Journal of Finance - Marketing Research; Vol. 16, Issue 6; 2025
p-ISSN: 1859-3690; e-ISSN: 3030-427X
DOI: https://doi.org/10.52932/jfmr.v16i6

Journal of Finance - Marketing Research
http://jfm.ufm.edu.vn

BUILD A CUSTOMER TOUCHPOINT PREDICTION MODEL BASED ON SHOPPING JOURNEY DATA

Thai Kim Phung[1*], Lam Thi Bich Ngan[1]
[1]College of Technology and Design, University of Economics Ho Chi Minh City, Vietnam

ARTICLE INFO

DOI:
10.52932/jfmr.v16i6.882

Received:
April 09, 2025
Accepted:
June 02, 2025
Published:
December 25, 2025

Keywords:
Predict touchpoint; Purchase journey; Recommendation system; Website.

JEL codes:
C88, M31, D12, C45, C38, L81

ABSTRACT

This study focuses on building a predictive model for customer journey touchpoints and purchasing decisions on e-commerce websites, aiming to enhance personalized marketing strategies. The dataset includes online shopping journey data and customer demographic information in the tourism sector collected between May 2015 and October 2016. The research employs K-Means clustering to segment customers and identify target groups. Then, Collaborative Filtering with Low Rank Matrix Factorization is applied, followed by training a Neural Network to predict the frequency of customer touchpoints. Using the predicted frequency data, machine learning models such as Logistic Regression, Decision Tree, Random Forest, KNN, and XGBoost are implemented to predict purchase behavior. The results show that the Random Forest model outperforms others with the highest accuracy (96%), strong F1-score, and ROC-AUC metrics. The study contributes theoretically by integrating process mining and recommendation systems for journey prediction and offers a practical model applicable to businesses seeking data-driven insights into customer behavior. Future research is encouraged to expand prediction to all customer segments and incorporate additional contextual factors such as access devices and interaction duration to improve personalization.

*Corresponding author:
Email: phungthk@ueh.edu.vn

Tạp chí Nghiên cứu Tài chính - Marketing
http://jfm.ufm.edu.vn

XÂY DỰNG MÔ HÌNH DỰ ĐOÁN ĐIỂM CHẠM CỦA KHÁCH HÀNG DỰA TRÊN DỮ LIỆU HÀNH TRÌNH MUA SẮM TRỰC TUYẾN

Thái Kim Phụng[1*], Lâm Thị Bích Ngân[1]
[1]Trường Công nghệ và Thiết kế, Đại học Kinh tế Thành phố Hồ Chí Minh

THÔNG TIN

DOI:
10.52932/jfmr.v16i6.882

Ngày nhận bài:
09/04/2025
Ngày chấp nhận:
02/06/2025
Ngày đăng:
25/12/2025

Từ khóa:
Dự đoán điểm chạm; Hành trình mua sắm; Hệ thống đề xuất; Website.

Mã JEL:
C88, M31, D12, C45, C38, L81

TÓM TẮT

Nghiên cứu này tập trung xây dựng mô hình dự đoán điểm chạm trong hành trình mua sắm và dự đoán quyết định mua hàng của khách hàng trên website, góp phần cải thiện chiến lược marketing và cá nhân hóa trải nghiệm người dùng. Bộ dữ liệu sử dụng là hành trình mua sắm trực tuyến của khách hàng trong ngành du lịch, chứa thông tin về các điểm chạm lẫn đặc trưng nhân khẩu học, thu thập từ tháng 5/2015 đến 10/2016. Quy trình phân tích khởi đầu bằng việc áp dụng thuật toán K-Means nhằm phân cụm và xác định các phân khúc khách hàng mục tiêu. Trên cơ sở này, nghiên cứu lần lượt triển khai lọc cộng tác kết hợp phân rã ma trận hạng thấp (Low Rank Matrix Factorization) và huấn luyện Neural networks để dự đoán tần suất xuất hiện của từng điểm chạm trong tương lai. Từ tập dữ liệu dự đoán đó, các mô hình học máy bao gồm Logistic Regression, Decision Tree, Random Forest, KNN và XGBoost được huấn luyện để dự đoán quyết định mua hàng. Kết quả thực nghiệm cho thấy rằng, mô hình Random Forest là phương án vượt trội, đạt độ chính xác 96% cùng các chỉ số F1-score và ROC-AUC cao nhất. Về mặt học thuật, nghiên cứu đóng góp vào lĩnh vực khai phá hành trình khách hàng bằng cách tích hợp hệ thống đề xuất và khai thác quyết định, đồng thời đưa ra mô hình dự đoán thực tiễn có khả năng áp dụng cho nhiều loại hình doanh nghiệp. Định hướng nghiên cứu tương lai đề xuất mở rộng phân tích cho toàn bộ phân khúc khách hàng, đồng thời bổ sung các thuộc tính tương tác như thiết bị truy cập và thời gian truy cập nhằm gia tăng mức độ cá nhân hóa và độ chính xác của hệ thống khuyến nghị.

1. Giới thiệu

Ngày nay, các công ty thu thập tất cả các loại dữ liệu từ sản phẩm hoặc dịch vụ và khách hàng của họ, nhằm sử dụng để phân tích hành trình mua sắm của khách hàng (Habets, 2020). Việc nghiên cứu và phân tích hành trình mua sắm của khách hàng sẽ mang lại cho doanh nghiệp và khách hàng với những định hướng có thể tối ưu hóa các sản phẩm, dịch vụ cung cấp (Skulimowski & Kacprzyk, 1997). Hành trình khách hàng thể hiện các bước khách hàng tương tác với công ty trực tiếp hay gián tiếp qua

*Tác giả liên hệ:
Email: phungthk@ueh.edu.vn

các nền tảng. Mỗi bước được gọi là điểm chạm, được xác định là sự tương tác của khách hàng với sản phẩm hoặc dịch vụ của công ty (Bernard & Andritsos, 2017). Những hành trình của khách hàng được ánh xạ vào bản đồ hành trình khách hàng (Customer Journey Maps – CJM) để thực hiện phân tích (Habets, 2020).

