Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á
Năm thứ 36, Số 11 (2025), 71-88

www.jabes.ueh.edu.vn

JABES
Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á

**Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á**
http://www.emeraldgrouppublishing.com/services/publishing/jabes/

# Dự đoán khả năng mua lại của khách hàng trong thương mại điện tử bằng kỹ thuật học máy trên dữ liệu hành vi rời rạc

LƯU YẾN NHI [a], LÊ THỊ KIM HIỀN [a], HỒ TRUNG THÀNH [*, a]

[a] *Trường Đại học Kinh tế - Luật, Đại học Quốc gia Thành phố Hồ Chí Minh*

THÔNG TIN

*Ngày nhận: 13/11/2025*
*Ngày nhận lại: 31/12/2025*
*Duyệt đăng: 05/01/2026*

**Mã phân loại JEL:**
C38; C41; L81.

**Từ khóa:**
Dữ liệu hành vi rời rạc;
Dự đoán thời điểm mua hàng;
Học máy;
Ngành thời trang;
Tiếp thị cá nhân hóa;
Xác suất mua lại.

**Keywords:**
Discrete behavioral data;
Purchase timing prediction;
Machine learning;
Fashion industry;

TÓM TẮT

Do vòng đời sản phẩm ngắn và xu hướng biến động nhanh, thương mại điện tử trong ngành thời trang đòi hỏi khả năng dự đoán chính xác thời điểm và khả năng mua lại của khách hàng nhằm nâng cao hiệu quả chiến lược tiếp thị cá nhân hóa. Tuy nhiên, các nghiên cứu hiện nay chủ yếu tập trung vào dự báo mua lặp lại trong dài hạn và còn hạn chế trong việc xem xét khả năng mua hàng ngắn hạn từ dữ liệu hành vi rời rạc. Nghiên cứu này đề xuất mô hình dự đoán khả năng mua lại của khách hàng, được lượng hóa bằng xác suất trong ngắn hạn (30 ngày), dựa trên dữ liệu hành vi rời rạc thông qua việc kết hợp các đặc trưng hành vi khách hàng (Recency-Frequency-Monetary – RFM), sự đa dạng danh mục sản phẩm và các chỉ số tương tác trong phiên truy cập. Mô hình được thực nghiệm bằng hai thuật toán học máy LightGBM và XGBoost trên tập dữ liệu giao dịch trực tuyến trong ngành thời trang. Kết quả cho thấy các mô hình đề xuất đạt hiệu năng phân loại tốt với chỉ số ROC-AUC xấp xỉ 0,8830. Đồng thời, nghiên cứu xác định lần mua hàng gần đây (R) và tần suất mua hàng (F) là hai yếu tố có tác động mạnh nhất đến quyết định mua hàng trong ngắn hạn. Những phát hiện này góp phần mở rộng hướng nghiên cứu về dự đoán thời điểm mua hàng và hỗ trợ triển khai các chiến lược tái tiếp thị hiệu quả.

**Abstract**

Due to short product life cycles and rapidly changing trends, e-commerce in the fashion industry requires accurate prediction of the timing and probability of repurchase to enhance the effectiveness of personalized marketing strategies. However, a research gap persists as

[*] Tác giả liên hệ.
Biên tập viên: Nguyễn Lương Tâm.
Email: nhily22411ca@st.uel.edu.vn (Lưu Yến Nhi); hienltk@uel.edu.vn (Lê Thị Kim Hiền); thanhht@uel.edu.vn (Hồ Trung Thành).
Trích dẫn bài viết: Lưu Yến Nhi, Lê Thị Kim Hiền, & Hồ Trung Thành. (2025). Dự đoán khả năng mua lại của khách hàng trong thương mại điện tử bằng kỹ thuật học máy trên dữ liệu hành vi rời rạc. *Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á, 36*(11), 71-88.
https://doi.org/10.24311/jabes/2025.36.11.05

Personalized marketing; Repurchase probability.

most studies prioritize long-term repurchase prediction, often overlooking short-term purchase likelihood, especially when dealing with discrete behavioral data. This study proposes a model to predict short-term repurchase likelihood, quantified as a probability over a 30-day period by integrating Recency-Frequency-Monetary (RFM) features with product category diversity and session interaction metrics. Validated using LightGBM and XGBoost algorithms on an online fashion dataset, the proposed models achieved strong classification performance, yielding an ROC-AUC score of approximately 0.8830. Furthermore, R and F were identified as the most influential predictors of short-term purchasing behavior. These findings not only extend the literature on purchase timing prediction but also enable businesses to identify high-potential customers for implementing more effective remarketing strategies.

## 1. Giới thiệu

Thị trường thương mại điện tử (TMĐT) đang duy trì đà tăng trưởng mạnh, trở thành một trong những lĩnh vực năng động nhất của nền kinh tế số. Theo báo cáo của VnEconomy (Hoang, 2025), tổng doanh thu TMĐT Việt Nam trong nửa đầu năm 2025 đạt 202.300 tỷ đồng (tương đương 7,96 tỷ USD), tăng 41,52% so với cùng kỳ năm trước, vượt xa mức tăng trưởng chung của ngành bán lẻ. Cùng với đó, hơn 1,9 tỷ sản phẩm đã được tiêu thụ qua các sàn TMĐT, phản ánh xu hướng dịch chuyển mạnh mẽ của người tiêu dùng sang môi trường mua sắm trực tuyến.

Mặc dù thị trường TMĐT đang tăng trưởng mạnh mẽ và hành vi mua sắm trực tuyến của người tiêu dùng ngày càng đa dạng, nhiều doanh nghiệp vẫn gặp khó khăn trong việc xác định thời điểm và nhóm khách hàng có khả năng mua lại, điều này làm suy giảm hiệu quả các chiến dịch tái tiếp thị (Remarketing) và cá nhân hóa (Gomes và cộng sự, 2023). Ngoài ra, Heinisch và cộng sự (2022) cũng chỉ ra rằng, nhiều doanh nghiệp triển khai chiến dịch gửi khuyến mãi, thông báo trên ứng dụng (Push Notification) hoặc tái tiếp thị theo chu kỳ cố định, thay vì dựa trên phân tích dữ liệu hành vi cá nhân và thời điểm sẵn sàng mua. Điều này dẫn đến chi phí tái tiếp thị cao nhưng tỷ lệ chuyển đổi thấp. Theo Li và cộng sự (2020), việc phân phối quảng cáo quá sớm so với thời điểm nhu cầu phát sinh có thể dẫn đến tỷ lệ mua hàng thấp và gây lãng phí ngân sách quảng cáo trực tuyến. Bên cạnh đó, nghiên cứu của Zhou và Hudin (2024) chỉ ra rằng, việc mã hóa dấu thời điểm sự kiện kết hợp mô hình chú ý (Attention Mechanism) và mạng nơ-ron có thể nắm bắt tốt hành vi tuần tự khi dữ liệu sự kiện theo thời gian đủ dày và liên tục. Tuy nhiên, Vallarino (2023) nhấn mạnh rằng, các mô hình dự đoán thời điểm (Time-to-event) thường gặp hạn chế khi dữ liệu rời rạc và không liên tục, khiến việc áp dụng và triển khai trong thực tiễn TMĐT trở nên khó khăn.

Trên cơ sở những vấn đề và khoảng trống nêu trên, nghiên cứu này hướng đến việc xây dựng mô hình học máy dự đoán khả năng khách hàng mua hàng trong vòng 30 ngày tới, nhằm hỗ trợ doanh nghiệp tối ưu hóa thời điểm triển khai các chiến dịch tiếp thị cá nhân hóa. Việc lựa chọn khung thời gian 30 ngày được xem là hợp lý khi xét đến chu kỳ nhu cầu trong ngành thời trang. Theo Popowska và Sinkiewicz (2021), 17,4% người tiêu dùng mua quần áo nhiều hơn một lần mỗi tháng và 16,3% mua hàng theo chu kỳ một tháng, phản ánh một nhóm khách hàng có hành vi mua lặp lại thường

xuyên theo tháng. Bối cảnh Việt Nam cũng ghi nhận xu hướng tương tự: báo cáo từ Cốc Cốc (2024) cho thấy gần 50% người tiêu dùng thời trang mua sắm theo chu kỳ ngắn, trong đó 7,4% mua hàng hàng ngày, 16,3% hàng tuần, và 28,5% hàng tháng. Mức tần suất này cho thấy nhu cầu mua mới diễn ra liên tục và đặc biệt ổn định theo tháng, đồng thời trùng khớp với các giai đoạn khuyến mãi định kỳ (như 8/8, 9/9, và 10/10). Do đó, việc sử dụng khung thời gian 30 ngày cho mô hình dự báo không chỉ phù hợp với hành vi tiêu dùng chung trong ngành thời trang mà còn giúp doanh nghiệp triển khai chiến dịch tiếp thị cá nhân hóa đúng thời điểm và hạn chế sai lệch trong dự báo ngắn hạn.

