HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

TRẦN PHÚC ĐỊNH

NGHIÊN CỨU PHÂN TÍCH HÀNH VI NGƯỜI DÙNG BỎ GIỎ HÀNG TRÊN CÁC TRANG THƯƠNG MẠI ĐIỆN TỬ

LUẬN VĂN THẠC SĨ KỸ THUẬT
(Theo định hướng ứng dụng)

Hà Nội - 2021

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

TRẦN PHÚC ĐỊNH

NGHIÊN CỨU PHÂN TÍCH HÀNH VI NGƯỜI DÙNG BỎ GIỎ HÀNG TRÊN CÁC TRANG THƯƠNG MẠI ĐIỆN TỬ

CHUYÊN NGÀNH: **KHOA HỌC MÁY TÍNH**
MÃ SỐ: **8.48.01.01**

LUẬN VĂN THẠC SĨ KỸ THUẬT

NGƯỜI HƯỚNG DẪN KHOA HỌC:

PGS. TS. Trần Đình Quế

Hà Nội - 2021

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn này là công trình nghiên cứu độc lập của riêng tôi dưới sự hướng dẫn của PGS.TS Trần Đình Quế. Các số liệu, kết quả nêu trong luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác. Các nguồn tài liệu trích dẫn được liệt kê trong danh mục tài liệu tham khảo theo đúng quy định.

Tôi xin chịu trách nhiệm về tính chính xác và trung thực của luận văn này.

Học viên

Trần Phúc Định

LỜI CẢM ƠN

Lời đầu tiên, xin trân trọng cảm ơn thầy đã hướng dẫn tôi là PGS. TS. Trần Đình Quế, thầy đã tận tình hướng dẫn tôi trong quá trình nghiên cứu cũng như trong việc hoàn thành luận văn.

Xin chân thành cảm ơn các Thầy, Cô thuộc Khoa Đào tạo Sau đại học – Học Viện Công Nghệ Bưu Chính Viễn Thông đã tận tình giảng dạy cho tôi trong thời gian học tập.

Do giới hạn kiến thức và khả năng lý luận của bản thân còn nhiều thiếu sót và hạn chế, kính mong sự chỉ dẫn và đóng góp của các Thầy, Cô để bài luận văn của tôi được hoàn thiện hơn. Xin chân thành cảm ơn!

Hà Nội, ngày 18 tháng 05 năm 2021

Trần Phúc Định

MỤC LỤC

MỞ ĐẦU **1**

CHƯƠNG I: TỔNG QUAN VỀ HÀNH VI NGƯỜI DÙNG TRONG THƯƠNG MẠI ĐIỆN TỬ **4**

1.1. Giới thiệu về khái niệm, sự ra đời, hình thành và phát triển của thương mại điện tử 4

1.1.1. Khái niệm về thương mại điện tử 4

1.1.2. Sự ra đời, hình thành và phát triển của thương mại điện tử 5

1.2. Tiềm năng khai thác dữ liệu người dùng trong thương mại điện tử 6

1.2.1. Phân tích giỏ hàng điện tử 6

1.2.2. Dự đoán nhu cầu thị trường 8

1.2.3. Đánh giá phân khúc thị trường 9

1.2.4. Phòng chống gian lận thương mại 10

1.3. Giới thiệu giỏ hàng điện tử và hành vi bỏ rơi giỏ hàng 10

1.3.1. Khái niệm giỏ hàng điện tử 10

1.3.2. Khuynh hướng sử dụng giỏ hàng điện tử của người tiêu dùng 11

1.3.3. Bỏ rơi giỏ hàng trong mua sắm trực tuyến 12

1.4. Kết luận 13

CHƯƠNG 2: PHÂN TÍCH HÀNH VI BỎ RƠI GIỎ HÀNG **14**

2.1. Các yếu tố chính quyết định bỏ rơi giỏ hàng 14

2.1.1. Trải nghiệm người dùng không tốt 14

2.1.2. Chi phí vận chuyển cao, đơn hàng không minh bạch 15

2.1.3. Nhận thức rủi ro từ người dùng trực tuyến 17

2.2. Thuật toán cây quyết định và rừng ngẫu nhiên 18

2.2.1. Thuật toán cây quyết định 18

2.2.2. Thuật toán rừng ngẫu nhiên 20

2.3. Ứng dụng học máy trong dự đoán người dùng bỏ rơi giỏ hàng 22

2.3.1. Dữ liệu và bối cảnh thử nghiệm 23

2.3.2. So sánh kết quả thuật toán 25

2.4. Kết luận.....27
CHƯƠNG 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ.....28
3.1. Phát biểu bài toán.....30
3.1.1. Bài toán phân tích và dự đoán phân luồng trực tiếp.....30
3.1.2. Bài toán thống kê dữ liệu phân luồng gián tiếp.....31
3.1.3. Tổng hợp bài toán và trình tự phân tích.....32
3.2. Cấu trúc hệ thống và dữ liệu.....34
3.2.1. Cấu trúc trang thương mại điện tử và dịch vụ phân tích.....34
3.2.2. Cấu trúc dữ liệu.....36
3.3. Thử nghiệm và đánh giá.....39
3.3.1. Thống kê và phân tích khuôn mẫu dữ liệu.....39
3.3.2. Thử nghiệm thực tế.....42
3.4. Kết luận.....45
KẾT LUẬN.....47
DANH MỤC TÀI LIỆU THAM KHẢO.....48

DANH MỤC BẢNG BIỂU

Bảng 2.1: Bảng mô tả và chỉ mục dữ liệu dùng cho so sánh 4 thuật toán 24

Bảng 3.1: Bảng chỉ mục dữ liệu của phân luồng trực tiếp 37

Bảng 3.2: Bảng chỉ mục dữ liệu của phân luồng gián tiếp 38

DANH MỤC HÌNH VẼ

Hình 2.1: 7 lý do chính khách hàng bỏ rơi giỏ hàng trực tuyến tại Hoa Kỳ 16

Hình 2.2: 10 lý do người tiêu dùng bỏ rơi giỏ hàng tại bước thanh toán 17

Hình 2.3: Mô hình thuật toán cây quyết định 19

Hình 2.4: Mô hình thuật toán rừng ngẫu nhiên 21

Hình 2.5: Biểu đồ so sánh độ chính xác của 4 thuật toán 25

Hình 2.6: Biểu đồ so sánh tốc độ xử lý của 4 thuật toán 26

Hình 3.1: Sơ đồ tiến trình phân tích dữ liệu và dự báo bỏ rơi giỏ hàng 29

Hình 3.2: Sơ đồ trình tự phân tích lịch sử và hoạt động phiên mua sắm 32

Hình 3.3: Cấu trúc trang thương mại điện tử và dự đoán bỏ rơi giỏ hàng 34

Hình 3.4: Kết quả phân tích với biến số thay đổi là kích thước thử nghiệm 40

Hình 3.5: Kết quả phân tích với biến số thay đổi là số lượng ước tính 40

Hình 3.6: Kết quả phân tích với biến số thay đổi là trạng thái ngẫu nhiên 41

Hình 3.7: Kết quả dự đoán trong ứng dụng dự đoán thực tế 42

Hình 3.8: Mức độ quan trọng của các thuộc tính trong phiên mua sắm 43

Hình 3.9: So sánh thời gian trung bình phiên mua sắm thành công và bỏ rơi 44

MỞ ĐẦU

Trong những năm gần đây, cuộc cách mạng khoa học công nghệ 4.0 đã và đang đem lại nhiều thay đổi tích cực trên mọi mặt đời sống của con người. Nhu cầu mua bán, giao dịch trực tuyến qua mạng Internet dần trở nên phổ biến và ngày càng phát triển. Bên cạnh những thuận lợi và tiện nghi không thể phủ nhận, thương mại điện tử vẫn bộc lộ nhiều vấn đề cho khách mua hàng như lòng tin với sản phẩm, bảo mật dữ liệu, sự phụ thuộc vào đơn vị vận chuyển, hay sự sẵn sàng của sản phẩm. Đối với người bán hàng, thách thức lớn nhất không thể không đề cập là vấn đề người dùng bỏ rơi giỏ hàng khi mua sắm trực tuyến.

Bên cạnh những thuận lợi và tiện nghi không thể phủ nhận, thương mại điện tử vẫn bộc lộ nhiều vấn đề cho khách mua hàng như lòng tin với sản phẩm, bảo mật dữ liệu, sự phụ thuộc vào đơn vị vận chuyển, hay sự sẵn sàng của sản phẩm. Đối với người bán hàng, thách thức lớn nhất không thể không đề cập là vấn đề người dùng bỏ rơi giỏ hàng khi mua sắm trực tuyến. Theo một thống kê độc lập từ học viện Baymard, 69% tổng số lượng giỏ hàng bị bỏ rơi và không được thanh toán trong các giao dịch mua sắm thương mại điện tử [2]. Tỉ lệ bỏ rơi giỏ hàng cao tất yếu sẽ giảm tỉ lệ chuyển đổi đơn hàng, từ đó giảm doanh thu cho người bán nói riêng và cho trang thương mại điện tử nói chung, gia tăng chi phí lưu kho sản phẩm cùng nhiều tác động tiêu cực khác. Không chỉ vậy, người mua hàng sẽ có định kiến tiêu cực và nảy sinh thái độ thận trọng trước tần suất giao dịch hạn chế của người dùng đó tại sàn thương mại điện tử.

Nhận thấy vấn đề còn tồn tại này trong mua hàng trực tuyến, luận văn “Nghiên cứu phân tích hành vi người dùng bỏ giỏ hàng trên các trang thương mại điện tử” ứng dụng khả năng khai thác và phân tích dữ liệu của máy tính để dự đoán tỉ lệ bỏ rơi giỏ hàng trực tuyến. Trên cơ sở đó, hình thành khuôn mẫu tiêu dùng mua sắm đặc trưng của khách hàng tại trang thương mại điện tử. Đồng thời, nghiên cứu cũng đưa đưa ra những yếu tố dẫn đến hành vi bỏ rơi giỏ hàng cùng một số phân tích dựa trên dữ liệu thống kê.

Trong mua sắm trực tuyến, hành vi bỏ rơi giỏ hàng điện tử có thể được hiểu là "một giao dịch mua hàng có kế hoạch rõ ràng nhưng lại không bao giờ được thanh toán" [6]. Một nghiên cứu thống kê cho thấy, cứ mỗi 5 giao dịch mua bán trực tuyến thì có đến 4 giao dịch bị bỏ ngỏ hoặc lãng quên bởi người tiêu dùng vì rất nhiều lý do khác nhau [15]. Ở khía cạnh tâm lý con người, hành vi bỏ rơi giỏ hàng của người dùng đã chỉ ra rằng giỏ hàng trực tuyến được sử dụng không chỉ thuần túy như những giỏ hàng ở siêu thị; mà nó còn được sử dụng như một danh sách mua hàng, một cách để xem trước các phụ phí (thuế, phí vận chuyển), hay đơn thuần chỉ vì mục đích giải trí [7].

Một số công trình nghiên cứu liên quan khác về khai thác và phân tích cách thức sử dụng giỏ hàng trực tuyến dựa trên kĩ thuật học sâu của Köhn, Dennis & Lessmann, Stefan & Schaal, Markus [6] hay khám phá vấn nạn bỏ rơi giỏ hàng - vấn đề rủi ro và danh tiếng của Moore S and Mathews S [15] đã phần nào khai thác được khuynh hướng tiêu dùng trong thương mại điện tử, hành vi mua sắm người tiêu dùng và hành vi sử dụng giỏ hàng trực tuyến. Tuy nhiên, những nghiên cứu này được thực hiện độc lập, dựa trên một lượng dữ liệu có sẵn và không đổi theo thời gian với mục đích chủ yếu để tham chiếu và lý luận cho các nghiên cứu tiếp theo. Chính vì vậy, luận văn sẽ thực hiện nghiên cứu và phân tích trên tập dữ liệu động có thể thay đổi theo thời gian sử dụng thực tế của khách hàng, nâng cao tính ứng dụng của kết quả nghiên cứu đồng thời phản ánh được những tiêu chí đặc trưng dẫn đến hành vi bỏ rơi giỏ hàng phù hợp với từng đối tượng thương mại điện tử.

Với định hướng nghiên cứu ứng dụng, luận văn sẽ sử dụng phương pháp thu thập, tổng hợp và thống kê các hoạt động, lịch sử mua sắm và giao dịch thương mại của người dùng tại trang thương mại điện tử. Từ dữ liệu tổng hợp, việc dự đoán người dùng bỏ rơi giỏ hàng sẽ được chia thành hai luồng xử lý chính là luồng xử lý trực tiếp ứng dụng thuật toán rừng ngẫu nhiên và luồng xử lý gián tiếp xử dụng kết quả thống kê từ lịch sử mua hàng. Luận văn sẽ xây dựng một hệ thống là sự kết hợp giữa trang thương mại điện tử và dịch vụ phân loại dự đoán hành vi bỏ rơi giỏ hàng sử dụng dữ hoạt động của người dùng theo thời gian thực. Trên các cơ sở thông tin và dữ liệu đã

được phân tích, tổng hợp kết quả giữa hai luồng phân tích và đưa ra kết quả cuối cùng. Không những vậy, luận văn sẽ đưa ra các yếu tố có tác động mạnh mẽ đến hành vi bỏ rơi giỏ hàng và so sánh giá trị trung bình của yếu tố đó giữa phiên mua sắm thành công và phiên mua sắm có giỏ hàng bị bỏ rơi.

Nội dung luận văn tập trung vào tiềm năng khai thác dữ liệu khách hàng tại trang thương mại điện tử và ứng dụng thuật toán rừng ngẫu nhiên để dự đoán và phân tích hành vi bỏ rơi giỏ hàng. Cấu trúc luận văn được chia thành 3 chương như sau:

CHƯƠNG 1 - Tổng quan về hành vi người dùng trong thương mại điện tử: Trình bày tổng quan về thương mại điện tử và tiềm năng khai thác dữ liệu người dùng.

CHƯƠNG 2 - Phân tích hành vi bỏ rơi giỏ hàng: Khái niệm bỏ rơi giỏ hàng và ứng dụng thuật toán rừng ngẫu nhiên trong phân loại dữ liệu.

CHƯƠNG 3 - Thử nghiệm và đánh giá: Ứng dụng thuật toán rừng ngẫu nhiên để dự đoán và đưa ra những phân tích về hành vi bỏi rơi giỏ hàng.

CHƯƠNG I: TỔNG QUAN VỀ HÀNH VI NGƯỜI DÙNG TRONG THƯƠNG MẠI ĐIỆN TỬ

1.1. Giới thiệu về khái niệm, sự ra đời, hình thành và phát triển của thương mại điện tử

1.1.1. Khái niệm về thương mại điện tử

Thương mại điện tử được tiếp cận và khái quát hóa dưới nhiều hình thức và quy cách khác nhau. Việc định nghĩa thương mại điện tử phụ thuộc vào nhiều yếu tố như chủ thể đưa ra định nghĩa (tổ chức, quốc gia, vùng lãnh thổ hoặc cá nhân), mục đích (hình thành khái niệm, áp dụng vào văn kiện luật pháp) hoặc ngữ cảnh thực tiễn (trong hội nghị của tổ chức thế giới, trong kỳ họp quốc gia hay trong một luận văn thạc sĩ). Do đó, việc định nghĩa khái niệm phù hợp và chính xác đóng một vai trò thiết yếu nhằm khái quát hóa nội dung cũng như tối đa khả năng truyền đạt thông tin đến người đọc. Tổ chức Thương mại thế giới (WTO) đưa ra định nghĩa: "*Thương mại điện tử bao gồm việc sản xuất, quảng cáo, bán hàng và phân phối sản phẩm được mua bán và thanh toán trên mạng Internet, nhưng được giao nhận một cách hữu hình, cả các sản phẩm giao nhận cũng như những thông tin số hoá thông qua mạng Internet*".

Trong phạm vi và giới hạn nghiên cứu của luận văn này, thương mại điện tử có thể được hiểu là "*hành động mua hoặc bán sản phẩm dựa trên các dịch vụ trực tuyến điện tử thông qua mạng Internet*". Trong đó, các doanh nghiệp thương mại điện tử có thể trao đổi thương mại dựa trên một hoặc nhiều các khía cạnh sau:

- Bán lẻ trực tiếp đến người tiêu dùng qua các trang mạng hoặc ứng dụng điện thoại, giao tiếp thương mại qua hệ thống chăm chóc khách hàng trực tuyến, robot chat tự động (chatbot) và các trợ lý ảo.
- Cung cấp hoặc tham gia vào thị trường thương mại trực tuyến, thực hiện các giao dịch từ bên thứ ba từ doanh nghiệp với khách hàng (business-to-customer) hoặc khách hàng đến khách hàng (customer-to-customer).

- Mua và bán hàng theo mô hình doanh nghiệp với doanh nghiệp (business-to-business).
- Trao đổi chéo thông tin trực tuyến giữa các doanh nghiệp thương mại điện tử.
- Cung cấp dịch vụ trao đổi tài chính và tỷ giá ngoại tệ.
- Thị trường hóa các phân khúc khách hàng qua hệ thống thư điện tử, tin nhắn, fax hoặc cuộc gọi thoại.
- Thu thập và phân tích thông tin dân số qua các thông tin mua sắm và mạng xã hội.
- Tham gia vào quá trình marketing để đưa thông tin và bán của các sản phẩm và dịch vụ.

1.1.2. Sự ra đời, hình thành và phát triển của thương mại điện tử

Năm 1971-1972, giao dịch thương mại điện tử đầu tiên được ghi nhận không chính thức giữa các sinh viên của Phòng thí nghiệm trí tuệ nhân tạo Stanford và Viện Công nghệ Massachusetts qua hệ thống mạng băng thông rộng nội bộ ARPANET. Giao dịch này được coi là khởi nguồn của giao thương trực tuyến qua mạng Internet và sau đó được mô tả là "hành động chính yếu của thương mại điện tử" trong cuốn sách *What the Dormouse Said của John Markoff* .

Năm 1991, mạng Internet lần đầu tiên được mở ra cho mục đích sử dụng thương mại bởi nhà khoa học máy tính Tim Berners-Lee. Sự kiện này là cột mốc quan trọng đánh dấu sự phát triển thương mại đầu tiên của ngành dịch vụ thương mại điện tử [20]. Năm 1994, trình duyệt Netscape 1.0 với phương thức mã hóa SSL lần đầu tiên được ra mắt và giới thiệu đến người sử dụng. Sự ra đời của phương thức mã hóa SSL cải thiện và nâng cao tính bảo mật của các giao dịch tài chính khi mua bán tạo các hệ thống thương mại điện tử.