Phân tích hành trình khách hàng là một trong những chủ đề đáng chú ý trong bộ phận truyền thông Marketing và thương mại điện tử, giúp hiểu hành vi khách hàng, hỗ trợ đưa ra quyết định trong các chiến lược, chiến dịch (Hernandez và cộng sự, 2017). Nhiều công ty dựa vào các nền tảng dịch vụ uy tín để quản lý hành trình khách hàng của mình, chẳng hạn như Adobe Experience (Adobe Experience Cloud) và Google Analytics (Google Marketing Platform). Tuy nhiên, các nền tảng của bên thứ ba này có thể không phù hợp với doanh nghiệp vì nhiều lý do, bao gồm chi phí sử dụng cao, không thể thích ứng của mô hình theo dữ liệu kinh doanh tuỳ doanh nghiệp vì theo nền tảng chung và phân tích đơn giản hóa quá mức (Dam và cộng sự, 2021). Để hỗ trợ phân tích hành trình của khách hàng, các kỹ thuật khai phá quy trình, bao gồm khám phá quy trình, phân cụm theo dõi và khai phá quyết định đã được sử dụng (Terragni & Hassani, 2018). Những kỹ thuật đó được tích hợp vào hệ thống để hỗ trợ doanh nghiệp phân tích tổng thể hành vi của khách hàng thông qua khám phá các nhóm hành trình phổ biến nhất và điều tra các đặc điểm của khách hàng ảnh hưởng đến các lựa chọn tương tác trong hành trình (Dam và cộng sự, 2021). Một hệ thống đề xuất (recommendation system) là một loại hệ thống lọc thông tin dự đoán 'đánh giá' hoặc 'sở thích' mà một khách hàng sẽ đưa ra cho một mục có thể là sản phẩm, website, điểm chạm (Pacuk và cộng sự, 2016).

Khả năng dự đoán chính xác và chủ động điểm tiếp xúc tiếp theo của từng khách hàng đóng vai trò quan trọng trong việc điều chỉnh hành trình khách hàng cá nhân, từ đó nâng cao trải nghiệm khách hàng và cuối cùng là thúc đẩy doanh thu (Singh & Singh, 2010). Tuy nhiên, việc thiếu các phương pháp chính thức và công cụ phân tích toàn diện để dự đoán và điều hướng hành trình khách hàng một cách tự động và có mục tiêu vẫn là một rào cản lớn đối với việc triển khai các giải pháp này trong thực tế (de Leoni và cộng sự, 2015).

Mục tiêu của nghiên cứu là phân tích, khám phá hành trình mua sắm của khách hàng trên website dựa trên dữ liệu các điểm chạm và thông tin nhân khẩu học. Phân cụm khách hàng và chọn phân khúc khách hàng mục tiêu, ứng dụng hệ thống đề xuất để dự đoán tần suất điểm chạm trong mỗi hành trình mua sắm khách hàng mục tiêu và dự đoán quyết định mua của phân khúc khách hàng tiềm năng này.

2. Cơ sở lý thuyết

2.1. Hành trình mua sắm của khách hàng

Bản đồ hành trình khách hàng (Customer Journey Map – CJM) là một biểu diễn tuyến tính, dựa trên thời gian các giai đoạn chính mà một khách hàng trải qua khi tương tác với một công ty hoặc dịch vụ (Mangiaracina và cộng sự, 2009). Trong phân tích hành trình khách hàng, các công ty tập trung vào cách khách hàng tương tác với nhiều điểm chạm, từ giai đoạn cân nhắc, tìm kiếm, mua hàng, sau mua hàng, sử dụng và tương tác hoặc mua lại trong tương lai. Mục tiêu chính của việc theo dõi trải nghiệm khách hàng là để tìm ra cách cải thiện các trải nghiệm (Lemon & Verhoef, 2016). Mô tả hành trình và hiểu các lựa chọn của khách hàng đối với các điểm chạm trong nhiều giai đoạn mua hàng khác nhau dẫn đến có nhiều biến thể của CJM (Halvorsrud và cộng sự, 2016).

Theo Bernard và Andritsos (2018), các thành phần chính thường có của một CJM bao gồm:

Khách hàng – người liên quan được hưởng lợi từ một dịch vụ, sản phẩm của doanh nghiệp.

Hành trình – một CJM chứa ít nhất một hành trình, là con đường điển hình mà một khách hàng tương tác với doanh nghiệp.

Mục tiêu – doanh nghiệp xác định khi lập bản đồ hành trình khách hàng, ví dụ như tìm điểm quyết định mua hàng, giảm tỷ lệ khách hàng rời bỏ.

Điểm chạm – điểm tiếp xúc tại một thời điểm mà khách hàng tương tác với công ty thông qua một sản phẩm hoặc dịch vụ, ví dụ khi một khách hàng truy cập website tìm kiếm một chuyến bay hoặc liên hệ với dịch vụ khách hàng.

Dòng thời gian – mô tả thời lượng và các mốc thời gian của hành trình trong một khoảng thời gian từ điểm chạm đầu tiên cho đến điểm chạm cuối cùng.

Kênh – là phương thức mà khách hàng chọn để tương tác với điểm chạm.

Các bản đồ hành trình khách hàng (CJM) có đặc điểm là cấu trúc phi tuyến tính, phản ánh các động lực nhận thức, cảm xúc và hành vi (Wolny & Charoensuksai, 2014). Để khám phá đúng nhu cầu và hành vi của các phân khúc khách hàng, đặc biệt là các phân khúc khách hàng mục tiêu, nghiên cứu sử dụng phân cụm khách hàng để quan sát. K-means Clustering được đề xuất là một trong những thuật toán phổ biến và điển hình trong phân cụm (Macqueen, 1967).

2.2. Hệ thống đề xuất

Thông thường, khai phá quy trình được xem là một ứng dụng độc lập để khám phá và đánh giá các quy trình, nhưng nó cũng có thể được sử dụng để hỗ trợ các đề xuất (Schonenberg và cộng sự, 2008). Áp dụng khai phá quy trình ngay lập tức, tức là bằng cách nhìn vào một hành trình (tập hợp các thực thi đã hoàn thành đầy đủ các điểm chạm) và một phần hành trình (một hoặc một vài điểm chạm đang tiếp tục được thực thi), và dự đoán tương lai của hành trình (các điểm chạm tương lai) (Rozinat & van Der Aalst, 2008). Cuối cùng, hệ thống đề xuất gửi lại cho khách hàng một danh sách các bước tiếp theo được đề xuất để hỗ trợ quyết định của khách hàng, tối ưu hóa thời gian tương tác hiệu quả hoặc giảm thiểu tỷ lệ rời bỏ *(xem Phụ lục 1 online)*.

Hệ thống đề xuất sử dụng phản hồi từ khách hàng làm đầu vào để cung cấp các gợi ý cá nhân hóa, có thể liên quan đến các quyết định trong nhiều quy trình (Ricci và cộng sự, 2010). Các thuật toán gợi ý này được thiết kế tùy thuộc vào lĩnh vực và các đặc điểm đặc trưng của dữ liệu có sẵn, thường ghi lại chất lượng của các tương tác tại điểm chạm của khách hàng (Melville & Sindhwani, 2010). Các tương tác này thường được gọi là phản hồi và có thể phân biệt thành phản hồi rõ ràng và tiềm ẩn (Aggarwal, 2016).