Cụ thể, nghiên cứu hiện tại đặt mục tiêu phát triển một khung dự đoán khả năng mua hàng trong khung thời gian cố định, sử dụng dữ liệu hành vi giao dịch thực tế như lần mua hàng gần đây (Recency – R) - tần suất mua hàng (Frequency – F) - giá trị của giao dịch (Monetary – M), đa dạng danh mục, đồng thời đánh giá và so sánh hiệu năng của các nhóm mô hình học máy thông qua các chỉ số đo lường độ chính xác và khả năng phân biệt. Song song với quá trình xây dựng mô hình, nghiên cứu cũng hướng đến việc phân tích các đặc trưng hành vi có ảnh hưởng lớn đến xác suất mua hàng, từ đó cung cấp những gợi ý hữu ích cho hoạt động cá nhân hóa trải nghiệm khách hàng và quản trị dữ liệu trong TMĐT. Từ những mục tiêu trên, kết quả bài báo sẽ đóng góp vào ba vấn đề chính bao gồm: (1) bổ sung hướng nghiên cứu về dự đoán hành vi mua hàng có yếu tố thời gian, nhấn mạnh khía cạnh thời điểm khách hàng có khả năng phát sinh mua lại trong khung ngắn - trung hạn, qua đó mở rộng khung lý thuyết cho lĩnh vực dự đoán mua hàng trong TMĐT; (2) đề xuất khung ứng dụng kỹ thuật học máy đơn giản, có thể áp dụng cho dữ liệu giao dịch rời rạc của doanh nghiệp TMĐT, kết hợp các biến hành vi mua hàng (Recency-Frequency-Monetary – RFM), biến đa dạng danh mục và phản ứng khuyến mãi để dự đoán khả năng mua hàng trong thời gian ngắn; và (3) cung cấp công cụ hỗ trợ ra quyết định giúp doanh nghiệp TMĐT, đặc biệt trong ngành thời trang trực tuyến, xác định nhóm khách hàng có khả năng mua cao trong ngắn hạn, từ đó tối ưu hóa hoạt động tái tiếp thị cá nhân hóa nhằm nâng cao hiệu quả tỷ lệ chuyển đổi.

Phần tiếp theo của bài báo được cấu trúc như sau: Phần 2 trình bày cơ sở lý thuyết và tổng quan các nghiên cứu liên quan. Phần 3 mô tả phương pháp nghiên cứu, quy trình xử lý dữ liệu, xây dựng biến đặc trưng và huấn luyện mô hình dự đoán. Phần 4 trình bày kết quả thực nghiệm và thảo luận, đồng thời phân tích hàm ý quản trị. Cuối cùng, Phần 5 đưa ra kết luận, tóm tắt các đóng góp chính, nêu hạn chế nghiên cứu và đề xuất hướng phát triển.

## 2. Cơ sở lý thuyết và tổng quan nghiên cứu

### *2.1. Mô hình RFM và các yếu tố hành vi khách hàng trong TMĐT*

Theo Cheng và Chen (2009), mô hình RFM (Recency-Frequency-Monetary) được giới thiệu đầu tiên bởi Hughes (1994), là một trong những công cụ phân tích hành vi khách hàng phổ biến nhất, được sử dụng rộng rãi trong thương mại điện tử để đánh giá mức độ gắn kết và giá trị của khách hàng. Ba thành phần của mô hình là RFM phản ánh tần suất, giá trị và mức độ gần đây của giao dịch, qua đó giúp doanh nghiệp xác định nhóm khách hàng tiềm năng cho các chiến dịch tái tiếp thị hoặc duy trì lòng trung thành (Wong và cộng sự, 2024). Trong bối cảnh TMĐT hiện đại, một số nghiên cứu đã mở rộng mô hình RFM truyền thống bằng cách bổ sung các yếu tố hành vi khác như đa dạng danh mục mua sắm, thời gian giữa các lần mua hoặc phản ứng với khuyến mãi, nhằm phản ánh tốt hơn động lực mua hàng của khách hàng trong môi trường trực tuyến (Gholamveisy và cộng sự, 2024; Jalal

& Elmaghraby, 2024). Đối với ngành thời trang trực tuyến, các đặc trưng này là quan trọng do hành vi mua mang tính chu kỳ và nhạy cảm với xu hướng, như Hoàng Nguyễn Thu Huyền và cộng sự (2023) đã chỉ ra trong nghiên cứu của mình. Việc kết hợp mô hình RFM mở rộng với dữ liệu hành vi thực tế (chẳng hạn như lượt xem sản phẩm, thêm giỏ hàng, hoặc tương tác với thư điện tử khuyến mãi) giúp doanh nghiệp xây dựng nền tảng phân tích hành vi khách hàng mạnh mẽ hơn.

### *2.2. Kỹ thuật học máy trong dự đoán hành vi mua hàng*

Trong bối cảnh dữ liệu người tiêu dùng ngày càng phong phú và phi cấu trúc, các kỹ thuật học máy đã trở thành công cụ chủ đạo trong việc dự đoán hành vi mua hàng trên nền tảng thương mại điện tử. Theo Segun-Falade và cộng sự (2024), học máy cho phép mô hình hóa các mối quan hệ phi tuyến giữa hành vi giao dịch và xác suất phát sinh mua hàng vượt trội hơn so với các phương pháp thống kê truyền thống. Các mô hình thường được áp dụng bao gồm hồi quy logistic, mô hình tổ hợp cây quyết định (Tree-based Ensemble) và mạng nơ-ron nhân tạo (Artificial Neural Network – ANN), giúp dự đoán khả năng mua lại dựa trên các đặc trưng hành vi RFM, lịch sử giao dịch hoặc hành vi duyệt web (Liu và cộng sự, 2024). Gần đây, hướng học sâu (Deep Learning) và mô hình chuỗi thời gian (Sequence Modeling) được chú trọng nhằm khai thác đặc trưng động trong hành vi người dùng (Liu và cộng sự, 2024). Điển hình là nghiên cứu của Zhou và Hudin (2024) đã chỉ ra rằng việc áp dụng mô hình chuỗi thời gian dựa trên kiến trúc biến đổi (Transformer-based Time-series Modeling) kết hợp với mạng nơ-ron đồ thị (Graph Neural Network – GNN) cho phép mô hình nắm bắt mối quan hệ giữa các sản phẩm và sự thay đổi hành vi theo thời gian, từ đó cải thiện đáng kể độ chính xác trong dự đoán ý định mua của người dùng. Tuy nhiên, hầu hết các nghiên cứu vẫn tập trung vào phân loại khả năng có mua hay không, thay vì dự đoán thời điểm mua hàng, do đó chưa phản ánh chính xác chu kỳ mua sắm thực tế của khách hàng.

### *2.3. Yếu tố thời gian trong dự đoán hành vi mua hàng*

Trong nghiên cứu dự đoán hành vi mua hàng, yếu tố thời gian giữa các lần giao dịch được xem là chỉ báo quan trọng phản ánh chu kỳ mua sắm và mức độ duy trì tương tác của khách hàng (Lismont và cộng sự, 2018). Theo nhóm tác giả này, việc mô hình hóa khoảng thời gian giữa hai lần mua giúp xác định xác suất khách hàng quay lại trong một khung thời gian cụ thể, từ đó hỗ trợ doanh nghiệp tối ưu thời điểm triển khai chiến dịch tái tiếp thị. Gần đây, Vallarino (2023) đề xuất ứng dụng các mô hình học máy sinh tồn (Survival Machine Learning) để trực tiếp dự đoán thời điểm khách hàng có khả năng mua lại, cho thấy hiệu quả cao hơn so với các mô hình phân loại nhị phân truyền thống. Tuy nhiên, phần lớn các nghiên cứu hiện nay vẫn yêu cầu dữ liệu dày và liên tục, trong khi dữ liệu TMĐT thực tế thường rời rạc và thiếu tính thời gian thực. Do đó, việc tích hợp yếu tố thời gian vào mô hình học máy ngắn hạn (như khung 30 ngày) được xem là hướng tiếp cận phù hợp.