Năm 1995, Amazon.com chính thức xuất hiện trên mạng Internet bởi Jeff Bezos; cùng năm, trang đấu giá trực tuyến eBay cũng được thành lập bởi lập trình viên máy tính Piere Omidyar [20]. Sự thành lập và phát triển chính thức của hai cây

đại thụ trong thương mại điện tử đã đánh dấu một bước ngoặt lớn trong mục tiêu và thị trường bán hàng của các doanh nghiệp. Sau năm 1995, tỉ lệ mua hàng trực tuyến của người tiêu dùng tăng trưởng vượt bậc khiến ngày càng nhiều công ty chú trọng hơn vào hình ảnh trực tuyến trên mạng Internet. Đến năm 2002-2003, eBay và Amazon lần đầu tiên báo lãi trong dịch vụ mua bán trực tuyến; chính thức khẳng định vai trò thiết yếu và tính khả thi của thương mại điện tử.

Những năm gần đây, thương mại điện tử đã và đang đóng góp một thị phần không hề nhỏ vào tổng giá trị giao dịch thương mại trên toàn thế giới. Năm 2017, thương mại điện tử trên thị trường bán lẻ thế giới đạt 2.304 nghìn tỷ đô-la, tăng 24,8 phần trăm so với cùng kỳ năm ngoái [26]. Cuối năm 2019, đầu năm 2020, đại dịch COVID-19 hoành hành trên toàn thế giới khiến việc giao thương mua bán trực tiếp gặp nhiều khó khăn và hạn chế; chính sự khó khăn này là đà thúc đẩy thương mại điện tử phát triển mạnh mẽ hơn với tăng trưởng 27,6% trên toàn thế giới. Riêng tại Hoa kỳ, bộ tài chính và công thương báo cáo thương mại điện tử tăng trưởng 44% trong năm 2020 với một phần 3 trong số đó là tăng trưởng của Amazon.com. [30]

1.2. Tiềm năng khai thác dữ liệu người dùng trong thương mại điện tử

1.2.1. Phân tích giỏ hàng điện tử

Khác với mua sắm truyền thống, người tiêu dùng không thật sự cảm nhận, cầm nắm trực tiếp sản phẩm khi mua hàng trực tuyến tại các trang thương mại điện tử. Thay vào đó, mọi hành động, đối tượng phục vụ cho việc mua sắm đều được đại diện bằng một chức năng nhất định khi mua hàng trực tuyến. Trong đó, chức năng quan trọng nhất không thể không đề cập đó là giỏ hàng điện tử ứng với mỗi phiên sử dụng của người dùng. Tương tự như hành động mua sắm tại siêu thị, người tiêu dùng thông thường sẽ lấy theo một xe đẩy hàng hoặc một giỏ hàng để giữ tạm các sản phẩm muốn mua. Giỏ hàng điện tử cũng phục vụ hành vi quen thuộc này cho khách mua hàng trực tuyến, tuy nhiên toàn bộ dữ liệu được lưu trữ tại hệ thống của trang thương mại điện tử. Chính yếu tố lưu trữ dữ liệu này đã giúp cho khai thác và phân tích dữ liệu

giỏ hàng điện tử trở thành tác vụ thiết yếu trong khai thác dữ liệu người dùng. Các số liệu và đánh giá từ giỏ hàng điện tử sẽ giúp cho chủ cửa hàng có cái nhìn tổng quan và rõ ràng về nhu cầu mua sắm của tập khách hàng tại chính trang thương mại điện tử. Không những vậy, nhu cầu về sản phẩm, số lượng đặt hàng, thời gian mua sắm, và nhiều thông tin quan trọng khác cũng được tổng hợp và đưa ra báo cáo. Qua đó, chủ cửa hàng hoặc các doanh nghiệp có thể đánh giá được khách hàng tiềm năng, cân nhắc đưa ra các khuyến mãi về miễn phí vận chuyển, giảm giá đơn hàng hoặc thậm chí liên lạc trực tiếp để nhắc khách hàng hoàn tất giao dịch trong trường hợp giỏ hàng bị bỏ rơi.

Khai thác dữ liệu giỏ hàng cũng chính là cách gã khổng lồ bán lẻ Amazon hoàn thiện phương thức quảng cáo và bán hàng chéo cho khách hàng mua sắm trực tuyến. Bằng cách giới thiệu đến với khách hàng các sản phẩm mua sắm theo bộ, quảng cáo bán chéo dựa trên các biểu ngữ "khách hàng đã mua những mặt hàng này cũng đã mua", "đề xuất cho bạn" hay "những sản phẩm thường xuyên mua cùng nhau", Amazon không những kích thích nhu cầu mua hàng mà còn giúp khách hàng cảm thấy được quan tâm, nâng tầm trải nghiệm người dùng [31]. Khi nhận được đề xuất và gợi ý mua hàng, khách hàng không chỉ thấy mình được hệ thống chú ý như mua hàng trực tiếp tại siêu thị, mà còn có thể tùy ý chấp nhận hoặc từ chối gợi ý vì biết rằng đây là một phần chức năng tự động. Qua quá trình phân tích dữ liệu giỏ hàng, Amazon thu được một lượng lớn dữ liệu thông qua các phương pháp xử lý khác nhau mà sau đó họ sử dụng để tùy chỉnh và cá nhân hóa các ưu đãi và khuyến mại cho khách hàng. Việc bán chéo sản phẩm và đề xuất mua hàng này đã được chứng minh là có thể tăng quy mô đặt hàng trung bình của một giỏ hàng điện tử lên hơn 35% [32]. Theo Báo cáo khảo sát người bán hàng năm lần thứ 8 của nhóm e-tailing (gồm 190 giám đốc điều hành thương mại điện tử), trong đó có hơn 55% nhà bán lẻ cùng đồng thuận với ý kiến rằng: giỏ hàng điện tử của người tiêu dùng nhất định phải bao gồm chức năng bán kèm và bán chéo [31]. Như vậy có thể thấy, thương mại điện tử cung cấp một mỏ dữ liệu vô cùng to lớn và tiềm năng dành cho các nhà bán lẻ, phân tích

thị trường và chủ doanh nghiệp. Tận dụng và khai thác triệt để tiềm năng dữ liệu này không chỉ đem lại lợi nhuận, doanh thu bán hàng cho chính người bán lẻ, mà còn kích thích nhu cầu tiêu dùng xã hội nói chung, nâng tầm trải nghiệm dịch vụ thương mại tại các hệ thống bán lẻ điện tử.

1.2.2. Dự đoán nhu cầu thị trường

Trong kinh doanh thương mại, các công ty bán lẻ, phân tích, đầu tư và tiếp thị luôn luôn có mong muốn dự đoán cung cầu thị trường một cách chính xác nhất. Trong quá khứ, khi việc giao thương, mua bán chỉ dừng lại ở mô hình cửa hàng tạp hóa, siêu thị bán lẻ và trung tâm mua sắm, việc phỏng đoán nhu cầu thị trường trong tương lai gần thường chỉ dựa vào lịch sử mua sắm của đại đa số người dân và các sự kiện lớn trong năm sắp tới. Quy luật cung cầu truyền thống này đã và đang là kim chỉ nam của nhiều doanh nghiệp kinh doanh truyền thống trên thế giới, là dự đoán nhu cầu tiêu dùng trong nhiều thập kỷ trong quá khứ. Trong thời điểm hiện tại, với sự phát triển của thương mại điện tử cùng sự thay đổi trong quy cách và thói quen mua sắm, việc dự đoán cung cầu theo cách truyền thống đã không còn phản ánh được nhu cầu tiêu dùng dựa trên mỗi khách mua hàng, làm mất đi sự tùy biến và cá nhân hóa thông tin tiêu dùng [22].

Nhận thấy sự cần thiết trong dự đoán nhu cầu thị trường dựa trên từng khách hàng, rất nhiều doanh nghiệp bán lẻ, các sàn thương mại điện tử đang tối ưu các thuật toán phân tích báo cáo hành vi tiêu dùng để có thể tạo nên một biểu đồ dự đoán nhu cầu mua sắm dựa trên từng nhu cầu mua sắm của khách hàng. Không những vậy, nhà bán lẻ cũng có được những thông tin quan trọng về nhu cầu tổng quan của cả thị trường nói chung và của từng khách hàng cá nhân nói riêng, từ đó có thể đưa ra nhiều chương trình chăm sóc, ưu đãi cho khách hàng thân thiết. Có thể thấy, thương mại điện tử đã và đang đem đến những trải nghiệm mua sắm không chỉ mới lạ mà còn rất cá nhân hóa dựa trên từng khách hàng, người mua hàng được hưởng lợi ích chăm sóc

chu đáo tận tình, người bán hàng có những dự đoán và báo cáo hữu ích về thị trường tại chính sàn giao dịch trực tuyến.

1.2.3. Đánh giá phân khúc thị trường

Phân khúc thị trường là quá trình đánh giá và phân loại các nhóm khách hàng để giúp nhà quản lý, nhà bán lẻ thực hiện các chiến dịch tiếp thị đúng mục tiêu. Với vai trò thiết yếu và không thể bỏ qua trong tiến trình mở rộng và quảng bá thương hiệu, nhiều nhà bán lẻ đã dành sự quan tâm không nhỏ để khả năng phân khúc thị trường qua các kỹ thuật khai thác dữ liệu từ khách hàng thực tế. Dựa trên báo cáo phân khúc thị trường của trang bán hàng trực tuyến, các nhà quản lý nắm rõ hơn về đối tượng khách hàng hiện có, nhu cầu sản phẩm hoặc dịch vụ có thể cung cấp, từ đó làm tiền đề để triển khai các chiến dịch quảng cáo, tiếp thị tập trung phân khúc khách hàng. Những chiến dịch quảng bá có mục tiêu và phân khúc khách hàng rõ ràng không chỉ đem lại sự nổi tiếng cho nhãn hiệu mà còn thu hút thêm nhiều khách hàng cùng phân khúc, góp phần cải thiện doanh thu trong tương lai.

Trong thời điểm hiện tại, bộ công cụ phân tích Google Analytics đang được nhiều trang thương mại trực tuyến tin dùng để báo cáo về phân khúc khách hàng truy cập, mua sắm và giao dịch [10]. Với Google Analytics, nhà bán lẻ sẽ có được những nhóm thông tin cơ bản từ khách hàng truy cập trang thương mại điện tử:

- *Nhân khẩu học*: Bao gồm các thuộc tính của người dùng như tuổi, giới tính và sở thích.
- *Vị trí*: Vị trí địa lý mà trang web được truy cập bao gồm mọi thứ từ lục địa, quốc gia, thành phố, quận, huyện, xã.
- *Hành vi*: Mức độ tương tác của người dùng với trang web dựa trên phiên hoạt động, số lần quay lại, số lần tìm kiếm.
- *Thiết bị*: Thiết bị sử dụng trong phiên hoạt động đó; bao gồm điện thoại di động, máy tính để bàn hoặc máy tính bảng.

- *Kênh truy cập*: Kênh truy cập khởi nguồn của người dùng, ví dụ: truy cập trực tiếp, giới thiệu, mạng xã hội, tìm kiếm chính xác, v.v.

1.2.4. Phòng chống gian lận thương mại

Các kỹ thuật khai thác dữ liệu người dùng hiện tại không chỉ đơn thuần phục vụ mục đích phân tích, dự đoán và phân khúc thị trường, mà còn bảo vệ người tiêu dùng khỏi những mối nguy hiểm về mạo danh trực tuyến, phòng chống gian lận thương mại. Trong quá trình xử lý và phân tích dữ liệu, các thuật toán khai thác thông tin liên tục tạo ra các khuôn mẫu để tham chiếu dữ liệu một cách tổng quát nhất. Các khuôn mẫu dữ liệu này sẽ được tổng hợp và xử lý trong giai đoạn cuối cùng để đưa ra một biểu đồ phân tích thói quen mua sắm dựa theo những tiêu chí đã được định nghĩa sẵn. Do đó, tiềm năng vô cùng lớn trong khai thác và phân tích dữ liệu người dùng trong thương mại điện tử là không thể phủ nhận, những thông tin khai thác từ nguồn dữ liệu này không chỉ phục vụ nhu cầu tiêu dùng đơn thuần mà còn giúp cho các giao dịch trực tuyến an toàn hơn, dịch vụ mua bán được bảo đảm hơn và hệ thống thương mại ít rủi ro hơn.

1.3. Giới thiệu giỏ hàng điện tử và hành vi bỏ rơi giỏ hàng

1.3.1. Khái niệm giỏ hàng điện tử

Giỏ hàng trên trang thương mại điện tử là một phần mềm có giao diện, hỗ trợ người dùng mua sắm các sản phẩm hoặc dịch vụ. Giỏ hàng điện tử chấp nhận thanh toán của khách hàng và xử lý thanh toán và phân phối thông tin đến nhà bán lẻ, các đơn vị chấp nhận thanh toán và các bên liên quan.

Giỏ hàng thu hẹp khoảng cách giữa mua sắm và mua hàng, vì vậy việc có phần mềm giỏ hàng tốt nhất là điều cực kỳ quan trọng trên trang web bán lẻ. Đối với người mới sử dụng các dịch vụ mua sắm trực tuyến, khái niệm giỏ hàng trực tuyến và cách thức sử dụng sẽ có phần xa lạ. Tuy nhiên đối với hầu hết người tiêu dùng đã từng ít nhất mua sắm trực tuyến, giỏ hàng lại là một chức năng không thể thiếu để hoàn thành

một đơn đặt hàng. Đa số người dùng đều không nhận thức được đầy đủ những chức năng và nhu cầu cần thiết của một giỏ hàng. Vậy nên, giỏ hàng điện tử thường có ba khía cạnh sử dụng phổ biến [22]:

- Lưu trữ thông tin sản phẩm
- Quản lý đơn hàng, danh mục và khách hàng
- Hiển thị dữ liệu sản phẩm, danh mục và thông tin trang web để người dùng hiển thị

Một cách tổng quan, giỏ hàng điện tử có mục đích sử dụng tương tự như giỏ hàng hữu hình trong siêu thị, nhưng lại có nhiều chức năng hơn nhờ vào tính trực tuyến và số hóa. Trong đó, mọi thông tin mua hàng, đặt hàng của người tiêu dùng được lưu trữ tại hệ thống thương mại điện tử của nhà bán lẻ, tất cả liên kết, xử lý và tính toán đều được thực hiện theo thời gian thực.

1.3.2. Khuynh hướng sử dụng giỏ hàng điện tử của người tiêu dùng

Khuynh hướng sử dụng bất kỳ một chức năng hay hệ thống nào của con người phụ thuộc rất nhiều vào hoàn cảnh, nhóm người và điều kiện cho phép sử dụng. Vậy nên khuynh hướng sử dụng giỏ hàng điện tử của người tiêu dùng cũng phụ thuộc vào ngoại cảnh, nhóm đối tượng và điều kiện tiên quyết của từng trang thương mại điện tử. Trong luận văn này, thói quen sử dụng giỏ hàng điện tử sẽ chỉ tập trung vào đối tượng là người dân Việt Nam cùng điện kiện là người dùng đã có kiến thức trong mua bán trực tuyến và thực hiện trên các trang thương mại điện tử nội địa.

Tuy chưa có một nghiên cứu chính thức nào về thói quen sử dụng giỏ hàng trong mua sắm trực tuyến tại các trang thương mại điện tử Việt Nam, tuy nhiên, khuynh hướng này có thể được chia làm hai dạng chính:

- *Lưu trữ sản phẩm*: khách hàng sẽ sử dụng giỏ hàng đơn thuần như một "kho hàng hóa" cho lượt tiêu dùng tiếp theo trong tương lai. Đối với thói quen này, giỏ hàng của người dùng luôn có ít nhất một sản phẩm được lưu trữ. Mục đích

lưu trữ cũng rất đa dạng từ sở thích, nhu cầu mua sắm lần tới hoặc chỉ đơn giản là lần mua sắm này chưa đủ tài chính.

- *Kiểm tra cước vận chuyển và khuyến mại*: Cước vận chuyển là vấn đề cân nhắc nhiều nhất và cũng là một trong những nguyên nhân hàng đầu dẫn đến người dùng từ bỏ mua hàng. Do vậy, trước khi thực hiện một đơn hàng, người tiêu dùng sẽ có những cân nhắc dựa trên phí vận chuyển có trên giỏ hàng điện tử. Đối với những người dùng này, mức phí vận chuyển sẽ đóng vai trò quyết định cho việc giỏ hàng này có thể trở thành đơn hàng. Ngoài ra, việc kiểm tra mã khuyến mại, mã giảm giá, điểm thưởng cũng được thực hiện qua việc sử dụng giỏ hàng điện tử. Đây cũng là hành vi thường thấy trong các chiến dịch giảm giá lớn của các sàn thương mại điện tử hiện nay.

1.3.3. Bỏ rơi giỏ hàng trong mua sắm trực tuyến

Trong thương mại điện tử, bỏ rơi giỏ hàng là khi một khách hàng tiềm năng bắt đầu quy trình thanh toán cho một đơn đặt hàng trực tuyến nhưng lại bỏ qua quy trình này trước khi hoàn tất mua hàng. Bất kỳ mặt hàng nào được đặt vào giỏ hàng trực tuyến nhưng không bao giờ hoàn tất được quy trình thanh toán và chuyển đổi đơn hàng sẽ được coi là mặt hàng bị bỏ rơi, hay rộng hơn là giỏ hàng đã bị bỏ rơi. Theo thống kê từ Viện công nghệ Baymard, có đến 69% số lượng giỏ hàng bị bỏi rơi trong các giao dịch tiêu dùng trực tuyến [2]. Tỷ lệ bỏ rơi giỏ hàng được tính bằng cách chia tổng số giao dịch đã hoàn thành cho tổng số giao dịch đã được bắt đầu. Tỷ lệ này sẽ xác định phần trăm người dùng của trang web báo hiệu ý định mua hàng bằng cách thêm một mặt hàng vào giỏ hàng nhưng không hoàn tất việc thanh toán và chuyển đổi giỏ hàng trực tuyến thành đơn hàng.

Tỷ lệ bỏ qua giỏ hàng là một số liệu quan trọng để các trang web thương mại điện tử theo dõi hành vi người dùng vì tỷ lệ bỏ qua cao có thể báo hiệu trải nghiệm người dùng kém hoặc kênh bán hàng liên tục gặp vấn đề. Giảm thiểu tình trạng bỏ rơi giỏ hàng trực tiếp không chỉ thúc đẩy doanh số bán hàng và doanh thu nhiều hơn

mà còn tối ưu hóa quy trình thanh toán và đặt hàng, vốn là lĩnh vực trọng tâm của nhiều doanh nghiệp bán lẻ trực tuyến.