Phản hồi rõ ràng (explicit rating): thường được thực hiện thông qua các đánh giá, như hệ thống đánh giá năm sao. Phản hồi tiềm ẩn (implicit feedback): dễ dàng thu thập hơn vì việc thu thập hoàn toàn không đòi hỏi khách hàng hành động thêm. Sở thích của khách hàng được suy ra từ các hoạt động của họ thay vì các đánh giá được chỉ định một cách rõ ràng. Tuy nhiên, trong các đánh giá tiềm ẩn, không có thông tin nào cho biết nếu khách hàng không thích một điểm chạm, việc không mua hoặc không xem qua một điểm chạm không phải lúc nào cũng chỉ thể hiện sự không thích (Aggarwal, 2016).

2.3. Dự đoán điểm chạm

Dự đoán đồng thời điểm chạm và ngữ cảnh là một bài toán phân loại đa nhãn (và đa lớp), mang lại những thách thức phức tạp riêng (Habets, 2020). Chuỗi các điểm chạm của một khách hàng duy nhất tạo thành một hành trình khách hàng, được ghi nhận riêng biệt. Mặc dù tương tác tại các điểm chạm trước có thể không được cung cấp trực tiếp, nhưng nhờ vào việc các hành trình được liên kết qua mã nhận dạng duy nhất và mỗi bước đều được ghi lại, có thể suy ra các điểm chạm trong hành trình của khách hàng.

Phương pháp lọc cộng tác (*collaborative filtering*) dựa trên một phép phân rã ma trận (*matrix factorization*) (Hu và cộng sự, 2008).

R là ma trận đánh giá khách hàng của hành trình của khách hàng, được xác định trong công thức:

$$R = |C| \times |P| \quad (2)$$

trong đó, C là tập hợp các id trường hợp và P là tập hợp các điểm chạm website đang được phân tích.

Ma trận này chứa các tương tác $r_{c,p}$ đã xảy ra giữa id trường hợp c và tập hợp các trang p. Như vậy, hành trình của khách hàng có thể được coi là một phản hồi ngụ ý cho các trang đã được truy cập trong suốt hành trình. Cụ thể, một phần tử $r_{c,p}$ của ma trận R có thể đại diện cho một tương tác giữa khách hàng c và trang p theo các cách sau:

Boolean: $r_{c,p}$ = (True or False) nếu khách hàng c đã truy cập trang p hoặc không.

Tần suất truy cập: $r_{c,p}$ bằng số lần khách hàng c đã truy cập trang p trong hành trình hiện tại.

Thời lượng truy cập: $r_{c,p}$ bằng tổng thời lượng khách hàng đã truy cập trang p. Điều này có thể tính dễ dàng bằng cách trừ timestamp t_{i+1} của trang tiếp theo được truy cập và timestamp ti.

Các định nghĩa này dựa trên giả định logic rằng một khách hàng thích một trang sẽ dành nhiều thời gian hơn trên trang đó hoặc truy cập nó thường xuyên hơn (Aggarwal, 2016). Do đó, ngay cả khi không có bằng chứng rõ ràng về việc khách hàng thích hoặc không thích điểm chạm nào, vẫn có thể sử dụng những yếu tố này để suy luận. Phụ lục 2 trình bày ví dụ ma trận đánh giá khách hàng, đánh giá trực tiếp (trái) và đánh giá ngầm định (phải) *(xem Phụ lục 2 online)*.

Nhật ký các sự kiện trong hành trình khách hàng tại Bảng 1. Mỗi hành trình được ghi nhận với một ID riêng biệt, các mốc thời gian điểm chạm xảy ra sẽ có ghi nhận timestamp. Trong đó, điểm chạm ở ví dụ này được ghi nhận là đường dẫn URL tại trang website với hành động cụ thể của khách hàng tại điểm chạm ghi nhận thu thập tại cột Action.

Phụ lục 3 *(xem Phụ lục 3 online)* là một ví dụ về ma trận đánh giá khách hàng, Từ ma trận này, có thể thấy rằng, dựa trên hành vi của hành trình 1 và 2, cụ thể tần suất Trang Bài viết 456, các Trang sản phẩm ABC và DEF có thể được đề xuất cho hành trình 3, người vừa truy cập Trang Bài viết 456. Đây là một ví dụ tốt về việc cung cấp các đề xuất cá nhân dựa trên hành vi quá khứ của khách hàng. Tuy nhiên, ma trận đánh giá khách hàng thường rất lớn và do đó cần các thuật toán phức tạp hơn để xử lý. Hành trình của khách hàng có thể được coi là một phản hồi ngầm.

Bảng 1. Ma trận đánh giá khách hàng xây dựng trên số liệu bảng 1

	Article Page 123	Article Page 456	Product Page ABC	Product Page DEF
Case id 1	1	1	1	0
Case id 2	0	1	0	1
Case id 3	0	1	0	0

Nguồn: Aggarwal (2016)

Sử dụng mô hình Neural networks để xử lý đa lớp phức tạp với ma trận mã hóa mỗi hành trình mua sắm của khách hàng và tần suất điểm chạm, mô hình bao gồm một lớp nút đầu vào và một hoặc nhiều lớp ẩn và lớp đầu ra, các nút được nối với nút khác với trọng số và ngưỡng liên quan, là một công cụ để xác định các hệ thống không tuyến tính và có thể tự thích ứng bằng việc thay đổi các hệ số trong môi trường, tính toán tốt trên dữ liệu hơn do cấu trúc phân tán song song khi học huấn luyện (Sharkawy, 2020).

2.4. Mô hình dự đoán

Khai thác quyết định (decision mining) là quá trình làm giàu mô hình bằng cách áp dụng khai thác dữ liệu trong quá trình khai thác quy trình (Rozinat & van der Aalst, 2008). Bằng cách phát hiện điểm quyết định trong mô hình quy trình, việc khai thác quyết định có thể được chuyển đổi thành vấn đề phân loại để áp dụng các thuật toán học máy, chẳng hạn như cây quyết định. Phụ lục 4 *(xem Phụ lục 4 online)*

minh họa một ví dụ về khai thác quyết định trong phân tích hành trình của khách hàng.

Mô hình hồi quy logistic được đề xuất là một trong những phương pháp hồi quy dự đoán tỉ lệ xác suất của một biến phân loại nhị phân [0,1], ở nghiên cứu này là mua và không mua, dựa trên biến là tần suất của các điểm chạm trên hành trình mua sắm của khách hàng trên website (Cox, 1958).

Mô hình cây quyết định (Decision Tree) sử dụng thuật toán sử dụng phương pháp chia để trị (divide-and-conquer) xác định thuộc tính có tương quan và quan trọng nhất, phân chia các tập dữ liệu con đến khi không chia được nữa hoặc kết quả đạt được độ chính xác kỳ vọng (Breiman và cộng sự, 1984).

