Hệ thống dự báo khách hàng rời bỏ dịch vụ ngân hàng trên nền tảng học máy

The customer churn prediction system for banking services on machine learning platform

Nguyễn Quốc Hùng[1*], Lê Thành Trung[1], Nguyễn Thị Xuân Đào[1], Nguyễn Quang Trường[2]

[1]Đại học Kinh tế Thành phố Hồ Chí Minh, Thành phố Hồ Chí Minh, Việt Nam
[2]Ngân hàng Quốc tế VIB, Thành phố Hồ Chí Minh, Việt Nam
[*]Tác giả liên hệ, Email: hungngq@ueh.edu.vn

THÔNG TIN

DOI:10.46223/HCMCOUJS.econ.vi.21.1.3474.2026

Ngày nhận: 05/06/2024

Ngày nhận lại: 17/06/2025

Ngày duyệt đăng: 31/07/2025

Mã phân loại JEL:

C45; G21; M31

Từ khóa:

CRM; dữ liệu hành vi giao dịch; dự đoán rời bỏ; hành vi khách hàng; học máy; khách hàng rời bỏ; Random Forest; tài chính - ngân hàng

TÓM TẮT

Sự phát triển nhanh chóng của công nghệ số, đặc biệt là trí tuệ nhân tạo, đã làm thay đổi căn bản phương pháp phân tích và khai thác dữ liệu trong ngành ngân hàng. Tuy nhiên, các mô hình truyền thống vẫn chủ yếu dựa trên dữ liệu định danh tĩnh, chưa tận dụng hiệu quả dữ liệu hành vi giao dịch để dự báo rủi ro khách hàng rời bỏ dịch vụ. Nghiên cứu này đề xuất một mô hình học máy ứng dụng dữ liệu giao dịch tài khoản DEBIT nhằm dự đoán khả năng khách hàng ngừng sử dụng dịch vụ ngân hàng. Dữ liệu được thu thập từ hệ thống CASA của ba ngân hàng thương mại tại khu vực phía Nam, thông qua hợp tác chính thức phục vụ mục đích nghiên cứu học thuật và đảm bảo các nguyên tắc bảo mật thông tin. Trên cơ sở dữ liệu này, nhóm tác giả tiến hành huấn luyện và so sánh hiệu quả dự báo giữa các mô hình học máy phổ biến. Kết quả thực nghiệm cho thấy mô hình Random Forest đạt độ chính xác cao nhất (96%) và có tiềm năng ứng dụng vào hệ thống quản trị quan hệ khách hàng (CRM). Mô hình đề xuất giúp nhận diện sớm các khách hàng có nguy cơ rời bỏ dịch vụ, từ đó hỗ trợ ngân hàng xây dựng các chiến lược giữ chân phù hợp. Nghiên cứu góp phần bổ sung cơ sở thực nghiệm cho lĩnh vực phân tích hành vi khách hàng trên nền tảng học máy trong bối cảnh chuyển đổi số tại các tổ chức tài chính - ngân hàng.

ABSTRACT

The increasing integration of artificial intelligence and the Internet has revolutionized data analysis and processing, particularly in customer behavior prediction. However, in the banking sector, machine learning techniques have not been widely adopted for churn analysis, with existing approaches often relying on outdated, static demographic information. This study proposes a machine learning-based model for predicting customer churn using historical transactional data from DEBIT accounts. The dataset was collected through formal agreements with three commercial banks in southern Vietnam, utilizing the

CASA transaction information system. Data collection and usage complied with strict confidentiality protocols, ensuring legal validity and data reliability. After a comprehensive data preprocessing pipeline, several machine learning algorithms were evaluated, with Random Forest achieving the highest performance with 96% accuracy. The proposed model has been integrated into the banks' Customer Relationship Management (CRM) systems to assist in early identification of potential churners and to inform retention strategies. The findings contribute theoretically by demonstrating the importance of dynamic behavioral indicators over static demographics, and practically by offering a scalable solution to improve customer retention in the banking sector.

Keywords:

CRM; transaction behavioral data; churn prediction; customer behavior; machine learning; customer churn; Random Forest; banking and finance

1. Đặt vấn đề

Khách hàng luôn là nguồn lực quan trọng đối với tổ chức, trong bối cảnh cạnh tranh hiện nay, khách hàng có xu hướng dịch chuyển sang các đối thủ cạnh tranh trong môi trường cạnh tranh cao, do đó việc giữ chân khách hàng dường như là một yêu cầu cơ bản, thiết yếu và có tầm quan trọng đối với bất kỳ một tổ chức nào. Nhiều công ty gặp phải vấn đề nghiêm trọng về việc khách hàng bỏ đi, do sự cạnh tranh khốc liệt do thị trường bão hòa, điều kiện thị trường năng động và liên tục đưa ra các dịch vụ cạnh tranh mới. Các ngân hàng cũng không nằm ngoài quy luật này, với số lượng khách hàng của các ngân hàng và công ty tài chính ngày càng tăng, các ngân hàng lớn thường có hàng chục triệu khách hàng trong danh mục kinh doanh của họ và điều này khiến các ngân hàng ý thức được chất lượng dịch vụ mà họ cung cấp. Hiện tượng rời đi của khách hàng, được định nghĩa là "rời bỏ dịch vụ" hay có sự chuyển đổi từ nhà cung cấp dịch vụ này sang nhà cung cấp dịch vụ khác xảy ra do các lý do như sự sẵn có của công nghệ mới nhất, nhân viên ngân hàng thân thiện với khách hàng, lãi suất thấp, vị trí địa lý gần, dịch vụ đa dạng được cung cấp.

Nhu cầu sử dụng sản phẩm và dịch vụ của khách hàng là yếu tố quan trọng trong sự hình thành và phát triển của bất kỳ thị trường hay doanh nghiệp nào. Việc thu hút các công ty nhằm đáp ứng nhu cầu đó đã thúc đẩy quá trình phát triển sản phẩm và dịch vụ mới. Một số nghiên cứu đã chỉ ra rằng chi phí để có được một khách hàng mới thường cao gấp 05 đến 06 lần so với chi phí giữ chân một khách hàng hiện có (Colgate & Danaher, 2000). Do tầm quan trọng của khách hàng và chi phí thu hút khách hàng mới cao hơn so với việc duy trì khách hàng hiện tại, các ngân hàng và các ngành phụ thuộc vào khách hàng khác phải có khả năng tự động hóa quá trình dự đoán hành vi của khách hàng bằng cách sử dụng dữ liệu của khách hàng trong cơ sở dữ liệu của họ. Trong khi đó, khách hàng rời đi là một trong những vấn đề quan trọng nhất đối với ngân hàng, làm giảm các khoản thu nhập khác nhau và thu nhập từ phí (tiền gửi không kỳ hạn, phí chuyển tiền, …). Và quan trọng hơn, tiền gửi không kỳ hạn của khách hàng là nguồn thu nhập chính của một ngân hàng. Khi không duy trì được hai nguồn thu nhập này, cùng với khả năng gia tăng rủi ro về uy tín, có thể dẫn ngân hàng đến bờ vực phá sản. Quản lý rời bỏ đã trở thành một phần của quản lý quan hệ khách hàng vì thách thức nghiêm trọng của việc khách hàng rời bỏ trong lĩnh vực ngân hàng. Quản lý rời bỏ nhấn mạnh sự cần thiết của các ngân hàng để thực hiện các bước để ngăn chặn hoặc giảm thiểu sự rời bỏ của khách hàng thông qua một số chương trình giữ chân khách hàng. Điều này cũng giúp thiết lập mối quan hệ lâu dài với khách hàng và tối đa hóa giá trị cơ sở khách hàng của họ. Khách hàng rời bỏ đặt ra mối quan tâm

nghiêm trọng đối với các ngân hàng vì nó gây ra tổn thất doanh thu cho ngành. Đây cũng là những lý do mà các ngân hàng rất muốn xác định những khách hàng có khả năng hủy đăng ký dịch vụ của họ cao nhất. Dự đoán rời bỏ cho phép sử dụng hồ sơ giao dịch của khách hàng để xác định khả năng khách hàng từ bỏ dịch vụ trước khi khách hàng thật sự rời bỏ. Tầm quan trọng của việc hiểu được sự rời bỏ của khách hàng đã được nhấn mạnh trong một số nghiên cứu gần đây như tỷ lệ duy trì tăng 1% cho thấy giá trị công ty tăng trung bình 5% có trong nghiên cứu của Ashraf (2024). Việc giảm tỷ lệ rời bỏ 5% đã được chứng minh là giúp tăng gấp đôi lợi nhuận trong một số ngành trong nghiên cứu của Liu và cộng sự (2024); Li và Yan (2025).