### *2.4. Khoảng trống nghiên cứu*

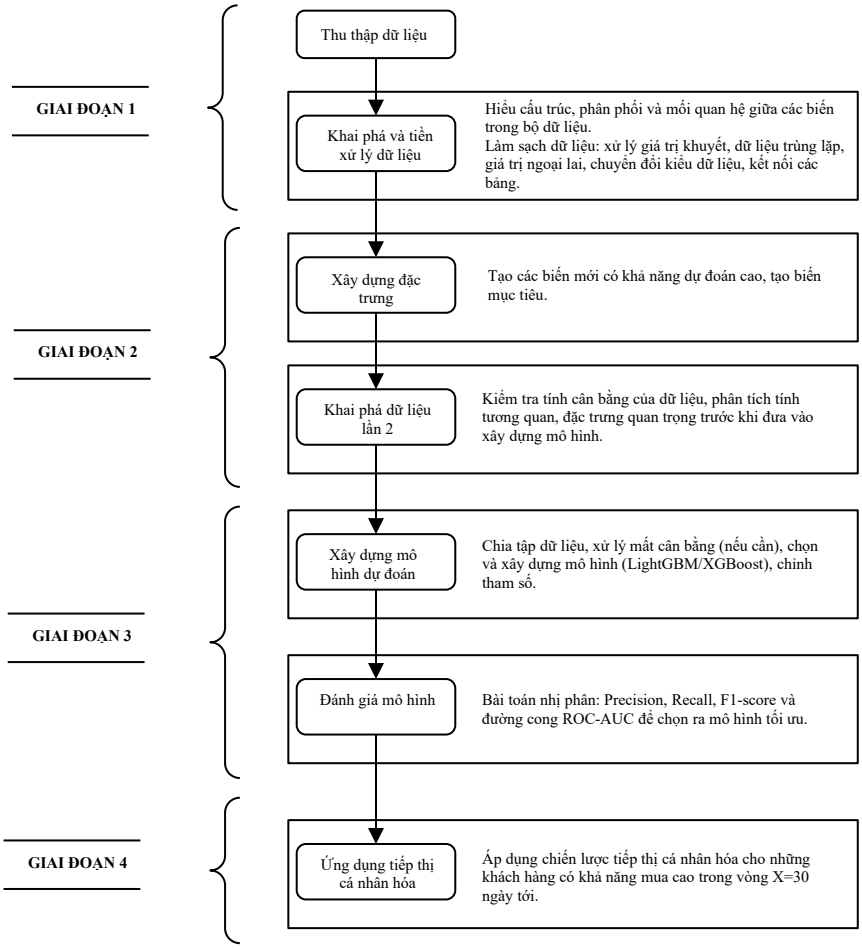
Trong những nghiên cứu gần đây, nhiều hướng tiếp cận tập trung mô hình hóa tín hiệu thời gian liên tục để dự đoán ý định mua, ví dụ như cơ chế chú ý theo thời điểm (Attention-To-Timestamp) hoặc khai thác chuỗi sự kiện đa cấp (Zhou & Hudin, 2024). Một nhánh khác tiếp cận theo hướng ước lượng thời gian mua (Time-to-Purchase) dạng ước lượng thời gian sống (Survival)/thời gian mua lại (Horizon), nhưng thường hướng tới các khoảng thời gian dự báo dài hơn và đòi hỏi chuỗi dữ liệu dày liên tục (Vallarino, 2023). Tuy nhiên, một số các nghiên cứu vẫn triển khai trên dữ liệu đa ngành,

chưa áp dụng trên các lĩnh vực đặc thù, trong khi ngành thời trang là một lĩnh vực có hành vi mua lặp lại ngắn, xoay vòng theo tháng, được chứng minh chịu tác động mạnh của R và F trong phân nhóm khách hàng (Verma và cộng sự, 2025; Vallarino, 2023). Từ đó, vẫn còn thiếu một hướng tiếp cận tập trung vào dự đoán khả năng mua trong một khung thời gian ngắn như 30 ngày, khung thời gian vốn phù hợp với chu kỳ tái tiếp thị và đo lường ngân sách tiếp thị trong ngành thời trang trực tuyến. Nghiên cứu này tập trung lấp đầy vào khoảng trống được phân tích và nhận định trên.

## 3. Phương pháp nghiên cứu

### *3.1. Quy trình nghiên cứu*

Nghiên cứu tập trung đề xuất mô hình để dự đoán khả năng mua lại của khách hàng được lượng hóa bằng xác suất trong khung thời gian ngắn (30 ngày). Quy trình nghiên cứu gồm bốn giai đoạn chính, được thể hiện trong Hình 1 dưới đây.




**Hình 1.** Quy trình và phương pháp nghiên cứu

*Giai đoạn 1: Thu thập và tiền xử lý dữ liệu đầu vào*

Ở giai đoạn đầu, nghiên cứu tiến hành thu thập và tiền xử lý dữ liệu. Dữ liệu được lọc theo ngành hàng thời trang nhằm đảm bảo tính đặc thù và nhất quán. Tiếp đó, các bước phân tích khám phá dữ liệu (Exploratory Data Analysis – EDA) và tiền xử lý được thực hiện để làm sạch, chuẩn hóa, tách (Parse) các trường dữ liệu phức hợp, xử lý giá trị khuyết, trùng lặp, và ngoại lệ. Kết quả là bộ dữ liệu đã được chuẩn hóa và sẵn sàng cho giai đoạn trích xuất đặc trưng hành vi khách hàng.

*Giai đoạn 2: Xây dựng và lựa chọn đặc trưng dữ liệu*

Dựa trên mô hình RFM mở rộng, nghiên cứu bổ sung thêm các biến về đa dạng danh mục mua và mức độ tương tác với khuyến mãi, nhằm phản ánh toàn diện hơn hành vi mua sắm trong môi trường TMĐT. Sau đó, nghiên cứu gán nhãn mục tiêu nhị phân: khách hàng được gán giá trị là 1 nếu phát sinh mua hàng trong vòng 30 ngày tiếp theo, và 0 nếu không mua. Bước cuối cùng là phân tích tương quan và lựa chọn đặc trưng bằng các phương pháp thống kê và kỹ thuật lọc (Feature Selection), đảm bảo giảm nhiễu và tăng khả năng khái quát của mô hình.

*Giai đoạn 3: Huấn luyện và đánh giá mô hình*

Dữ liệu sau khi tiền xử lý được chia thành tập huấn luyện và kiểm định theo tỷ lệ 80:20. Tỷ lệ này đảm bảo được mô hình có đủ dữ liệu để học (80%) trong khi vẫn giữ lại một phần độc lập (20%) để đánh giá khách quan hiệu quả dự đoán của mô hình. Đồng thời áp dụng lấy mẫu phân tầng (Stratified Sampling) để giữ tỷ lệ lớp và đánh giá mô hình một cách khách quan, từ đó hạn chế hiện tượng quá khớp dữ liệu (Overfitting). Nghiên cứu áp dụng hai mô hình XGBoost và LightGBM, đại diện cho nhóm các phương pháp tổ hợp mô hình dựa trên cây quyết định (Tree-based Ensemble Methods), nhằm dự đoán khả năng khách hàng phát sinh mua hàng trong 30 ngày tới. Các mô hình được tinh chỉnh siêu tham số và đánh giá hiệu năng thông qua các chỉ số độ chính xác (Precision), độ bao phủ (Recall), chỉ số F1 (F1-score) và diện tích dưới đường cong ROC (ROC-AUC), từ đó lựa chọn mô hình có khả năng dự đoán tối ưu cho ứng dụng trong tiếp thị cá nhân hóa.

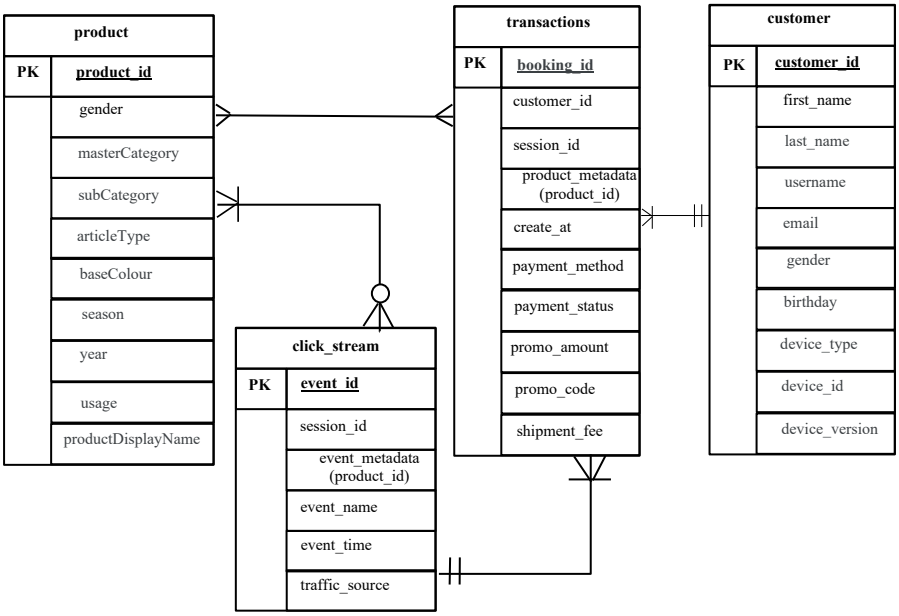
*Giai đoạn 4: Ứng dụng mô hình trong tiếp thị cá nhân hóa*

Kết quả dự đoán được sử dụng để nhận diện nhóm khách hàng có xác suất mua cao trong 30 ngày tới, phục vụ cho việc triển khai chiến dịch quảng cáo tái tiếp thị cá nhân hóa, giúp doanh nghiệp tối ưu chi phí quảng cáo và nâng cao tỷ lệ chuyển đổi.