Mô hình rừng ngẫu nhiên (Random Forest) sử dụng thuật toán học có giám sát (supervised learning) nhiên được đề xuất dựa vào kỹ thuật tập hợp mô hình (ensemble learning) với nhiều cây quyết định được khởi tạo từ các tập dữ liệu con được chọn ngẫu nhiên của tập dữ liệu huấn luyện, xây dựng một bộ sưu tập lớn các cây không tương quan (decorrelated), sau đó tính trung bình của chúng (Breiman, 2002).

Mô hình KNN là một trong những thuật toán supervised-learning đơn giản nhất, còn được gọi là thuật toán học dựa trên trường hợp hoặc dựa trên ghi nhớ dữ liệu (Instance-based or Memory-based learning) (Fix & Hodges, 1951).

Mô hình tăng cường độ dốc cực đại (extreme gradient boosting – XG Boost) là thuật toán tập hợp mô hình (ensemble learning) kết hợp mô hình cây quyết định để tối ưu mô hình dự đoán bằng một cây quyết định tổng hợp qua các vòng lặp (Chen & Guestrin, 2016).

3. Phương pháp nghiên cứu

Chọn dữ liệu và tiền xử lý dữ liệu

Thu thập và lựa chọn các cột thông tin xử lý, thống kê và trực quan hóa bằng các thư viện: Pandas, Matplotlib, Seaborn để hiểu dữ liệu. Tìm và thay thế các dữ liệu thiếu bằng giá trị trung bình qua Simple Imputer. Xử lý dữ liệu mất cân bằng trong dữ liệu huấn luyện để mô hình có kết quả chính xác không bị lệch bằng Synthetic Minority Over-sampling (SMOTE), tạo mẫu nhằm gia tăng kích thước mẫu của nhóm thiểu số trong trường hợp xảy ra mất cân bằng mẫu.

Phân cụm và xác định phân khúc khách hàng

Sử dụng thuật toán K Means để phân cụm khách hàng thành các phân khúc có đặc trưng chung dựa theo các thông tin nhân khẩu học. Thống kê mô tả để trực quan phân bổ của các khách hàng mỗi phân khúc ở các thông tin nhân khẩu học. Từ đó, đánh giá các đặc trưng của các phân khúc khách hàng nhằm lựa chọn phân khúc khách hàng tiềm năng, là khách hàng mục tiêu để huấn luyện mô hình.

Dự đoán điểm chạm

Kiểm tra tần suất truy cập điểm chạm của mỗi lần truy cập mua sắm với ma trận utility matrix. Tiến hành áp dụng mô hình lọc cộng tác (collaborative filtering) được sử dụng để tạo ra các gợi ý điểm chạm dựa trên phản hồi tiềm ẩn từ tần suất điểm chạm của khách hàng tại mỗi hành trình mua sắm. Kết hợp sử dụng phương pháp phân rã ma trận (low rank matrix factorization) để tạo ra các nhúng (embeddings) cho các hành trình mua sắm và điểm chạm, biểu diễn dưới dạng các vector có số chiều thấp. Sau đó, thông qua mô hình mạng lưới nơron (Neural Network) sử dụng Keras, mô hình học cách dự đoán xem mỗi hành trình mua sắm của khách hàng, cụ thể sẽ có điểm chạm cụ thể nào dựa trên sự tương tự giữa nhúng của hành trình tại mỗi lần mua hàng và các nhúng của điểm chạm.

Huấn luyện mô hình dự đoán quyết định mua hàng

Tiến hành mã hóa nhãn cho các giá trị hạng mục trong dữ liệu nhân khẩu học của khách hàng, gán nhãn số (integer) cho các cột dữ liệu định dạng chữ (string) và danh sách (list) để huấn luyện các mô hình. Chuẩn hóa dữ liệu bằng MinMaxScaler và Standardization. Chia tập dữ liệu huấn luyện và kiểm tra. Các mô

hình học máy được huấn luyện trong nghiên cứu được tác giả thực hiện lập trình bằng ngôn ngữ Python để phân tích dữ liệu: Logistic Regression, Decision Tree, Random Forest, KNN, XGBoost.

Đo lường kết quả

Sau khi có kết quả huấn luyện mô hình, nhằm đánh giá độ chính xác của các mô hình học máy, nghiên cứu sử dụng ma trận nhầm lẫn (confusion matrix), kết hợp các chỉ số precision, recall, đường cong AUC-ROC và F1-Score (Tharwat, 2020). Ma trận nhầm lẫn (confusion matrix) biểu diễn số lượng các dự đoán và số lượng quan sát thực tế ở mỗi quyết định (mua hoặc không mua). Từ đó, xác định được phương án vượt trội, phù hợp với dữ liệu và ngữ cảnh hiện tại của nghiên cứu.

Bảng 2. Ma trận nhầm lẫn với tập dữ liệu có 2 lớp được gán nhãn

Thực tế/Dự đoán	Lớp dương (Positive)	Lớp âm (Negative)
Lớp dương (Positive)	True Positive (TP)	False Negative (FN)
Lớp âm (Negative)	False Positive (FP)	True Negative (TN)

Nguồn: Tharwat (2020)

Precision (Tỷ lệ lớp dương đoán đúng): Trong tất cả các dự đoán lớp dương (Positive) được đưa ra, bao nhiêu dự đoán là chính xác.

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

Recall (Tỷ lệ lớp dương thực): Trong tất cả các trường hợp Positive, bao nhiêu trường hợp đã được dự đoán chính xá.

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

Accuracy (Độ chính xác): tỉ lệ dự đoán quyết định (mua/không mua) là chính xác.

$$\text{Accuracy} = \frac{TN + TP}{TN + TP + FP + FN} \quad (2.4)$$

F1 là số dung hòa Recall và Precision giúp ta có căn cứ để lựa chọn mô hình tốt nhất. F1 càng cao mô hình càng tốt.

$$\text{F1} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (2.5)$$

Đường cong ROC (Receiver operating characteristic): Thể hiện sự tương quan giữa Precision và Recall khi thay đổi giá trị ngưỡng (threshold). Đường cong AUC-ROC (Area Under the ROC): Là vùng nằm dưới ROC, vùng này càng lớn thì mô hình lựa chọn càng tốt theo (Fawcett, 2006).