Hiện nay, việc phân tích và dự đoán tỷ lệ khách hàng rời bỏ ngân hàng nhận được sự quan tâm đáng kể từ giới nghiên cứu, với nhiều mô hình học máy được ứng dụng như Random Forest, Logistic Regression, Decision Tree, XGBoost, … Tuy nhiên, mỗi mô hình đều tồn tại những hạn chế nhất định. Chẳng hạn, theo nghiên cứu của Genuer và cộng sự (2020), mô hình Random Forest có thể dễ bị ngưng hoạt động khi vượt quá giới hạn xử lý hoặc tài nguyên hệ thống. Trong khi đó, Decision Tree lại phụ thuộc nhiều vào đặc điểm của tập dữ liệu; chỉ một thay đổi nhỏ trong dữ liệu đầu vào cũng có thể làm thay đổi hoàn toàn cấu trúc cây quyết định (Rokach & Maimon, 2005). Mô hình Logistic Regression, theo Kleinbaum và cộng sự (2002), yêu cầu các điểm dữ liệu đầu vào phải độc lập với nhau, điều này trở nên khó đảm bảo trong các hệ thống dữ liệu thực tế của ngân hàng, nơi hành vi khách hàng thường có tính liên kết theo thời gian hoặc theo nhóm. Do đó, các tổ chức tài chính và ngân hàng cần những mô hình phân tích có khả năng xử lý hiệu quả trên tập dữ liệu lớn và phi tuyến tính, đồng thời thích ứng với tính phụ thuộc tiềm ẩn giữa các quan sát.

Trong bài báo này tập trung vào xây dựng mô hình học máy dựa trên hành vi của quá khứ và hiện tại nhằm xác định khách hàng rời bỏ trong tương lai để cho ngân hàng. Kết quả mô hình đề xuất có thể đưa ra các chiến lược và hành động phù hợp để giảm thiểu lượng khách hàng rời bỏ và đưa ra những ngưỡng tiêu chí để phân loại khách hàng nào cần giữ lại. Đóng góp chính của bài báo là đã tiến hành đánh giá so sánh các phương pháp học máy như Logistic Regression, Random Forest, Decision Tree, XGBoost, Naïve Bayes dựa trên nguồn dữ liệu thực tế tại một số ngân hàng phía Nam để chọn ra phương pháp phù hợp và tích hợp vào hệ thống quản trị quan hệ khách hàng CRM nhằm dự báo chính xác số lượng khách hàng rời bỏ dịch vụ trong tương lai.

Bài báo được bố cục thành các phần như sau: phần I giới thiệu về tính cấp thiết khi áp dụng mô hình học máy vào bài toán dự báo khách hàng rời bỏ dịch vụ, phần II là tổng quan nghiên cứu trình bày các nghiên cứu liên quan và cơ sở lý thuyết khi về xác định khách hàng rời bỏ và một số phương pháp máy học. Phần III là mô hình đề xuất và phương pháp thực hiện. Phần IV là một số kết quả đánh giá và bình luận. Cuối cùng là phần kết luận và các tài liệu tham khảo liên quan.

2. Tổng quan nghiên cứu và cơ sở lý thuyết

2.1. Các nghiên cứu liên quan

Phần tổng quan nghiên cứu trình bày có hệ thống các công trình tiêu biểu liên quan đến dự báo hành vi rời bỏ dịch vụ trong các lĩnh vực ngân hàng, truyền thông và tài chính. Trong đó, có thể kể đến các nghiên cứu của Vo và cộng sự (2018), Brownlow và cộng sự (2018), Devriendt và cộng sự (2021), và đặc biệt là nghiên cứu của Huang và cộng sự (2012) với việc xây dựng hệ thống các độ đo định lượng hành vi rời bỏ của khách hàng. Các nghiên cứu này là cơ sở quan trọng giúp phân loại các yếu tố đầu vào thành ba nhóm chính: (1) thông tin nhân khẩu học, (2) hành vi giao dịch, và (3) mức độ tương tác giữa khách hàng với tổ chức tài chính. Trên nền tảng

đó, khung lý thuyết và mô hình dự báo ứng dụng học máy được hình thành nhằm giải quyết bài toán nhận diện và dự đoán khả năng rời bỏ của khách hàng trong bối cảnh thực tiễn.

Nhằm nâng cao hiệu quả dự báo hành vi rời bỏ khách hàng, Huang và cộng sự (2012) đã phát triển hệ thống chỉ số định lượng dựa trên dữ liệu lớn trong ngành viễn thông, gồm các tỷ lệ: rời bỏ tiêu chuẩn, duy trì ròng và rời bỏ theo doanh thu định kỳ (MRR churn). Mô hình đề xuất cho phép xác định khách hàng rời bỏ dựa trên chuỗi hành vi không tương tác trong khoảng thời gian nhất định (30 - 90 ngày), với độ chính xác đạt 87% tại ngưỡng 60 ngày. Kết quả thử nghiệm cho thấy việc kết hợp cả ba chỉ số đem lại hiệu quả dự báo cao hơn 15 - 18% so với sử dụng riêng lẻ từng chỉ số. Trong khi đó, Vo và cộng sự (2018) áp dụng kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) để khai thác nhật ký cuộc gọi - một dạng dữ liệu phi cấu trúc - nhằm nâng cao độ chính xác của mô hình dự báo rời bỏ trong lĩnh vực tài chính. Liao và Chueh (2011) sử dụng kỹ thuật mờ (fuzzy) để phân tích dữ liệu chiến dịch tiếp thị, xây dựng mô hình giúp doanh nghiệp tối ưu hóa chiến lược tiếp cận từng nhóm khách hàng.

Brownlow và cộng sự (2018) đề xuất khung lấy mẫu đồng bộ kết hợp trọng số theo số dư tài khoản nhằm khắc phục mất cân bằng dữ liệu tài chính, và ứng dụng thành công trong thực tế để phát hiện khách hàng rời bỏ có giá trị cao. Mô hình này giúp tích hợp dữ liệu không đồng nhất và thu hẹp nhóm mục tiêu trong các chiến dịch marketing. Silveira và cộng sự (2021) xây dựng mô hình học máy tích hợp trực tiếp vào hệ thống CRM để giám sát và quản trị rủi ro mất khách hàng trong ngành ngân hàng Brazil. Nghiên cứu nhấn mạnh vai trò của phân tích dự báo trong việc duy trì khách hàng, giảm thiểu chi phí và rủi ro danh tiếng. Devriendt và cộng sự (2021) đề xuất thước đo "mức nâng lợi nhuận tối đa" (maximum profit uplift) nhằm tối ưu hóa chiến dịch giữ chân khách hàng, tập trung vào nhóm "có thể thuyết phục được" (persuadables). Các thuật toán tăng cường như XGBoost được chứng minh có hiệu quả cao trong việc phát hiện nhóm này.

Lee và cộng sự (2018) phát triển quy trình dự báo hành vi rời bỏ trong ngành game online, tính đến lợi nhuận kỳ vọng trên mỗi người dùng. Kết quả cho thấy việc giới hạn mô hình cho nhóm khách hàng trung thành có thể tăng lợi nhuận 10 - 30% so với áp dụng đại trà. Li và cộng sự (2021) áp dụng dữ liệu lớn trong ngành truyền hình cáp để dự báo hành vi rời bỏ dựa trên cường độ xem, thói quen chi tiêu và sở thích khách hàng. Việc kết hợp mô hình dự báo và chiến lược giữ chân góp phần tối ưu hóa hiệu quả tiếp thị. Rosa (2019) nhấn mạnh sự cần thiết của cách tiếp cận chủ động trong việc dự đoán rời bỏ, thông qua tích hợp công cụ phân tích dữ liệu và hệ thống Business Intelligence (BI) vào quy trình quyết định. Cách tiếp cận này không chỉ giảm tỷ lệ rời bỏ mà còn nâng cao hiệu quả quản trị mối quan hệ khách hàng. Vélez và cộng sự (2020) đề xuất phương pháp lựa chọn biến dựa trên kỹ thuật Weight of Evidence (WOE) nhằm tăng cường khả năng diễn giải của mô hình hồi quy logistic. Cách tiếp cận này đặc biệt phù hợp trong dự báo rời bỏ, khi tính minh bạch và khả năng giải thích mô hình đóng vai trò thiết yếu.

Qua tổng quan các nghiên cứu, có thể nhận diện rõ quy trình xây dựng mô hình học máy dự đoán hành vi rời bỏ khách hàng trong các lĩnh vực như viễn thông, truyền hình và tài chính, với những điểm mạnh và hạn chế sau:

- **Ưu điểm:** Các nghiên cứu mô tả chi tiết quy trình xây dựng mô hình dự báo và ứng dụng các phương pháp xử lý dữ liệu hiện đại như kỹ thuật NLP để khai thác dữ liệu phi cấu trúc và chuẩn hóa biến đầu vào bằng kỹ thuật Weight of Evidence (WOE). Cách tiếp cận này góp phần tăng cường khả năng giải thích của mô hình, đồng thời cải thiện độ chính xác trong dự báo hành vi rời bỏ.