1.4. Kết luận

Trong chương một, luận văn đã trình bày một cách tổng quan về khái niệm, sự ra đời, hình thành và phát triển của thương mại điện tử trên thế giới. Dựa trên cơ sở đó, luận văn đề cập những tiềm năng khai thác dữ liệu và hành vi tiêu dùng trên các trang thương mại điện tử. Tiềm năng khai thác dữ liệu được thể hiện qua bốn phương diện chính đó là: tiềm năng về dữ liệu giỏ hàng điện tử, dự đoán nhu cầu thị trường, đánh giá phân khúc thị trường và phòng chống gian lận thương mại. Bốn tiềm năng trên là bốn khía cạnh mà khai thác dữ liệu thể hiện được vai trò chủ yếu của mình trong việc tìm ra khuôn mẫu và phân tích hành vi sử dụng của khách hàng. Từ những khái niệm và tiềm năng khai thác dữ liệu, chương một cũng giới thiệu đến khái niệm giỏ hàng điện tử trong mua sắm trực tuyến. Không những vậy, khuynh hướng sử dụng giỏ hàng điện tử tại Việt Nam cùng hành vi bỏ rơi giỏ hàng được nêu lên để làm cơ sở nghiên cứu trong chương hai.

CHƯƠNG 2: PHÂN TÍCH HÀNH VI BỎ RƠI GIỎ HÀNG

Bỏ rơi giỏ hàng là một trong những vấn đề phổ biến nhất tại các hệ thống bán lẻ trực tuyến. Một nghiên cứu thống kê cho thấy, cứ mỗi 5 giao dịch mua bán trực tuyến thì có đến 4 giao dịch bị bỏ ngỏ hoặc lãng quên bởi người tiêu dùng [15]. Theo kết quả nghiên cứu của Ouellet (2010), khoảng 70% số lượng giỏ hàng bị bỏ rơi mỗi ngày trong giao dịch mua bán tại các trang thương mại điện tử [20], cho thấy rằng vấn đề bỏ rơi giỏ hàng là thách thức không hề nhỏ đối với các nhà bán lẻ trực tuyến và xử lý vấn đề này là mối quan tâm hàng đầu của hầu hết doanh nghiệp dựa vào thương mại điện tử.

Đã có nhiều nghiên cứu về hành vi bỏ rơi giỏ hàng của người tiêu dùng trong quá trình mua sắm trực tuyến, trong đó, tùy theo mục đích nghiên cứu và đối tượng, vấn đề bỏ rơi giỏ hàng được tiếp cận và định nghĩa với một khái niệm khác nhau. Một nghiên cứu định nghĩa rằng hành động bỏ rơi giỏ hàng là khi người mua hàng bắt đầu quy trình thanh toán nhưng không hoàn thành [20]. Một công trình khác lại đưa ra khái niệm bỏ rơi giỏ hàng là khi người mua sắm đặt các mặt hàng vào giỏ hàng trực tuyến của họ để thu thập thông tin nhưng quyết định từ bỏ giỏ hàng trước khi tiến đến giai đoạn mua hàng cuối cùng [15]. Tuy nhiên, các khái niệm đều nêu lên rất rõ hai yếu tố chính là sản phẩm được chọn để thêm vào giỏ hàng và giỏ hàng điện tử không được chuyển đổi thành đơn hàng. Vậy trong luận văn này, bỏ rơi giỏ hàng có thể được hiểu là hành vi khi người tiêu dùng đặt (một hoặc nhiều) mặt hàng vào giỏ hàng trực tuyến của họ mà không cần mua bất kỳ mặt hàng nào trong phiên mua sắm trực tuyến đó [21].

2.1. Các yếu tố chính quyết định bỏ rơi giỏ hàng

2.1.1. Trải nghiệm người dùng không tốt

Một giao diện trực quan cùng một trải nghiệm phù hợp với khuynh hướng người sử dụng sẽ là sự khởi đầu thuận lợi đối với mọi người dùng trực tuyến. Nếu như trong mua sắm trực tuyến, khách hàng được mở cửa đón tiếp bởi nhân viên bán

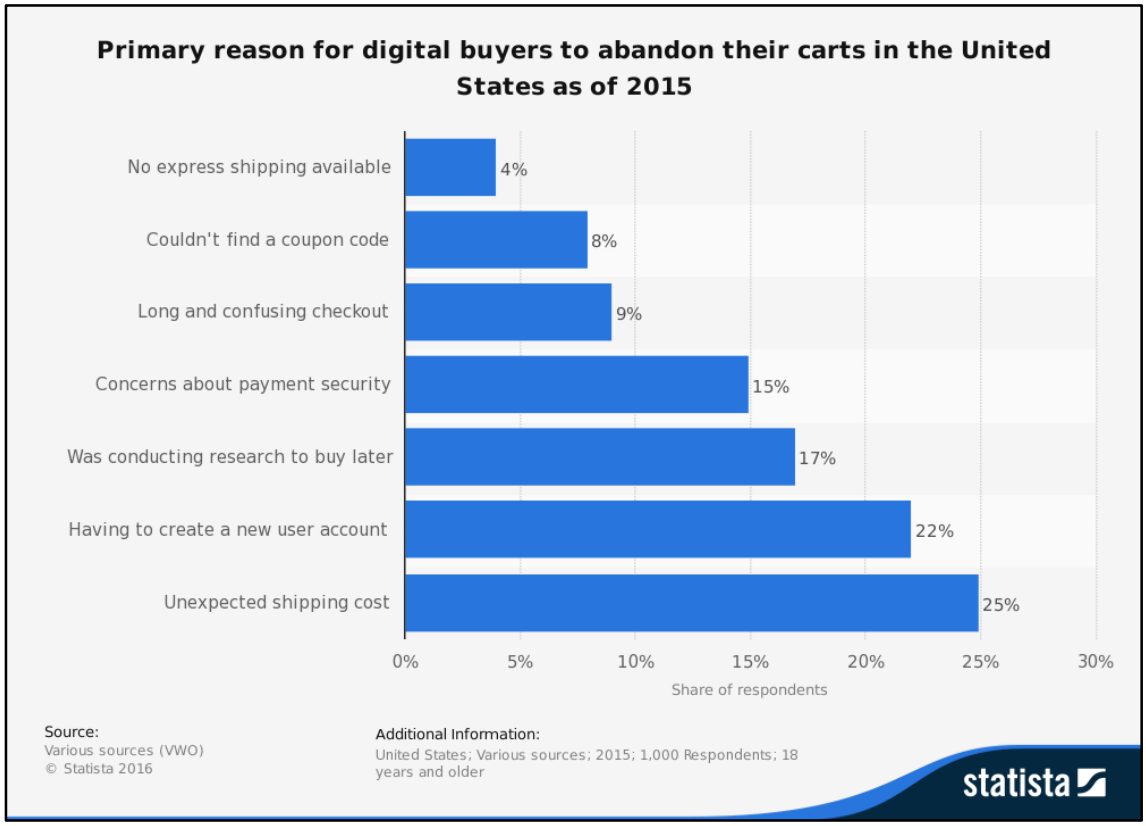
hàng, được giới thiệu và tư vấn sản phẩm thì trong mua sắm trực tuyến, giao diện và trải nghiệm người dùng sẽ thay lời chào hỏi của nhà bán lẻ đến với người tiêu dùng. Một nghiên cứu chỉ ra rằng, trải nghiệm mua sắm trực tuyến không chỉ nằm ở chất lượng sản phẩm mà còn ở chính những tương tác của người dùng với trang thương mại điện tử [15]. Do vậy, một trang thương mại điện tử với giao diện thiếu trực quan và đổi mới, thao tác người dùng không phù hợp tâm lý đa số người sử dụng sẽ khiến khách hàng bỏ đi ngay từ những giai đoạn mua sắm đầu tiên.

Một trong những trải nghiệm gây ám ảnh cả người tiêu dùng lẫn nhà bán lẻ không thể không nhắc đến là thời gian chờ tương tác của trang thương mại điện tử. Thật vậy, tốc độ tải trang và thời gian tương tác là một trong những tiêu chí hàng đầu đánh giá mức độ trải nghiệm của người dùng đối với bất kỳ một trang web nào. Người dùng nói chung và người tiêu dùng trực tuyến nói riêng sẽ có khuynh hướng mất dần kiên nhẫn khi bất kỳ tương tác nào có biểu hiện chậm trễ hoặc có dấu hiệu trục trặc [33]. Ngoài ra, rất nhiều sàn thương mại điện tử hiện nay thường xuyên tổ chức các đợt giảm giá định kỳ hàng tháng, trong những ngày đó, thời gian mua sắm của người dùng sẽ là "hữu hạn" trong một khoảng nhất định nhằm đạt được chương trình khuyến mại. Do vậy, sự chậm trễ trong thương gian tương tác mua sắm và khiến người tiêu dùng lỡ mất khuyến mại sẽ dẫn đến tỉ lệ bỏ rơi giỏ hàng gần như tuyệt đối. Không chỉ ảnh hưởng trong các ngày có chương trình giảm giá, người tiêu dùng nói chung sẽ gần như mất đi động lực mua sắm khi thời gian trung bình giữa các tương tác lớn hơn 10 giây [22].

2.1.2. Chi phí vận chuyển cao, đơn hàng không minh bạch

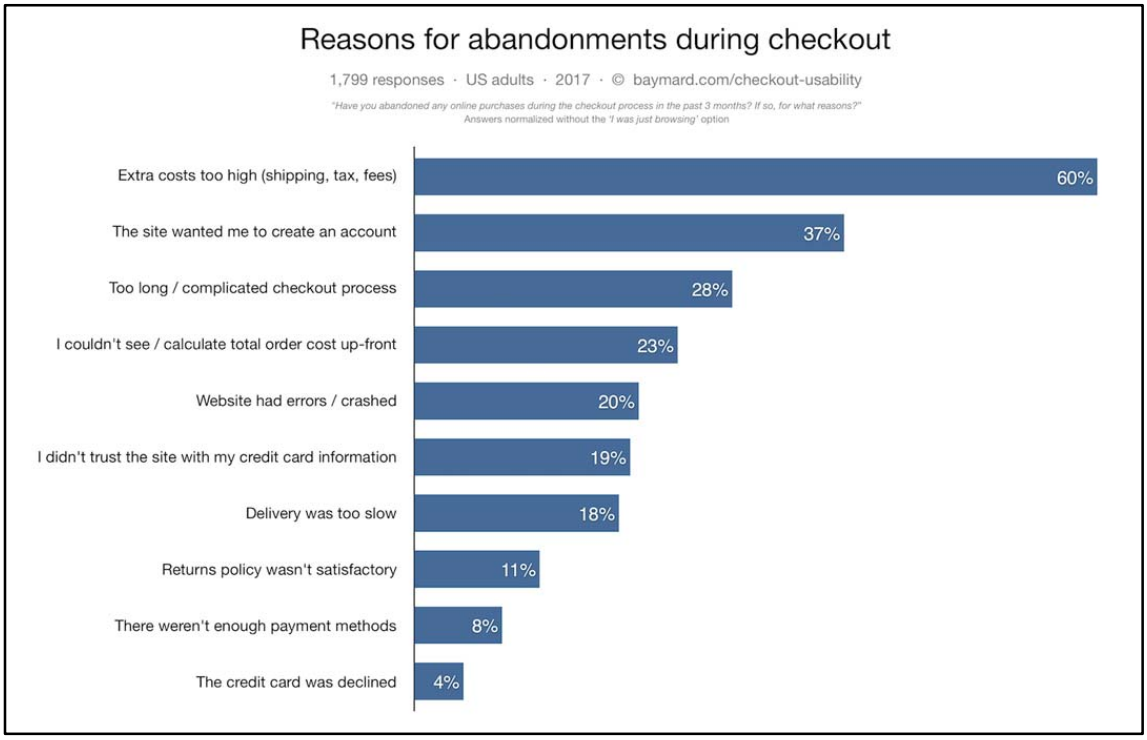
Chi phí vận chuyển cao, đơn hàng có nhiều thuế phí không minh bạch là nguyên nhân hàng đầu dẫn đến người tiêu dùng bỏ rơi giỏ hàng trực tuyến. Theo thống kê từ Statista, trong năm 2015 có đến 25 % số giỏ hàng trực tuyến bị bỏ rơi tại Mỹ vì chi phí vận chuyển quá cao và đắt đỏ đối với người tiêu dùng. Chính lý do này làm những khách hàng tiêu dùng nhỏ lẻ muốn trải nghiệm dịch vụ của trang thương

mại điện tử càng thêm tâm lý dè chừng trong mua sắm vì chi phí vận chuyển quá tốn kém [22]. Hơn nữa, chi phí vận chuyển cao đối với một đơn hàng có số lượng sản phẩm hạn chế cùng thời gian vận chuyển lâu hơn 2 ngày sẽ khiến giỏ hàng có tỷ lệ bị bỏ rơi gần như tuyệt đối. Tuy nhiên, những suy luận trên không hoàn toàn đúng đối với một số loại sản phẩm đặc biệt như đồ nội ngoại thất, một số trang thiết bị đặc biệt vì các loại hình sản phẩm này cần có dịch vụ vận chuyển chuyên nghiệp. Nhưng nếu đứng trên phương diện tiêu dùng hàng ngày, khi người tiêu dùng trực tuyến phải đối mặt với giỏ hàng có chi phí vận chuyển cao, thuế và phí dịch vụ quá nhiều, khách hàng sẽ tất yếu bỏ ngỏ ý định mua sắm và tìm kiếm các giao dịch thay thế tốt hơn.



Hình 2.1: 7 lý do chính người tiêu dùng bỏ rơi giỏ hàng trực tuyến tại Hoa Kỳ

(Nguồn: Báo cáo dữ liệu tiêu dùng Hoa Kỳ bởi Statista năm 2015)



Hình 2.2: 10 lý do người tiêu dùng bỏ rơi giỏ hàng tại bước thanh toán

(Nguồn: Tổng hợp dữ liệu tiêu dùng bởi học viện Baymard tại Hoa Kỳ năm 2017[2])

2.1.3. Nhận thức rủi ro từ người dùng trực tuyến

Xây dựng lòng tin khách hàng là ưu tiên hàng đầu của tất cả công ty, doanh nghiệp tư nhân trong nền kinh tế thị trường hiện nay. Lòng tin của người sử dụng không chỉ đem lại một nguồn doanh thu ổn định mà còn giúp cho danh tiếng của công ty ngày một vững vàng, tăng sức cạnh tranh của doanh nghiệp. Tương tự, các nhà bán lẻ cũng cần xây dựng một lòng tin đối với người tiêu dùng trực tuyến trong chính hoạt động giao dịch thương mại điện tử. Nhiều công trình nghiên cứu đã kết luận rằng sự nhận thức về rủi ro của người dùng có tác động trực tiếp đến hành vi mua sắm trực tuyến của người tiêu dùng [23]. Đồng thời, chỉ ra mối liên hệ chặt chẽ giữa hành vi bỏ rơi giỏ hàng trực tuyến và những dấu hiệu khiến người tiêu dùng có cảm giác bất an khi thực hiện các hoạt động đặt hàng và mua sắm.

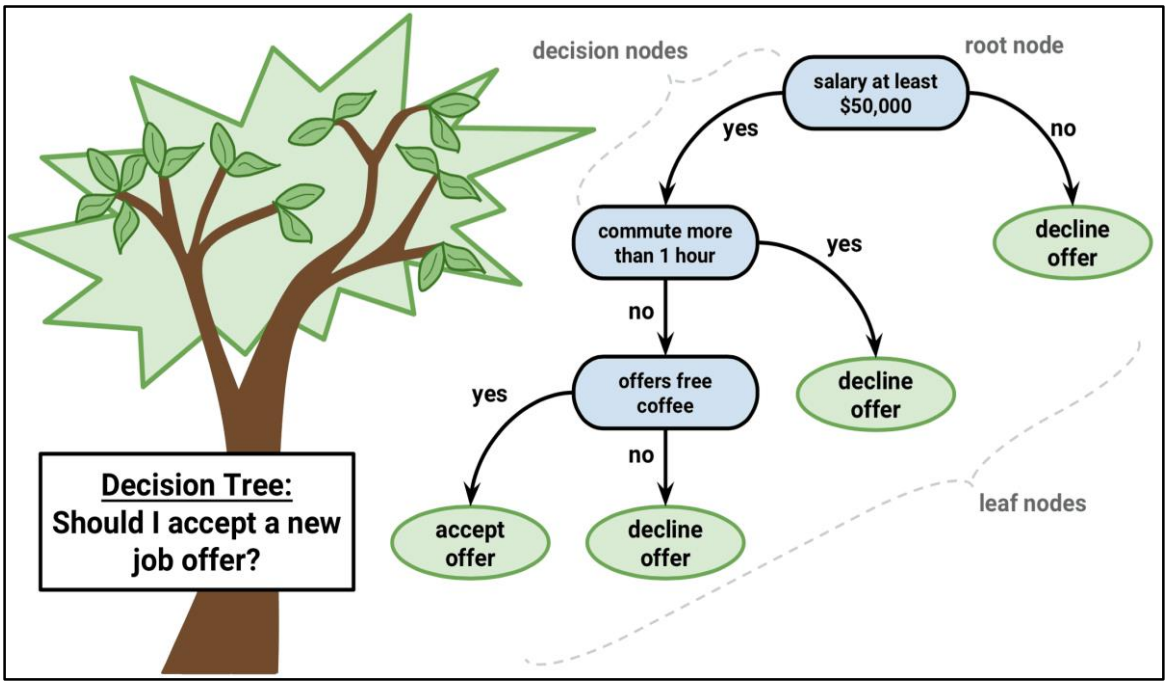
Nghiên cứu của Cheon, Cho và Kang năm 2006 công bố rằng cảm giác an toàn đóng một vai trò quan trọng trong quyết định tiếp tục mua sắm của người tiêu dùng

trực tuyến. Một khi khách hàng thấy được những dấu hiệu lừa đảo, cảm giác bất an sẽ tác động tiêu cực đến quyết định đặt hàng của người tiêu dùng, từ đó làm gia tăng số lượng giỏ hàng bị bỏ rơi. Những dấu hiệu mất an toàn trực tuyến của các trang thương mại điện tử được bộ lộ rõ ở giai đoạn cuối cùng của chuỗi hành vi mua sắm khi người tiêu dùng thực hiện thanh toán và đặt hàng. Trong giai đoạn này, khi quá trình thanh toán đòi hỏi người dùng phải cung cấp hoặc đăng nhập vào tài khoản ngân hàng, bị điều hướng thanh toán quá nhiều hay bị yêu cầu cung cấp nhiều thông tin cá nhân hoặc thậm chí những dấu hiệu rất cơ bản như website không có chứng chỉ bảo mật SSL, v.v... sẽ ngay lập tức khiến khách hàng cân nhắc việc mua sắm và đặt hàng. [10]. Do vậy, việc xây dựng một quá trình thanh toán minh bạch, cung cấp một trải nghiệm an toàn sẽ giảm thiểu đáng kể tâm lý dè chừng khi mua sắm của người tiêu dùng, từ đó hạn chế tỷ lệ giỏ hàng bị bỏ ngỏ ở những giai đoạn mua sắm cuối cùng.

