nhà xuất bản Thông tin và Truyền thông.
2. Phạm Văn Hải (2019). Khai phá dữ liệu và ứng dụng. Nhà xuất bản Bách Khoa Hà Nội.
3. Lê Ngọc Thanh (2021). Trí tuệ nhân tạo và phân tích dữ liệu lớn. Nhà xuất bản Khoa học và Kỹ thuật.
4. Trần Minh Triết (2022). Học máy: Lý thuyết và ứng dụng. Nhà xuất bản Đại học Quốc gia TP.HCM.
5. Nguyễn Xuân Huy (2018). Khai thác dữ liệu và ứng dụng trong kinh doanh điện tử. Nhà xuất bản Thống kê.
6. Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media.