### *3.2. Dữ liệu*

Dữ liệu được thu thập từ cơ sở dữ liệu E-commerce App Transactional trên Kaggle[1], bao gồm thông tin về hành vi người dùng, lịch sử giao dịch và thuộc tính sản phẩm trong giai đoạn 2016–2022. Bộ dữ liệu được sử dụng trong nghiên cứu này được xem như môi trường thực nghiệm có kiểm soát, nhằm phục vụ việc đánh giá và so sánh hiệu năng của khung mô hình đề xuất. Các thực thể dữ liệu chính và ý nghĩa được mô tả chi tiết trong Bảng 1, đồng thời mối quan hệ giữa các đối tượng nghiên cứu và đặc trưng dữ liệu được minh họa qua sơ đồ tại Hình 2. Sau khi thu thập, nghiên cứu tiến hành lọc chọn các giao dịch thuộc ngành hàng thời trang (Apparel) để đảm bảo tính tập trung và phù hợp với phạm vi nghiên cứu.

[1] Aditya Bagus Pratama (2023), https://www.kaggle.com/datasets/bytadit/transactional-ecommerce, truy cập ngày 11/07/2025.




**Hình 2.** Sơ đồ mối quan hệ giữa các khách thể, đối tượng nghiên cứu và các đặc trưng

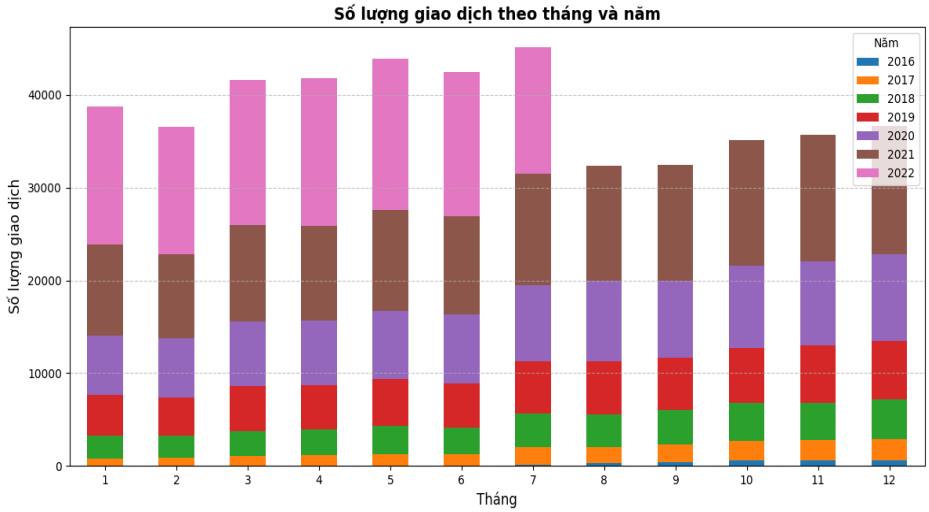
*Ghi chú:* Ký hiệu các biến được mô tả trong Bảng 1.

**Bảng 1.**

Các thực thể trong bộ dữ liệu

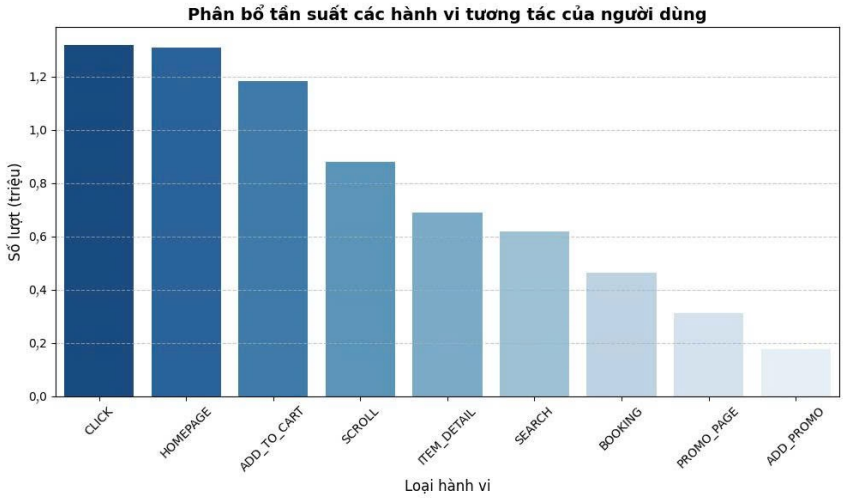
| Thực thể | Ý nghĩa | Ý nghĩa các biến quan trọng |
|---|---|---|
| Khách hàng (Customer) | Chứa thông tin chi tiết của người dùng đã đăng ký trên ứng dụng TMĐT, bao gồm mã khách hàng, giới tính, và tên. | - customer_id: mã khách hàng |
| Sản phẩm (Product) | Lưu trữ dữ liệu về sản phẩm được bán trên nền tảng, gồm mã sản phẩm, danh mục, và thương hiệu. | - product_id: mã sản phẩm<br>- articleType: phân loại sản cụ thể (như Shirts, Dress, Jeans) |
| Giao dịch (Transactions) | Ghi nhận thông tin về từng giao dịch hoặc đơn hàng mà khách hàng thực hiện. Mỗi khách hàng có thể có nhiều giao dịch với nhiều sản phẩm khác nhau, bao gồm các trường như mã đơn hàng, mã khách hàng, ngày giao dịch, tổng giá trị, số lượng, và trạng thái thanh toán. | - created_at: thời điểm phát sinh giao dịch<br>- total_amount: tổng chi tiêu của giao dịch<br>- num_unique_products: số sản phẩm khác nhau trong đơn |
| Hành vi (Clickstream) | Ghi lại hành vi sử dụng ứng dụng và các sự kiện tương tác của người dùng trong mỗi phiên (Session), như xem sản phẩm, thêm vào giỏ hàng, và thao tác thanh toán; phản ánh chuỗi hành vi dẫn đến quyết định mua hàng. | - session_id: mã phiên truy cập<br>- event_name: tên sự kiện (như ADD_TO_CART, BOOKING)<br>- event_time: thời điểm xảy ra sự kiện |

*3.2.1. Phân tích khách phá dữ liệu (EDA)*




**Hình 3.** Số lượng giao dịch qua các năm

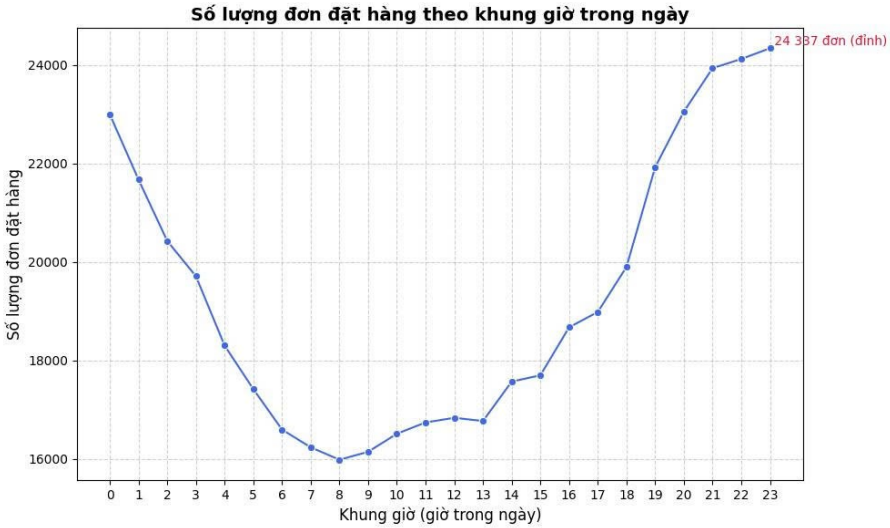
Kết quả ở Hình 3 cho thấy, khối lượng giao dịch tăng mạnh qua các năm, đạt đỉnh vào năm 2022 và thấp nhất vào năm 2016. Dữ liệu thể hiện rõ tính mùa vụ, khi số lượng giao dịch tập trung nhiều hơn trong nửa cuối năm (tháng 7–12). Mặc dù dữ liệu 2016 và 2022 chưa đầy đủ, xu hướng chung cho thấy hoạt động mua sắm tăng trưởng nhanh, phản ánh sự mở rộng của thị trường thương mại điện tử.