4. Kết quả nghiên cứu

4.1. Kết quả thu thập và tiền xử lý dữ liệu

Dữ liệu hành trình mua sắm trực tuyến trên website du lịch ứng dụng trong nghiên cứu cũng đã được dùng trong đề tài nghiên cứu Exploring Customer Journey Mining and RPA: Prediction of Customers' Next Touchpoint (Wiethölter và cộng sự, 2023). Dữ liệu sau xử lý, được mẫu dữ liệu nghiên cứu ví dụ ở Phụ lục 6 *(xem Phụ lục 6 online)* gồm 134.037 dòng bao gồm thông tin nhân khẩu học của khách hàng đã được các điểm chạm trên hành trình mua sắm của khách hàng trên website từ 5/2015 đến tháng 10/2016.

4.2. Kết quả phân cụm và xác định phân khúc khách hàng mục tiêu

Phân cụm khách hàng vào các phân khúc khách hàng có đặc trưng tương tự dựa trên các thuật toán phân cụm KMeans và KModes, theo các trường dữ liệu nhân khẩu học: 'Gender', 'Age Range', 'Region', 'Size of Household', 'Occupation', 'Gross income per year in €',

'Number of Children', 'Social class', 'Education (completed)', 'Lifestage', được kết quả số phân khúc khách hàng và số lượng khách hàng ở mỗi phân khúc trình bày ở bảng 3.

Bảng 3. Kết quả phân cụm khách hàng

Cụm	0	1	2	3	4
Số lượng dòng dữ liệu (khách hàng)	22338	23610	42930	25005	20154

Tiến hành trực quan dữ liệu của các cụm khách hàng theo trường dữ để đánh giá sự phân bổ của các đặc trưng nhân khẩu học (vùng, độ tuổi, giới tính, số thành viên trong gia đình, trình độ học vấn đã tốt nghiệp, nghề nghiệp, thu nhập, trạng thái, số con và tầng lớp xã hội) của khách hàng ở từng cụm khách hàng và lựa chọn cụm khách hàng phù hợp, là tập khách hàng đặc trưng, có số lượng lớn và là khách hàng tiềm năng của doanh nghiệp *(xem Phụ lục 7 và 8 online)*. Với số lượng lớn để có thể huấn luyện mô hình, phân khúc 2 có những đặc trưng chung về giới tính, vùng của toàn bộ dữ liệu, và có các đặc tính phù hợp với sản phẩm du lịch mà doanh nghiệp hướng đến cho phân khúc khách hàng mục tiêu. Trong đó, độ tuổi tập trung 35-54 tuổi, tình trạng có gia đình và có số thành viên trong gia đình phù hợp, trình độ, thu nhập và nghề nghiệp cũng phù hợp với giá và sản phẩm du lịch của doanh nghiệp.

4.3. Kết quả dự đoán điểm chạm trên hành trình mua sắm

Tần suất điểm chạm của mỗi lần truy cập mua sắm với ma trận utility matrix trình bày tại bảng 4. Với cột đầu tiên là thông tin hành trình khách hàng theo từng ID. Hàng trên cùng là thông tin các điểm chạm từ 1-16. Tương ứng tần suất của mỗi điểm chạm trong mỗi hành trình được thống kê cụ thể để sử dụng huấn luyện mô hình.

Bảng 4. Tần suất điểm chạm ở mỗi hành trình (utility matrix)

Touch	1	2	3	4	5	6	7	8	9	10	12	13	14	15	16
Purchase ID															
12	2,0	0	0	0	0	0	1,0	0	0	0	0	0	0	0	3,0
13	1,0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	5,0	0	2,0	1,0	0	0	5,0	0	3,0	0	0	0	0	0	3,0
15	0	0	0	1,0	0	0	1,0	0	0	0	0	0	0	0	0
17	1,0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
28981	0	0	0	1,0	0	0	0	0	0	0	0	0	0	0	0
28982	3,0	0	0	3,0	0	0	0	0	0	0	0	0	0	0	0
28994	1,0	0	0	1,0	0	0	0	0	0	0	0	0	0	0	0
29005	1,0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29006	1,0	0	0	0	0	0	1,0	0	0	0	0	0	0	0	1,0
6248 rows × 15 columns															

Sau khi thử và tìm kiếm các chỉ số tối ưu, mô hình bao gồm 3 tầng chính *(i)* tầng đầu tiên có 20 neuron với hàm kích hoạt rectified linear unit – ReLU và sử dụng regularization L1 với hệ số 0,0001; *(ii)* tầng bỏ qua một vài đơn vị (dropout) với tỷ lệ dropout là 0,4 để ngăn chặn

kết quả huấn luyện bị quá khớp; *(iii)* tầng có 1 neuron với ReLU. Tầng dropout thứ hai với tỷ lệ dropout là 0,5.

Huấn luyện mô hình với 120 epochs và batch size là 1000, sử dụng 20% dữ liệu huấn luyện để làm tập xác thực.

Kết quả thu được:

Epoch 1/120: 13,8057 – val_loss: 14,1167;

Epoch 120/120: 5,9277 – val_loss: 5,7892;

Loss: 5,5989;

Test loss: 5,789198398590088.

Từ epoch đầu tiên đến epoch cuối cùng, độ mất mát (loss) trên tập huấn luyện đã giảm từ 13,8057 xuống 5,9277, còn độ mất mát trên tập xác thực (validation loss) cũng giảm từ 14,1167 xuống 5,7892, biểu diễn biểu đồ mất mát của mô hình *(xem Phụ lục 9 online)*. Điều này cho thấy mô hình đã học được mối quan hệ phức tạp giữa dữ liệu đầu vào và đầu ra trên tập huấn luyện.

Độ mất mát của tập xác thực thấp hơn so với độ mất mát trên tập huấn luyện, cho thấy mô hình đã được huấn luyện, và có khả năng dự đoán tốt trên dữ liệu mới *(xem Phụ lục 9 online)*. Trục hoành (epoch) là số vòng lặp huấn luyện toàn bộ dữ liệu. Mỗi epoch tương ứng với một lần Neural networks đi hết tập huấn luyện, cập nhật trọng số nhờ thuật toán tối ưu. Trục tung (loss) là giá trị hàm mất mát đo sai số giữa dự đoán của mô hình và kết quả thật. Loss càng thấp càng chứng tỏ mô hình khớp dữ liệu càng tốt. Đường xanh lá thể hiện loss trên tập huấn luyện, đường xanh dương là loss trên tập kiểm thử. Tiến trình học dần theo thời gian, càng về bên phải, mạng càng trải qua nhiều chu kỳ huấn luyện, cho ra kết quả loss tốt hơn, cả hai tập huấn luyện và kiểm thử giảm từ khoảng 14 đến khoảng 5,7 cho thấy mô hình học được thông tin dù không quá rõ ràng nhưng có khả năng bị quá khớp (overfitting) khi huấn luyện thêm epoch hoặc tăng số lượng lớn phức tạp mô hình. Kiểm tra dự đoán hành trình mua sắm và điểm chạm ngẫu nhiên:

Chọn ngẫu nhiên ta có PurchaseID: 5615, và lựa chọn touch: 1 để tiến hành dự đoán.