- **Hạn chế:** Nhiều nghiên cứu vẫn chưa khai thác đầy đủ các thuật toán học máy tiên tiến như XGBoost hay Random Forest, dẫn đến thiếu cơ sở so sánh toàn diện. Ngoài ra, phần lớn mô hình mới dừng lại ở việc xác định khách hàng đã rời bỏ mà chưa dự báo được thời điểm rời bỏ trong tương lai - yếu tố quan trọng trong chiến lược giữ chân khách hàng. Bên cạnh đó, chưa có nghiên cứu nào thiết lập ngưỡng định lượng rõ ràng cho giá trị vòng đời khách hàng (Customer Lifetime Value - CLV) để hỗ trợ quyết định duy trì hay loại bỏ, gây khó khăn trong việc tối ưu hóa chi phí.

Tổng hợp từ các nghiên cứu trên, có thể thấy việc dự báo khách hàng rời bỏ dịch vụ ngân hàng phụ thuộc vào ba nhóm yếu tố chính: *(i) Thông tin nhân khẩu học như tuổi, giới tính, thời gian giao dịch với ngân hàng; (ii) Hành vi giao dịch định lượng như số lượng sản phẩm tín dụng, doanh số gửi/rút tiền, số dư tài khoản; và (iii) Mức độ tương tác với ngân hàng trong thời gian gần đây*. Các yếu tố này được sử dụng để hình thành khung lý thuyết nhằm xây dựng mô hình học máy có khả năng dự đoán chính xác hành vi rời bỏ. Do đó, phần 2.2 sẽ trình bày chi tiết các lý thuyết liên quan để đo lường và xác định hành vi rời bỏ của khách hàng, đóng vai trò làm nền tảng cho mô hình đề xuất.

2.2. Cở sở lý thuyết về xác định khách hàng rời bỏ và các độ đo, xây dựng khung đánh giá mức độ rời bỏ của khách hàng dựa trên dữ liệu giao dịch và tương tác

Phần này kế thừa các kết quả nghiên cứu trước đã trình bày, nhằm xây dựng cơ sở lý luận cho việc xác định hành vi rời bỏ dịch vụ của khách hàng.

Dựa trên nghiên cứu của Huang và cộng sự (2012), một số chỉ số đánh giá mức độ rời bỏ của khách hàng đã được xác lập, đặt nền tảng cho việc định lượng hành vi rời bỏ. Nghiên cứu này đóng vai trò quan trọng trong việc thiết kế mô hình lý thuyết dựa trên các tỷ lệ rời bỏ, tỷ lệ duy trì và các phương pháp đo lường thực tiễn, qua đó hỗ trợ việc lượng hóa khả năng rời bỏ của khách hàng trong mô hình học máy.

2.2.1. Độ đo đánh giá khách hàng rời bỏ

- *Tỷ lệ khách hàng* rời bỏ là tỷ lệ tính số khách hàng (quan sát) rời bỏ trên tổng số lượng khách hàng (tổng quan sát) tại cùng một thời điểm. Tỷ lệ khách hàng rời bỏ tiêu chuẩn thường được gọi là tỷ lệ khách hàng rời bỏ vì nó đề cập đến tình trạng ngừng hoạt động hoàn toàn của một chủ tài khoản có thể có nhiều đăng ký. Vì vậy, đối với tỷ lệ khách hàng rời bỏ tiêu chuẩn, một chủ tài khoản hủy một đăng ký nhưng vẫn giữ một đăng ký khác không được coi là bỏ qua.

$$\text{Tỷ lệ khách hàng} = \frac{Khách\ hàng\ rời\ bỏ}{Tổng\ khách\ hàng} \quad (1)$$

- *Tỷ lệ duy trì khách hàng* trái ngược với tỷ lệ khách hàng rời bỏ ta có tỷ lệ duy trì (tỷ lệ giữ chân khách hàng), được tính bằng khách hàng ở lại trên tổng số lượng khách hàng (tổng quan sát) tại cùng một thời điểm.

$$\text{Tỷ lệ duy trì} = \frac{Khách\ hàng\ ở\ lại}{Tổng\ khách\ hàng} \quad (2)$$

- *Mối quan hệ giữa tỷ lệ duy trì và tỷ lệ khách hàng rời bỏ*: Đây là một thực tế quan trọng về tỷ lệ khách hàng rời bỏ và tỷ lệ duy trì: chúng có liên quan theo một cách rất chính xác và là hai mặt của cùng một đồng tiền.

$$\text{Tỷ lệ rời bỏ} + \text{tỷ lệ duy trì} = 100\% \quad (3)$$

2.2.2. Các phương pháp xác định tỷ lệ rời bỏ khách hàng

Ba chỉ số chính thường được sử dụng để đo lường hành vi rời bỏ bao gồm:

- **Tỷ lệ rời bỏ ròng (Net Churn Rate):** Được phản ánh phần doanh thu định kỳ bị mất sau khi đã tính đến những khách hàng duy trì nhưng có thay đổi về giá trị sử dụng. Chỉ số này phù hợp trong các mô hình dịch vụ có nhiều gói sản phẩm hoặc giá linh hoạt.

- **Tỷ lệ rời bỏ tiêu chuẩn (Standard Churn Rate):** Là tỷ lệ khách hàng hoàn toàn hủy bỏ dịch vụ trong một khoảng thời gian xác định, không bị ảnh hưởng bởi các yếu tố như giảm giá, bán thêm hay thay đổi gói cước. Đây là chỉ số đơn giản nhưng hữu ích khi toàn bộ khách hàng trả cùng mức giá hoặc sử dụng dịch vụ miễn phí. Việc xác định khách hàng rời bỏ thường dựa trên chuỗi hành vi gián đoạn vượt quá một ngưỡng thời gian nhất định (ví dụ: không tương tác trong 60 ngày).

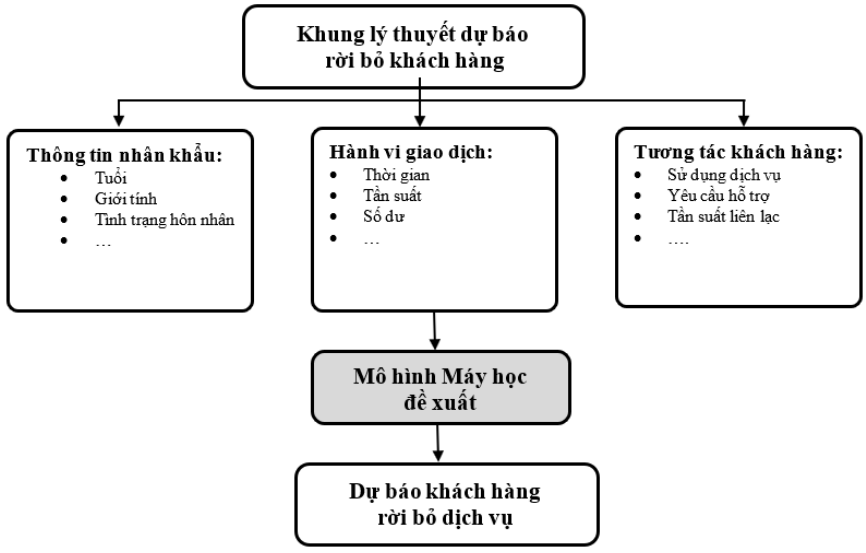
- **Tỷ lệ rời bỏ theo doanh thu định kỳ (MRR Churn Rate):** Là chỉ số phức tạp nhưng chính xác nhất trong các mô hình có nhiều cấp độ sản phẩm hoặc phân khúc khách hàng trả mức phí khác nhau. MRR churn loại trừ ảnh hưởng từ các giao dịch gia tăng (upselling) và tập trung vào phần doanh thu mất đi do khách hàng rời bỏ hoặc hạ cấp dịch vụ. Chỉ số này đặc biệt hữu ích để đánh giá tác động tài chính thực tế từ hành vi rời bỏ.

2.2.3. Mô hình khung lý thuyết dự báo rời bỏ khách hàng

Để tổng hợp các yếu tố có ảnh hưởng đến hành vi rời bỏ dịch vụ ngân hàng, nhóm tác giả đề xuất khung lý thuyết gồm 03 nhóm chính: *(i) thông tin nhân khẩu học, (ii) hành vi giao dịch, và (iii) tương tác khách hàng*. Khung lý thuyết này làm cơ sở cho việc xây dựng mô hình dự báo trên nền tảng học máy, đồng thời là điểm nối kết giữa lý thuyết và ứng dụng thực tiễn trong triển khai CRM.