**Hình 4.** Phân bổ tần suất các hành vi tương tác của người dùng

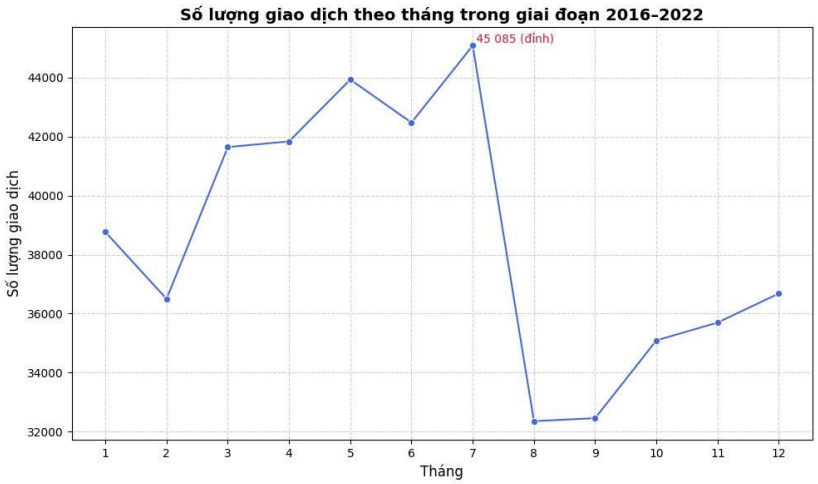
Biểu đồ phân tích dữ liệu Hành vi (Clickstream) (xem Hình 4) cho thấy người dùng tương tác chủ yếu qua các sự kiện CLICK và HOMEPAGE, thể hiện mức độ truy cập cao. Sự kiện ADD_TO_CART xuất hiện nhiều, phản ánh ý định mua hàng rõ rệt, trong khi các hành động BOOKING, PROMO_PAGE và ADD_PROMO ít hơn đáng kể, gợi ý tồn tại "nút thắt chuyển đổi" khi người dùng thường dừng lại ở bước thêm sản phẩm vào giỏ hàng mà chưa hoàn tất mua hàng. Điều này gợi mở

hướng phân tích sâu hơn về hành trình chuyển đổi của khách hàng và các yếu tố ảnh hưởng đến quyết định mua hàng.




**Hình 5.** Số lượng các đơn đặt hàng theo khung giờ trong ngày

Bên cạnh đó, việc quan sát thời điểm thực hiện giao dịch trong Hình 5 cho thấy, các hoạt động đặt hàng tập trung chủ yếu vào buổi tối và rạng sáng (20g00–2g00), trong khi thấp nhất vào ban ngày (6g00–10g00). Kết quả này cũng phản ánh thói quen mua sắm ngoài giờ làm việc, là đặc trưng của nhóm khách hàng trên các sàn thương mại điện tử.




**Hình 6.** Số lượng giao dịch theo tháng trong giai đoạn 2016–2022

Ngoài ra, Hình 6 cho thấy số lượng giao dịch cao nhất vào tháng 7 và có xu hướng giảm rõ rệt trong giai đoạn từ tháng 8 đến tháng 9, trước khi tăng nhẹ trở lại vào cuối năm. Điều này phản ánh tính mùa vụ trong hành vi mua sắm, khi nhu cầu tăng mạnh vào giữa năm, đó cũng là thời điểm thường diễn ra nhiều chiến dịch khuyến mãi lớn trong ngành thời trang.

*3.2.2. Tiền xử lý dữ liệu*

Sau khi lọc dữ liệu thuộc ngành hàng thời trang và trực quan hóa cấu trúc dữ liệu thì tiến hành xóa các biến không cần thiết và không phục vụ mục tiêu dự đoán hành vi mua hàng. Bảng 2 dưới đây thể hiện các biến được giữ lại sau khi xử lý.

**Bảng 2.**

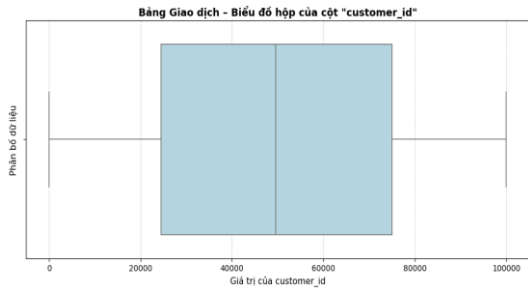
Các biến còn lại trong bộ dữ liệu

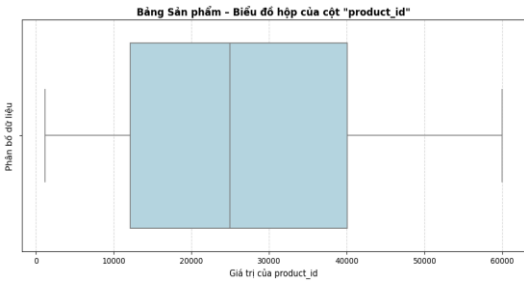
| Tên bảng dữ liệu | Các biến còn lại | Mô tả |
|---|---|---|
| apparel_transactions.csv (Giao dịch) | created_at, customer_id, booking_id, session_id, product_metadata, payment_status, promo_amount, total_amount | Lưu trữ thông tin chi tiết của từng giao dịch |
| apparel_products.csv (Sản phẩm) | product_id, gender, masterCategory, subCategory, articleType | Chứa dữ liệu mô tả sản phẩm thời trang |
| apparel_customers.csv (Khách hàng) | customer_id, gender, birthdate, first_join_date | Chứa thông tin cơ bản của khách hàng |
| apparel_clickstream.csv (Hành vi) | session_id, event_name, event_time, event_metadata | Ghi nhận chuỗi hành vi người dùng trong mỗi phiên truy cập (Session) |

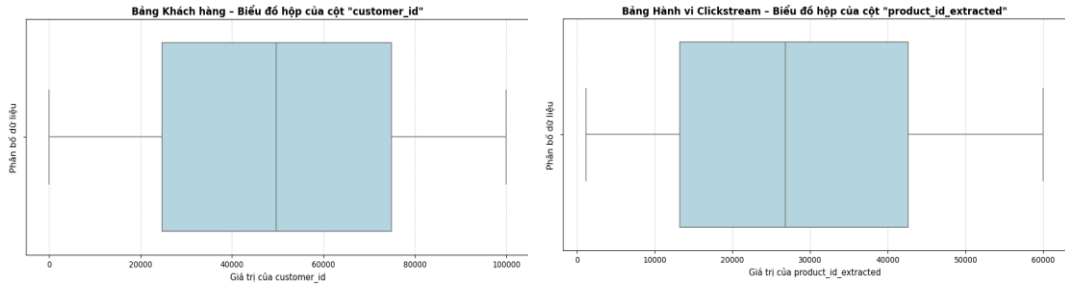
*Ghi chú*: Ký hiệu các biến được mô tả trong Bảng 1.

Tiếp theo, dữ liệu sẽ được tách chuỗi các trường chứa thông tin phức hợp nhằm chuẩn hóa cấu trúc dữ liệu, ví dụ như tách biến event_metadata trong bảng Hành vi (Clickstream) để lấy biến product_id_extracted, hoặc tách biến product_metadata trong bảng Giao dịch (Transactions) thành các cột các biến num_items, num_unique_products, và avg_item_price.

Cuối cùng, dữ liệu được kiểm tra và xử lý giá trị khuyết (Null Values), loại bỏ bản ghi trùng lặp (Duplicates) và xử lý các giá trị ngoại lai (Outliers) để đảm bảo tính toàn vẹn và chất lượng trước khi trích xuất đặc trưng (xem Hình 7).






**Hình 7.** Kiểm tra các giá trị ngoại lai

### *3.3. Xây dựng đặc trưng*

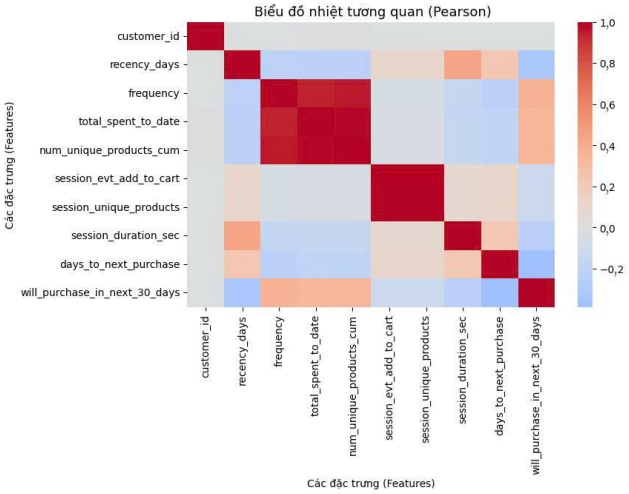
Kỹ thuật xây dựng đặc trưng được thực hiện trong nghiên cứu này nhằm chuyển đổi dữ liệu thô thành các đặc trưng (Features) phản ánh hành vi và giá trị của khách hàng trong quá trình mua sắm. Các biến đặc trưng này (xem Bảng 3) được lựa chọn và xây dựng dựa trên lý thuyết hành vi tiêu dùng trong thương mại điện tử, đặc biệt là mô hình RFM, kết hợp với hành vi tương tác trên ứng dụng.