2.2. Thuật toán cây quyết định và rừng ngẫu nhiên

2.2.1. Thuật toán cây quyết định

Cây quyết định (Decision Tree) là một thuật toán học máy có giám sát lần đầu được giới thiệu bởi J. Ross Quinlan tại Đại học Sydney và đồng thời cũng được trình bày trong cuốn sách “Machine Learning” của ông vào năm 1975. Cây quyết định là một trong những thuật toán phân loại đầu tiên có khả năng dự đoán rất mạnh mẽ, được ứng dụng trong nhiều lĩnh vực của trí tuệ nhân tạo. Thuật toán được biểu diễn dưới dạng một lưu đồ có cấu trúc hình cây, trong đó mỗi nút bên trong biểu thị một phép thử trên một thuộc tính cần đánh giá, mỗi nhánh cây biểu thị kết quả của phép thử, và mỗi nút lá (nút đầu cuối của mỗi nhánh) chứa một nhãn phân loại. Qua đó, thuật toán được chia thành 2 mô hình là phân loại và hồi quy tuân theo cách tiếp cận đệ quy từ trên xuống, trong đó cây nhị phân sẽ phân vùng không gian dự báo sử dụng các biến số để thành các tập con để huấn luyện cho thuật toán. Các biến số độc lập không kết nối đến tiến trình dự báo sẽ được tách rời và đồng nhất với kết quả [21].



Hình 2.3: Mô hình thuật toán cây quyết định

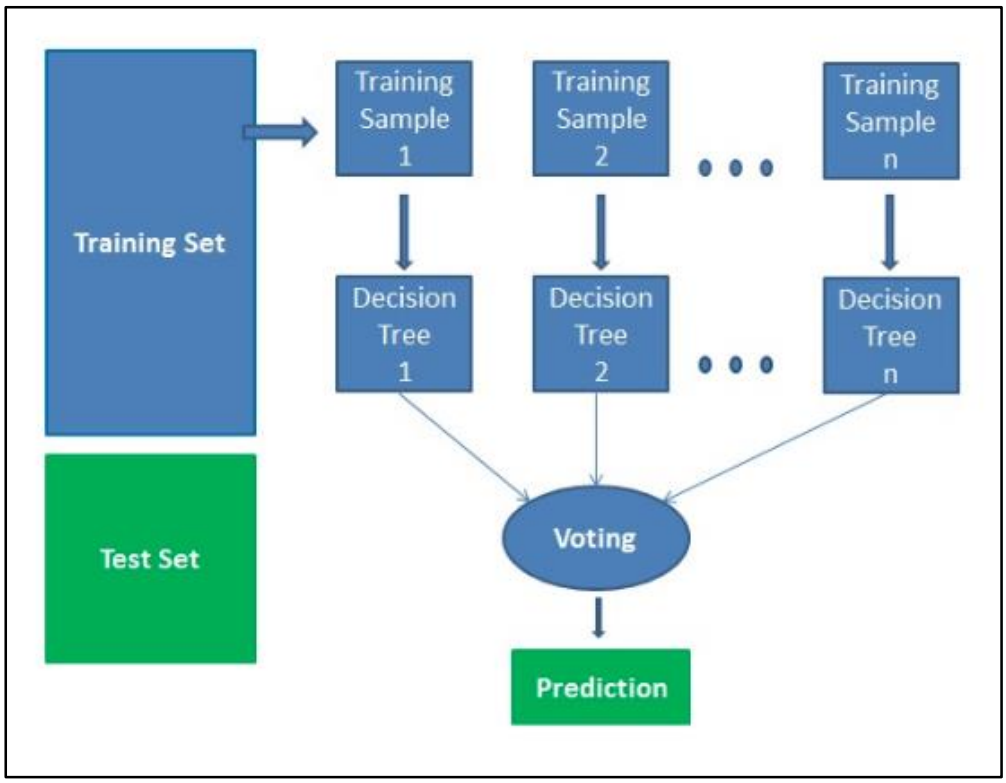
(Nguồn: Sưu tầm trên Internet)

Một cây có thể "học" hay “huấn luyện” bằng cách tách tập dữ liệu chính thành các tập con dựa trên các phép thử thuộc tính. Quá trình này được lặp đi lặp lại trên từng tập con dẫn xuất theo nguyên tắc đệ quy, trong trường thuật toán này là phân vùng đệ quy. Quá trình đệ quy sẽ hoàn thành khi tập hợp con tại một nút có cùng giá trị của biến số mục tiêu hoặc khi việc tách nhỏ không còn thêm giá trị vào các phép thử dự đoán. Việc xây dựng bộ phân loại cây quyết định không yêu kiến thức cao về kỹ thuật hoặc thiết lập tham số, bộ phân loại cũng rất dễ đọc và dễ hiểu cho người dùng. do đó thích hợp cho việc khai phá kiến thức và khuôn mẫu sẵn có của dữ liệu [33]. Ngoài ra, thuật toán cây quyết định có thể xử lý dữ liệu có chiều sâu mà vẫn có độ chính xác khá tốt, là một cách tiếp cận quy nạp điển hình để khai phá kiến thức về phân loại dữ liệu. Cây quyết định phân loại các cá thể bằng cách sắp xếp theo cây từ gốc đến một số nút lá, giúp tối ưu việc phân loại cá thể. Một cá thể được phân loại bằng cách bắt đầu từ nút gốc của cây, kiểm tra thuộc tính được chỉ định bởi nút này, sau đó di chuyển xuống nhánh cây tương ứng với giá trị của thuộc tính. Quá trình này sau đó được lặp lại cho các cây con bắt nguồn từ nút mới từ cây cha phía trước. Ngoài

những điểm mạnh của thuật toán trên, cây quyết định cũng bộc lộ một số điểm yếu trong các bài toán phân loại với nhiều lớp và số lượng ví dụ huấn luyện tương đối nhỏ, từ đó dẫn đến vấn hiện tượng "quá phù hợp" (overfitting). Không chỉ vậy, thuật toán cây quyết định có khuynh hướng xác định những tính chất không có sự liên hệ hoặc thậm chí gây cản trở trong quá trình "trồng cây", do vậy gây tiêu tốn tài nguyên tính toán [9].

2.2.2. Thuật toán rừng ngẫu nhiên

Rừng ngẫu nhiên hoặc rừng quyết định ngẫu nhiên là một phương pháp học máy tổng hợp để phân loại và hồi quy bằng cách xây dựng rất nhiều cây quyết định tại thời điểm huấn luyện và xuất ra lớp cây trong đó chứa tham số trung bình dự đoán (khi dùng để hồi quy) và các phân hình lớp (khi dùng để phân loại). Rừng ngẫu nhiên được phát triển và xây dựng trên nguyên lý "trí thông minh của bầy đàn", trong đó quyết định sẽ được trao cho nhiều thành viên và quyết định nào nhận được sự ủng hộ cao nhất sẽ là quyết định của cả tập thể [26]. Các phương pháp học máy tổng hợp như rừng ngẫu nhiên sẽ khắc phục được hạn chế của đặc thù của thuật toán cây quyết định khi thuật toán này thường vấp phải hiện tượng "quá phù hợp". Việc kiến tạo nhiều cây quyết định dựa trên nhiều đặc tính khác nhau của tập dữ liệu và "trồng" cây một cách độc lập để lấy kết quả trung bình cao nhất sẽ mang lại độ chính xác cao hơn hẳn so với kết quả từ một cây quyết định [23].



Hình 2.4: Mô hình thuật toán rừng ngẫu nhiên

(Nguồn: Sưu tầm trên Internet)

Dựa vào đặc thù xây dựng cây quyết định của rừng ngẫu nhiên, có thể thấy thuật toán này tuân theo một xu hướng cải thiện độ chính xác là "bao hàm", được sử dụng rất phổ biến trong các thuật toán tiếp cận phương pháp hình cây. Bao hàm (bagging) là phương pháp phát triển các cây kế tiếp độc lập với các cây trước đó, tức là mỗi cây được xây dựng bằng cách sử dụng một mẫu dữ liệu ngẫu nhiên và sau quá trình "trồng cây", đa số phiếu được lấy để dự đoán quyết định [26]. Thuật toán rừng ngẫu nhiên thêm một lớp ngẫu nhiên bổ sung vào việc bao hàm và thay đổi cách cấu trúc các cây quyết định: trong cây quyết định tiêu chuẩn, mỗi nút được tách bằng cách sử dụng cách tách tốt nhất trong số tất cả các biến dự báo trong khi tại rừng ngẫu nhiên, các nút được tách bằng cách tốt nhất nhất trong một tập hợp con các yếu tố dự báo được chọn ngẫu nhiên tại nút đó [15].

Nhìn chung, các phương pháp dựa trên hình cây có hiệu quả tốt hơn các phương pháp tiếp cận đã được thiết lập khác trong nhiều nhiệm tác vụ phân loại khác nhau như phân loại luồng lưu lượng truy cập IP [30], dự đoán churn của khách hàng [33], hoặc dự đoán về ý định mua hàng trực tuyến [33]. Rừng ngẫu nhiên cho thấy sự vượt trội vì các phương pháp tổng hợp này có thể giảm cả độ lệch và phương sai của các thuật toán học đơn lẻ. Trong khi các mô hình riêng lẻ có thể bị mắc kẹt trong cực tiểu cục bộ, sự kết hợp có trọng số của một số cực tiểu cục bộ khác nhau - được tạo ra bằng phương pháp tổng hợp - có thể giảm thiểu rủi ro chọn giá trị tối thiểu cục bộ cho so sánh và dự đoán của cả thuật toán rừng ngẫu nhiên.

2.3. Ứng dụng học máy trong dự đoán người dùng bỏ rơi giỏ hàng

Hiện tượng người tiêu dùng bỏ rơi giỏ hàng không chỉ gây thiệt hại đáng kể về doanh thu cho các công ty bán lẻ mà còn suy yếu khả năng cạnh tranh trên thị trường thương mại điện tử trong khu vực. Do đó, rất nhiều nghiên cứu đã tiếp cận vấn đề này dựa những phân tích hành vi của người dùng nhằm tìm ra các yếu tố dẫn đến hiện tượng bỏ rơi giỏ. Dựa trên những nghiên cứu tâm lý này cùng nền tảng phát triển của khoa học công nghệ, các nghiên cứu tiếp theo đã có sự thay đổi trong cách tiếp cận và xử lý vấn đề, chuyển từ các phương pháp tâm lý sang các phân tích dựa trên dữ liệu người tiêu dùng, từ đó có những dự đoán về khả năng giỏ hàng bị bỏ rơi. Nhờ các tiếp cận mới này, các nghiên cứu không chỉ đưa ra được những phân tích hành vi mua sắm, sử dụng giỏ hàng trực tuyến của khách hàng mà còn có những dự đoán cụ thể về tỷ lệ giỏ hàng sẽ bị bỏ rơi từ việc xử lý dữ liệu người dùng.

Dữ liệu sử dụng cho những nghiên cứu và phân tích hành vi người tiêu dùng thường rất đa dạng từ thông tin các nhân khách hàng (ví dụ: giới tính, địa chỉ, ngày sinh, sở thích v...v), lịch sử mua sắm, mức độ tương tác, khuynh hướng hành vi trong từng phiên giao dịch [30]. Chính sự đa dạng và phong phú về dữ liệu hành vi tiêu dùng khiến thương mại điện tử là một nguồn dữ liệu tiềm năng trong ứng dụng học

máy. Do đó, có rất nhiều thuật toán và hệ thống học máy có thể được ứng dụng vào khai thác tiềm năng dữ liệu mua sắm từ người tiêu dùng.

Đối với vấn đề bỏ rơi giỏ hàng, việc dự đoán khả năng người tiêu dùng có thực hiện giao dịch mua sắm có thể được coi là một một tác vụ phân loại nhị phân. Phân loại nhị phân là thuật toán học máy có giám sát nhằm phân loại các phần tử của một tập hợp thành hai nhóm riêng biệt dựa trên cơ sở đặc tính của phần tử. Thuật toán phân loại nhị phân có thể được trình như sau:

Cho $\{y_k, x_k\}_{k=1}^{N}$ là tập dữ liệu huấn luyện, trong đó
K = 2 (vì là thuật toán phân loại nhị phân)
$y_k \in \{0,1,2,\ldots,K-1\}$là một phần tử của tập hợp
$x_k = R^n$ là vector của các giá trị dự đoán
Thuật toán sẽ được huấn luyện để dự đoán nhãn y_k từ x_k

Bốn thuật toán phân loại nhị phân tiêu biểu là: *cây quyết định, rừng ngẫu nhiên, máy hỗ trợ vector* và *K hàng xóm gần nhấ*t sẽ được thử nghiệm và so sánh hiệu năng trong khai thác và dự đoán giỏ hàng bị bỏ rơi.

2.3.1. Dữ liệu và bối cảnh thử nghiệm

Trong thử nghiệm nhằm so sánh hiệu năng phân loại của bốn thuật toán: cây quyết định, rừng ngẫu nhiên, máy hỗ trợ vector và K hàng xóm gần nhất; dữ liệu người dùng được thu thập từ các tệp nhật ký máy chủ của một công ty phân phối và bán lẻ thời trang hàng đầu tại Đức. Dữ liệu được tạo bằng cách trích xuất các hoạt động mua sắm trực tuyến theo trình tự thời gian của khách hàng từ các tệp lưu trữ lịch sử hành động mua sắm. Mỗi tệp nhật ký ghi lại các hoạt động mua sắm và tiêu dùng trong một phiên sử dụng của người dùng, ví dụ các hành động như đăng nhập, thanh toán giỏ hàng, xem thông tin chi tiết sản phẩm, thêm sản phẩm vào giỏ hàng v...v. Dữ liệu bao gồm 3.511.037 phiên mua sắm trong khoảng thời 3 tháng gian từ

ngày 1 tháng 2 năm 2019 đến ngày 30 tháng 4 năm 2019 [24]. Từ dữ liệu thu thập được, 18 tiêu chí dữ liệu được lựa chọn và trình bày trong bảng mô tả sau.

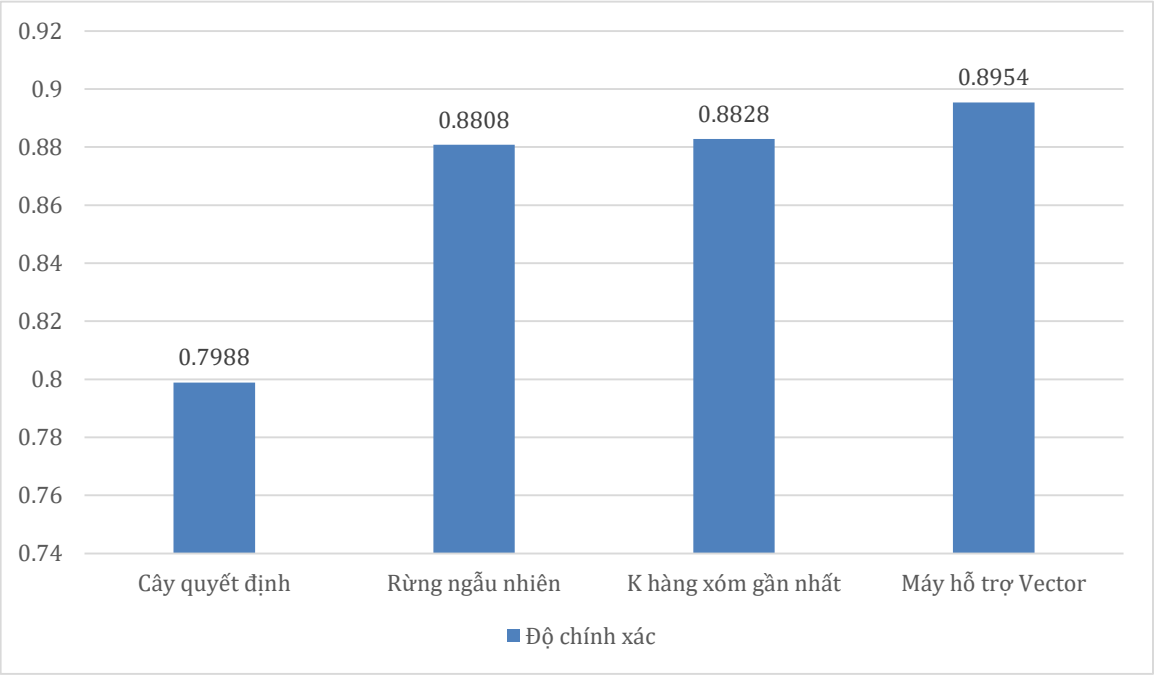
Bảng 2.1: Bảng mô tả và chỉ mục dữ liệu dùng cho so sánh 4 thuật toán

Chỉ mục dữ liệu	Mô tả dữ liệu
SCA	Phân loại giỏ hàng bị bỏ rơi
BASKETS_BB	Số lượng giỏ hàng được chuyển đổi thành đơn hàng
BASKETS	Số lượng giỏ hàng được khởi tạo trong phiên mua sắm
LOGS	Số lần đăng nhập của khách hàng
LOGS_CUST_STEP2	Số lượng khách hàng thực hiện giao dịch bước thanh toán
LOGS_NEWCUST_STEP2	Số lượng khách hàng mới thực hiện giao dịch đến bước thanh toán
PIS	Số lượt xem tổng thể
PIS_AP	Số lượt xem giỏ hàng
PIS_DV	Số lượt xem sản phẩm
PIS_PL	Số lượt xem danh mục sản phẩm
PIS_SDV	Số lượt xem sản phẩm qua tìm kiếm
PIS_SR	Số lượt tìm kiếm sản phẩm
POSITIONS	Số loại sản phẩm trong giỏ hàng
QUANTITY	Số lượng sản phẩm trong giỏ hàng
VALUE_BB	Tổng giá trị của giỏ hàng
NEW_CUST	Khách hàng mới
WEB_CUST	Khách hàng sử dụng máy tính cá nhân
MOBILE_CUST	Khách hàng sử dụng thiết bị di động

(Nguồn: [24] Rausch, Theresa & Derra, Nicholas & Wolf, Lukas (2020), "Predicting online shopping cart abandonment with machine learning approaches" International Journal of Market Research.)