Hình 1

Khung Lý Thuyết Dự Báo Hành Vi Rời Bỏ Dịch Vụ Ngân Hàng, được Xây Dựng Dựa Trên Tổng Hợp các Yếu Tố Nhân Khẩu Học, Hành Vi Giao Dịch và Tương Tác Khách Hàng (Tham Khảo từ Huang và Cộng Sự, 2012; Gupta và Cộng Sự, 2004)



Ghi chú. Dữ liệu từ "Customer churn prediction in telecommunications" bởi B. Huang, M. T. Kechadi, & B. Buckley, 2012, *Expert Systems with Applications, 39*(1), pp. 1414-1425 (https://doi.org/10.1016/j.eswa.2011.08.024). Dữ liệu từ "Valuing customers" bởi S. Gupta, D. R. Lehmann, & J. A. Stuart, 2004, *Journal of Marketing Research, 41*(1), pp. 7-18 (https://doi.org/10.1509/jmkr.41.1.7)

Tóm lại: trong các nghiên cứu về hành vi rời bỏ dịch vụ, lý thuyết vòng đời khách hàng (Customer Lifecycle Theory) và hành vi khách hàng (Customer Behavior Analytics) là những nền tảng chính giúp định hình khung lý thuyết. Lý thuyết này chỉ ra rằng hành vi rời bỏ có thể được phát hiện sớm qua các tín hiệu như giảm tần suất giao dịch, biến động số dư, hoặc giảm tương tác. Ngoài ra, mô hình RFM (Recency - Frequency - Monetary) cũng đóng vai trò nền tảng trong việc lượng hóa hành vi khách hàng. Từ đó, bài báo xây dựng mô hình khung lý thuyết dự báo hành vi rời bỏ gồm ba nhóm yếu tố: nhân khẩu học, hành vi giao dịch và mức độ tương tác với ngân hàng.

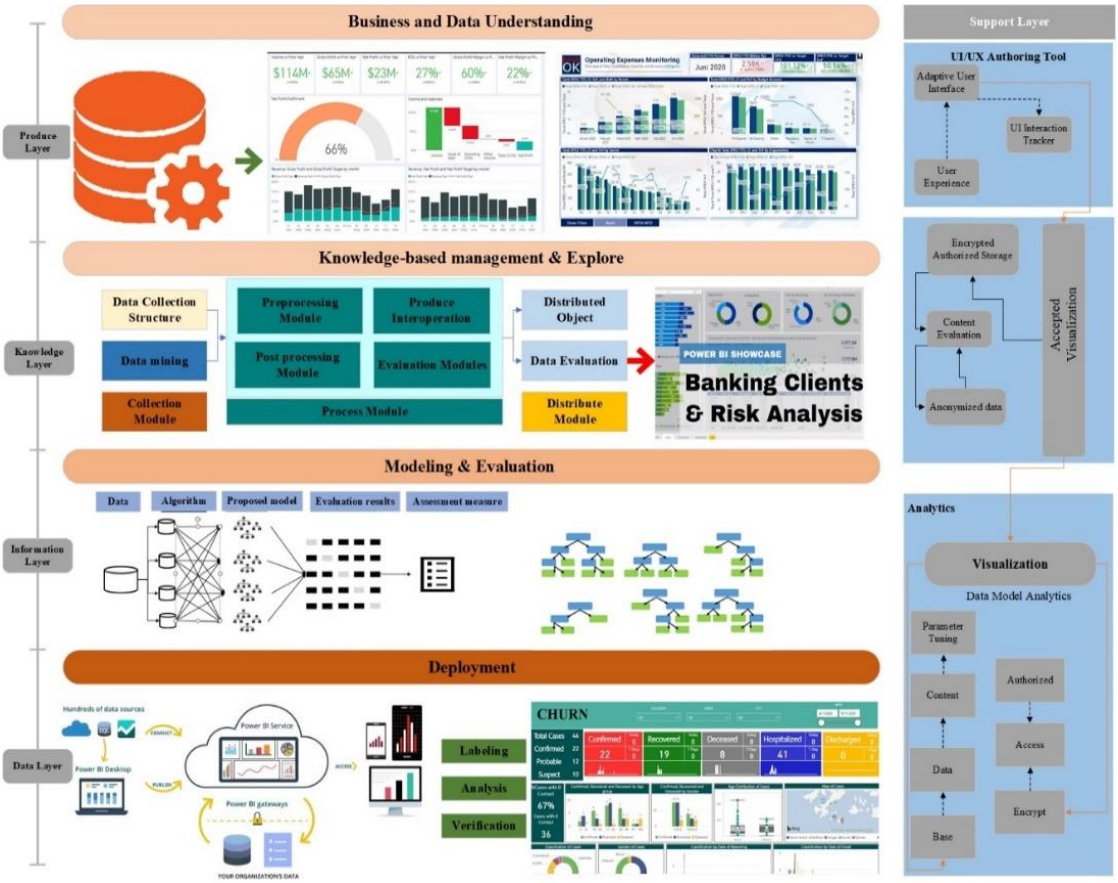
3. Phương pháp nghiên cứu

Trong bài báo này, nhóm tác giả đề xuất một khung làm việc tổng quát dựa trên phương pháp CRISP-DM của Chapman và cộng sự (1999) và các mô hình khai phá dữ liệu trong lĩnh vực ngân hàng được tổng hợp từ các nghiên cứu như Huang và cộng sự (2012), Devriendt và cộng sự (2021), Verbeke và cộng sự (2012). Khung làm việc này hướng đến việc áp dụng học máy vào dự báo rời bỏ khách hàng trong bối cảnh thực tế tại ngân hàng thương mại Việt Nam.

3.1. Đề xuất khung làm việc tổng quát

Hình 2

Hệ Thống Máy Học Dự Báo Khách Hàng Rời Bỏ Dịch Vụ Ngân Hàng



Ghi chú. Tác giả đề xuất

Trong đó:

(1) **Xây dựng mục tiêu khai thác dữ liệu:** Dựa trên mục tiêu kinh doanh của ngân hàng, xác định cụ thể vấn đề rời bỏ khách hàng và các biến liên quan đến hành vi khách hàng.

(2) **Tiền xử lý dữ liệu:** Áp dụng các bước như làm sạch, chuẩn hóa, mã hóa biến rời rạc, xử lý thiếu dữ liệu, ... nhằm chuẩn bị dữ liệu đầu vào phù hợp cho các thuật toán học máy (Kotsiantis & ctg., 2006). Giai đoạn này chiếm khoảng 70 - 90% thời gian toàn bộ quy trình khai phá dữ liệu.

(3) **Mô hình hóa và đánh giá**: Sử dụng các mô hình thống kê và học máy (Logistic Regression, Random Forest, TabNet, RealTabR) để xây dựng hệ thống dự báo, sau đó đánh giá hiệu suất qua các chỉ số như AUC, F1-score, Precision, Recall của Jiawei và Micheline (2006).

(4) **Triển khai**: Đưa mô hình vào môi trường ứng dụng thử nghiệm tại doanh nghiệp, theo hướng tích hợp với hệ thống CRM và dashboard phân tích rời bỏ khách hàng.

3.2. Tiền xử lý dữ liệu

3.2.1. Thu thập dữ liệu và định nghĩa các tiêu chí

Dữ liệu được thu thập từ hệ thống quản lý giao dịch và thông tin khách hàng (CASA) tại một số ngân hàng thương mại khu vực phía Nam Việt Nam trong giai đoạn từ tháng 01 năm 2020 đến tháng 12 năm 2023. Bộ dữ liệu ban đầu bao gồm hơn 10,000 khách hàng cá nhân, có lịch sử sử dụng dịch vụ ngân hàng liên tục trong ít nhất 12 tháng, nhằm đảm bảo tính liên tục và độ tin cậy trong việc phân tích hành vi. Các khách hàng tổ chức, hồ sơ chưa đầy đủ định danh, hoặc có trên 50% dữ liệu bị thiếu đã được loại bỏ trong bước xử lý trước. Sau khi làm sạch và chuẩn hóa, dữ liệu đầu vào sử dụng cho mô hình học máy gồm 9,382 bản ghi với 24 biến đầu vào định lượng. Việc khai thác dữ liệu được thực hiện thông qua các biên bản hợp tác nghiên cứu với các ngân hàng, đảm bảo tuân thủ nghiêm ngặt các quy định pháp lý và chuẩn mực đạo đức trong nghiên cứu. Dữ liệu đã được mã hóa và loại bỏ toàn bộ thông tin định danh để đảm bảo bảo mật và phục vụ mục đích nghiên cứu học thuật. Tập khách hàng rời bỏ được xác định dựa trên lịch sử giao dịch ghi nhận từ hệ thống kho dữ liệu (Data Warehouse) của ngân hàng, với một số chỉ tiêu chính phản ánh hành vi sử dụng dịch vụ:

- Số lượng sản phẩm tín dụng đang sử dụng;
- Tuổi của khách hàng;
- Thời gian gắn bó với ngân hàng (tháng);
- Số dư bình quân tài khoản;
- Tổng doanh số tiền vào và tiền ra quy đổi;
- Tỷ lệ chi phí chuyển khoản;
- Tổng doanh số giao dịch qua tài khoản.