**Bảng 3.**

Các biến đặc trưng và biến mục tiêu

| Nhóm | Tên biến | Ý nghĩa |
|---|---|---|
| Biến đặc trưng | recency_days | Số ngày kể từ lần mua hàng gần nhất của khách hàng, phản ánh độ "mới" trong hành vi mua sắm. |
| | frequency | Tần suất mua hàng của khách hàng trong giai đoạn quan sát (số lượng giao dịch). |
| | total_spent_to_date | Tổng giá trị chi tiêu tích lũy của khách hàng cho đến thời điểm quan sát. |
| | num_unique_products_cum | Số lượng sản phẩm duy nhất mà khách hàng đã từng mua. |
| | session_evt_add_to_cart | Tổng số lần khách hàng thêm sản phẩm vào giỏ hàng trong các phiên truy cập. |
| | session_unique_products | Số lượng sản phẩm khác nhau mà khách hàng đã xem trong một phiên truy cập. |
| | session_duration_sec | Thời lượng trung bình mỗi phiên truy cập (tính bằng giây), phản ánh mức độ quan tâm của khách hàng. |
| Biến mục tiêu | will_purchase_in_next_30_days | Biến mục tiêu, nhận giá trị 1 nếu khách hàng có phát sinh giao dịch trong 30 ngày kế tiếp, ngược lại là 0. |

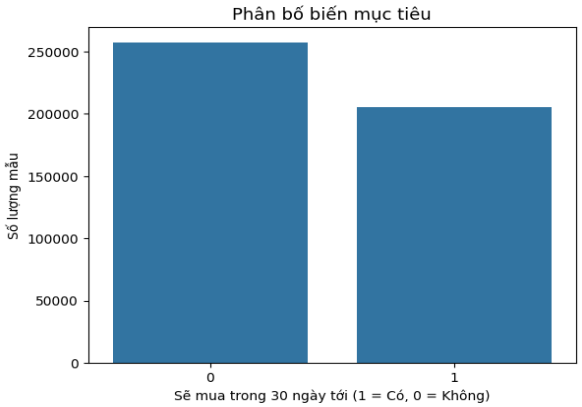
Biến mục tiêu will_purchase_in_next_30_days được xây dựng dựa trên hai khung thời gian: khung quan sát (T_obs) ghi nhận hành vi mua hàng trong quá khứ của từng khách hàng, và khung dự đoán (T_pred = 30 ngày) dùng để kiểm tra khả năng phát sinh giao dịch trong tương lai gần. Nếu khách hàng có ít nhất một giao dịch trong 30 ngày sau T_obs, biến mục tiêu được gán giá trị 1, ngược lại là 0. Cách tiếp cận này giúp mô hình học máy học được mối quan hệ giữa hành vi lịch sử và xác suất mua hàng ngắn hạn.




**Hình 8.** Độ tương quan giữa các biến đặc trưng với biến mục tiêu

*Ghi chú:* Ký hiệu biến được mô tả trong Bảng 1 và Bảng 3.

Tiếp theo, nghiên cứu tiến hành kiểm tra mối tương quan giữa các biến độc lập và biến mục tiêu bằng hệ số tương quan Pearson (xem Hình 8). Kết quả cho thấy biến recency_days có tương quan nghịch mạnh nhất với xác suất mua hàng trong 30 ngày tới (r = -0,3864), nghĩa là khách hàng vừa mua gần đây có khả năng quay lại mua cao hơn. Hai biến frequency (r = 0,3737) và total_spent_to_date (r = 0,3412) thể hiện mối tương quan thuận rõ rệt, phản ánh rằng khách hàng mua thường xuyên và chi tiêu cao có xu hướng mua lại nhiều hơn. Ngược lại, các biến hành vi như session_duration_sec, session_evt_add_to_cart và session_unique_products có tương quan yếu hoặc âm, cho thấy việc duyệt sản phẩm nhiều hoặc thêm giỏ hàng không nhất thiết dẫn đến hành động mua.



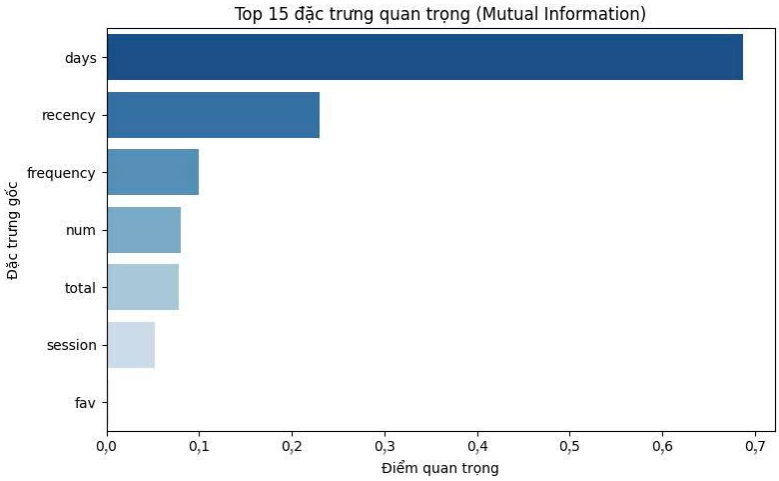

**Hình 9.** Độ cân bằng của biến mục tiêu

Hình 9 biểu diễn phân phối của biến mục tiêu được kiểm tra nhằm đánh giá độ cân bằng dữ liệu. Kết quả cho thấy hai lớp (0: không mua, 1: có mua) có tỷ lệ tương đối cân bằng, trong đó nhóm khách hàng không mua trong 30 ngày tới chiếm tỷ lệ cao hơn nhẹ.

### *3.4. Xây dựng mô hình dự đoán*

Trong nghiên cứu này, hai thuật toán LightGBM và XGBoost được lựa chọn để huấn luyện mô hình dự đoán khả năng phát sinh mua hàng trong 30 ngày tới. Cả hai thuộc nhóm mô hình tổ hợp cây quyết định, nổi bật với khả năng xử lý dữ liệu phi tuyến tính, tốc độ huấn luyện nhanh và hiệu quả cao trong các bài toán phân loại nhị phân.

LightGBM sử dụng cơ chế tăng trưởng theo lá (Leaf-wise Growth) giúp giảm lỗi huấn luyện nhanh hơn so với cơ chế tăng trưởng theo tầng (Level-wise Growth) truyền thống, nhờ đó tối ưu hiệu năng khi dữ liệu có đặc trưng phức tạp và nhiều chiều (Ke và cộng sự, 2017). XGBoost được lựa chọn vì khả năng tổng quát hóa tốt và cơ chế chuẩn hóa mô hình (Regularization) giúp kiểm soát hiện tượng quá khớp (Overfitting), phù hợp với dữ liệu TMĐT có sự chênh lệch giữa hai nhóm hành vi mua và không mua (Chen & Guestrin, 2016).




**Hình 10.** Các đặc trưng quan trọng

Trước khi huấn luyện, mức độ quan trọng của đặc trưng (Feature Importance) được đánh giá bằng phương pháp thông tin hỗ trợ (Mutual Information). Hình 10 cho thấy các biến liên quan đến yếu tố thời gian như days_to_next_purchase và recency_days có ảnh hưởng lớn nhất đến khả năng mua lại, tiếp theo là các biến frequency và total_spent_to_date, phản ánh đúng bản chất hành vi tiêu dùng trong ngành thời trang – phụ thuộc mạnh vào tần suất và độ gần của lần mua gần nhất.

Dữ liệu được chia thành tập huấn luyện và kiểm thử theo tỷ lệ 80:20 bằng phương pháp lấy mẫu phân tầng (Stratified Sampling). Hai mô hình LightGBM và XGBoost được huấn luyện với cùng bộ siêu tham số đồng nhất để đảm bảo tính so sánh công bằng, bao gồm tốc độ học (Learning Rate) 0,05, độ sâu tối đa (Max Depth) 6 và số lượng cây (n_estimators) 200.

## 4. Kết quả nghiên cứu và thảo luận

### *4.1. Kết quả thực nghiệm*

**Bảng 4.**

So sánh chỉ số ROC-AUC giữa 2 mô hình

| Mô hình | ROC-AUC |
|---|---|
| LightGBM | 0,8831 |
| XGBoost | 0,8830 |

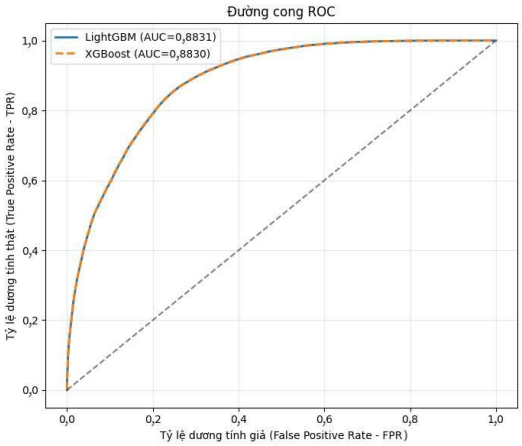
**Bảng 5.**

Các chỉ số hiệu suất của mô hình học máy

| | Độ chính xác | Độ nhạy | Điểm F1 | Hỗ trợ |
|---|---|---|---|---|
| 0 | 0,8451 | 0,7799 | 0,8112 | 51.408 |
| 1 | 0,7489 | 0,8212 | 0,7834 | 41.098 |
| Tỷ lệ dự đoán đúng | | | 0,7983 | 92.506 |
| Hiệu suất trung bình vĩ mô | 0,7970 | 0,8005 | 0,7973 | 92.506 |
| Trung bình có trọng số | 0,8024 | 0,7983 | 0,7988 | 92.506 |

Kết quả huấn luyện cho bài toán dự đoán khả năng khách hàng phát sinh mua hàng trong 30 ngày tới cho thấy cả hai mô hình LightGBM và XGBoost đều đạt hiệu năng ổn định và tương đồng (xem Bảng 4). Mô hình LightGBM đạt độ chính xác tổng thể Accuracy = 0,7983 và ROC-AUC = 0,8831, trong khi đó XGBoost cũng đạt Accuracy = 0,7983 và ROC-AUC = 0,8830, thể hiện khả năng dự đoán chính xác và độ phân biệt cao giữa hai nhóm khách hàng (có và không có khả năng mua hàng).