Dự đoán tần suất từ mô hình được huấn luyện, touch 1 cho PurchaseID 5615 có tần suất là 2,480213165283203.

Sau đó kiểm tra tần suất thực tế của touch 1 cho PurchaseID 5615 là 2,00.

Kết quả hiệu chuẩn cho thấy giá trị dự đoán của mô hình (2,48) xấp xỉ giá trị quan sát thực tế (2,00), hàm ý sai lệch tuyệt đối chỉ 0,48 đơn vị. Khoảng chênh lệch nhỏ này chứng thực năng lực khái quát hóa của mô hình và củng cố tính khả thi khi triển khai thực tế nhằm ước lượng tần suất của các điểm chạm trong hành trình mua sắm trực tuyến.

4.4. Kết quả huấn luyện và đánh giá mô hình dự đoán quyết định

Huấn luyện các mô hình đã liệt kê dự đoán quyết định mua hàng dựa theo điểm chạm, và tần suất điểm chạm tại mỗi hành trình mua sắm của khách hàng phân khúc 2, nghiên cứu được kết quả theo bảng 5. Các chỉ số đánh giá Precision, Recall, F1-score cho hai quyết định không mua hàng và mua hàng, sau đó đánh giá chỉ số Accuracy và ROC-AUC.

Bảng 5. Kết quả huấn luyện mô hình dự đoán quyết định mua hàng

Mô hình	Dự đoán không mua hàng			Dự đoán mua hàng			Accuracy	ROC-AUC
	Precision	Recall	F1-score	Precision	Recall	F1-score		
Logistic Regression	82%	69%	75%	43%	60%	50%	66%	64%
Decision Tree	97%	97%	97%	92%	93%	92%	96%	94,9%
Random Forest	97%	97%	97%	92%	93%	92%	96%	95%
Kneighbors Classifier	96%	99%	97%	96%	89%	93%	96%	94%
XG Boost	94%	93%	93%	82%	86%	84%	91%	89%

Dựa vào kết quả ở bảng 9, mô hình Decision Tree và Random Forest cho kết quả tốt nhất so với các mô hình còn lại. Random Forest có ROC-AUC score tốt hơn một chút và dự đoán đúng nhiều trường hợp hơn so với Decision Tree. Nghiên cứu chọn mô hình Random Forest và biểu diễn ma trận nhầm lẫn (confusion matrix) ở biểu đồ 2 để so sánh kết quả thực tế và dự đoán *(xem Phụ lục 10 online)*. Số lượng dự đoán không mua sai với thực tế là 164 trường hợp (so với 5982 trường hợp dự đoán đúng), đối với dự đoán mua, mô hình dự đoán sai với thực tế 202 trường hợp (so với 2238 trường hợp dự đoán đúng).

Mô hình dự đoán quyết định mua hàng đạt hiệu quả cao với độ chính xác 96% và ROC-AUC 0.95. Số mẫu được dự đoán đúng ở cả hai quyết định (mua và không mua) đều cao, trong khi số nhầm lẫn thấp. Precision và Recall cho quyết định không mua lần lượt là 97% và 97%, cao hơn một chút so với quyết định mua (92% và 93%). F1-score cho cả hai nhóm quyết định mua và không mua đều cao (97% và 92%), cho thấy mô hình cân bằng tốt giữa độ chính xác và khả năng bao phủ.

4.5. Thảo luận kết quả nghiên cứu

So với nghiên cứu của Wiethölter và cộng sự (2023), vốn cũng sử dụng dữ liệu hành trình khách hàng và phân tích điểm chạm để dự đoán hành vi, nghiên cứu này mở rộng hướng tiếp cận bằng cách kết hợp phân cụm khách hàng bằng K-Means và áp dụng Neural networks để dự đoán tần suất điểm chạm trước khi dự đoán hành vi mua hàng. Điểm mới của nghiên cứu là mô hình hóa song song hai giai đoạn – dự đoán điểm chạm và dự đoán quyết định mua – thay vì chỉ dự đoán điểm chạm tiếp theo như các nghiên cứu trước.

So sánh với mô hình Cây Quyết định cho thấy Random Forest, nhờ cơ chế bagging, đã triệt tiêu đáng kể phương sai, qua đó mang lại độ ổn định cao hơn trong các thử nghiệm triển khai thực tế (A/B test). Độ chính xác 96 % của Random Forest cao hơn 1,1 % so với kết quả công bố của Wiethölter và cộng sự (2023) trên cùng bộ dữ liệu, từ đó khẳng định hiệu quả của quy trình xử lý hai phần lần lượt, từ dự đoán điểm chạm đến dự đoán hành vi mua hàng.

Ngoài ra, phương pháp xử lý dữ liệu mất cân bằng bằng SMOTE cũng góp phần cải thiện hiệu năng mô hình, điều mà nhiều nghiên cứu trước chưa chú trọng. Kết quả cũng phù hợp với các phát hiện của Aggarwal (2016) về vai trò của phản hồi ngầm định trong hệ thống đề xuất.

Kết quả nghiên cứu cho thấy mô hình Random Forest có hiệu suất dự đoán cao với độ chính xác đạt 96% và ROC-AUC 0.95, vượt trội hơn so với các mô hình khác như Decision Tree, Logistic Regression hay KNN. Điều này khẳng định hiệu quả của phương pháp kết hợp hệ thống đề xuất (recommendation system) với học máy trong dự đoán hành vi mua hàng.

Khi bóc tách ma trận nhầm lẫn của mô hình Random Forest, có thể nhận thấy hai điểm nổi bật. Đầu tiên, tỷ lệ True Positive cao, cho thấy hệ thống nhận diện dự đoán chính xác khách hàng mua, giúp marketing không lãng phí ngân sách remarketing vào nhóm đã chuyển đổi. Bên cạnh đó, tỷ lệ False Negative thấp, mô hình hiếm khi bỏ sót khách hàng tiềm năng, nhờ đó tối đa hóa doanh thu tiềm ẩn.

Phân tích độ quan trọng của thuộc tính, cũng cho thấy được hai yếu tố dẫn đến hành vi mua hàng mạnh nhất. Rõ ràng nhất, tần suất của điểm chạm 7 (trang thông tin về đại lý, công ty du lịch đối thủ) càng nhiều lượt xem sản phẩm, xác suất mua càng cao, từ đó có thể ưu tiên cá nhân hóa trang đích và tập trung cải thiện tối ưu thông tin tại trang. Về nhân khẩu học, thông tin vùng và thu nhập cho thấy khả năng chi trả và vị trí địa lý cũng tác động đến quyết định rõ rệt, với dữ liệu này, nên tập trung phân bổ quảng cáo và tối ưu thông tin cho phân khúc phù hợp.