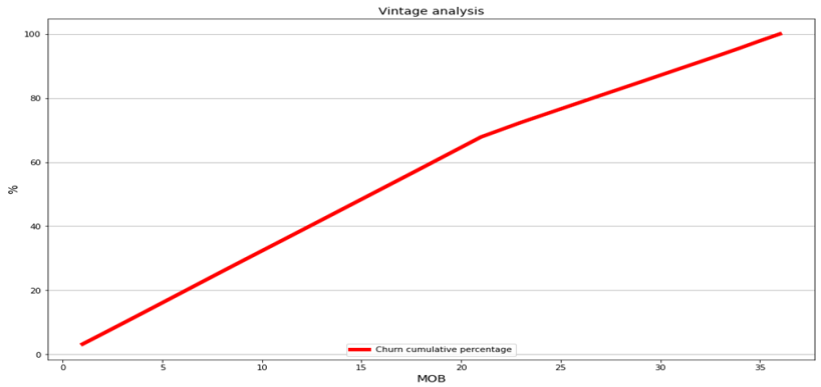
Các biến này phản ánh trực tiếp hành vi tài chính của khách hàng, là nền tảng quan trọng cho việc xây dựng mô hình dự báo rời bỏ với độ chính xác và khả năng ứng dụng thực tiễn cao trong hệ thống CRM của ngân hàng. Chi tiết các tiêu chí được định nghĩa và tiền xử lý trong Bảng 1.

3.2.2. Xác định mối quan hệ giữa biến phụ thuộc và biến độc lập

Mô hình dự đoán khách hàng rời bỏ được phát triển bằng cách sử dụng lý thuyết "kết quả trong tương lai được phản ánh bằng hành vi trong quá khứ" có trong nghiên cứu của Agarwal và cộng sự (2022). Dựa trên phân tích Vintage, hoạt động của các tài khoản khách hàng đã mở trước đó được phân tích để dự đoán hoạt động của các tài khoản trong tương lai có trong Hình 3:

Hình 3

Phân Tích Vintage nhằm Xác Định Hiệu Năng Mô Hình Đề Xuất



Ghi chú. Tác giả đề xuất

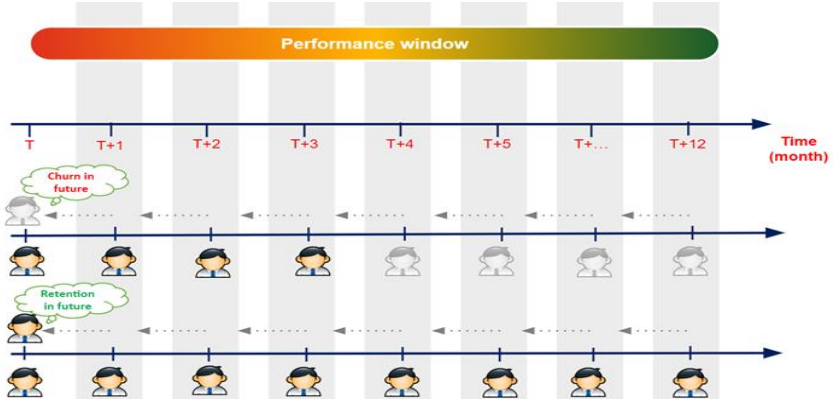
Quá trình phân tích Vintage là phân tích dựa trên thông tin nếu là khách hàng rời bỏ thì bao lâu khách hàng rời bỏ được tính trên số cộng dòng tích lũy. Trong đó MOB (Month on Book: Thời gian có trên hệ thống) là 12 tháng và trục tung chỉ số là 30%, có nghĩa là nếu khách hàng rời bỏ ngân hàng trong 12 tháng sẽ có 30% số khách hàng rời bỏ ngân hàng cho hiệu năng mô hình đề xuất là 12 tháng là số hợp lý để ngân hàng có thể lên kế hoạch ứng phó với vấn đề khách hàng rời bỏ ngân hàng. Như thường được thực hành trong tính điểm hành vi, để tạo ra một biến mục tiêu để phát triển mô hình, người ta sẽ chọn một ngày quan sát đủ lâu trong quá khứ (hơn 12 tháng), và sau đó quan sát hiệu suất thanh toán của các khách hàng được mở trong 12 tháng kể từ ngày quan sát ngày để xem liệu một sự kiện rời bỏ dịch vụ đã xảy ra hay chưa.

Để thực hiện phân tích này, dữ liệu được thu thập từ các tài khoản được mở trong một khung thời gian cụ thể, sau đó theo dõi hiệu suất của chúng trong một khoảng thời gian cụ thể khác để xác định xem chúng tốt (retention) hay xấu (rời bỏ dịch vụ). Dữ liệu được thu thập (các tiêu chí) cùng với phân loại tốt/xấu (mục tiêu) tạo thành mẫu phát triển mà từ đó thẻ điểm được phát triển. Đối với thẻ điểm hành vi, điều tương tự cũng được thực hiện cho các tài khoản hiện có, nơi chúng ta xem xét các tài khoản tại một thời điểm và theo dõi hành vi thanh toán của họ trong một khoảng thời gian được chỉ định để xác định mục tiêu.

Quá trình này có thể được lặp lại với nhiều ngày quan sát, để đảm bảo rằng thông tin mặc định không thiên về một khoảng thời gian 12 tháng cụ thể.

Hình 4

Minh Họa Xác Định Rời Bỏ Dịch Vụ trong Tương Lai (trong 12 Tháng Tới)



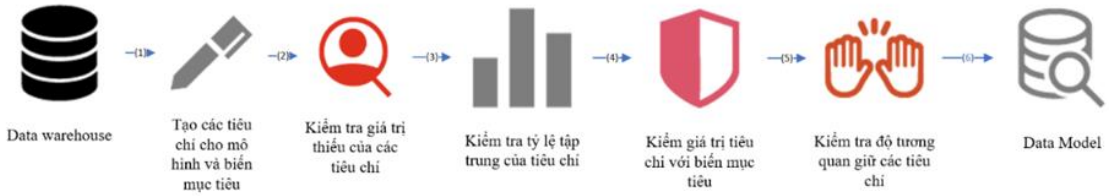
Ghi chú. Tác giả đề xuất

Quy trình trên đảm bảo rằng mọi trường hợp đơn lẻ đều được liên kết với biến rời bỏ dịch vụ mục tiêu dựa trên khoảng thời gian hoạt động 12 tháng. Trong trường hợp như vậy, chúng ta sẽ tính số lượng tất cả các sự kiện rời bỏ dịch vụ xảy ra trong vòng 12 tháng, cho một nhóm tài khoản không có rời bỏ dịch vụ vào ngày quan sát và chúng ta sẽ bao gồm các sự kiện rời bỏ dịch vụ lại sau. Như vậy sau khi thống nhất với các chuyên gia ngân hàng, chúng ta có thể xác định được biến mục tiêu rời bỏ dịch vụ trong vòng 12 tháng tới của khách hàng tính từ thời điểm quan sát.

3.3. Tiền xử lý dữ liệu và phân tích đơn biến

Hình 5

Sơ Đồ Quy Trình Tiền Xử Lý Dữ Liệu và Phân Tích Đơn Biến



Ghi chú. Tác giả đề xuất

Trong đó:

1. Từ kho dữ liệu của ngân hàng chúng ta tiến hành tạo các biến mục tiêu và biến phụ thuộc cho mô hình máy học đã thu được 104 tiêu chí và biến mục tiêu.

2. Kiểm tra chất dữ liệu từng tiêu chí nếu tiêu chí thiếu chiếm 50% trong tổng số quan sát sẽ loại ra khỏi danh sách tiêu chí.

3. Kiểm tra tỷ lệ tập trung tại một giá trị nếu giá trị đó chiếm trên 90% sẽ loại ra khỏi danh sách tiêu chí. Dựa trên 02 ngưỡng của (2) và (3) ta thu được 86 tiêu chí thỏa điều kiện.

4. Đánh giá mức độ ảnh hưởng của từng tiêu chí với biến mục tiêu thông qua chỉ số IV. Nếu tiêu chí có chỉ số IV theo nghiên cứu của tác giả (Howard, 1966)) thì chỉ số IV < 2%. Điều này chứng tỏ tiêu chí có ảnh hưởng rất ít với biến mục tiêu, bộ dữ liệu thu thập được 67 tiêu chí thỏa điều kiện.

5. Kiểm tra độ tương quan của từng tiêu chí với nhau nếu 02 tiêu chí có độ tương quan trên 70% chúng ta xem xét tiêu chí có chỉ số IV cao hơn. Chúng ta đã thu được 20 tiêu chí thỏa mãn.