Trong tổng số hơn 3,5 triệu phiên mua sắm của người tiêu dùng, có 520.653 (63,41% tổng số giỏ hàng) trường hợp người tiêu dùng bỏ rơi giỏ hàng và 300.395 (36,59% tổng số giỏ hàng) giỏ hàng được chuyển đổi thành công thành đơn hàng. Do vậy, tập dữ liệu có sự phân lập rõ rệt về tỷ lệ giỏ hàng thành công và giỏ hàng bị bỏ rơi, giúp đánh giá độ chính xác của các thuật toán dễ dàng hơn.

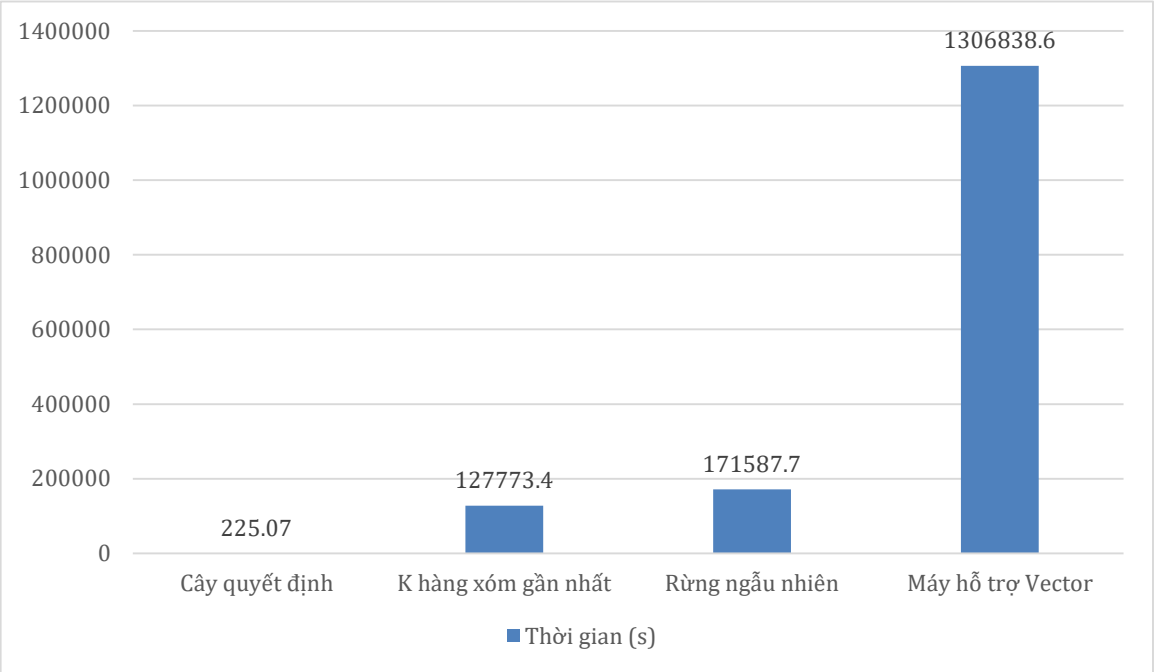
2.3.2. So sánh kết quả thuật toán



Hình 2.5: Biểu đồ so sánh độ chính xác của 4 thuật toán [24]

Từ thống kê độ chính xác của 4 thuật toán, có thể thấy thuật toán cây quyết định cho độ chính xác thấp nhất trong tổng số 4 thuật toán. Điều này không quá bất ngờ vì thuật toán cây quyết định thường mất đi độ chính xác khi các tiêu chí phân loại phần tử và lượng dữ liệu tương đối lớn. Hai thuật toán Máy hỗ trợ Vector và K hàng xóm gần nhất có hiệu năng tương đối đồng đều ở mức 0,8808 và 0,8828 khi chỉ

có sự chênh lệch khoảng 0,002. Dẫn đầu là thuật toán rừng ngẫu nhiên với mức độ chính xác khoảng 0,8954 cao hơn 0.1 so với thuật toán có độ chính xác thấp nhất là cây quyết định. Rừng ngẫu nhiên cho thấy khả năng phân loại chính xác vượt trội hơn vì thuật toán đã khắc phục được những nhược điểm về dữ liệu và tiêu chí phân loại lớn của cây quyết định. Đồng thời rừng ngẫu nhiên có khả năng đưa ra được những tiêu chí phân loại quan trọng giúp quyết định giá trị của mục tiêu phân loại.



Hình 2.6: Biểu đồ so sánh tốc độ xử lý của 4 thuật toán [24]

Đối với tốc độ xử lý phân loại, thuật toán máy hỗ trợ Vector có thời gian xử lý chậm rõ rệt với hơn một 1.300.000 giây (khoảng hơn 350 giờ) để khai thác và phân loại hơn 3.500.000 bản ghi. Đối lập với tốc độ của máy hỗ trợ Vector, thuật toán cây quyết định chỉ mất khoảng 225 giây để xử lý lượng thông tin đó và cho ra kết quả phân loại, nhưng độ chính xác lại thấp nhất trong cả 4 thuật toán. Rừng ngẫu nhiên và K hàng xóm gần nhất có kết quả gần tương quan nhau với cách biệt khoảng 50.000 giây trong tốc độ xử lý. Tuy nhiên, thuật toán rừng ngẫu nhiên lại đem lại độ chính xác cao hơn đáng kể so với thuật toán K hàng xóm gần nhất.

Từ những so sánh và phân tích về hiệu năng, tốc độ xử lý của 4 thuật toán tiêu biểu trong phân loại nhị phân, có thể thấy rừng ngẫu nhiên là thuật toán có sự cân bằng giữa tốc độ xử lý và độ chính xác trong phân loại hành vi người tiêu dùng. Điều này không chỉ giúp cho việc phân loại và dự đoán người tiêu dùng bỏ rơi giỏ hàng được thuận lợi và nhanh chóng mà kết quả dự đoán có khả năng chính xác cao.

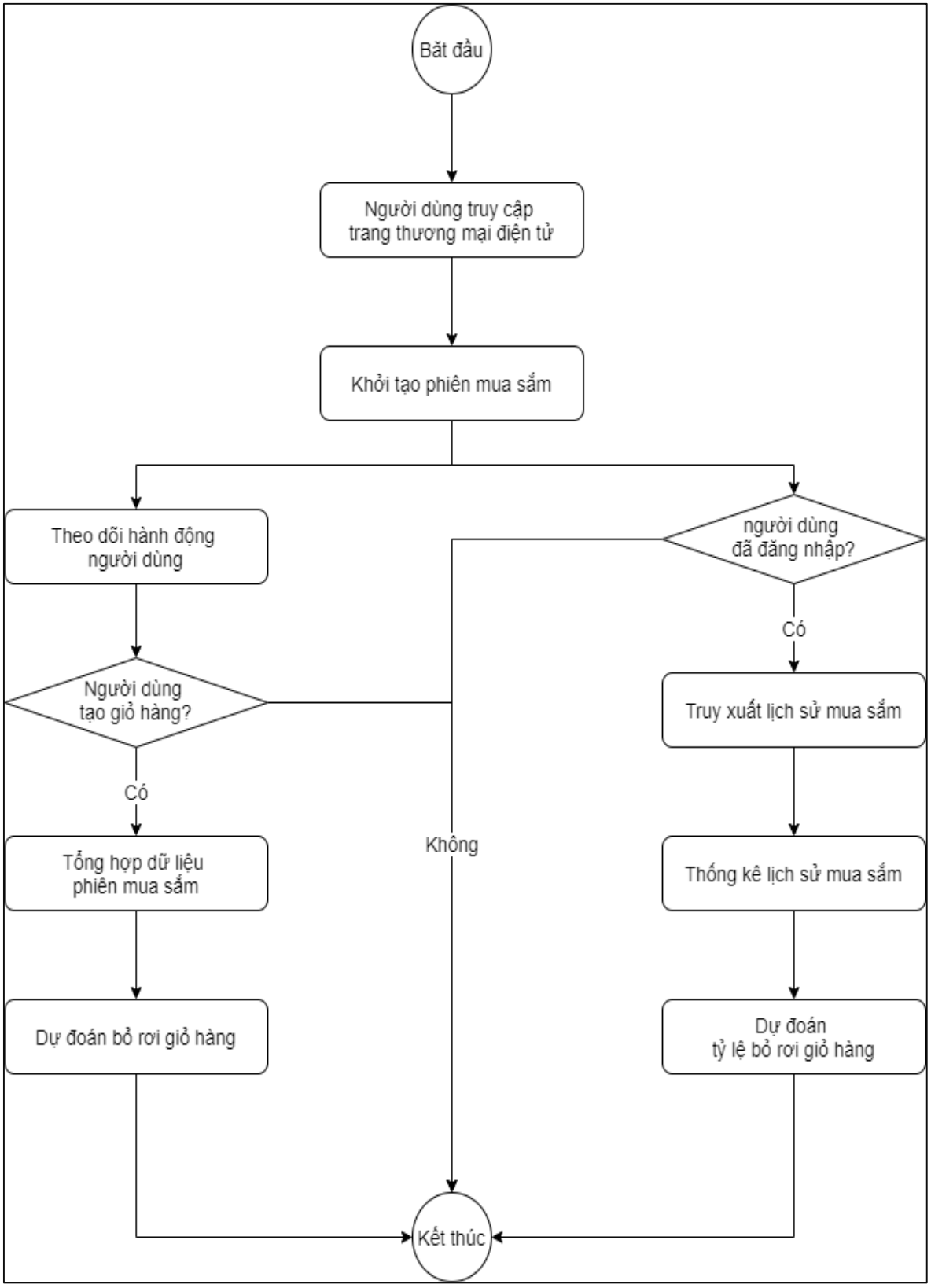
2.4. Kết luận

Từ hành vi bỏ rơi giỏ hàng trong chương một, chương hai của luận văn đã làm rõ hơn các nguyên nhân chính dẫn đến hiện tượng người dùng bở rơi giỏ hàng điện tử. Trong đó, luận văn đề cập đến ba yếu tố chính là trải nghiệm người dùng, chi phí vận chuyển cao, không minh bạch và khả năng nhận thức rủi ro của người dùng. Đồng thời, thuật toán rừng ngẫu nhiên cũng được giới thiệu và đề cập cùng thuật toán cây quyết định là đơn vị căn bản và cốt lõi của thuật toán rừng ngẫu nhiên. Một số thuật toán phổ biến trong phân loại dữ liệu cũng được đề cập và so sánh tốc độ và hiệu năng xử lý tác vụ phân tích dữ liệu. Trên những cơ sở so sanh đó, thuật toán rừng ngẫu nhiên thể hiện được sự hài hòa giữa thời gian xử lý cũng như độ chính xác của phân loại, từ đó cho thấy lý do ứng dụng rừng ngẫu nhiên trong phân loại và dự đoán hành vi bỏ rơi giỏ hàng.

CHƯƠNG 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ

Với mục đích ứng dụng thuật toán rừng ngẫu nhiên để phân tích thói quen mua sắm của người tiêu dùng và dự đoán hành vi bỏ rơi giỏ hàng, dữ liệu người dùng thực tế đã được thu thập từ một trang thương mại điện tử chuyên cung cấp, buôn bán vật liệu xây dựng và thiết bị vệ sinh phòng tắm. Toàn bộ hoạt động của khách hàng từ đăng nhập, đăng ký tài khoản mới, tìm kiếm, đánh giá, nhận xét sản phẩm đến thêm sản phẩm vào giỏ hàng, thanh toán giỏ hàng v...v sẽ được lưu lại trong cơ sở dữ liệu. Trong quá trình lưu trữ, mỗi bản ghi hoạt động cũng được gán tương ứng với một phiên mua sắm của khách hàng tại một thời điểm nhất định. Trong bối cảnh này, một phiên mua sắm là khoảng thời gian duyệt web liên tục hoặc một chuỗi các lần xem trang của người dùng cho đến khi người dùng thoát khỏi cửa hàng trực tuyến [21].

Tiến trình phân tích dữ liệu người dùng và dự đoán bỏ rơi giỏ hàng được chia thành hai luồng tương ứng với hai loại dữ liệu là *phân luồng trực tiếp* và *phân luồng gián tiếp* tương ứng với hai bài toán là thống kê dữ liệu ở phần luồng gián tiếp và phân tích hành vi ở phân luồng trực tiếp. Phân luồng trực tiếp sử dụng dữ liệu là lịch sử hoạt động của người dùng và có khả năng phân tích theo thời gian thực dựa trên những biến số thay đổi trong hành vi của người mua sắm. Luồng phân tích trực tiếp sẽ phù hợp với những khách hàng mới, chưa có tài khoản tại hệ thống thương mại điện tử hoặc đối với phiên mua sắm mà khách hàng không thực hiện đăng nhập. Phân luồng gián tiếp dựa trên lịch sử mua sắm người dùng sẽ sử dụng dữ liệu lịch sử giao dịch đã được tổng hợp và lưu trữ trong hệ thống thương mại điện tử. Luồng thống kê gián tiếp mặc dù không có khả năng phân tích theo thời gian nhưng lại có thể đưa ra một tỷ lệ tương đối chính xác dựa trên các thống kê lịch sử giao dịch và hành động tiêu dùng trước đó của khách hàng.



Hình 3.1: Sơ đồ tiến trình phân tích dữ liệu và dự báo bỏ rơi giỏ hàng

3.1. Phát biểu bài toán

3.1.1. Bài toán phân tích và dự đoán phân luồng trực tiếp

Trong điều kiện lý tưởng nhất, người mua hàng đã là người dùng có sẵn của trang thương mại điện tử và cũng đã có một lịch sử mua sắm tương đối. Do đó, hệ thống trang thương mại điện tử và dịch vụ phân tích giỏ hàng sẽ tổng hợp dữ kết quả từ hai phân luồng dữ liệu trực tiếp và gián tiếp. Mỗi phân luồng sẽ có tỷ trọng 50% trong tổng hợp kết quả dự đoán. Tuy nhiên, trong thực tế thử nghiệm và sử dụng bởi người tiêu dùng, phân luồng trực tiếp được sử dụng để phân tích và dự đoán hành vi bỏ rơi giỏ hàng chiếm đa số do thói quen tiêu dùng trực tuyến không cần đăng nhập của khách hàng. Chính vì vậy, bài toán phân tích và dự đoán phân luồng trực tiếp sẽ phù hợp cho những khách hàng vãng lai không có tài khoản hoặc những khách hàng chỉ đăng nhập để hoàn tất quá trình mua sắm.

Bài toán phân tích và dự đoán phân luồng trực tiếp nhận đầu vào là các dữ liệu tổng hợp hoạt động của khách hàng trong các phiên mua sắm nhất định tại trang thương mại điện tử. Tập dữ liệu chủ yếu tổng hợp những hoạt động xem sản phẩm và sử dụng giỏ hàng của người tiêu dùng. Đây sẽ là tập dữ liệu được ứng dụng thuật toán rừng ngẫu nhiên để phân loại và dự đoán hành vi bỏ rơi giỏ hàng. Kết quả của bài toán là một giá trị dự đoán nhị phân 1 hoặc 0 với giá trị bằng 1 là người dùng sẽ hoàn tất giỏ hàng và 0 là người dùng sẽ bỏ rơi giỏ hàng.

Mục đích của bài toán phân tích và dự đoán phân luồng trực tiếp là sử dụng dữ liệu trong phiên mua sắm hiện tại để đối chiếu với khuôn mẫu mua sắm đã xây dựng từ trước nhằm đưa ra phân loại bỏ rơi giỏ hàng. Bài toán sẽ thử nghiệm khả năng tổng hợp hoạt động trong phiên mua sắm và gửi yêu cầu dự đoán đến dịch vụ phân tích của trang thương mại điện tử. Đồng thời, kiểm tra khả năng dự đoán hành động bỏ rơi giỏ hàng trong phiên mua sắm hiện tại của dịch vụ phân tích dựa trên khuôn mẫu hành vi đã được xây dựng trước đó. Từ những kết quả trên, đưa ra đánh giá về tỷ lệ chính xác của dự đoán bỏ rơi giỏ hàng so với kết quả thực tế từ khách hàng trực tuyến.