5. Kết luận và hàm ý quản trị

5.1. Kết luận

Nghiên cứu đã xây dựng thành công mô hình dự đoán điểm chạm và quyết định mua hàng

của khách hàng dựa trên dữ liệu hành trình mua sắm trực tuyến trong lĩnh vực du lịch. Mô hình đề xuất gồm hai phần, phần 1 dự đoán tần suất điểm chạm thông qua kết hợp Low-Rank Matrix Factorization và Neural networks, phần 2 dự đoán quyết định mua hàng qua các thuật toán học máy. Phương pháp tiếp cận kết hợp giữa phân cụm khách hàng, hệ thống đề xuất (recommendation system) và các thuật toán học máy như Neural Network, Random Forest, XGBoost đã giúp tăng độ chính xác trong việc nhận diện hành vi khách hàng và hỗ trợ doanh nghiệp định hướng chiến lược marketing hiệu quả hơn.

Trên bộ dữ liệu hành trình mua sắm trực tuyến trong ngành du lịch (05/2015 – 10/2016), mô hình Random Forest đạt hiệu năng vượt trội, đồng thời chứng minh khả năng tổng hợp thông tin nhân khẩu học và hành vi để ước lượng chính xác xác suất mua. Kết quả này góp phần mở rộng về khai phá hành trình khách hàng, cung cấp bằng chứng thực nghiệm cho giá trị của việc tích hợp hệ thống đề xuất với phân tích quyết định.

Về mặt lý thuyết, nghiên cứu đóng góp bằng việc tích hợp khai phá quy trình (process mining) và mô hình dự đoán điểm chạm – một hướng tiếp cận còn ít được khai thác trong bối cảnh thương mại điện tử. Về mặt thực tiễn, mô hình cho thấy tiềm năng cao trong việc hỗ trợ các doanh nghiệp tối ưu hóa hành trình khách hàng và cá nhân hóa trải nghiệm mua sắm trên nền tảng số.

5.2. Hàm ý quản trị

Doanh nghiệp cần tối ưu ngân sách đa kênh đối với phân khúc khách hàng tiềm năng phù hợp với định hướng doanh nghiệp và định vị thương hiệu. Từ đó tổ chức hệ thống dữ liệu, thu thập các thông tin hành trình mua sắm của khách hàng đáp ứng mục tiêu phân tích. Thông qua phân tích, tiến hành kết nối với website để đo lường và đề xuất điều chỉnh kịch bản cá nhân hóa theo thời gian thực (real-time), tối ưu hoặc kết hợp các hành động tư vấn và tăng tương tác tại điểm chạm quan trọng trong hành trình. Dựa vào kết quả phân tích và đề xuất, doanh nghiệp cân nhắc thiết kế tối ưu giao diện và hành trình khách hàng tốt hơn, có thể mở rộng hoặc rút ngắn hành trình khách hàng tại các điểm mấu chốt, lược bớt hoặc gộp các điểm chạm không cần thiết.

Theo lộ trình phát triển dài hạn, doanh nghiệp phân tích và phân bổ đầu tư cho lưu trữ dữ liệu, thời gian truy xuất nhanh chóng hỗ trợ trong mô hình học máy và các kỹ thuật xử lý tối ưu. Mở rộng ngoài nghiên cứu có thể phân tích thêm các thông tin ngữ cảnh như thiết bị dùng, thời gian truy cập để có đánh giá cụ thể cho từng hành trình và hành vi.

5.3. Hạn chế và hướng nghiên cứu tương lai

Tuy nhiên, nghiên cứu vẫn còn một số hạn chế. Đầu tiên là hạn chế về thời gian và lĩnh vực, dữ liệu hiện tại được sử dụng đang giới hạn chỉ trong ngành du lịch và áp dụng xử lý tập trung một phân khúc khách hàng mục tiêu, chưa phản ánh được sự đa dạng hành vi trên nhiều lĩnh vực khác nhau cũng như kiểm tra được độ chính xác của mô hình đối với dữ liệu đa dạng hơn của những phân khúc khách hàng khác nhau. Thứ hai, nghiên cứu hạn chế thiếu thuộc tính ngữ cảnh, mô hình hiện tại tập trung vào tần suất điểm chạm mà chưa khai thác sâu các yếu tố như thời lượng tương tác, thứ tự và thời gian thao tác, thiết bị sử dụng hay mức độ phản hồi của người dùng. Các nghiên cứu sau đề xuất nghiên cứu các thuộc tính thời gian, thứ tự điểm chạm và mức độ tương tác tại các điểm chạm nhằm xác định rõ hơn các yếu tố ảnh hưởng trực tiếp đến quyết định mua hàng của khách hàng.

Định hướng nghiên cứu tương lai đề xuất mở rộng dữ liệu sang nhiều ngành hàng khác nhau để kiểm nghiệm tính linh hoạt của mô hình. Mở rộng nghiên cứu các thuộc tính nhân khẩu học như thiết bị và ngữ cảnh để nâng cao tính cá nhân hóa trong huấn luyện mô hình. Ngoài ra, việc ứng dụng các kỹ thuật học sâu theo chuỗi thời gian như Long Short-Term Memory (LSTM) hoặc Transformer cũng là một hướng đi tiềm năng nhằm nâng cao độ

chính xác trong dự đoán hành vi khách hàng theo thời gian thực. Trong đó, LSTM là một biến thể nâng cao của mạng nơ-ron hồi quy (Recurrent Neural Network, RNN), được thiết kế để học các phụ thuộc dài hạn trong chuỗi dữ liệu với các cơ chế cho phép điều chỉnh lưu giữ hay loại bỏ thông tin trong quá trình lan truyền. Transformer là một kiến trúc Neural networks dựa hoàn toàn trên cơ chế tự chú ý (self-attention), thực hiện tính toán song song trên toàn bộ chuỗi, nhờ đó giảm thời gian huấn luyện và cải thiện khả năng mô hình hóa quan hệ phụ thuộc ở mọi khoảng cách. Các nghiên cứu tương lai khi lựa chọn mô hình phải gắn liền đặc tính chuỗi sự kiện trong hành trình và giới hạn tài nguyên của từng ngành để lựa chọn mô hình và kỹ thuật phù hợp.