Sau khi qua trình phân tích của dữ liệu ban đầu đã được mô tả từng bước theo Hình 4 ta có kết quả sau khi tiền xử lý dữ liệu và phân tích đơn biến như:

Bảng 1

Danh Sách Số Lượng Biến cho Xây Dựng Mô Hình Máy Học

STT	Tên tiêu chí	Mô tả
1	AB10	Tổng số dư bình quân của các tài khoản tại tháng báo cáo/Tổng số dư của các tài khoản tại tháng báo cáo
2	AB21	Tỷ lệ Tổng doanh số gửi tiền mặt ra tài khoản trên Tổng doanh số giao dịch chuyển tiền

STT	Tên tiêu chí	Mô tả
3	AB22	Tỷ lệ Tổng doanh số tiền ra tài khoản quy đổi (-) trên Tổng doanh số tiền vào tài khoản quy đổi (-) tháng trước
4	AB32	Tỷ lệ Tổng số dư của các tài khoản tại tháng báo cáo trên Tổng số dư của các tài khoản tại tháng báo cáo tháng trước
5	AB33	Tỷ lệ Tổng số dư bình quân của các tài khoản tại tháng trên Tổng số dư bình quân của các tài khoản tại tháng trước
6	AB37	Tổng doanh số giao dịch qua tài khoản thanh toán quy đổi trung bình trong 03 tháng gần đây
7	AB42	Tổng số dư bình quân của các tài khoản tại tháng báo cáo/Tổng số dư của các tài khoản tại tháng báo cáo trung bình trong 03 tháng gần đây
8	AB48	Tỷ lệ tổng chi phí chuyển qua tài khoản trên Tổng doanh số giao dịch chuyển tiền trung bình trong 03 tháng gần đây
9	AB53	Tỷ lệ Tổng doanh số gửi tiền mặt ra tài khoản trên Tổng doanh số giao dịch chuyển tiền trung bình trong 03 tháng gần đây
10	AB57	Tỷ lệ Tổng doanh số tiền vào tài khoản quy đổi (-) trên Tổng doanh số giao dịch qua tài khoản thanh toán quy đổi tháng trước trung bình trong 03 tháng gần đây
11	AB64	Tỷ lệ Tổng số dư của các tài khoản tại tháng báo cáo trên Tổng số dư của các tài khoản tại tháng báo cáo tháng trước trung bình trong 03 tháng gần đây
12	AB65	Tỷ lệ Tổng số dư bình quân của các tài khoản tại tháng trên Tổng số dư bình quân của các tài khoản tại tháng trước trung bình trong 03 tháng gần đây
13	AB66	Tổng số dư của các tài khoản tại tháng báo cáo trung bình trong 06 tháng gần đây
14	AB73	Tổng số dư bình quân của các tài khoản tại tháng báo cáo/Tổng số dư của các tài khoản tại tháng báo cáo trung bình trong 06 tháng gần đây
15	AB79	Tỷ lệ tổng chi phí chuyển qua tài khoản trên Tổng doanh số giao dịch chuyển tiền trung bình trong 06 tháng gần đây
16	AB89	Tỷ lệ Tổng doanh số tiền vào tài khoản quy đổi (-) Tổng doanh số tiền vào tài khoản quy đổi (-) tháng trước trung bình trong 06 tháng gần đây
17	AB95	Tỷ lệ Tổng số dư của các tài khoản tại tháng báo cáo trên Tổng số dư của các tài khoản tại tháng báo cáo tháng trước trung bình trong 06 tháng gần đây
18	NUM_NO_CREDIT	Số sản phẩm tín dụng khách hàng sử dụng
19	Z1	Tuổi khách hàng
20	Z6	Thời gian quan hệ với ngân hàng

Ghi chú. Tác giả đề xuất

Nhận xét: Từ kết quả tiền xử lý dữ liệu và phân tích đa biến ta được một bảng dữ liệu có 20 tiêu chí và 11,6011 quan sát (với 10,0548 quan sát Retention_next_12M hay còn gọi là quan sát tốt và 15,463 quan sát Rời bỏ dịch vụ_next_12M hay còn gọi là quan sát xấu).

Để tiến hành xây dựng mô hình máy học dự đoán khách hàng trung thành rời bỏ ngân hàng từ đó chúng ta tiến hành chia dữ liệu thành 02 phần là huấn luyện (70%) tương ứng 81,207 quan sát và kiểm thử (30%) tương ứng 34,804 quan sát.

Dùng hai tập dữ liệu nói trên tiến hành xây dựng mô hình với các thuật toán như Logistic Regression, Decision Tree, RandomForest, Xgboost, Naïve Bayes.

4. Kết quả và thảo luận

Để đảm bảo tính thuyết phục của mô hình, nhóm tác giả thực hiện đối chiếu giữa các mô hình học máy khác nhau: Logistic Regression (mô hình tuyến tính truyền thống), MLP (mô hình mạng nơ-ron phổ biến), TabNet và RealTabR (mô hình sâu chuyên biệt cho dữ liệu bảng). Kết quả benchmark được trình bày trong Bảng 2, trong đó TabNet đạt AUC = 0.863 và RealTabR đạt F1-score cao nhất = 0.741, vượt trội so với Logistic Regression (F1-score = 0.639). Các kết quả này cho thấy khả năng dự báo vượt trội của các mô hình sâu trong bối cảnh dữ liệu thực tế.

4.1. Đánh giá so sánh

Kết quả nghiên cứu này cho thấy sự vượt trội về hiệu quả mô hình so với một số công trình trước đó. Cụ thể, nghiên cứu của Vo và cộng sự (2018) sử dụng dữ liệu phi cấu trúc từ nhật ký cuộc gọi của khách hàng để dự báo rời bỏ, tuy đạt được kết quả nhất định nhưng độ chính xác vẫn dưới 90%, đồng thời chưa xác lập rõ ràng khung thời gian dự báo. Trong khi đó, bài báo hiện tại sử dụng dữ liệu hành vi tài chính định lượng và thiết lập khung thời gian dự báo cụ thể là 12 tháng, giúp mô hình đạt độ chính xác cao lên đến 94%, từ đó nâng cao khả năng triển khai trong thực tiễn ngân hàng số. Bên cạnh đó, khác với các nghiên cứu như của Devriendt và cộng sự (2021) và Silveira và cộng sự (2021), vốn chủ yếu tập trung vào việc đánh giá hiệu năng mô hình, bài báo này còn đề xuất một quy trình tích hợp mô hình dự báo rời bỏ vào hệ thống quản trị quan hệ khách hàng (CRM), qua đó góp phần nâng cao tính khả thi và hiệu quả trong ứng dụng quản trị thực tế.

Bảng 2

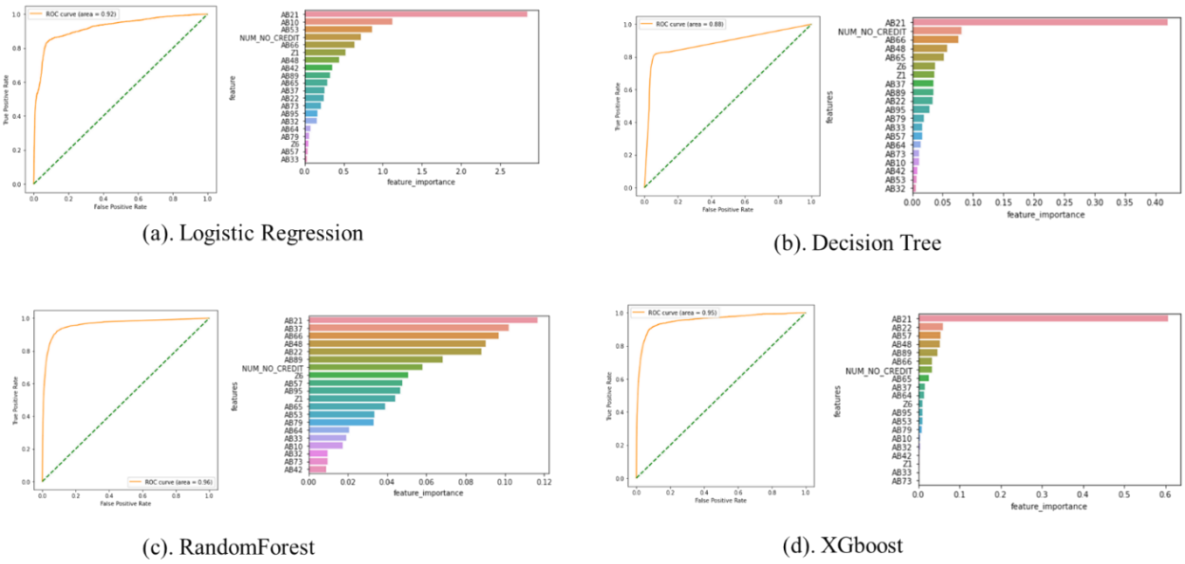
Kết Quả Đánh Giá Định Lượng các Mô Hình Máy Học

	Mô hình	Precision	Recall	f1-score	Accuracy	ROC
Retention trong 12 tháng tới	Logistic Regression	0.93	0.98	0.95	0.92	0.92
	Random Forest	0.96	0.97	0.97	0.94	0.96
	Decision Tree	0.96	0.96	0.96	0.93	0.88
	XGBoost	0.96	0.97	0.97	0.94	0.95
	Naïve Bayes	0.95	0.94	0.94	0.90	0.89
Rời bỏ dịch vụ trong 12 tháng tới	Logistic Regression	0.79	0.54	0.64	0.92	0.92
	Random Forest	**0.82**	**0.75**	**0.78**	**0.94**	**0.96**
	Decision Tree	0.75	0.73	0.74	0.93	0.88
	XGBoost	0.81	0.74	0.77	0.94	0.95
	Naïve Bayes	0.62	0.67	0.65	0.90	0.89