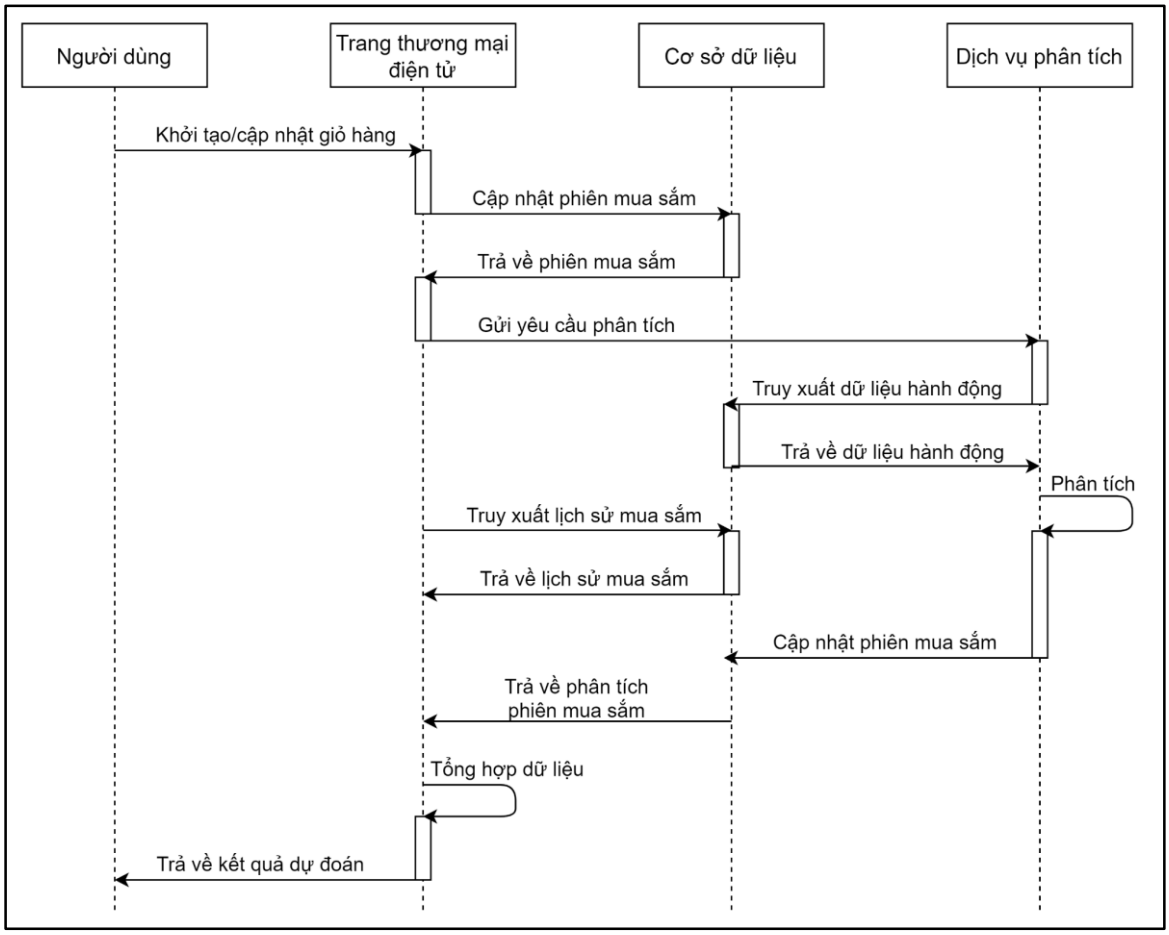
3.1.2. Bài toán thống kê dữ liệu phân luồng gián tiếp

Một trong những trở ngại lớn trong ứng dụng các thuật toán học có giám sát như rừng ngẫu nhiên là thuật toán cần dữ liệu và thời gian để tìm ra các khuôn mẫu hành vi của người tiêu dùng. Do đó, bài toán thống kê dữ liệu phân luồng gián tiếp sẽ nhằm mục đích giải quyết trở ngại ban đầu khi ứng dụng thuật toán vào trang thương mại điện tử. Dữ liệu tại phân luồng gián tiếp sẽ được tổng hợp từ lịch sử mua sắm của tập khách hàng sẵn có ví dụ: tổng số lượng đơn hàng công, tổng số lượng giỏ hàng, tổng giá trị giao dịch trên toàn hệ thống. Phân luồng dữ liệu gián tiếp không chỉ thích hợp cho việc phân tích lịch sử mua hàng của tập khách hàng nội tại mà còn phù hợp để đánh giá tiềm năng và khuynh hướng mua sắm của tập khách hàng tiềm năng.

Bài toán thống kê dữ liệu phân luồng gián tiếp nhận đầu vào là các dữ liệu được thống kê và tổng hợp từ các lịch sử mua sắm, những hoạt động và tương tác của khách hàng . Từ dữ liệu đã được tổng hợp, bài toán sẽ cho ra kết quả là một tỷ lệ phần trăm cho khả năng thực hiện mua hàng sắm tới của từng khách hàng hiện tại trong hệ thống. Kết quả dự đoán chủ yếu dựa trên số lượng giỏ hàng hoàn thành và số lượng đơn hàng thành công của khách hàng tại trang thương mại điện tử.

Mục tiêu của bài toán thống kê dữ liệu phân luồng gián tiếp là tổng hợp toàn bộ thông tin giao dịch mua sắm thành công và thất bại cũng như những đánh giá và nhận xét của tập khách hàng sẵn có tại trang thương mại điện tử. Những thống kê này một phần hỗ trợ dịch vụ phân tích đưa ra dự đoán bỏ rơi giỏ hàng trong thời gian hình thành khuôn mẫu hành vi, phần khác thử nghiệm khả năng tổng hợp và truy vấn dữ liệu thống kê của trang thương mại điện tử. Ngoài mục tiêu hỗ trợ và thử nghiệm, kết quả của bài toán cũng đánh giá khuynh hướng tiêu dùng và thái độ tin tưởng của khách hàng đối với trang thương mại điện tử dựa trên lịch sử mua sắm. Những đánh giá trên góp phần hoàn thiện mục tiêu chung của luận văn là tổng hợp khuôn mẫu hành động mua sắm của khách hàng.

3.1.3. Tổng hợp bài toán và trình tự phân tích



Hình 3.2: Sơ đồ trình tự phân tích lịch sử và hoạt động trong phiên mua sắm

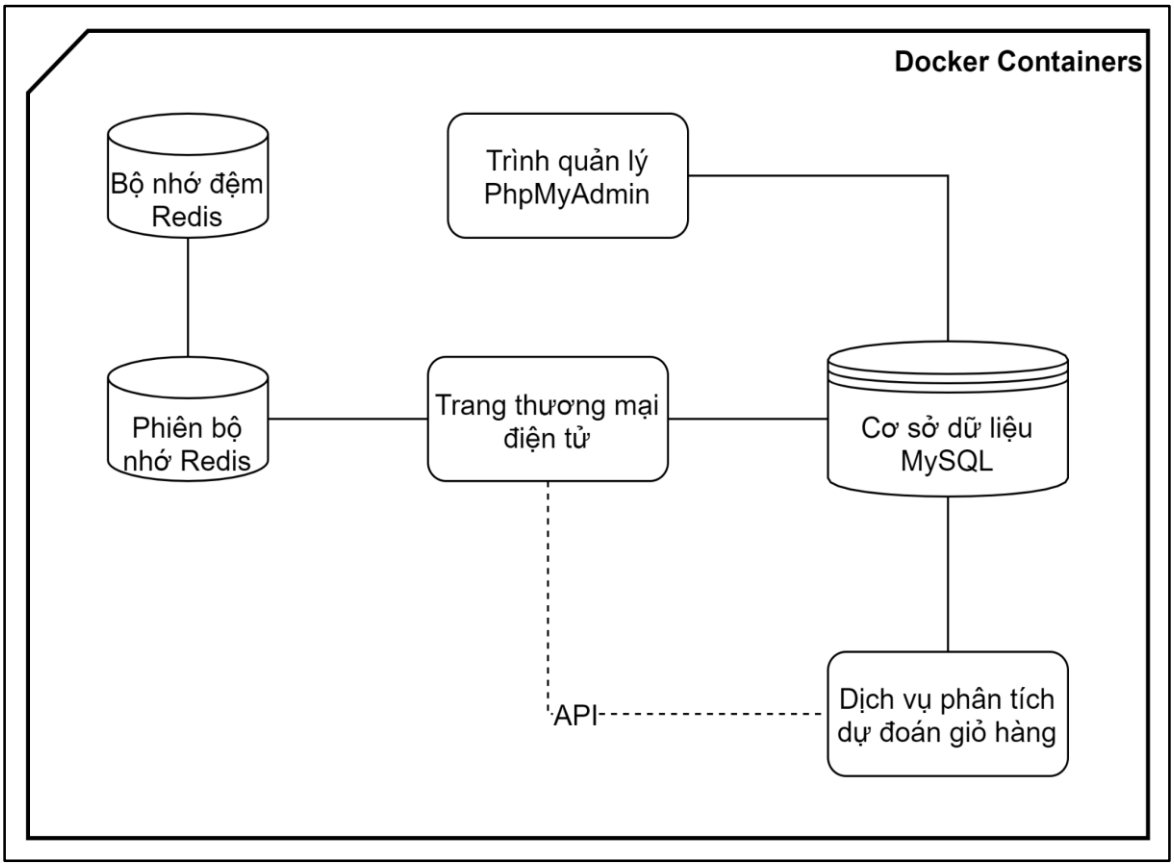
Quá trình phân tích và tổng hợp dữ liệu được bắt đầu ngay sau khi người dùng thêm sản phẩm đầu tiên vào giỏ hàng điện tử, hay nói cách khác là ngay sau khi người dùng khởi tạo giỏ hàng. Khi đó, trang thương mại điện tử sẽ cập nhật thông tin giỏ hàng đến phiên mua sắm và đồng thời gửi yêu cầu phân tích phiên mua sắm đến dịch vụ phân tích. Yêu cầu sẽ được gửi bằng tín hiệu API theo phương thức bất đồng bộ để tối ưu hóa quá trình xử lý hoạt động mua sắm tại trang thương mại điện tử. Việc gửi tín hiệu theo phương thức bất đồng bộ không chỉ tiết kiệm thời gian chờ xử lý cho trang thương mại điện tử mà còn tránh được những gián đoạn không đáng có khi dịch vụ phân tích gặp lỗi hoặc trục trặc trong quá trình hoạt động. Ngoài ra, đối với người dùng đã đăng nhập, lịch sử mua sắm sẽ được truy xuất để tổng hợp với kết quả dự đoán bỏ rơi giỏ hàng.

Ngay sau khi nhận được yêu cầu phân tích từ trang thương mại điện tử, dịch vụ phân tích sẽ truy xuất lịch sử hoạt động dựa theo định danh phiên mua sắm nhận từ yêu cầu. Các thông tin cần thiết trong phiên hoạt động sẽ được truy xuất trực tiếp từ cơ sở dữ liệu và tổng hợp để chuẩn bị cho tác vụ dự đoán người dùng bỏ rơi giỏ hàng. Tiếp đến, dịch vụ phân tích sẽ truy xuất lại khuôn mẫu dữ liệu đã được huấn luyện và chuẩn bị trước đó, ứng dụng vào phiên hoạt động mới và dự đoán kết quả bỏ rơi giỏ hàng của người tiêu dùng. Kết quả dự đoán sẽ là một giá trị nhị phân, trong đó giá trị đúng (1) là người dùng thực hiện thanh toán và hoàn thành giỏ hàng, giá trị sai (0) là người dùng không thực hiện thanh toán, bỏ rơi giỏ hàng. Sau khi hoàn thành dự đoán, dịch vụ phân tích sẽ cập nhật kết quả vào phiên mua sắm dựa trên định danh được gửi từ yêu cầu phân tích.

Trong quá trình dịch vụ phân tích thực hiện dự đoán, trang thương mại điện tử sẽ liên tục kiểm tra kết quả dự đoán trong phiên mua sắm của người dùng tại thời điểm đó. Ngay sau khi kết quả dự đoán được cập nhật, trang thương mại điện tử sẽ tổng hợp kết quả dự đoán bỏ rơi giỏ hàng cùng với lịch sử mua sắm đối với người dùng đã đăng nhập. Giá trị dự đoán người dùng bỏ rơi giỏ hàng cuối cùng sẽ được tổng hợp từ trung bình cộng của hai kết quả dự đoán và thống kê lịch sử mua sắm.

3.2. Cấu trúc hệ thống và dữ liệu

3.2.1. Cấu trúc trang thương mại điện tử và dịch vụ phân tích



Hình 3.3: Cấu trúc trang thương mại điện tử và dự đoán bỏ rơi giỏ hàng

Nhằm giảm thời gian cấu hình và cài đặt đồng thời tăng khả năng tương thích với mọi môi trường máy chủ, trang thương mại điện tử xây dựng trên nền tảng Magento 2 cùng các thành phần ngoại vi và dịch vụ phân tích giỏ hàng sẽ được quản lý và phân bổ tài nguyên bởi nền tảng ảo hóa Docker. Nền tảng ảo hóa Docker không chỉ hỗ trợ việc cài đặt và cấu hình trong lần chạy ban đầu mà còn giúp cho các dịch vụ thành phần được kết nối và hoạt động một cách đồng nhất, tránh rủi ro về khác biệt phiên bản, hệ điều hành hay thiếu sót các ứng dụng cơ bản.

Cấu trúc hệ thống trang thương mại điện tử cùng dịch vụ phân tích và dự đoán bỏ rơi giỏ hàng được được chia nhỏ thành từng khối hệ thống và hoạt động dựa trên trình quản lý Docker. Hệ thống sẽ gồm 3 thành phần chính là trang thương mại điện

tử sử dụng nền tảng Magento 2, hệ quản trị cơ sở dữ liệu MySQL, và dịch vụ phân tích và dự đoán bỏ rơi giỏ hàng. Trang thương mại điện tử và dịch vụ phân tích sẽ được kết nối trực tiếp với hệ cơ sở dữ liệu để tối ưu hóa tốc độ ghi đọc và quản lý dữ liệu. Việc kết nối dịch vụ phân tích giỏ hàng điện tử trực tiếp với cơ sở dữ liệu MySQL sẽ giảm gánh nặng xử lý logic và dữ liệu đối với trang thương mại điện tử. Không chỉ vậy, kết nối trực tiếp cũng giúp cho việc đọc ghi dữ liệu của dịch vụ phân tích nhanh hơn và không phải đi đường vòng qua hệ thống nguồn của trang thương mại điện tử. Bên cạnh đó, dịch vụ phân tích cũng sẽ mở cổng kết nối API để trang thương mại điện tử có thể gửi tín hiệu yêu cầu dự đoán bỏ rơi giỏ hàng. Trang thương mại điện tử và dịch vụ phân tích sẽ hoạt động song song trên hệ cơ sở dữ liệu, khi cần kích hoạt tác vụ phân tích và thu thập kết quả, trang thương mại điện tử sẽ kết nối với dịch vụ phân tích qua cổng API đã được mở từ trước. Dịch vụ phân tích sẽ nhận định danh phiên mua sắm cần phân tích do trang thương mại điện tử gửi đến, từ đó phân tích dữ liệu, đưa ra dự đoán và cập nhật vào cơ sở dữ liệu. Sau quá trình cập nhật, trang thương mại điện tử cũng sẽ nhận được kết quả dự đoán bỏ rơi giỏ hàng mỗi lần truy xuất thông tin của phiên mua sắm tại thời điểm đó.

Ngoài các thành phần chính của hệ thống, một số thành phần phụ trợ như trình quản lý bộ nhớ đệm Redis và Redis Session được sử dụng nhằm nâng cao trải nghiệm người dùng tại trang thương mại điện tử. Việc sử dụng một khối thành phần chuyên biệt cho bộ nhớ đệm như Redis sẽ giúp trang thương mại điện tử bảo toàn được hiệu năng tải trang và đồng thời giảm đi băng thông khi lưu bộ nhớ đệm tại cơ sở dữ liệu. Không chỉ vậy, giao diện quản lý cơ sở dữ liệu PhpMyAdmin được sử dụng để giúp quản trị viên hệ thống dễ dàng quản lý cơ sở dữ liệu và sao lưu hệ thống khi cần thiết. Giao diện quản lý cơ sở dữ liệu PhpMyAdmin không thật sự quá cần thiết trong các hệ thống thực tế vì có rất nhiều phần mềm khác có chức năng tương tự có thể sử dụng ở máy tính người dùng mà không tiêu tốn tài nguyên của chính máy chủ.

3.2.2. Cấu trúc dữ liệu

Đối với phân luồng trực tiếp, dữ liệu thử nghiệm trong luận văn sẽ bao gồm tất cả các phiên mua sắm của 188 khách hàng đăng nhập và các khách hàng vãng lai khác được tổng hợp từ một trang thương mại điện tử cung cấp vật liệu xây dựng và thiết bị vệ sinh phòng tắm trong khoảng thời gian từ ngày 01 tháng 12 năm 2020 đến ngày 15 tháng 02 năm 2021 - khoảng 2,5 tháng. Trong thời gian đó, Gần 10.000 hoạt động tiêu dùng trực tuyến trên trang thương mại điện tử được ghi lại và tổng hợp thành 1.173 phiên mua sắm khác nhau. Các hoạt động tiêu dùng được ghi lại bằng cách lưu nhật ký yêu cầu tài nguyên (yêu cầu xem sản phẩm, xem giỏ hàng, thêm sản phẩm vào giỏ hàng, thanh toán, v.v) của người dùng đến trang thương mại điện tử.

Với ưu thế xây dựng và phát triển dựa trên nền tảng Magento 2, nhật ký yêu cầu tài nguyên sẽ được trích xuất qua việc ứng dụng cơ chế "sự kiện – người quan sát". Trong đó, khi người dùng khởi tạo yêu cầu tài nguyên, một sự kiện sẽ được phát đi trong toàn bộ hệ thống kèm theo toàn bộ thông tin về ngữ cảnh hoạt động, tình trạng hệ thống, trạng thái phiên đăng nhập, v.v. Bằng cách khởi tạo một "người quan sát" để bắt sự kiện yêu cầu tài nguyên của người dùng, việc truy xuất các dữ liệu phục vụ cho phân luồng trực tiếp hoàn toàn có thể đạt được qua quá trình phần tách và tổng hợp các dữ liệu đính kèm sự kiện. Đồng thời, trong quá trình lưu nhật ký hành động, dữ liệu của phiên mua sắm cũng được tổng hợp và tính toán để luôn luôn ở trạng thái mới nhất. Các dữ liệu dạng chuỗi (*device_type, origin*) và dạng số thập phân (total_cart_value, average_interval) cũng được tinh chỉnh và chuyển đổi sang dạng số nguyên để chuẩn hóa dữ liệu đầu vào cho thuật toán rừng ngẫu nhiên.

Chỉ mục dữ liệu của phân luồng trực tiếp được tham khảo dựa trên nghiên cứu dự đoán bỏ rơi giỏ hàng của *Rausch, Theresa & Derra, Nicholas & Wolf, Lukas (2020)* [24]. Ngoài ra, một số dữ liệu về loại thiết bị truy cập (device_type), nguồn gốc khách hàng (origin) và thời gian trung bình giữa các hành động (average_interval) cũng được bổ sung vào phân luồng trực tiếp do đặc tính sẵn có của hệ thống. Bảng 3.1 mô tả chi tiết các chỉ mục dữ liệu có trong một bản ghi phiên mua sắm.