Phân tích chi tiết các chỉ số precision, recall và F1-score trong Bảng 5 cho thấy cả hai mô hình đều duy trì mức cân bằng tốt giữa độ chính xác và khả năng phát hiện khách hàng tiềm năng. Cụ thể, nhóm khách hàng có khả năng mua hàng đạt precision = 0,7489 và recall = 0,8212; phản ánh rằng mô hình có thể nhận diện hiệu quả nhóm người mua thực sự trong ngắn hạn mà không bỏ sót quá nhiều trường hợp.




**Hình 11.** Đường cong ROC và giá trị AUC thể hiện khả năng phân biệt của mô hình

Kết quả đường cong ROC (xem Hình 11) cho thấy khả năng phân loại tương đương giữa hai mô hình, với đường cong của LightGBM và XGBoost gần như trùng nhau, chỉ ra rằng hiệu năng dự đoán ổn định và khả năng tổng quát hóa tốt. Việc cả hai mô hình đều đạt chỉ số AUC gần bằng nhau (AUC ≈ 0,8830) chứng tỏ rằng chúng phù hợp cho các ứng dụng trong thực tế thương mại điện tử, đặc biệt trong việc xác định nhóm khách hàng có xác suất mua cao để triển khai chiến dịch tái tiếp thị hoặc khuyến mãi mục tiêu.

### *4.2. Thảo luận kết quả và hàm ý quản trị*

Kết quả thực nghiệm cho thấy mô hình tổ hợp cây quyết định (LightGBM, XGBoost) đạt hiệu năng tốt trong việc dự đoán khả năng khách hàng mua trong 30 ngày, với các biến recency và frequency là những đặc trưng có đóng góp lớn nhất cho khả năng phân biệt khách hàng có/không mua lại. Quan sát này tương hợp với những công trình gần đây nhấn mạnh vai trò của các chỉ số RFM (Recency-Frequency-Monetary) trong phân khúc khách hàng và dự báo hành vi mua lại, khi RFM vẫn được coi là "bộ chỉ báo lõi" cho tiếp thị dự đoán (Predictive Marketing) (Verma và cộng sự, 2025).

So sánh với các hướng nghiên cứu khác: Zhou và Hudin (2024) trình bày một khuôn khổ kết hợp mã hóa dấu thời điểm sự kiện (Event-based Timestamp Encoding) và mô hình áp dụng cơ chế chú ý (Attention Model) trên chuỗi thời gian kèm mạng nơ-ron đồ thị để nắm bắt mẫu hành vi tuần tự của người dùng; tác giả báo cáo rằng cách tiếp cận này nâng cao hiệu quả dự đoán khi dữ liệu sự kiện theo thời gian có tính tiếp nối (Zhou & Hudin, 2024). Ngược lại, Vallarino (2023) khảo sát và so sánh các phương pháp học máy sinh tồn (Survival Machine Learning) hay mô hình dự đoán thời gian sự kiện (Time-to-event Models), và chỉ rõ rằng những mô hình này hữu ích để ước lượng khi nào một giao dịch sẽ xảy ra, nhưng đồng thời nêu ra thách thức thực tiễn khi dữ liệu ghi nhận theo thời gian không dày hoặc không đủ số lượng quan sát (Vallarino, 2023).

Với kết quả này, doanh nghiệp TMĐT có thể áp dụng mô hình dự đoán để xác định nhóm khách hàng có khả năng mua cao trong 30 ngày tới và ưu tiên các chiến dịch nhắc giỏ hàng, ưu đãi mã giảm giá hoặc gợi ý sản phẩm cho nhóm này, từ đó tăng tỷ lệ chuyển đổi và giảm chi phí quảng cáo bị lãng phí vào nhóm khách hàng không có ý định mua. Ngoài ra, việc theo dõi chỉ số recency và frequency có thể được sử dụng như tín hiệu cảnh báo sớm để kích hoạt chiến dịch tiếp thị khi hành vi mua gần đây có dấu hiệu giảm.

Kết quả nghiên cứu mang lại hàm ý quản trị rõ ràng cho doanh nghiệp TMĐT tại Việt Nam, đặc biệt trong ngành thời trang: doanh nghiệp có thể sử dụng xác suất mua trong 30 ngày làm "thang điểm ưu tiên" để phân bổ nguồn lực tiếp thị. Nhóm xác suất cao được ưu tiên khuyến mãi/ ưu đãi mạnh để tối đa hóa chuyển đổi, nhóm trung bình dùng ưu đãi nhẹ để thúc đẩy tiếp cận, trong khi nhóm thấp đưa vào các chương trình nuôi dưỡng nhằm duy trì mức độ tương tác dài hạn. Đồng thời, hai tín hiệu hành vi R và F vốn đã được chứng minh là ổn định và có sức dự báo mạnh (Verma và cộng sự, 2025) có thể dùng làm như một kích hoạt (Trigger) đơn giản trong quản trị mối quan hệ khách hàng (Customer Relationship Management – CRM) để kích hoạt chiến dịch tái tiếp thị theo thời gian thực. Ngoài ra, doanh nghiệp có thể triển khai thêm thử nghiệm A/B để đo mức cải thiện hiệu quả (Uplift), tối ưu ngân sách quảng cáo và thiết lập quy trình cập nhật mô hình định kỳ cho chiến dịch tái tiếp thị dựa trên mô hình dự đoán. Tóm lại, điểm mạnh của cách tiếp cận này là vừa khả thi triển khai ngay với dữ liệu giao dịch tiêu chuẩn, vừa mang lại giá trị thực tiễn trong cá nhân hóa, thay vì đòi hỏi hạ

tầng sự kiện liên tục và các mô hình chuỗi thời gian (Time-series) phức tạp như các hướng tiếp cận dựa trên dấu thời gian (Timestamp-based) gần đây.

### *4.3. Hạn chế của nghiên cứu*

Mặc dù đạt được kết quả dự đoán khả quan, nghiên cứu vẫn tồn tại một số hạn chế. *Thứ nhất*, dữ liệu sử dụng mang tính lịch sử và giới hạn trong ngành hàng thời trang, nên khả năng tổng quát hóa sang các lĩnh vực khác còn hạn chế. *Thứ hai*, một số biến hành vi có thể vẫn chứa dấu hiệu rò rỉ thông tin (Data Leakage) do khoảng thời gian quan sát và dự đoán chưa được tách biệt tuyệt đối, dẫn đến mô hình có thể đánh giá quá cao năng lực dự đoán. *Thứ ba*, nghiên cứu mới tập trung vào các mô hình tổ hợp cây quyết định (LightGBM, XGBoost) vốn có đặc tính tương tự nhau nên chưa phản ánh rõ sự khác biệt về hiệu năng giữa các nhóm thuật toán. Cuối cùng, yếu tố ngoại cảnh như chiến dịch tiếp thị, mùa vụ hay hành vi xã hội chưa được tích hợp, khiến mô hình còn mang tính tĩnh. Việc mở rộng hướng thực nghiệm mô hình đề xuất trên dữ liệu thực tế, dữ liệu thời gian thực và mô hình học sâu (Deep Learning) có thể nâng cao hiệu năng và tính ứng dụng thực tế của mô hình dự đoán.

## 5. Kết luận

Nghiên cứu này tập trung phát triển mô hình dự đoán khả năng mua lại xác suất khách hàng phát sinh hành vi mua hàng trong 30 ngày tới, dựa trên dữ liệu giao dịch và hành vi sử dụng nền tảng thương mại điện tử trong ngành thời trang. Kết quả thực nghiệm với các thuật toán học máy như LightGBM và XGBoost cho thấy hiệu năng dự đoán cao, với độ chính xác trung bình đạt khoảng 0,7983 và chỉ số ROC-AUC trên 0,8830. Trong đó, LightGBM thể hiện khả năng khái quát tốt hơn, đồng thời có ưu thế trong xử lý dữ liệu dạng bảng (Tabular Data) với số lượng đặc trưng lớn, phù hợp với bối cảnh thương mại điện tử đa hành vi (Ke và cộng sự, 2017). Phân tích tương quan cho thấy các biến hành vi như số ngày kể từ lần mua gần nhất (Recency) và tần suất mua hàng (Frequency) có ảnh hưởng mạnh nhất đến khả năng mua lại, củng cố cho các kết luận trong các nghiên cứu trước đây về giá trị dự báo của mô hình RFM đối với hành vi mua hàng lặp lại (Verma và cộng sự, 2025).