Tài liệu tham khảo

Aggarwal, C. C. (2016). *Recommender systems: The textbook* (1st ed.). Springer. https://doi.org/10.1007/978-3-319-29659-3

Bernard, G. & Andritsos, P. (2017). A process mining based model for customer journey mapping. In *Forum and doctoral consortium papers presented at the 29th International Conference on Advanced Information Systems Engineering (CAiSE 2017)* (Vol. 1848, pp. 49-56). CEUR Workshop Proceedings. https://api.unil.ch/iris/server/api/core/bitstreams/4a163d97-833d-454a-a9db-5380c2973948/content

Bernard, G., & Andritsos, P. (2018). CJM-ab: Abstracting customer journey maps using process mining. In J. Mendling & H. Mouratidis (Eds.), *Information systems in the big data era* (CAiSE 2018, Lecture Notes in Business Information Processing, Vol. 317, pp. 65–80). Springer. https://doi.org/10.1007/978-3-319-92901-9_5

Breiman, L. (2002). *Manual on setting up, using, and understanding random forests v3.1*. Statistics Department University of California Berkeley.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. J. (1984). *Classification and regression trees* (1st ed.). Wadsworth, Inc.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). https://doi.org/10.1145/2939672.2939785

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, *20*(2), 215-232. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x

Dam, N. A. K., Dinh, T. L., & Menvielle, W. (2021). Towards a conceptual framework for customer intelligence in the era of big data. *International Journal of Intelligent Information Technologies (IJIIT)*, *17*(4), 64-80. https://doi.org/10.4018/IJIIT.289968

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Fix, E., & Hodges, J.L. (1951). *Discriminatory analysis: nonparametric discrimination: Consitency properties* (Technical Report No. 4). USAF School of Aviation Medicine.

Habets, S. (2020). *Predicting a customer's next touch point from customer journey data*. Eindhoven University of Technology, Netherlands. https://pure.tue.nl/ws/portalfiles/portal/174215023/Habets_S..pdf

Halvorsrud, R., Kvale, K., & Følstad, A. (2016). Improving service quality through customer journey analysis. *Journal of Service Theory and Practice*, *26*(6), 840-867. https://doi.org/10.1108/JSTP-05-2015-0111

Hernandez, S., Alvarez, P., Fabra, J., & Ezpeleta, J. (2017). Analysis of users' behavior in structured e-commerce websites. *IEEE Access, 5,* 11941–11958. https://doi.org/11941-11958. 10.1109/ACCESS.2017.2707600

Hu, Y., Koren, Y., & Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In *Proceeding of the 2008 Eighth IEEE International Conference on Data Minin*g (pp. 263-272). IEEE. https://doi.org/10.1109/ICDM.2008.22

Kabir, S., Mudur, S. P., & Shiri, N. (2012). Capturing browsing interests of users into web usage profiles. In *Intelligent Techniques For Web Personalization And Recommender Systems, AAAI Workshop* (pp.19-25). AAAI Press.

Lemon, K. N., & Verhoef, P. C. (2016). Understanding customer experience throughout the customer journey. *Journal of Marketing*, *80*(6), 69-96. https://doi.org/10.1509/jm.15.0420

de Leoni, M., Maggi, F. M. & van der Aalst, W. M. P. (2015). An alignment-based framework to check the conformance of declarative process models and to preprocess event-log data. *Information Systems*, *47*, 258-277. https://doi.org/10.1016/j.is.2013.12.005

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. L. Cam, J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281-297). http://projecteuclid.org/euclid.bsmsp/1200512992

Mangiaracina, R., Brugnoli, G., & Perego, A. (2009). The ecommerce customer journey: A model to assess and compare the user experience of the ecommerce websites. *Journal of Internet Banking and Commerce*, *14*(3), 1-11.

Melville, P. & Sindhwani, V. (2010). Recommender systems. In: C. Sammut, G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 829-838). https://doi.org/10.1007/978-0-387-30164-8_705

Pacuk, A., Sankowski, P., Węgrzycki, K., Witkowski, A., & Wygocki, P. (2016). RecSys Challenge 2016: Job recommendations based on preselection of offers and gradient boosting. *Proceedings of the Recommender Systems Challenge (RecSys Challenge '16)*. Association for Computing Machinery. *https://doi.org/10.1145/2987538.2987544*

Ricci, F., Rokach, L., & Shapira, B. (2010). Introduction to recommender systems handbook. *Recommender systems handbook* (pp. 1-5). Springer. https://doi.org/10.1007/978-0-387-85820-3_1

Rozinat, A., & van der Aalst, W. M. P. (2008). Conformance checking of processes based on monitoring real behavior. *Information Systems*, *33*(1), 64-95. https://doi.org/10.1016/j.is.2007.07.001

Schonenberg, H., Weber, B., van Dongen, B. & van Der Aalst, W. (2008). Supporting flexible processes through recommendations based on history. In M. Dumas, M. Reichert, M. C. Shan (Eds.), *Business Process Management: 6th International Conference*, BPM 2008, Milan, Italy, September 2-4, 2008 (pp. 51-55), Springer. https://doi.org/10.1007/978-3-540-85758-7_7

Singh, H. B. & Singh, H. K. (2010). Web Data Mining research: A survey. In the *2010 IEEE International Conference on Computational Intelligence and Computing Research* (pp. 1–10). IEEE. https://doi.org/10.1109/ICCIC.2010.5705856

Sharkawy, A.-N. (2020). Principle of neural network and its main types: Review. *Journal of Advances in Applied & Computational Mathematics*, *7*, 8-19. https://doi.org/10.15377/2409-5761.2020.07.2

Skulimowski, A. & Kacprzyk, J. (1997). Knowledge, information and creativity support systems: Recent trends, advances and solutions. *Proceedings of the KICSS2013-8th International Conference on Knowledge, Information, and Creativity Support Systems*, *November 7-9, 2013, Kraków, Poland*. Springer. https://doi.org/10.1007/978-3-319-19090-7

Terragni, A. & Hassani, M. (2018). Analyzing customer journey with process mining: From discovery to recommendations. *2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud)* (pp. 224-229). IEEE. https://doi.org/10.1109/FiCloud.2018.00040

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168-192. https://doi.org/10.1016/j.aci.2018.08.003

Wiethölter, J., Salingré, J., Feldmann, C., Schwanitz, J., & Niessing, J. (2023). Exploring customer journey mining and RPA: Prediction of customers' next touchpoint. In J. Köpke, O. López-Pintado, R. Plattfaut, J.-R. Rehse, K. Gdowska, F. Gonzalez-Lopez, J. Munoz-Gama, K. Smit, & J. M. E. M. van der Werf (Eds.), *Business process management: Blockchain, robotic process automation and educators forum (BPM 2023)* (Lecture Notes in Business Information Processing, Vol. 491, pp. 181–196). Springer. https://doi.org/10.1007/978-3-031-43433-4_12

Wolny, J. & Charoensuksai, N. (2014). Mapping customer journeys in multichannel decision-making. *Journal of Direct, Data and Digital Marketing Practice*, *15*, 317-326. https://doi.org/10.1057/dddmp.2014.24