Bảng 3.1: Bảng chỉ mục dữ liệu của phân luồng trực tiếp

STT	Chỉ mục dữ liệu	Định nghĩa
1	customer_id	ID định danh khách hàng trong trường hợp khách hàng đăng nhập vào hệ thống
2	cart_id	ID định danh giỏ hàng
3	device_type	Loại thiết bị truy cập trong phiên mua sắm. Phân loại thiết bị truy cập: ● *Máy tính* ● *Điện thoại* ● *Máy tính bảng*
4	origin	Nguồn gốc truy cập. Phân loại nguồn gốc truy cập: Quảng cáo: người dùng truy cập từ các trang quảng cáo. ● *Trực tiếp*: người dùng truy cập trực tiếp địa chỉ. ● *Giới thiệu*: người dùng truy cập trực tiếp địa chỉ cùng với mã giới thiệu của một cá nhân hay tổ chức khác. ● *Điều hướng*: người dùng truy cập từ các đường dẫn điều hướng khác. ● *Khác*: người dùng truy cập thông qua các phương thức khác.
5	total_view	Tổng số lượt xem
6	total_product_view	Tổng số lượt xem sản phẩm
7	total_cart_view	Tổng số lượt xem giỏ hàng
8	total_category_view	Tổng số lượt xem danh mục sản phẩm
9	total_search	Tổng số lượt tìm kiếm sản phẩm
10	total_item_qty	Tổng số lượng mặt hàng trong giỏ hàng

11	total_cart_value	Tổng giá trị giỏ hàng
12	average_interval	Trung bình thời gian giữa các hành động

Đối với phân luồng gián tiếp, dữ liệu sẽ được thu thập và tổng hợp ngay sau khi dịch vụ dự đoán bỏ rơi giỏ hàng được tích hợp vào trang thương mại điện tử. Trong đó, các thành phần ngoại vị của dịch vụ dự đoán sẽ yêu cầu truy xuất lịch sử mua sắm, các nhận xét và đánh giá của từng tài khoản khách hàng trong trang thương mại điện tử. Dữ liệu từng khách hàng sẽ được tổng hợp và tính toán để đưa ra các tỷ lệ về tổng số lượng giao dịch thành công, thất bại cũng như thái độ tiêu dùng của khách mua hàng qua các nhận xét và đánh giá về sản phẩm. Số lượng bản ghi tổng hợp sẽ bằng với số lượng khách hàng hiện có và có thể tăng theo thời gian trong trường hợp trang thương mại điện tử có thêm khách hàng đăng ký tài khoản mới. Thống kê dữ liệu này sẽ không bao gồm các giao dịch của khách hàng vãng lai do các bản ghi sẽ thiếu đi định danh khách hàng. Bảng 3.2 mô tả chi tiết các chỉ mục dữ liệu có trong một bản ghi tổng hợp lịch sử mua sắm của khách hàng.

Bảng 3.2: Bảng chỉ mục dữ liệu của phân luồng gián tiếp

STT	Chỉ mục	Định nghĩa
1	customer_id	Định danh của khách hàng
2	total_order	Tổng số lượng đơn hàng
3	total_complete_order	Tổng số lượng đơn hàng hoàn thành
4	life_time_sale	Tổng giá trị đơn hàng trên hệ thống
5	total_cart	Tổng số lượng giỏ hàng
6	total_complete_cart	Tổng số lượng giỏ hàng thành công
7	total_abandon_cart	Tổng số lượng giỏ hàng bị bỏ rơi
8	average_cart_total	Trung bình giá trị giỏ hàng
9	total_rating	Tổng số lượng đánh giá sản phẩm

10	total_review	Tổng số lượng nhận xét sản phẩm
11	average_rating_score	Trung bình điểm đánh giá sản phẩm
12	average_review_length	Trung bình độ dài nhận xét sản phẩm
13	total_wishlist	Tổng số lượng sản phẩm mong muốn
14	total_login	Tổng số lần đăng nhập
15	total_product_view	Tổng số lần xem sản phẩm
16	total_product_search	Tổng số lần tìm kiếm sản phẩm
17	cart_abandon_rate	Tỷ lệ bỏ rơi giỏ hàng
18	average_order_total	Trung bình giá trị đơn hàng

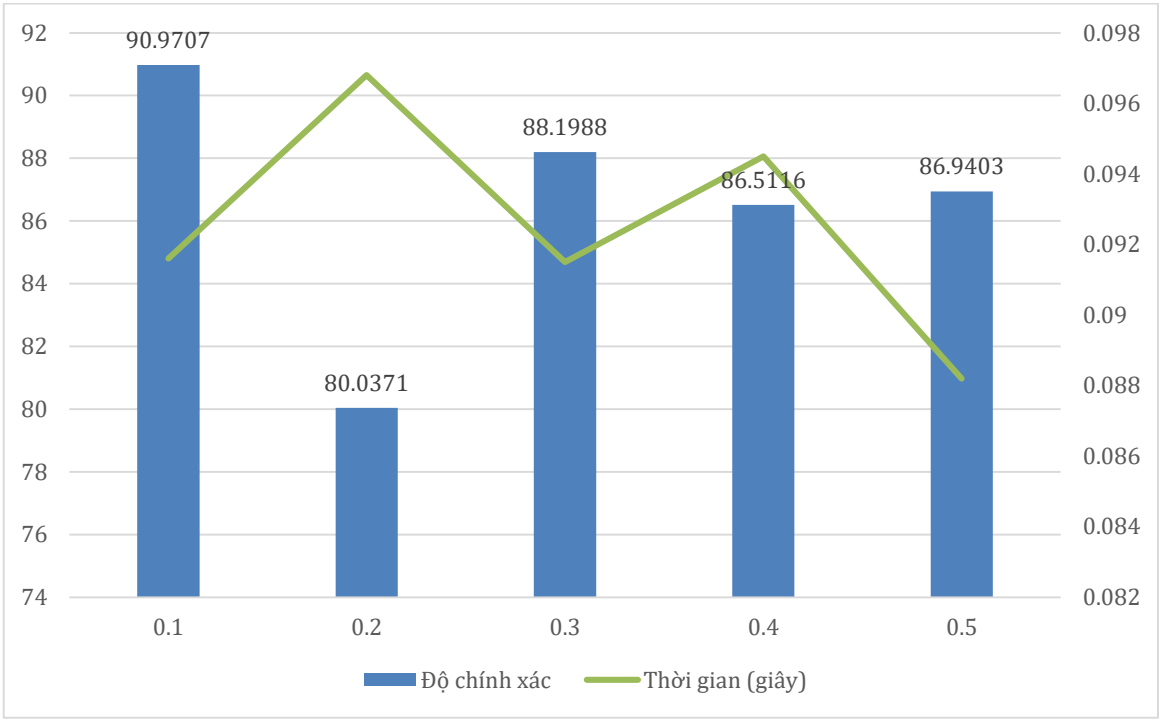
3.3. Thử nghiệm và đánh giá

3.3.1. Thống kê và phân tích khuôn mẫu dữ liệu

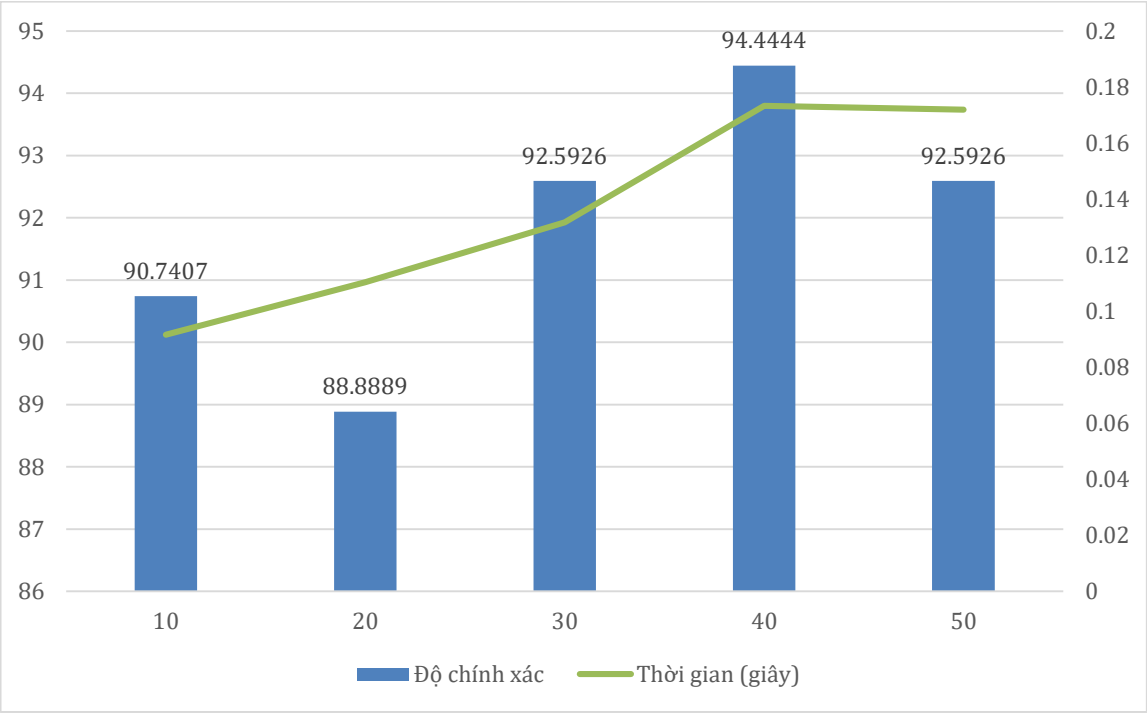
Dữ liệu hoạt động và các hành vi mua sắm của người tiêu dùng trên trang thương mại điện tử sau quá trình thu thập và xử lý sẽ được sử dụng làm dữ liệu huấn luyện cho thuật toán rừng ngẫu nhiên để phân loại khuôn mẫu hành vi. Quá trình huấn luyện đòi hỏi dữ liệu được phân chia thành các nhóm dữ liệu khác nhau nhằm phục vụ mục đích huấn luyện, thử nghiệm và so sánh độ chính xác của dự đoán phân loại. Do vậy, ba đối số là *kích thước thử nghiệm (test size), trạng thái ngẫu nhiên (random state) và số lượng ước tính (number of estimators)* được cân nhắc để thuật toán không chỉ đưa ra được kết quả với độ chính xác cao mà thời gian thực hiện phải tối ưu nhất. Trong đó, các đối số có vai trò như sau:

- *Kích thước thử nghiệm*: tỉ lệ phân chia giữa liệu giữa tập dữ liệu huấn luyện và tập dữ liệu thử nghiệm.
- *Trạng thái ngẫu nhiên*: sự nhất quán trong mỗi lần xáo trộn dữ liệu để đạo tạo.
- *Số lượng ước tính*: số lượng các cây quyết định được khởi tạo trước khi đưa ra kết quả phân loại cuối cùng.

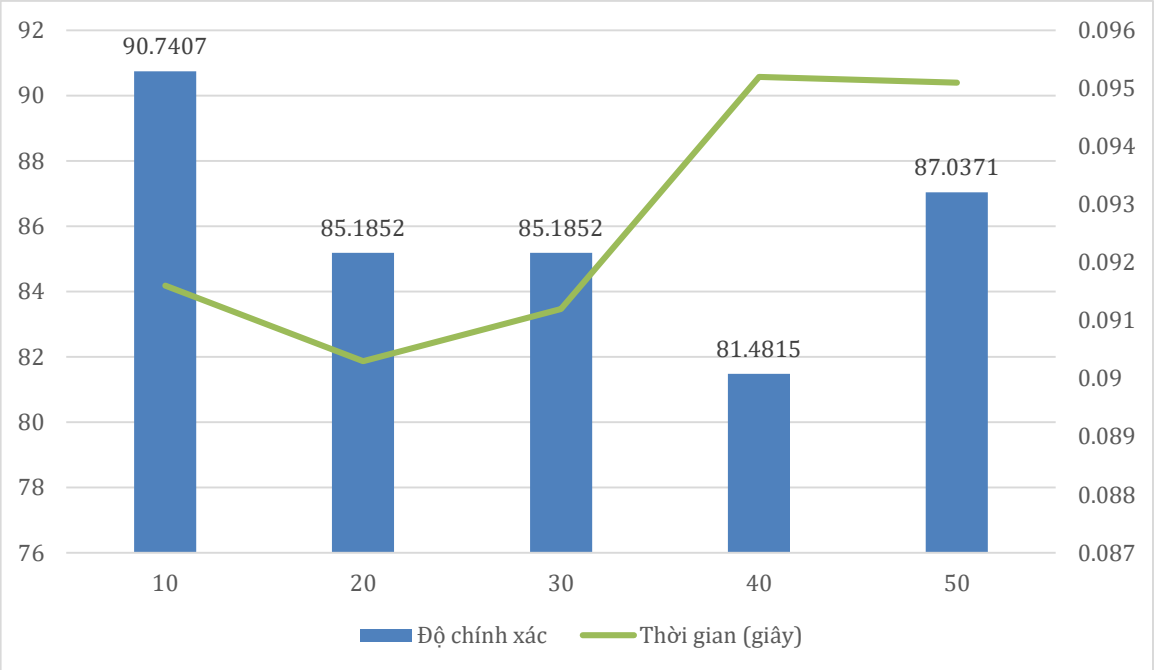
Tập dữ liệu phân tích bao gồm 1.173 bản ghi là nhật ký các phiên mua sắm của người dùng, trong đó, với sự thay đổi của các đối số, tổng thời gian thực hiện phân tích và dự đoán giao động từ 0,08 đến 0,2 giây cùng với mức giao động độ chính xác từ 80% đến 95%. Các đối số ban đầu được khởi tạo với các giá trị lần lượt là: kích thước thử nghiệm – 0.1; trạng thái ngẫu nhiên - 10 và số lượng ước tính - 10. Qua mỗi lần phân tích và dự đoán, giá trị của các đối số sẽ được tăng thêm 10 đơn vị cho đến mức tối đa là: kích thước thử nghiệm – 0.5; trạng thái ngẫu nhiên - 50 và số lượng ước tính – 50. Các đối số không có sự thay đổi về giá trị sẽ có giá trị cơ sở như lúc khởi tạo ban đầu. Sau đó, kết quả phân tích dữ liệu, độ chính xác của dự đoán và thời gian thực hiện được so sánh tương quan và biểu thị trong hình 3.4, hình 3.5 và hình 3.6.



Hình 3.4: Kết quả phân tích với biến số thay đổi là kích thước thử nghiệm



Hình 3.5: Kết quả phân tích với biến số thay đổi là số lượng ước tính

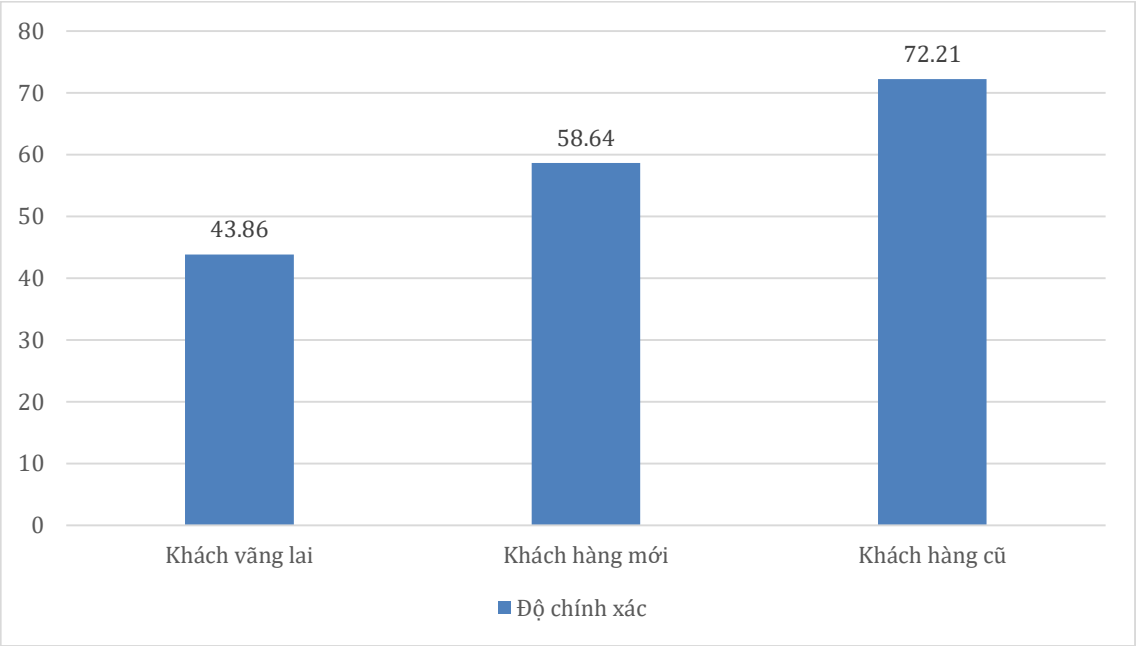


Hình 3.6: Kết quả phân tích với biến số thay đổi là trạng thái ngẫu nhiên

Qua kết quả độ chính xác tương quan với thời gian thực hiện phân tích và dự đoán, có thể thấy khi tăng giá trị của số lượng ước tính, thời gian thực hiện phân tích tăng lên một cách rõ rệt. Kết quả trên có thể dự đoán được vì việc tăng giá trị của số lượng ước tính sẽ gia tăng số lượng cây quyết định trong quá trình huấn luyện, từ đó việc thực hiện quyết định và tổng hợp kết quả của rừng ngẫu nhiên sẽ lâu hơn và tốn nhiều tài nguyên hơn. Ở chiều ngược lại, việc thay đổi kích cỡ thử nghiệm và trạng thái ngẫu nhiên không ảnh hưởng nhiều đến thời gian thực hiện tác vụ phân loại khi mà sự thay đổi giá trị chỉ làm thời gian thực hiện giao động trong 1/100 giây. Từ bảng so sánh kết quả chính xác và thời gian thực hiện tác vụ, có thể kết luận rằng với các đối số có giá trị lần lượt là: kích thước thử nghiệm – 0.1; trạng thái ngẫu nhiên – 10; số lượng ước tính – 10 sẽ cho kết quả tối ưu nhất với độ chính xác khoảng 90% và thời giạn thực nhiện trong khoảng 0,09 giây. Do vậy, ba giá trị trên sẽ được sử dụng trong quá trình ứng dụng thực tế việc phân loại người tiêu dùng bỏ rơi giỏ hàng.