Ghi chú. Tác giả đề xuất

Hình 6

Kết Quả Đánh Giá Định Tính dựa trên Sơ Đồ ROC các Phương Pháp Máy Học



(a). Logistic Regression

(b). Decision Tree

(c). RandomForest

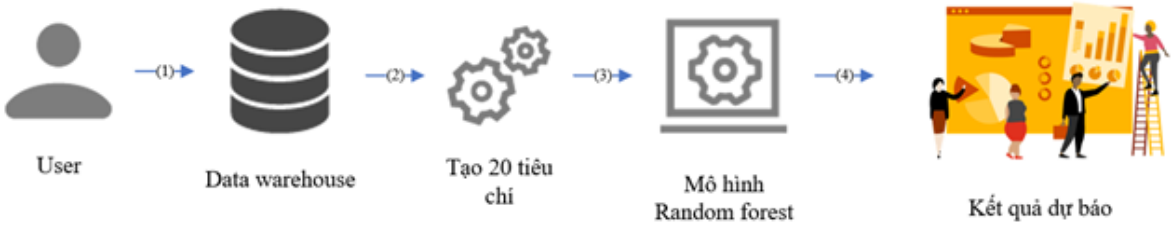
(d). XGboost

Ghi chú. Tác giả đề xuất

4.2. Xây dựng Hệ thống dự đoán khách hàng rời bỏ ngân hàng tích hợp hệ thống CRM

Hình 7

Sơ Đồ Vận Hành Hệ Thống Mô Hình Dự Báo Khách Hàng Rời Bỏ



Ghi chú. Tác giả đề xuất

Bài báo đã cung cấp bằng chứng thực nghiệm có giá trị, cho thấy dữ liệu hành vi có thể nâng cao hiệu quả dự báo trong CRM, đồng thời làm rõ vai trò của các yếu tố hành vi trong mô hình học máy. Cụ thể:

- Từ yêu cầu người dùng (dự đoán danh sách khách hàng có khả năng rời bỏ trong giai đoạn từ 01/12/2022 đến 31/12/2022), hệ thống truy vấn đến kho dữ liệu dùng chung (Data Warehouse) của ngân hàng.

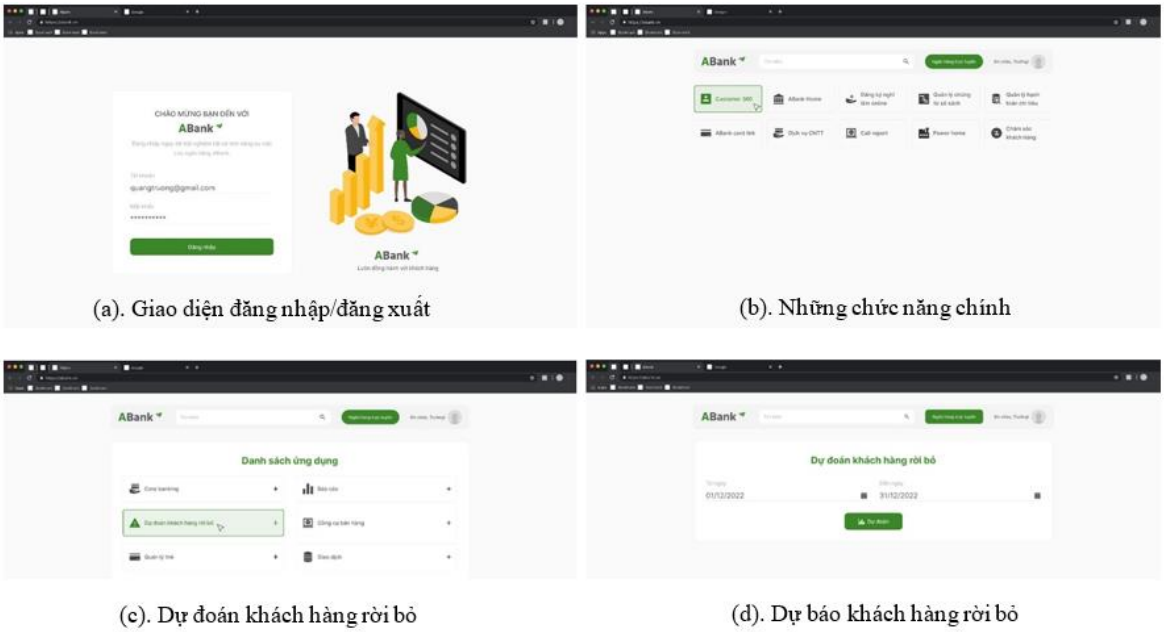
- Từ kho dữ liệu này, hệ thống tiến hành trích xuất dữ liệu hành vi phù hợp với 20 tiêu chí phân tích đã được xác lập (tham khảo Bảng 1), áp dụng cho toàn bộ tập khách hàng mục tiêu trong khung thời gian từ 01/12/2022 đến 31/12/2022.

- Kết quả mô hình dự báo cho thấy dữ liệu hành vi đóng vai trò then chốt trong việc cải thiện độ chính xác, khả năng nhận diện rủi ro, cũng như hỗ trợ các mô hình học máy phản ánh đúng hơn xu hướng thay đổi trong hành vi tài chính của khách hàng.

- Người dùng (các nhà quản trị hoặc bộ phận chăm sóc khách hàng) có thể sử dụng đầu ra từ hệ thống để thiết lập các chiến lược giữ chân phù hợp, từ đó giảm tỷ lệ rời bỏ và nâng cao giá trị vòng đời của khách hàng trong hệ thống CRM.

Hình 8

Một Số Giao Diện Tương Tác của Hệ Thống Dự Báo Khách Hàng Rời Bỏ



(a). Giao diện đăng nhập/đăng xuất (b). Những chức năng chính

(c). Dự đoán khách hàng rời bỏ (d). Dự báo khách hàng rời bỏ

Ghi chú. Tác giả đề xuất

5. Kết luận

Bài báo này cung cấp cơ sở thực nghiệm nhằm giải quyết bài toán dự báo khả năng rời bỏ dịch vụ ngân hàng của khách hàng cá nhân thông qua ứng dụng các thuật toán học máy. Dữ liệu đầu vào bao gồm hơn 10,000 quan sát từ hệ thống giao dịch thực tế của một số ngân hàng thương mại khu vực phía Nam, trong giai đoạn từ năm 2020 đến năm 2023. Trên cơ sở đó, chúng tôi đã triển khai và so sánh hiệu quả giữa các mô hình học máy phổ biến, gồm Logistic Regression, Decision Tree, Naïve Bayes, XGBoost và Random Forest. Kết quả cho thấy mô hình Random Forest đạt độ chính xác cao nhất với độ chính xác tổng thể (Accuracy) 95%, Precision 82%, Recall 75%, F1-score 87% và ROC-AUC đạt 96%.

Đóng góp học thuật

Bài báo đã đóng góp vào nền tảng lý thuyết và học thuật trong lĩnh vực phân tích hành vi khách hàng và quản trị quan hệ khách hàng ngân hàng, cụ thể:

(i) Khai thác dữ liệu hành vi định lượng dạng động với chuỗi thời gian 12 tháng, thay thế cho dữ liệu tĩnh thường được sử dụng trong các nghiên cứu trước, qua đó phản ánh sát hơn hành vi thực tế của khách hàng;

(ii) Đề xuất khung tích hợp lý thuyết gồm CRISP-DM, phân tích hành vi và phương pháp Vintage nhằm kết nối giữa phân tích dữ liệu và triển khai thực tế trong CRM ngân hàng;

(iii) Thực hiện so sánh thực nghiệm giữa các thuật toán truyền thống và mô hình học sâu xử lý dữ liệu bảng (MLP, TabNet, RealTabR), từ đó làm rõ tính hiệu quả của từng phương pháp trong ngữ cảnh dữ liệu tài chính.

Đóng góp thực tiễn

Kết quả này có thể được tích hợp vào hệ thống CRM để hỗ trợ nhà quản trị xây dựng chiến lược giữ chân khách hàng, đặc biệt là các nhóm khách hàng có giá trị cao. Mô hình giúp

tăng cường khả năng dự đoán và ra quyết định, đóng góp vào nâng cao hiệu quả hoạt động kinh doanh ngân hàng.

Hạn chế và hướng phát triển

Bài báo này hiện tại chủ yếu dựa trên dữ liệu nội bộ ngân hàng, chưa tích hợp các nguồn dữ liệu bên ngoài như dữ liệu mạng xã hội hay tín dụng chéo. Đồng thời, các kỹ thuật học sâu (Deep Learning) vẫn chưa được triển khai. Trong tương lai, việc tích hợp dữ liệu từ các nền tảng như Zalo hoặc Trusting Social, kết hợp với các công cụ học máy nâng cao như Teachable Machine, có thể mở rộng tính ứng dụng và độ chính xác của mô hình trong bối cảnh ngân hàng số.