Về phương diện học thuật, nghiên cứu góp phần mở rộng ứng dụng của học máy trong bài toán dự đoán mua hàng tiếp tục có ràng buộc một khoảng thời gian ngắn hạn (một hướng nghiên cứu còn tương đối mới), đặc biệt trong ngành hàng thời trang, nơi hành vi tiêu dùng chịu tác động mạnh bởi xu hướng, mùa vụ và yếu tố cảm xúc (Zhou & Hudin, 2024). Khác với các nghiên cứu chỉ dự đoán khả năng "có mua hay không" trong tương lai, cách tiếp cận của đề tài nhấn mạnh yếu tố dự báo trong khung thời gian cụ thể ("khách hàng có mua trong tháng tới hay không"), qua đó mang lại góc nhìn thực tiễn hơn cho các hoạt động hoạch định chiến lược tiếp thị và quản trị khách hàng.

Về mặt thực tiễn, mô hình đề xuất đóng vai trò là một giải pháp khả thi để tích hợp trực tiếp vào hệ thống CRM hoặc nền tảng TMĐT, hỗ trợ doanh nghiệp xác định nhóm khách hàng có khả năng mua cao trong chu kỳ 30 ngày tiếp theo. Điều này cho phép minh họa quy trình triển khai các chiến dịch tiếp thị cá nhân hóa, tối ưu hóa chi phí truyền thông và nâng cao tỷ lệ chuyển đổi. Mặt khác cũng phù hợp với xu hướng tiếp thị hướng dữ liệu (Data-driven Marketing) đang được nhiều doanh nghiệp TMĐT lớn áp dụng.

Trong tương lai, để khắc phục các hạn chế của dữ liệu thử nghiệm hiện tại, nghiên cứu có thể được mở rộng theo ba hướng chính: (1) kiểm định mô hình trên tập dữ liệu thực tế từ doanh nghiệp và mở

rộng sang các ngành hàng khác như điện tử, mỹ phẩm, hàng tiêu dùng nhanh để kiểm định khả năng tổng quát hóa của mô hình; (2) tích hợp thêm các yếu tố ngữ cảnh động như khuyến mãi, mùa vụ, xu hướng tìm kiếm nhằm phản ánh chính xác hơn động lực mua hàng; và (3) ứng dụng các kỹ thuật học sâu như LSTM (Long Short-term Memory) hoặc mô hình học sâu áp dụng cơ chế chú ý của TabTransformer để khai thác đặc trưng chuỗi thời gian của hành vi người dùng (Liu và cộng sự, 2024). Những hướng phát triển này hứa hẹn nâng cao độ chính xác và giá trị ứng dụng của mô hình, hướng tới việc xây dựng hệ thống dự đoán và ra quyết định tiếp thị thông minh trong môi trường TMĐT hiện đại.

**Lời cảm ơn**

Nghiên cứu này được tài trợ bởi Trường Đại học Kinh tế - Luật, Đại học Quốc gia Thành phố Hồ Chí Minh, Việt Nam.

**Tài liệu tham khảo**

Cheng, C. H., & Chen, Y. S. (2009). Classifying the segmentation of customer value via RFM model and RS theory. *Expert Systems with Applications, 36*(3), 4176-4184. https://doi.org/10.1016/j.eswa.2008.04.003

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '16) (pp. 785-794). https://doi.org/10.1145/2939672.2939785

Cốc Cốc. (2024). *Ngành thời trang Việt Nam: Nhìn cơ hội từ sự đa dạng trong hành vi và thói quen tiêu dùng*. Truy cập tại https://qc.coccoc.com/vn/news/nganh-thoi-trang-viet-nam-nhin-co-hoi-tu-su-da-dang-trong-hanh-vi-va-thoi-quen-tieu-dung

Gholamveisy, S., Homayooni, S., Shemshaki, M., Sheykhan, S., Boozary, P., Tanhaei, H. G., & Akbari, N. (2024). Application of data mining technique for customer purchase behavior via Extended RFM model with focus on BCG matrix from a data set of online retailing. *Journal of Infrastructure Policy and Development*, *8*(7), 4426. https://doi.org/10.24294/jipd.v8i7.4426

Gomes, M. A., Wönkhaus, M., Meisen, P., & Meisen, T. (2023). TEE: Real-time purchase prediction using time extended embeddings for representing customer behavior. *Journal of Theoretical and Applied Electronic Commerce Research, 18*(3), 1404-1418. https://doi.org/10.3390/jtaer18030070

Heinisch, J. S., Gao, N., Anderson, C., Deldari, S., David, K., & Salim, F. (2022). Investigating the effects of mood & usage behaviour on notification response time. *arXiv*. Retrieved from https://doi.org/10.48550/arXiv.2207.03405

Hoang. A. (2025). E-commerce in upward trend. *Vietnam Economic Times – VnEconomy*. Retrieved from https://en.vneconomy.vn/e-commerce-in-upward-trend-1250945.htm

Hoàng Nguyễn Thu Huyền, Lê Ngọc Sơn, & Nguyễn Quốc Cường. (2023). Các yếu tố ảnh hưởng đến hành vi mua sắm sản phẩm thời trang trên ứng dụng di động của Gen Z tại Thành phố Hồ Chí Minh. *Tạp chí Khoa học và Công nghệ - Trường Đại học Công nghiệp TP.HCM*, *66*(6), 56-72. https://jst.iuh.edu.vn/index.php/jst-iuh/article/view/4989

Hughes, A. M. (1996). Boosting response with RFM. *Marketing Tools*, *3*(3), 4-10.

Jalal, M. E., & Elmaghraby, A. (2024). Analyzing the dynamics of customer behavior: A new perspective on personalized marketing through counterfactual analysis. *Journal of Theoretical and Applied Electronic Commerce Research, 19*(3), Article 81. https://www.mdpi.com/0718-1876/19/3/81

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017) – Proceedings of the 30th Conference* (pp. 3149-3157). https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

Li, J., Luo, X., Lu, X., & Moriguchi, T. (2020). *Boosting returns on E-Commerce retargeting campaigns*. American Marketing Association. Retrieved from https://www.ama.org/2020/11/12/boosting-returns-on-e-commerce-retargeting-campaigns/

Lismont, J., Ram, S., Vanthienen, J., Lemahieu, W., & Baesens, B. (2018). Predicting interpurchase time in a retail environment using customer-product networks: An empirical study and evaluation. *Expert Systems with Applications, 104*, 22-32. https://doi.org/10.1016/j.eswa.2018.03.016

Liu, D., Huang, H., Zhang, H., Luo, X., & Fan, Z. (2024). Enhancing customer behavior prediction in e-commerce: A comparative analysis of machine learning and deep learning models. *Applied and Computational Engineering, 55*(1), 181-195. https://doi.org/10.54254/2755-2721/55/20241475

Popowska, M., & Sinkiewicz, A. (2021). Sustainable fashion in Poland - Too early or too late?. *Sustainability, 13*(17), 9713. https://doi.org/10.3390/su13179713

Segun-Falade, O. D., Osundare, O. S., Kedi, W. E., Okeleke, P. A., Ijomah, T. I., & Abdul-Azeez, O. Y. (2024). Utilizing machine learning algorithms to enhance predictive analytics in customer behavior studies. *International Journal of Scholarly Research in Engineering and Technology, 4*(1), 1-18. https://doi.org/10.56781/ijsret.2024.4.1.0018

Vallarino, D. (2023). *Buy when? Survival machine learning model comparison for purchase timing*. arXiv. https://doi.org/10.48550/arXiv.2308.14343

Verma, R., Rathor, D., Kumar, S., Mishra, M., & Baranwal, M. (2025). Enhancing customer repurchase prediction: Integrating classification algorithms with RFM analysis for precision and actionable insights. *IIMB Management Review, 37*(2), 100574. https://doi.org/10.1016/j.iimb.2025.100574

Wong, C. G., Tong, G. K., & Haw, S. C. (2024). Exploring customer segmentation in e-commerce using RFM analysis with clustering techniques. *Journal of Telecommunications and the Digital Economy, 12*(3), 97-125. https://doi.org/10.18080/jtde.v12n3.978

Zhou, S., & Hudin, N. S. (2024). Advancing e-commerce user purchase prediction: Integration of time-series attention with event-based timestamp encoding and Graph Neural Network-Enhanced user profiling. *PLoS ONE*, *19*(4), e0299087. https://doi.org/10.1371/journal.pone.0299087