3.3.2. Thử nghiệm thực tế

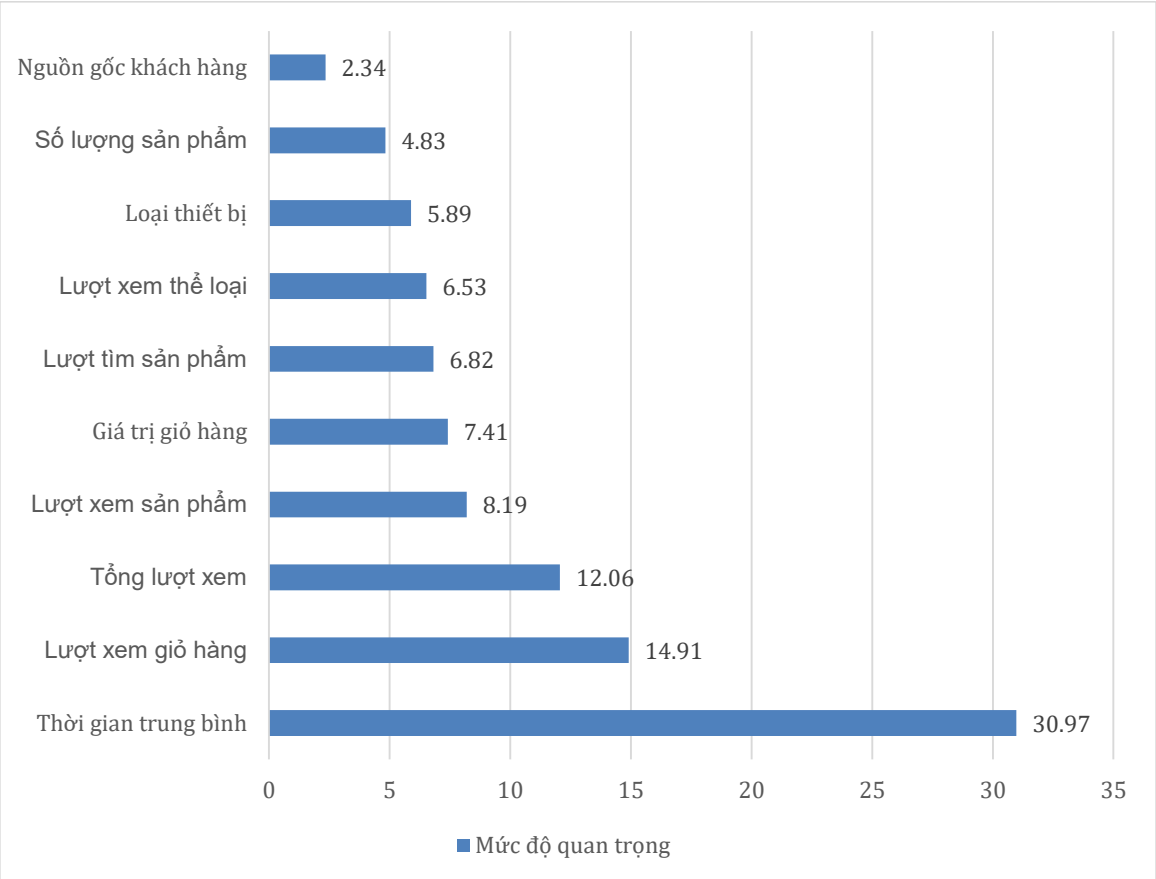
Quá trình ứng dụng và thử nghiệm trên người dùng thực tế dịch vụ phân loại và dự đoán hành vi bỏi rơi giỏ hàng được tiến hành trong thời gian khoảng 3 tuần. Với 230 phiên mua sắm trực tuyến của 3 nhóm khác hàng chính là khách hàng vãng lai, khách hàng mới và khác hàng cũ. Khác hàng vãng lai là tệp khách hàng lần đầu truy cập và sử dụng dịch vụ mua sắm tại trang thương mại điện tử. Tệp khách hàng này chủ yếu đến từ các chương trình khoảng cáo và tiếp thị, một số ít đến từ gợi ý của các công cụ tìm kiếm do có từ khóa liên quan. Khác hàng mới là những khách hàng đã có tài khoản và đã biết đến trang thương mại điện tử nhưng chưa từng có một đơn hàng giao dịch giỏ hàng thành công. Cuối cùng là khách hàng cũ hay tệp khách hàng đã có ít nhất một đơn hàng giao dịch thành công và đã có tài khoản tại trang thương mại điện tử.



Hình 3.7: Kết quả dự đoán trong ứng dụng dự đoán thực tế

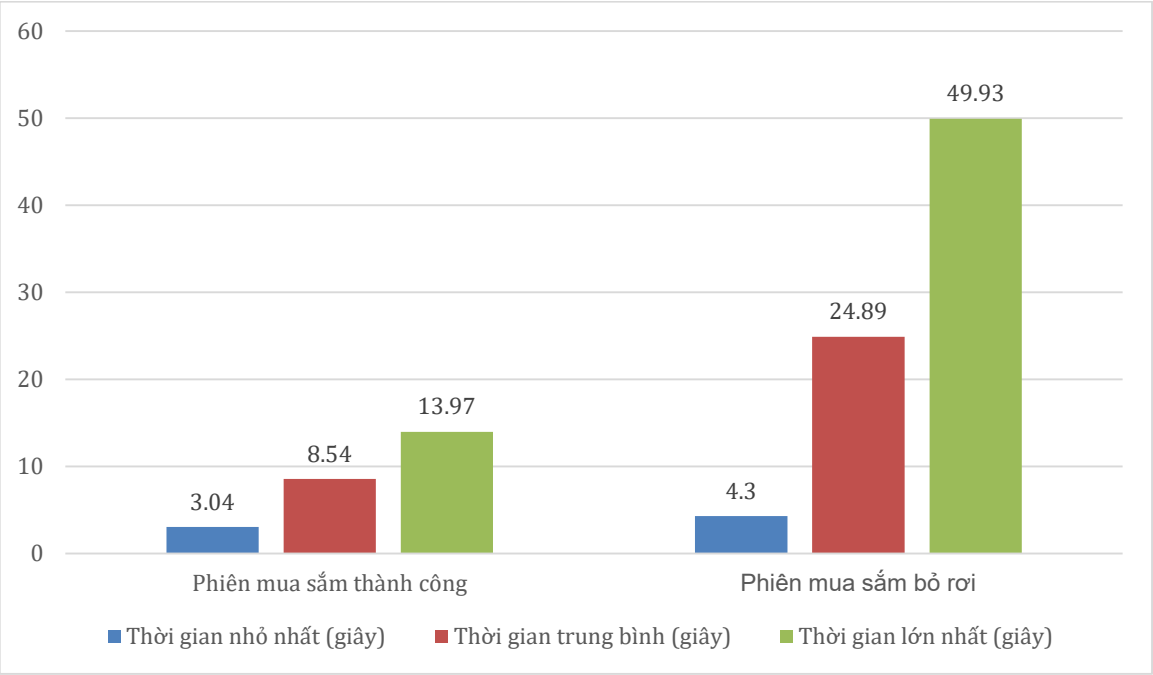
Từ biểu đồ kết quả dự đoán, có thể thấy độ chính xác với tệp khách hàng cũ cao hơn đáng kể so dữ liệu của khác hàng vãng lai và khách hàng mới. Nhóm khách hàng cũ không chỉ có nhiều thông tin cho quá trình phân loại mà họ đã có một niềm tin nhất định đối với trang thương mại điện tử nên việc hoàn tất giao dịch mua sắm hoàn toàn có thể dự đoán được.

Ngoài độ chính xác trong các tệp khác hàng, thuật toán rừng ngẫu nhiên còn đưa ra thống kê về mức độ quan trọng của các thuộc tính dữ liệu. Qua thống kê này, người dùng có thể ứng dụng để tối ưu hóa thuật toán trong quá trình phân loại cũng như tối ưu các chức năng của trang thương mại điện tử để giảm thiểu tỷ lệ bỏ rơi giỏ hàng, gia tăng số lượng đơn hàng được chuyển đổi thành công.



Hình 3.8: Mức độ quan trọng của các thuộc tính trong phiên mua sắm

Qua thống kê trong hình 3.8, thời gian trung bình giữa các hoạt động của người tiêu dùng là thuộc tính quan trọng nhất quyết định hành vi bỏ rơi giỏ hàng trong phiên mua sắm với mức độ quan trọng hơn 30%. Thời gian trung bình càng lâu thì tỷ lệ người tiêu dùng bỏi rơi giỏ hàng càng cao và ngược lại, thời gian càng ngắn thì tỷ lệ bỏ rơi giỏ hàng càng thấp. Bên cạnh đó, lượt xem giỏ hàng và tổng lượt xem trên toàn trang thương mại điện tử cũng là những thuộc tính có mức độ quan trọng tương đối rõ rệt, lần lượt đạt ở mức 15% và 13%. Số lượt xem giỏ hàng của người tiêu dùng tăng cao một phần thể hiện mức độ cân nhắc của người mua sắm trước khi thực hiện giao dịch.



Hình 3.9: So sánh thời gian trung bình của phiên mua sắm thành công và bỏ rơi

Biểu đồ so sánh hình 3.9 cho thấy sự khác biệt rõ rệt trong thời gian trung bình giữa các hành động của người dùng hoàn thành giỏ hàng và người dùng bỏ rơi giỏ hàng. Có thể thấy, thời gian trung bình của phiên mua sắm thành công dao động từ 3 giây đến khoảng 15 giây, với mức trung bình phổ biến ở 8,54 giây. Ở chiều ngược lại, phiên mua sắm có giỏ hàng bị bỏ rơi có mức dao động rất lớn từ 4 giây đến khoảng 50 giây, với mức trung bình là 25 giây. Như vậy có thể kết luận rằng, khi thời gian trung bình giữa các hành động của người dùng vượt qua khoảng 15 đến 20 giây, phiên mua sắm sẽ có tỷ lệ bỏ rơi giỏ hàng rất lớn do giá trị đã vượt qua ngưỡng tối đa của phiên mua sắm thành công và gần chạm mốc trung bình của phiên mua sắm có hiện tượng bỏ rơi giỏ hàng.

3.4. Kết luận

Trong chương ba, luận văn tập trung vào trình hai bài toán phân tích và thống kê dữ liệu tương ứng với hai phân luồng trực tiếp và dán tiếp trong dự đoán hành vi bỏ rơi giỏ hàng của người tiêu dùng. Dựa trên phát biểu của hai bài toán, chương ba cũng đã làm sáng tỏ trình tự phân tích và tổng hợp kết quả của hai phân luồng để đưa ra kết quả dự đoán cuối cùng. Đồng thời, luận văn đã trình bày cấu trúc hệ thống trang

thương mại điện tử và dịch vụ phân tích cùng mô tả chi tiết, kích cỡ và cách thu thập dữ liệu thử nghiệm. Trong thử nghiệm và đánh giá thực tế, luận văn đã tìm ra được giá trị tối ưu cho các đối số của thuật toán rừng ngẫu nhiên, từ đó đưa ra kết quả thử nghiệm thực tế với tập khách hàng cũ (khách hàng đã có tài khoản và đã từng có giao dịch thành công) có dự đoán bỏ rơi giỏ hàng chính xác nhất là 72,21%. Ngoài ra, thuật toán rừng ngẫu nhiên trong chương ba cũng đưa ra được tiêu chí thời gian trung bình giữa các hành động đóng vai trò quan trọng nhất trong quyết định bỏ rơi giỏ hàng của tiêu dùng. Trong đó, thời gian trung bình của phiên mua sắm thành công là 8,5 giây trong khi tại phiên mua sắm có hiện tượng bỏ rơi giỏ hàng, thời gian trung bình giữa các hành động cao hơn gần gấp 3 lần (24,9 giây).

KẾT LUẬN

Sau quá trình tìm hiểu, nghiên cứu và ứng dụng, luận văn "Nghiên cứu phân tích hành vi người dùng bỏ giỏ hàng trên các trang thương mại điện tử" đã cơ bản đáp ứng được các nội dung trình bày trong đề cương. Cụ thể, luận văn đã đạt được một số kết quả chính sau:

- Giới thiệu tổng quan về thương mại điện tử trên thế giới và ở Việt Nam trong những năm gần đây và làm rõ tiềm năng khai thác dữ liệu tiêu dùng và hoạt động mua sắm của khách hàng.
- Trình bày vấn đề bỏ rơi giỏ hàng điện tử trong mua sắm trực tuyến và những yếu tố chính dẫn đến hiện tượng này.
- Trình bày thuật toán học máy có giám sát cây quyết định và rừng ngẫu nhiên; đồng thời, so sánh hiệu năng rừng ngẫu nhiên với các thuật toán phân loại khác để làm sáng tỏ mức độ phù hợp trong tác vụ phân loại dự đoán bỏ rơi giỏ hàng.
- Ứng dụng thuật toán rừng ngẫu nhiên, xây dựng dịch vụ phân tích và tích hợp cùng hệ thống của trang thương mại điện tử để thử nghiệm phân tích và dự đoán hành vi bỏ rơi giỏ hàng.
- Trên kết quả thu được từ thử nghiệm thực tế, luận văn đưa ra thời gian trung bình giữa các hành động là yếu tố chính trong quyết định bỏ rơi giỏ hàng của người dùng trong mua sắm trực tuyến.

Trong tương lai, đề tài nghiên cứu và ứng dụng của luận văn có thể được mở rộng ở nhiều phương diện về thuật toán sử dụng cũng như nền tảng thương mại điện tử được áp dụng. Trong đó, hướng nghiên cứu phát triển tiếp theo có thể được cụ thể hóa như sau:

- Nghiên cứu thêm các thuật toán học máy mới để đa dạng hóa thuật toán sử dụng cũng như đối chiếu kết quả trong tác vụ dự đoán hành vi bỏ rơi giỏ hàng.
- Ứng dụng dịch vụ phân tích và dự đoán bỏ rơi giỏ hàng trên các nền tảng thương mại điện tử khác ví dụ: BigCommerce, Shopify, WooCommerce, v.v

DANH MỤC TÀI LIỆU THAM KHẢO

[1] Al Imran, Md Abdullah (2014), "A Study On Amazon: Information Systems, Business Strategies And E-Crm" *University of Liberal Arts Bangladesh.*

[2] Baymard Institute Research Team (2017), *41 Cart Abandonment Rate Statistics*, Baymard Institute. Available: https://baymard.com/lists/cart-abandonment-rate

[3] Breiman L (2001), *Random Forests. Machine Learning Vol. 45 No. 1*, pp. 5–32.

[4] Bucklin RE and Sismeiro C (2003), "A Model of Web Site Browsing Behavior Estimated on Clickstream Data" *Journal of Marketing Research* Vol. 40 No. 03, pp. 249–267.

[5] Chipman HA, George EI and McCulloch RE (1998), "Bayesian CART Model Research" *Journal of the American Statistical Association*, pp. 935–948.

[6] Cho J (2004), "Likelihood to abort an online transaction: influences from cognitive evaluations, attitudes, and behavioral variables" *Information & Management*, pp. 827-838.

[7] Close Scheinbaum, Angeline & Kukar-Kinney, Monika & Benusa, Kyle (2012), "Towards a Theory of Consumer Electronic Shopping Cart Behavior" *Motivations of E-Cart Use and Abandonment*, pp. 156-230.

[8] Dowling, G.R (1986), "Perceived risk: the concept and its measurement" *Psychology and Marketing, Vol. 3 No. 3*, pp. 193-210.

[9] Friedman J (2001), "Greedy Function Approximation: A Gradient Boosting Machine" *The Annals of Statistics*, pp. 1189–1232.

[10] Hasan, Layla & Morris, Anne & Probets, Steve (2009), "Using Google Analytics to Evaluate the Usability of E-Commerce Sites".

[11] Köhn, Dennis & Lessmann, Stefan & Schaal, Markus (2020), "Predicting Online Shopping Behaviour from Clickstream Data using Deep Learning". *Berlin Expert Systems with Applications Journal*.

[12] Leiner, Barry & Cerf, Vinton & Clark, David & Kahn, Robert & Kleinrock, et al (2009), "A Brief History of the Internet. Computer Communication Review", *ACM SIGCOMM Computer Communication Review*, pp. 22-31

[13] Liaw A and Wiener M (2002) Classification and Regression by randomForest. R News 2(3): 18–22.

[14] Montgomery AL, Li S, Srinivasan K, et al (2004), "Modeling Online Browsing and Path Analysis Using Clickstream Data". *Marketing Science Vol. 23 No. 04*, pp. 579–595.

[15] Moore S and Mathews S (2006), "An exploration of online shopping cart abandonment syndrome – a matter of risk and reputation" *Journal of Website Promotion*, pp. 71-88.

[16] Moro S, Cortez P and Rita P (2014), "A data-driven approach to predict the success of bank telemarketing" *Decision Support Systems & Technology Security Magazine*, pp. 22–31.

[17] Nayak, Richi (2003). *Data Mining for Web-Enabled Electronic Business Applications*, Queensland University of Technology, Brisbane, Australia.

[18] Nielson, J (1996), "Response times: the three important limits" *Neilson Norman Group Science Journal 2010*, pp. 34-55.

[19] Opitz D and Maclin R (1999), "Popular Ensemble Methods: An Empirical Study" *Journal of Artificial Intelligence Research Vol. 11*, pp. 169–198.

[20] Ouellet M (2010), "Recovering lost sales through an automated shopping cart abandonment strategy" *Listrak Information & Technology Journal*, pp. 18-24.

[21] Rajamma, R., Paswan, A., & Hossain, M (2009), "Why do shoppers abandon shopping carts? Perceived waiting time, risk, and transaction inconvenience" *Journal of Product & Brand Management*, pp. 188–197.

[22] Rajini, G., & Krithika, M (2017), "Risk Factors Discriminating Online Metropolitan Women Shoppers: A Behavioural Analysis" *International Journal of Cyber Behaviour, Psychology and Learning (IJCBPL)*, pp. 52-64.

[23] Ranganathan, C. and Ganapathy, S (2002), "Key dimensions of business-to-business websites*" Information and Management,Vol. 39*, pp. 457-65.

[24] Rausch, Theresa & Derra, Nicholas & Wolf, Lukas (2020), "Predicting online shopping cart abandonment with machine learning approaches*" International Journal of Market Research*.

[25] Rastegari, Hamid & Md Noor, Mohd. (2008), "Data mining and e-commerce : methods, applications, and challenges" *Islamic Azad University Annual Journal, pp. 15-24*.

[26] Shahriari, Shahrzad & Shahriari, Mohammadreza & gheiji, Saeid (2015), "E-Commerce And It Impacts On Global Trend And Market" *International Journal of Research*, pp. 49-55.

[27] Statista Research Team (2020), Available: https://www.statista.com/statistics/272391/us-retail-e-commerce-sales-forecast

[28] Szymanski, D.M. and Henard, D.H (2001), "Customer satisfaction: a meta-analysis of the empirical evidence" *Journal of the Academy of Marketing Science Vol. 29 No. 1*, pp. 16-35.

[29] Vafeiadis T, Diamantaras KI, Sarigiannidis G, et al (2015), "A comparison of machine learning techniques for customer churn prediction" *Simulation Modelling Practice and Theory Vol. 55*, pp. 1–9.

[30] Williams N, Zander S and Armitage G (2006), "A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification" *ACM SIGCOMM Computer Communication Review Vol. 36 No. 5*, pp. 5–16.

[31] Wolfinbarger, M. and Gilly, M.C (2001), "Shopping online for freedom, control and fun" *California Management Review, Vol. 43 No. 2*, pp. 34-55.

[32] Wu, Lihua & Deng, Tian (2016), "Web Data Mining and Its Implication in E-commerce" *International Conference on Education, Management, Computer and Society.*

[33] Zheng B and Liu B (2018), "A scalable purchase intention prediction system using extreme gradient boosting machines with browsing content entropy" *International Conference on Consumer Electronics (ICCE) 2018.*