ĐÓNG GÓP KHOA HỌC

Bài báo xác định rõ khoảng trống nghiên cứu; bài báo mở rộng hoặc bổ sung lý thuyết hiện có; bài báo đề xuất mô hình lý thuyết hoặc mô hình phân tích mới; bài báo phát triển phương pháp mới hoặc cải tiến phương pháp hiện có; bài báo cung cấp bộ dữ liệu mới hoặc bằng chứng thực nghiệm mới; bài báo có ý nghĩa thống kê và thực tiễn rõ ràng; bài báo đưa ra hàm ý chính sách, quản trị hoặc công nghệ; bài báo gợi mở các hướng nghiên cứu tiếp theo.

ĐÓNG GÓP CỦA TÁC GIẢ

CRediT: **Nguyễn Quốc Hùng**: Xây dựng ý tưởng, Thiết kế nghiên cứu, Xây dựng khung lý thuyết và mô hình nghiên cứu, Đề xuất phương pháp nghiên cứu, Giám sát thu thập dữ liệu, Xử lý và phân tích dữ liệu, Điều tra/Thí nghiệm, Phân tích chính thức, Diễn giải kết quả nghiên cứu, Viết bản thảo ban đầu, Viết bản chỉnh sửa; **Nguyễn Thị Xuân Đào**: Thu thập dữ liệu, Làm sạch và tiền xử lý dữ liệu, Quản lý dữ liệu, Xây dựng bộ biến đầu vào, Đóng góp phần phương pháp nghiên cứu; **Lê Thành Trung**: Điều phối và liên hệ thu thập dữ liệu, Xác thực dữ liệu, Đánh giá tính hợp lệ của dữ liệu, Thảo luận kết quả nghiên cứu dưới góc độ thực tiễn; **Nguyễn Quang Trường**: Thiết kế thử nghiệm, Triển khai mô hình học máy, Điều tra/Thí nghiệm, Đánh giá và so sánh hiệu năng mô hình, Phân tích và thảo luận kết quả thực nghiệm.

TÀI TRỢ

Nghiên cứu này được tài trợ bởi Đại học Kinh tế TP. HCM (UEH), Việt Nam theo. Mã số đề tài: CS-2023-12.

TUYÊN BỐ KHÔNG CÓ XUNG ĐỘT LỢI ÍCH

Các tác giả cam kết, tuyên bố không có bất kỳ xung đột lợi ích nào liên quan đến việc công bố bài báo này.

Tài liệu tham khảo

Agarwal, V., Taware, S., Yadav, S. A., Gangodkar, D., Rao, A., & Srivastav, V. (2022). Customer-churn prediction using machine learning. *6th International Conference on Electronics, Communication and Aerospace Technology, Coimbatore, India*, 1035-1040. https://doi.org/10.1109/ICECA55336.2022.10009093

Ashraf, R. (2024). Bank customer churn prediction using machine learning framework. *Journal of Applied Finance & Banking*, *14*(4), 1-5. https://doi.org/10.47260/jafb/1445

Brownlow, J., Chu, C., Fu, B., Xu, G., Culbert, B., & Meng, Q. (2018). Cost-sensitive churn prediction in fund management services. In J. Pei, Y. Manolopoulos, S. Sadiq, & J. Li (Eds.), *Database systems for advanced applications* (pp. 776-788). Springer International Publishing. https://doi.org/10.1007/978-3-319-91458-9_49

Colgate, M. R., & Danaher, P. (2000). Implementing a customer relationship strategy: The asymmetric impact of poor versus excellent execution. *Journal of the Academy of Marketing Science, 28*(3), 375-387. https://doi.org/10.1177/0092070300283006

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (1999). *CRISP-DM 1.0: Step-by-step data mining guide.* CRISP-DM Consortium. https://mineracaodedados.files.wordpress.com/2012/12/crisp-dm-1-0.pdf

Devriendt, F., Berrevoets, J., & Verbeke, W. (2021). Why you should stop predicting customer churn and start using uplift models. *Journal of Information Science*, *548*, 497-515. https://doi.org/10.1016/j.ins.2019.12.075

Genuer, R., Poggi, J.-M., Genuer, R., & Poggi, J.-M. (2020). *Random forests with R* (vol. 42). https://link.springer.com/book/10.1007/978-3-030-56485-8

Gupta, S., Lehmann, D. R., & Stuart, J. A. (2004). Valuing customers. *Journal of Marketing Research, 41*(1), 7-18. https://doi.org/10.1509/jmkr.41.1.7

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Tibshirani, R., & Friedman, J. (2009). Random forests. In *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed., pp. 587-604). Springer. https://doi.org/10.1007/978-0-387-84858-7_15

Howard, R. A. (1996). Information value theory. *IEEE Transactions on Systems Science and Cybernetics, 2*(1), 22-26. https://doi.org/10.1109/TSSC.1966.300074

Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. *Expert Systems with Applications, 39*(1), 1414-1425. https://doi.org/10.1016/j.eswa.2011.08.024

Jiawei, H., & Micheline, K. (2006). *Data mining: Concepts and techniques*. Morgan Kaufmann.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). *Logistic regression: A self-learning text*. Statistics for Biology and Health.

Kotsiantis, S., Zaharakis, I., & Pintelas, P. (2006). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review, 26*, 159-190. https://doi.org/10.1007/s10462-007-9052-3

Lee, E., Kim, B., Kang, S., Kang, B., Jang, Y., & Kim, H. K. (2018). Profit optimizing churn prediction for long-term loyal customers in online games. *IEEE Transactions on Games, 12*(1), 41-53. https://doi.org/10.1109/TG.2018.2871215.

Li, Y., Hou, B., Wu, Y., Zhao, D., Xie, A., & Zou, P. (2021). Giant fight: Customer churn prediction in traditional broadcast industry. *Journal of Business Research, vol. 131*, pp. 630-639. https://doi.org/10.1016/j.jbusres.2021.01.022.

Li, Y., & Yan, K. (2025). Prediction of bank credit customers churn based on machine learning and interpretability analysis. *Data Science in Finance and Economics, 5*(1), 19-34. https://doi.org/10.3934/dsfe.2025002

Liao, K.-H., & Chueh, H.-E. (2011). Applying fuzzy data mining to telecom churn management. In *Proceedings of the International Conference on Intelligent Computing and Information Science* (Vol. 134, pp. 259-264). Springer. https://doi.org/10.1007/978-3-642-18129-0_41

Liu, X., Xia, G., Zhang, X., & Ma, W. (2024). Customer churn prediction model based on hybrid neural networks. *Scientific Reports, 14*, 1-17. https://doi.org/10.1038/s41598-024-79603-9

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, É. (2025). *Naive Bayes - scikit-learn 1.4.dev0 documentation*. https://scikit-learn.org/dev/modules/naive_bayes.html#naive-bayes

Reichheld, F. F. (1996). Learning from customer defections. *Journal of Harvard Business Review, 74*(2), 56-67.

Rokach, L., & Maimon, O. (2005). Decision trees, decision trees. In O. Maimon & L. Rokach (Eds.), *Data mining and knowledge discovery handbook* (pp. 165-192). Springer. https://doi.org/10.1007/0-387-25465-X_9

Rosa, N. B. D. C. (2019). *Gauging and foreseeing customer churn in the banking industry: A neural network approach* [Doctoral dissertation, University of Lisboa]. https://core.ac.uk/reader/303765845

Shilong, Z. (2021). Machine learning model for sales forecasting by using XGBoost. In *IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), Guangzhou, China* (pp. 480-483). https://doi.org/10.1109/ICCECE51280 .2021.9342304

Silveira, L. J., Pinheiro, P. R., & Junior, L. S. D. M. (2021). A novel model structured on predictive churn methods in a banking organization. *Journal of Risk Financial Management, 14*(10), Article 481. https://doi.org/10.3390/jrfm14100481

Vélez, D., Ayuso, A., Perales-González, C., & Rodríguez, J. T. (2020). Churn and net promoter score forecasting for business decision-making through a new stepwise regression methodology. *Knowledge-Based Systems, 196*. https://doi.org/10.1016/j.knosys.2020 .105762

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research, 218*(1), 211-229. https://doi.org/10.1016/ j.ejor.2011.09.031

Vo, N. N., Liu, S., Brownlow, J., Chu, C., Culbert, B., & Xu, G. (2018). Client churn prediction with call log analysis. In J. Pei, Y. Manolopoulos, S. Sadiq, & J. Li (Eds.), *Database Systems for Advanced Applications. DASFAA 2018. Lecture Notes in Computer Science* (Vol. 10828). Springer. https://doi.org/10.1007/978-3-319-91458-9_47

Webb, G. I., Keogh, E., & Miikkulainen, R. (2011). *Naïve Bayes*. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 713-714). Springer. https://doi.org/10.1007/978-0-387-30164-8_576